A History of MOS Transistor Compact Modeling

Chih-Tang Sah

University of Florida
Department of Electrical and Computer Engineering
Florida Solid-State Electronics Laboratory
Gainesville, Florida, USA
6 major revisions 20050118-to-20050331

Abstract

Metal-Oxide-Silicon Field-Effect-Transistor (MOSFET) or MOS Transistor (MOST) is a three-dimensional electronic device. It operates on the conductivity modulation principle in a thin semiconductor layer using a controlling electric field to give amplifying and switching functions between two of the three electrical terminals (input, output and common) attached to the film. This principle was first proposed 80 years ago (1926) [1-4] by Lilienfeld. A review was given in 1988 on the evolution of the MOS transistor [5]. A detailed tutorial exposition of the MOST Compact Modeling (CM) development is planned [6]. Electrical characterization experiments and mathematical theory began 45 years ago (1959) when stable silicon oxides were grown on nearly perfect (crystalline, low defect) silicon semiconductor by Atalla, Tannenbaum and Scheibner at Bell Telephone Laboratories [7]. Simple analytical compact models of the MOS transistors are needed for computer-aided design of digital and analog integrated circuits containing thousands to millions transistors on a silicon chip, using circuit simulators such as SPICE. This keynote address describes an early history of MOS transistor compact modeling, from the **threshold voltage model** used in the first version of SPICE to the two latest advances under development, the **inversion charge model** and the original (one-piece) **surface potential model**.

I. Introduction

The purpose of compact modeling is to derive simple, fast and accurate analytical (mathematical equations) representations of the terminal electrical (DC, switching, and also small-signal) characteristics of MOS field-effect transistors (MOSFETs) or MOS transistors MOSTs. Compact transistor models are needed to compute numerically the transistor characteristics, rapidly enough, for use in circuit simulators to design and optimize the performance of silicon monolithic integrated circuits (or chips) containing thousands to millions of similar and dissimilar transistors for switching and analog applications.

Metal-Oxide-Silicon Field-Effect Transistor is inherently a two-dimensional (2-D) electronic (electron and hole) device {See Fig.640.1 on p.539 of [8], Fig.P610 on p.73 and Fig.940.1 on 148 of [10], or Fig.3.1 on p.70 and Fig.6.1 on p.231 of [11].} Its input voltage is applied to the **gate** conductor electrode to create an electric field in the x-direction perpendicular to a semiconductor layer in order to modulate the sheet conductance of the layer in the y-z plane. This applied gate voltage modulates the current passing through the layer in the y-direction between two contacts, **drain** and **source**. To make the 2-D problem tractable, device-physics-based decompositions of the 2-D problem into two coupled 1-D problems have been employed by all authors since day one (~1960).

The mathematical culprit is the nonlinear coupling due to the fractional-power/exponential dependence on the coupling variable (the surface potential, which is explained in the next paragraph) arisen from the spatially varying dopant impurity concentration in the basewell-channel region (known as the **bulk-charge**) of the transistor in the 1-D x-solution. The major effort of advanced compact modeling has been to find numerical algorithms and analytical linearization formulations of this coupling in order to provide fast and accurate extraction of the model parameters which are then used in the model to compute the characteristics of a wide-range of transistor designs for the thousands to millions of transistors in the integrated circuit chips using circuit simulation tools such as SPICE to predict the circuit performance.

The 1-D x-solution is known as the input **voltage equation**, which relates the input gate-terminal voltage to the electric potential at the semiconductor surface or the gate-insulator/semiconductor interface, which is known as the **surface potential**. (Literature has used Ψ_S, ψ_S, ψ_s, Φ_S, φ_s, u_s. We follow **strictly** the IEEE Standards of Symbols and Notations for Circuits and Devices [8-10]: $V_S(x=0,y,z)$ and $U_S=qV_S/kT$ for DC steady-state, respectively not-normalized and normalized to the thermal voltage, kT/q, while $u_S(x,y,z,t)$ for large signal transient, $u_s(x,y,z,t)$ for small signal transient, and $U_s(x,y,z,\omega)$ for sinusoidal steady-state. A voltage between two terminals G and B is represented by V_{GB} with B as the reference terminal.) The major task for the compact model developers has been to find a fast algorithm to invert the implicit dependence, $V_{GB}(U_S)$, to an explicit dependence, $U_S(V_{GB})$.

The 1-D y-solution is known as the output terminal **current equation**. It relates the output current passing into (or out of) the output terminal (drain or source terminal) to the surface potential, with the voltage applied between the output terminal and a reference terminal as the parameter, $I_D(U_S,V_{SB},V_{DB})$. Thus, the surface potential is the **independent variable** of this system of characterization equations of the MOST. It couples the current and voltage equations.

This paper describes an early history of MOS transistor compact modeling, including the **threshold voltage model** used in the first version of SPICE to the two latest advances under development, the **inversion charge model** and the **surface potential model**.

In **threshold voltage modeling**, a linear approximation was made between the surface potential and the applied drain (or channel) and gate voltages. This eliminates the surface potential and relates the

input gate voltage to the output drain current, giving a simple current-voltage equation which is parabolic in its simplest form and it was used in the initial version of the circuit simulator SPICE [11-15]. It defines a **gate threshold voltage** below which current ceases and a **drain saturation voltage** above which the drain current remains constant, independent of the "excess" drain voltage and depends only on the gate voltage. Below the gate threshold voltage, there is still a drain or channel current, not zero as the simplest threshold voltage model would predict. This subthreshold channel current is dominated by diffusion of the minority carriers (electrons on p-substrate) in the surface channel. The magnitude of this diffusion channel current depends on the diffusion barrier height at the Source/Channel boundary. The source p/n junction potential barrier height is lowered by the applied gate voltage which increases the channel current exponentially with the gate/source applied voltage, V_{GS}. This subthreshold current-voltage characteristic is similar to that of the BJT where the emitter p/n junction barrier height is lowered by the forward DC voltage applied to the emitter p/n junction. The main task of threshold voltage modeling has been to find **connection formulas** that will join smoothly the DCIV equations of the subthreshold and above-threshold ranges, defined and separated by the **threshold** and **subthreshold** gate voltages, including their slopes and second and higher derivatives, in order to accurately predict the digital switching waveform and delay, and the analog small-signal distortion and noise, near these joining voltages.

On **charge modeling**, the voltage equation [voltage-versus-surface-potential, $V_{JK}(U_S)$, where JK are internal or external-terminal node labels] is transformed into an equation of node-charge versus node-voltage drop, $q_j=C_{jk}v_{jk}$, where is the node charge and C_{jk} is the capacitance coefficient. Similarly, the current equation (current-versus-surface-potential and node voltages) can also be transformed into a current-versus-node-charge equation. Thus, the node charges are the independent variables. This is the approach used in the **inversion charge model** in both strongly and weakly inverted surface channel (electron channel on p-type semiconductor surface covered by an insulator and a gate-conductor electrode). Its node charge is the mobile charge in the inverted electron channel (inverted from p-type bulk with holes) at the SiO_2/Si interface, while still trying to include the bulk charge by linearization of its dependence on the voltage drop. The mobile charge then provides the terminal current, $i_J=q_J/T_j=C_{jk}v_K/T_j$, and also the lump-form (so-called quasi-static, that is, diffusion and propagation delays from distributed transmission-lines are not included) terminal and equivalent-circuit-element immittances, i.e. the differential or charge-control conductances $g_{jk}=\partial i_J/\partial v_K$ and capacitances, $C_{jk}=\partial q_J/\partial v_K$, which can be used to design switching and small-signal sinusoidal circuits for digital and analog applications. Diffusion, trapping, and dielectric relaxation delays are not represented by the charge-control immittances. It is claimed that the charge control model is more accurate and sufficient fast compared with the threshold voltage and the surface potential models to give the small-signal model for analog applications. However, charges cannot be measured with available instruments, so the charge control model is still calibrated by measurements of DC current-voltage characteristics and small-signal immittance versus DC voltage and signal frequency. Calibration measurements are usually made on many test transistors of different designs to provide the basic design data that can be interpolated for circuit simulations to give the optimized performance .

On **surface potential modeling**, surface potential and electric potentials are internal device and material parameter that cannot be measured as easily as the terminal currents and voltages. For example, floating or contact potential was measured by high impedance (10^{10}ohm) voltmeter on large area ($>\sim 1cm^2$) Ge surfaces with a parallel platinum electrode 1mm away, vibrating at 0.1mm amplitude in the 1949-1952 Brattain-Bardeen experiment. For the micron and submicron modern-transistor sizes, the measurement of surface potential at each region of the transistor would require extremely high input-resistance voltmeter with very small probe area, hundreds times smaller than the area to be measured. There are no charge measurement instruments to measure the charges at the terminals and in the internal

regions of the transistor. Regardless of lacking measurability, the surface potential approach to compact modeling has been in hibernation for ~ 30 years because the iterative computations for finding the theoretical surface potential value at a given input or gate terminal voltage would take too much computer time in the past when the CPU speed was slow. So, a linear approximation between surface potential and terminal voltages has been used since the first generation (~early 1970's) of computer-aided-design of integrated circuits using SPICE [11-15]. In this linear model, the surface potential is assumed to be proportional to the input voltage or charge, so the surface potential is replaced by the input voltage or charge. But, today, even desktop personal computer is sufficiently fast to quickly give numerical value of the surface potential from the implicit, exact, and 1-dimentional, nonlinear relationship between surface potential and applied gate voltage, $V_{GB}(U_S)$. Therefore, there are revivals of the surface potential modeling. This also avoid known and unknown errors arisen from the analytical approximations used to approximately invert $V_{GB}(U_S)$ to give $U_S(V_{GB})$, which is then substituted into $I_D(U_S,V_{GB},V_{DB},V_{SB})$ to give the MOST DCIV equations.

II. Analysis and History

In this section, we shall summarize the basic differential equations whose solutions characterize the electrical characteristics of MOSTs. The general theory and solution methods and results are described. It is then used to obtain the MOSC voltage equation and the MOSR current equation which are then combined to give the current-voltage characteristics of the transistor. The history is described on the use of these equations by the transistor compact model developers to characterize MOSC and MOST.

20 Theoretical Foundations

The response of an MOS transistor to an applied force (electrical, mechanical, thermal, optical, and particle) can be described mathematically at two levels based on its physical size: **macroscopic** level via ensemble average for a large transistor; **microscopic** level via time average for a small transistor.

For a large (**macroscopic**) transistor in all its geometric features (thickness, length, and width), there are enough number (billions, trillions or more) of atoms and electrons in each of its features so that ensemble average in a volume element of dxdydz is meaningful because the statistical fluctuations from the ensemble average over many particles in the volume element dxdydz are small. The ensemble average of a device or transistor is an average of measurements of an electrical characteristic (current, voltage, admittance and impedance; or the 1940 Bode general terms adpedance and the more popular immittance used in textbooks) taken in a small time interval at a specified point in time. Then, the mathematical description of the DC and AC steady-states and time-dependence (switching including delay, and small-signal frequency or analog including harmonic distortion and noise) can all be obtained accurately by solving the macroscopic transport equations either analytically via the governing **differential equations** or numerically via the **difference equations.** The latter, however, is too time-consuming and inaccurate for compact modeling.

For a geometrically very small (**microscopic**) transistor containing a small number of atoms (tens of thousands or less in a few atomic layers and small areas), fluctuations are large. The transistor characteristics are frequently obtained by time average, namely, shooting an electron into the transistor and tracing the electron trajectory coming out, then averaging many 1-electron trajectories which requires a long time. This is known as the Monte Carlo method, which is again too time-consuming and inaccurate for compact modeling.

The starting macroscopic equations are the six time-dependent **Shockley Equations** that include generation-recombination-trapping-tunneling (**GRTT**). {Eqs.(350.1)-(350.6), p.268 of [8].}

General Time-Dependent (Dynamic) Shockley Equations

$q(\partial n/\partial t) = + \nabla\cdot\mathbf{j}_N + q(g_N - r_N)$ Continuity Equation of Electrons (20.1)

$q(\partial p/\partial t) = - \nabla\cdot\mathbf{j}_P + q(g_P - r_P)$ Continuity Equation of Holes (20.2)

$\mathbf{j}_N = + q\mu_n n\mathbf{E} + qD_n\nabla n$ Current Density of Electrons (20.3)

$\mathbf{j}_P = + q\mu_p p\mathbf{E} + qD_p\nabla p$ Current Density of Holes (20.4)

$\nabla\cdot\varepsilon\mathbf{E} = \rho$ Poisson Equation (20.5)

$q(\partial n_T/\partial t) = + q(g_P - r_P) - q(g_N - r_N)$ **GRTT** Equation (20.6)

Consider the DC steady-state or static model for a semiconductor with some generation-recombination-trapping-tunneling (GRTT), $\partial n_T(\mathbf{r},t)/\partial t = 0$, $n_T(\mathbf{r},t) = N_T(r) \neq 0$; $0 \neq g_N(\mathbf{r},t) = G_N(r) \neq r_N(r,t) = R_N(r) \neq 0$; $0 \neq g_P(\mathbf{r},t) = G_P(r) \neq r_P(r,t) = R_P(r) \neq 0$, and with a fixed or immobile singly-charged acceptor N_{AA^-} and donor N_{DD^+} impurity concentrations, and an intrinsic carrier concentration n_i. Use the **exponential** or **Boltzmann "representation"** [16,17] for the mobile electron and hole concentrations defined by $N = n_i\exp(U-U_N)$ and $P = n_i\exp(U_P-U)$. Here the **quasi-Fermi** or **electrochemical potentials for electrons and holes** and the electric potential are normalized to the thermal voltage kT/q: $U_P=qV_P/kT$, $U_N=qV_N/kT$ and $U=qV/kT$. Use also the definition of the electric field vector $\mathbf{E}=-\mathbf{grad}V\equiv-\boldsymbol{\nabla}V=-(kT/q)\mathbf{grad}U$. Then, the general time-dependent Shockley Equations (20.1) to (20.6) are reduced to the simpler DC-Steady-State (static) working form listed below which have been employed as the starting point for the differential-equation solution of the DC current-voltage (DCIV) characteristics of semiconductor devices such as the MOS transistors. These five DC steady-state and finite-GRTT Shockley Equations in a semiconductor region with constant dielectric constant ε are [8,9]

General DC Steady-State (Static) Shockley Semiconductor Equations

$0 = + \nabla\cdot\mathbf{J}_N = - qD_N\nabla\cdot(N\nabla U_N)$ (20.7)

$0 = - \nabla\cdot\mathbf{J}_P = + qD_P\nabla\cdot(P\nabla U_P)$ (20.8)

$\mathbf{J}_N = + q\mu_n N\mathbf{E} + qD_n\nabla N = - q\mu_N N\nabla V_N = - qD_N N\nabla U_N$ (20.9)

$\mathbf{J}_P = + q\mu_p P\mathbf{E} + qD_p\nabla P = - q\mu_P P\nabla V_P = - qD_P P\nabla U_P$ (20.10)

$\nabla\cdot\varepsilon\mathbf{E} = - \varepsilon\nabla^2 V = - (\varepsilon kT/q)\nabla^2 U = \rho$ (20.11)

$\rho = q(P - N - P_{AA}\ominus + N_{DD}\oplus - N_T)$ Semiconductor Space-Charge Density (20.12)

$= qn_i[\exp(U_P-U) - \exp(U-U_N) - (P_{AA^-}-N_{DD^+}-N_T)/n_i]$ (20.13)

The Poisson Equation in the normalized forms are given by

$-\varepsilon(kT/q^2n_i)\nabla^2 U = \exp(U_P-U) - \exp(U-U_N) - P_{IM}/n_i - N_T/n_i$ (20.14)

$-2L_{Di}^2\nabla^2 U = \exp(U_P-U) - \exp(U-U_N) - P_{IM}/n_i - N_T/n_i$ (20.15)

The net ionized acceptor-like-charge or negative-charge **bulk impurity concentration** is defined by $P_{IM} \triangleq P_{AA^-} - N_{DD^+} > 0$ when applied to a p-type sample and L_{Di} is the intrinsic Debye length or the Debye length of an intrinsic semiconductor defined by $L_{Di}=(\varepsilon kT/2q^2n_i)^{1/2}$. Frequently, the extrinsic Debye length is used which for a p-type sample is $L_D=(\varepsilon kT/q^2P_{IM})^{1/2}$. Impurity deionization is usually neglected so that $P_{AA}\ominus = P_{AA} - P_A\otimes \approx P_{AA}$ and $N_{DD}\oplus = N_{DD} - N_D\otimes \approx N_{DD}$ where $P_A\otimes$ and $N_D\otimes$ are the concentrations of the

holes and electrons trapped at the negatively charged acceptor and positively charged donor ions respectively. Then $P_{IM} \approx P_{AA} - N_{DD}$.

The steady-state generation-recombination-trapping-tunneling rate is given by

$$0 = q(\partial n_T/\partial t) = + q(G_P - R_P) - q(G_N - R_N) \quad (20.16)$$

$$qG_{SS} = - qR_{SS} = + q(G_P - R_P) = + q(G_N - R_N) \text{ Steady-State GRTT Rate} \quad (20.17)$$

For a one-energy-level Shockley-Read-Hall center with electron and hole capture and emission rate-coefficients (under arbitrary nonequilibrium conditions) c_n and c_p (cm^3/s) and e_n and e_p (s^{-1}), and bulk concentration of N_{TT} (cm^{-3}) or interface concentration of N_{IT}(cm^{-2}), its steady-state GRTT rate is

$$R_{SS} = - G_{SS} = (c_n N c_p P - e_n e_p) N_{TT}/(c_n N + e_n + c_p P + e_p) \quad (20.18)$$

$$N_T/N_{TT} = (c_n N + e_p)/(c_n N + e_n + c_p P + e_p) \quad (20.19)$$

21 The MOSC Theory (The Voltage Equation)

Under the direction of Shockley, the mathematics underlying the analyses of the MOS Capacitance (MOSC) and MOS transistor began around 1950 [16,17]. The results were reported by Walter L. Brown in 1953 [18] in an analysis of the electrical measurements of the Ge n/p/n grown-junction bipolar transistor with the base layer surface exposed to the ambient. This **1953-Brown-Shockley** analysis contained the first description of several device-physics conceptions applied to the surface channel conduction (the electron channel on the p-Base of the n/p/n), most of which are still used today for semiconductor devices, including MOS transistor compact modeling. For examples: (1) the **quasi-Fermi potentials** to describe nonequilibrium from applied voltages to a p/n junction introduced by Shockley [16,17], (2) the two-layer surface space-charge layer model: a surface space-charge layer that is depleted of electrons and holes, the **depletion layer**, and a surface space-charge layer which has the conductivity type opposite to that of the bulk, the **inversion layer**, (3) the depletion-layer capacitance theory [16,17] and measurements, (4) the negative-number problem of the square of the surface electric field because the minority carrier was neglected in the electrical neutrality boundary condition, (5) the surface **inversion** or **electron channel** on the p-Base that connects the n-emitter and n-collector regions, (6) the electrical thickness of the electron surface channel, (7) the physical and electrical **pinch-off** of the surface channel, following that first used by Shockley for the p/n-junction-gate field-effect transistor [19,20] including (8) the local **pinch-off voltage equation**, and (9) the near equality of the pinch-off voltage and the **floating potential** of the emitter with the n-collector/p-base junction reverse biased, (10) the 1-D E(x) **surface energy-band diagram** and (11) the 2-D E(x,y) **electron potential energy surface** near and include the surface, and (12) the two solutions of the Poisson Equation, one by integration in space, and one by integration via quadrature in electric potential.

Two years later in 1955, Garrett and Brattain [21] completed and reported a comprehensive mathematical analysis of the MOSC or the Voltage Equation for the 1-D geometry. The 1955-Garrett analyses greatly extended and expanded the simple and intuitive 1953-Brown-Shockley solutions [18] including also the surface-channel conductance or MOST Current Equation first analyzed in 1953-Brown-Shockley [18]. The 1955-Garrett [21] analyses included nonequilibrium conditions represented by quasi-Fermi potentials, from exposure of germanium and later silicon to controlled ambient and light, and later, from electric field via voltage applied to a gate-electrode, first without and later with an insulator, finally

with a thermally grown oxide on silicon. The 1955-Garrett analyses extended the 1953-Brown-Shockley 2-layer/2-range semiconductor space-charge model to three layers and three ranges: three physical layers or three surface-potential ranges for the three energy-band-bending ranges. They were named by them as the **accumulation**, the **exhaustion** then **depletion** layer-and-range which he called the **parabolic-potential** range {See Garrett's Equations (12)-(14) on p.379 of [21].}, and the **inversion** layer-channel-and-range, {See Eq.(15) of [21].}

To further review the historical developments of the MOSC and its Voltage Equation or the x-solution, we shall provide a mathematical derivation of the formulas using the recent notation that would help to connect the current authors' solutions to the historical firsts. Three solutions, listed as I, II and III, can be obtained from integrating the **3-D, 2-D or 1-D Poisson Equation** (20.11) and (20.15). The two single-integrations are the volume variable $d\mho = dxdydz$ and the potential variable $dU = \nabla U \bullet d\mathbf{r}$. The one double-integration is along the volume $d\mathbf{x_r} \bullet d\mathbf{x_t}$ or length variable dx.

I. The volume integration gives the **Gauss Law** which states that the integration of the **Displacement Vector D** = $\varepsilon\mathbf{E}$ over a closed surface is equal to the charge enclosed.

$$\oint \rho d\mho = \oint \nabla \bullet \varepsilon \mathbf{E} d\mho = \oint \varepsilon \mathbf{E} \bullet d\mathbf{S} = \oint \mathbf{D} \bullet d\mathbf{S} \qquad \text{Gauss Law 3-D} \qquad (21.1)$$

$$\int [\partial(\varepsilon E_x)/\partial x]\,dx = \varepsilon_2 E_2(x_2) - \varepsilon_1 E_1(x_1)$$

$$= D_2(x_2) - D_1(x_1) = \int \rho dx = Q_{21}(x_{21}) \qquad \text{Gauss Law 1-D} \qquad (21.2)$$

Here $Q_{21}(x_{21})$ is the charge density (per unit area dydz) between the planes $x=x_1$ and $x=x_2$.

II. The above can be integrated again to give a relationship between the potential drop and the moment of the enclosed space-charge. For the 1-D case, integrating by part along x twice from $\xi=x_1$ to x_2, we get {See Equation (412.14) on p.140 in [9]} we get

$$v(x_2,t) - v(x_1,t) + (x_2-x_1)E_x(x_1,t) = -\int (x_2-\xi)\rho(\xi,t)d\xi/\varepsilon \qquad \text{General 1-D} \qquad (21.3G)$$

$$V(x_2) - V(x_1) + (x_2-x_1)E_x(x_1) = -\int (x_2-\xi)\rho(\xi)d\xi/\varepsilon \qquad \text{DC 1-D} \qquad (21.3D)$$

$$U(x_2) - U(x_1) - (x_2-x_1)\partial U(x_1)/\partial x = -\int (x_2-\xi)\rho(\xi)d\xi\,(kT/\varepsilon q) \qquad \text{DC 1-DN} \qquad (21.3DN)$$

III. The integration in electric potential, $\int(\ )dU = \int(\ )(dV/kT)$ gives the relationship between the **electric field and potential** {See Equation (412.4A) on p.138 of [9].}. Integrating by quadrature along the line U(x,y,z) from a smaller surface $U(x,y,z)=U(x_1,y_1,z_1)=U_1$ to a larger surface $U(x_2.y_2,z_2)=U_2 > U_1$, gives the difference of the square of the Displacement Vectors normal to the two surfaces. For a region of spatially independent dielectric constant, $\varepsilon \neq f(r)$, and denoting the radial and transverse components respectively by subscripts r and t, then

$$\int \rho(U)\,dU = \int \nabla \bullet \varepsilon \mathbf{E} dU = -(\varepsilon kT/2q) \int 2\nabla^2 U dU$$

$$= -(\varepsilon kT/2q)\left[(\nabla_r U_2)^2 - (\nabla_r U_1)^2 + (\nabla_t U_2)^2 - (\nabla_t U_1)^2\right] \qquad \text{3-D} \qquad (21.4)$$

$$= -(\varepsilon kT/2q)\left[(\nabla_x U_2)^2 - (\nabla_x U_1)^2\right] \qquad \text{1-D} \qquad (21.5)$$

$$\int \rho(V)\,dV = -(\varepsilon/2)\left[(\nabla_x V_2)^2 - (\nabla_x V_1)^2\right] = -(\varepsilon/2)\left[(E_{x2})^2 - (E_{x1})^2\right] \qquad \text{1-D} \qquad (21.6)$$

Application to MOSC (M/n+G/SiO_2/p-Si/M)

The general 1-D results, (21.2), (21.3DN) and (21.5) or (21.6), can now be applied to the MOSC between the gate and the body metal terminals of a polycrystalline-gate nMOST with a p-Si-basewell or p-Si-body, and the layer structure M/n+Gate/SiO_2/p-Si/M. The metal terminal is defined as one with zero space-charge, zero electric field, and a spatially constant electric potential, with the potential difference between two metal terminals equal to the voltage applied between the two metal terminals. The space-charge densities in the MOSC layers are as follows (ρ in Coulomb/cm^{-3}).

$$\rho(x,y) = 0 \qquad \text{Metal gate terminal } x<-X_M \qquad (21.7A)$$

$$\rho(x,y) = qN_M\delta(-X_M) \qquad \text{Metal/n+Si-Gate interface at } X=-X_M \qquad (21.7B)$$

$$\rho(x,y) = q(P - N + N_{GG}) \qquad \text{n+Si-Gate layer in } -X_{M+}<x<-X_{M+}+X_{n+G-} \qquad (21.7C)$$

$$\rho(x,y) = qN_{ITG}\delta(-X_{n+G}) \qquad \text{n+Gate/Gate-Oxide interface } -X_{n+G-}<x<-X_{n+G+} \qquad (21.7D)$$

$$\rho(x,y) = qN_{OT}(x,y) \qquad \text{Gate-Oxide layer } -X_{n+G-}<x<-0_- \qquad (21.7E)$$

$$\rho(x,y) = qN_{IT}\delta(0) \qquad \text{Gate-oxide/p-Base interface } -0_-<x<+0_+ \qquad (21.7F)$$

$$\rho(x,y) = q(P - N - P_{AA} + N_{DD}) \qquad \text{p-Basewell layer or p-Body } +0_+<x<+x_\infty \qquad (21.7G)$$

$$\rho(x,y) = 0 \qquad \text{p-Body/Metal interface (or } \rho_{SOI}) \quad x=+X_\infty \qquad (21.7H)$$

$$\rho(x,y) = 0 \qquad \text{metal body terminal } x>+X_\infty \qquad (21.7I)$$

Using these space-charge densities in (21.2), (21.3DN) and (21.5) we obtained the following 1-D MOSC voltage equations for a spatially constant net impurity concentration in the p-Si base-body, $P_{IM} = P_{AA} - N_{DD} \neq f(x)$ and for the boundary conditions of $\rho(X_\infty,y)=0$, $E_X(X_\infty,y) = -\partial V(x=X_\infty,y)/\partial x=0$, and $U(X_\infty,y)=U_\infty(y)=qV(X_\infty,y)/kT$. We also assume x-independent quasi-Fermi potentials $U_P(x,y)=U_P(y)$ and $U_N(x,y)=U_N(y)$ in order to get (21.12) to (21.15).

$$0 = Q_M + Q_{n+G} + Q_{ITG} + Q_{OT} + Q_{IT} + Q_S \qquad (21.7)$$

$$\triangleq C_o(V_{GB} - V_{FB} - V_S) + Q_S \qquad \text{(Definition of flatband voltage } V_{FB}) \qquad (21.8)$$

$$= C_o(V_{GB} - V_{FB} - V_S) - \varepsilon_S E_S \qquad (21.9)$$

$$Q_S = D_{X\infty} - D_{X0+} = +\varepsilon_S E_{X\infty} - \varepsilon_S E_{X0+} = +0 - \varepsilon_S E_{X0}$$

$$\equiv -\varepsilon_S E_S = -\varepsilon_S(-\partial U_S/\partial x) = +\varepsilon_S(\partial U_S/\partial x) \qquad (21.10)$$

$$Q_S = \int\rho(x)\,dx = q\int(P - N - P_{IM})\,dx \triangleq Q_P + Q_N + Q_{PIM} \qquad (21.11)$$

$$E^2(x,y) = (2kT/\varepsilon_S)\{[P - P_\infty(1 - U + U_\infty)] + [N - N_\infty(1 + U - U_\infty)]\} \qquad (21.12)$$

$$= (2kT/\varepsilon_S)[P - P_\infty + N - N_\infty + (P_\infty - N_\infty)(U - U_\infty)] \qquad (21.13)$$

$$E_X^2(x,y) = (kT/qL_{Di})^2F_X^2 \qquad (21.14)$$

$$F_X^2 = +[\exp(-U) + (+U - 1)\exp(-U_P + U_{P\infty})]\exp(+U_{P0}) \qquad (21.15)$$

$$+ \left[\exp(+U)+(-U-1)\exp(+U_N - U_{N\infty})\right]\exp(-U_{N0}) \qquad (21.15)$$

A better solution for physics-based approximation of the x-dependence of the quasi-Fermi potentials, in anticipation of solutions for the drain/base p/n junction space-charge region is to express the x-component of the electric field at a location x, with reference to that at the SiO_2/Si interface x=0 rather than x=∞. So, instead of (21.14) and (21.15), we have

$$E_S^2 = E_X^2 + (kT/qL_{Di})^2 F_{SX}^2 \qquad (21.14A)$$

$$F_{SX}^2 = + \left[\exp(+U_{P0}-U_S) - \exp(+U_P-U) + (+U_S - U)\exp(+U_{P\infty} - U_\infty)\right]$$
$$+ \left[\exp(+U_S-U_{N0}) - \exp(+U-U_N) - (+U_S - U)\exp(+U_\infty - U_{N\infty})\right] \qquad (21.15A)$$

Using the above with (21.9), we obtained the following **Voltage Equation**

$$U_{GB} - U_{FB} - U_S = U_{OX} = 2(U_{II})^{1/2} \times F_{SI}(U_S, U_{P0}, U_{P\infty}, U_{N0}, U_{N\infty}) \qquad (21.16)$$

where $(F_{SI})^2 = + \left[\exp(-U_S)+(+U_S - 1)\exp(-U_{P0} + U_{P\infty})\right]\exp(+U_{P0})$

$$+ \left[\exp(+U_S)+(-U_S - 1)\exp(+U_{N0} - U_{N\infty})\right]\exp(-U_{N0}) \qquad (21.17)$$

and $U_{II}=qV_{II}/kT=(q/kT)(\varepsilon_S qn_i/2C_O^2)$, $C_O=\varepsilon_O/X_O$, $U_{PX}=U_P(x,y)$, $U_{NX}=U_N(x,y)$, x=0=the SiO_2/Si interface, and $x=X_\infty\rightarrow\infty$=the remote boundary.

Historical Uses of the Voltage Equation - General

The terms in the three groups of the solution for $E^2(x,y)$ in (21.12) and (21.13) immediately show their relative importance, either in three internal space layers or three external applied gate voltage ranges. This grouping was recognized as a good teaching tool in 1991-Sah {See [8] equations (412.6) to (412.9) on page 348.} and 1993-Sah {See [9] (411.13) to (411.14) on page 128.}. The term $P-P_\infty$ dominates in the p-Si basewell-body where hole concentration is high, or in the negative range of the applied gate voltage which attracts holes to the p-Si surface or the SiO_2/p-Si interface. This is known as the majority-carrier accummulation for p-type semiconductor with $P_{IM}>0$. The term $N-N_\infty$ dominates in the p-Si basewell-body where electron concentration is high, or in the positive range of the applied gate voltage which attracts electrons to the p-Si surface or the SiO_2/p-Si interface, causing the surface conductivity on the p-Si to invert to n-type, i.e. $N_S \triangleq N(x=0_+,y) \geq P_S \triangleq P(x=0_+,y)$. It defines an electrical thickness of the electron channel or the inversion layer at the SiO_2/Si interface on a p-Semiconductor. {See another definition of electrical thickness based on charge-control differential capacitance given Equations (412.4) to (412.6A) and (412.10C1) to (412.10E) on pages 138 to 139 in 1993-Sah [9].} The term $(P_\infty-N_\infty)(U-U_\infty)$ dominates near flatband, $U-U_\infty\rightarrow 0$, or the voltage applied between the gate and base terminals is near the flatband voltage, $V_{GB} \cong V_{FB}$. This A1G {At One Glance. See page xii of the Preface of 1991-Sah [8].} three-layer three-potential-range model was first described in 1953-Brown-Shockley [18] which considered only the two-layer, two-surface-potential weak and strong inversion ranges, without the accummulation range. Details were given in 1955-Garrett [21] to compute the surface conductance versus surface potential for analyses of experimental surface field-effect measurements on Germanium and Silicon surfaces. The 1955-Garrett analyses were used to explain and analyze experimental results obtained by the previous [6,18,22,25a] and subsequent [18a-18c,21-25a] investigators during the 1953-

1960 period on the DC conductance, small-signal-capacitance, conductance relaxation, and pulsed field-effect on bare but clean (in high and mercury-pumped 'ultra-high'-vacuum quartz-tube capsule) Ge and Si surfaces, and later on gated bare Ge and gated oxidized Si surfaces. These investigations on bare Ge and Si surfaces ceased when stable thermal oxide was grown on Si in 1959 by Atalla, Tannenbaum and Scheibner at Bell Telephone Laboratories [7].

Historical Uses of the Voltage Equations – Applications to MOS Capacitors

The equilibrium solution of the 1-D Poisson Equation was derived for the silicon MOS capacitor during the late 1950's due to its voltage-controllable capacitance, known as the varactors. In addition to the Gate-Voltage versus Surface-Potential equation just discussed and given by (21.16) and (21.17), the Gate-Capacitance versus Surface-Potential equation was also derived via charge-control, giving the quasi-static differential capacitance, $C_{gb}=dQ_G/dV_{GB}$. John L. Moll of Stanford, who just left Bell Telephone Laboratories, proposed the MOS Varactor at the August 1959 IRE Wescon (Western Convention and Show) [23]. W. G. Pfann and C. G. B. Garrett of Bell Telephone Laboratories provided the first theoretical capacitance-voltage curves in a Letter to the Editor in the 1959 IRE Proceedings [24]. D. R. Frankl of General Telephone and Electronics made extensive calculations in 1960 for Ge, Si as well as InSb [25] who used the simple 1955 Kingston-Neustadter notation [26] u=qV/kT which greatly helped this author to learn the MOSC theory during 1959-1961. The equilibrium theory of the MOS capacitor was extended to the nonequilibrium by Lindner [27] of Bell Telephone Laboratories in 1961, using the three-layer approximation of the surface space-charge layer of the 1955-Garrett analyses [21], and the **charge-control theory** to compute the MOSC's capacitance-vs-DC-voltage curves under three ranges of minority carrier generation-recombination rates relative to the measurement time. The three ranges were: **(1)** The usual equilibrium or **low-frequency MOSCV curves (C_{lf} $-V_{GB}$)** which is the oxide capacitance $C_{ox}=\varepsilon_{ox}/X_{ox}$ in series with the semiconductor space-charge capacitance $C_S=-\partial Q_S/\partial V_S$ assuming no interface traps which would give an interface trapping capacitance $C_{IT}=-\partial Q_{IT}/\partial V_S$ in parallel with C_S. Here $Q_S=\int \rho dx$ (x=0 to ∞). The two nonequilibrium MOSCV curves obtained by Lindner [27] are as follows. **(2)** The **depletion MOSCV curves (C_{dep} $-V_{GB}$)** in which the oxide is so leaky that both majority (hole) and minority (electron) carriers are not generated or transported fast enough to accumulate in the semiconductor surface space charge layer under the SiO_2/Si interface. Thus, the mobile carriers are completely depleted in the semiconductor surface space-charge layer or P~0 and N~0 in $0 \le x \le X_d$. Only the majority carriers can be supplied and withdrawn fast enough to the edge of the depletion layer, $x=X_d$, during the measurement time. The electrical thickness of the depletion layer X_d can then be defined by the parallel-plate surface space-charge-layer capacitance $C_S = C_{dep} = -\partial Q_{dep}/\partial V_S \triangleq \varepsilon_{si}/X_d$. The depleted charge density is given by $Q_{dep} \triangleq \int q(P+N_\infty-P_\infty)dx$ (x=0 to X_d) and dV/dx is obtained with P=N=0. {See $X_{S\text{-}dep}$ given by Equation (412.10D) on page 139 of 1993-Sah [9].} **(3)** The **high-frequency MOSCV curves (C_{hf} – V_{GB})** in which the oxide is insulating so that DC steady-state concentration of majority and minority carriers can be accumulated at the interface, but the measurement time is so short (in the charge-control mode) or the sinusoidal small-signal measurement frequency is so high (a.c. or sinusoidal steady-state mode) that the minority carriers (electrons in p-Si basewell-body) cannot be generated and recombined fast enough in the measurement time interval of the switched charge-change or the sinusoidal frequency of small amplitude signal variation, so $Q_N \triangleq \int q(N-N_\infty)dx$ (x=0 to ∞) or N must be kept constant during the variation of ∂V_S to measure the capacitance from $C_S = C_P = -\partial Q_P/\partial V_S$ with $Q_P \triangleq \int q(P-P_\infty)dx$ =

$\int q(P-P_\infty)dV/(dV/dx)$ (x=0 to ∞ and V=V_S to 0.). The dV/dx in the integration is the DC steady-state value. This integration and its differentiation turned out to be rather nontrivial and in fact, extremely difficult to mathematically carry out, then numerically compute. So Gove and Sah [28, 29] tried a different way to compute the high-frequency MOSCV curves. A systematic equation was developed by Sah [29] which is defined by $C_S=\varepsilon_{si}/X_S$ where X_S is the space-charge layer thickness. This Sah **three-layer charge-sheet model** [29] was used to give $X_S=X_1+X_2+X_3$. X_1 is the inversion layer thickness, X_3 is the depletion layer thickness, and X_2 is the thickness of the transition layer between the inversion layer and depletion layer. The depletion layer thickness, X_3, also includes the thickness of the transition layer between depletion layer and the quasi-neutral p-Si basewell-body region.

Historical Uses of the Voltage Equations – Interface Traps from MOSCV

The capacitance-voltage characteristics were also strongly affected by the surface states located at the oxide/silicon interface. {See a survey of the history in [5].} These surface states or interface traps, deduced earlier by the distortion of the surface-conductance versus surface potential curves in the surface field-effect experiments for bare or not-thermally oxidized Ge and Si surfaces [18,18a,21,22,25a,26,27], were now obtained by the distortion of the capacitance-voltage characteristics compared with the theory on oxidized silicon surfaces. The earliest and most extensive study was made by Lewis M. Terman in his 1960 PhD thesis at Stanford [30] under the direction of John Moll at Stanford [23]. Terman developed two methods to measure the interface traps, (1) the frequency dependence of the MOSCV curves and (2) the distortion of the high-frequency CV due to slow trapping. The second is known as the **Terman Method** and it has been widely used in academic research and in manufacturing due to the ease of experimental measurement.

Historical Uses of the Voltage Equations – Negative E_S^2 Near Flatband

The compact model developments have focused on two approaches to give the fast and accurate I_D versus V_{GB} characteristics (also the immittance y_{jk} versus V_{GB}) by solving the two simultaneous device-physics-based implicit equations having the surface potential, U_S, as the independent variable, $I_D(U_S)$ and $y_{ik}(U_S)$ versus $V_{GB}(U_S,V_{DB},V_{SB})$ which are 2-D and 1-D integrals. These two approaches are: **(i)** analytical approximation to give fast and accurate explicit equations of $I_D(V_{GB},V_{DB},V_{SB})$ and $y_{ik}(V_{GB},V_{DB},V_{SB})$ equations, such as the **threshold-voltage** and **inversion charge models**, and **(ii)** design of fast numerical algorithm to compute $U_S(V_{GB,}V_{NP})$ for nMOST where $V_{SB}\le V_{NP}(Y_S\le y\le Y_L)\le V_{DB}$ or $U_S(V_{GB},V_{PN})$ for pMOST where $V_{SB}\le -V_{PN}(Y_S\le y\le Y_L)\le V_{DB}$ to calculate the single and double integrals, known as the **surface potential model**.

In search of fast numerical algorithms in the second approach, (ii) above, it was noticed that electric-field-square function, (21.17), was giving a negative value in a smaller range of U_S near flatband, $U_S=0$. This was infrequently noticed, and only appeared when the dU_S step-size in the computation is very small and near $U_S=0$. This violation of physical reality was communicated via email to Sah by Professor Xing Zhou of Nanyang Technological University in Singapore on November 30, 2004 [31] who made references to the mathematical algorithms designed by Professor Gennady Gildenblat of Pennsylvania State University to circumvent this error [32a, 32b] and analyzed by Dr. Colin McAndrew of Motorola-Freescale-Semiconductor Inc. to trace the error source empirically [32c]. [I shall call this the

2002-Gildenblat-McAndrew Correction.] After confirming that this difficulty also occurred in computations by Sah's former and current students and current postdoc Bin Jie, the cause of this error was identified within an hour while in San Diego during holidays by three means: aside from physical unreality, two mathematical tests were made, a Taylor series expansion of the old $F_S^2(U_S,U_P,U_N)$ near $U_S=0$, and a Taylor series expansion of the new and self-consistent $F_S^2(U_S,U_P,U_N)$ to show the latter is positive-real at all orders of U_S near $U_S=0$. One of the two sources of error previously encountered by programmers was the incorrect or not-self-consistent remote boundary condition used for the quasi-Fermi-potentials in the charge-neutrality condition that gave the old and erroneous $F_S^2(U_S,U_P,U_N)$. The second was from neglecting minority carriers in computing the quasi-Fermi levels at the far boundary, commonly made by all engineers. These are demonstrated and proven in the following derivations.

The charge-neutrality condition at the remote boundary is given by the following algebra which give both the quasi-Fermi potentials and the carrier concentrations at $x=X_S \rightarrow \infty$. This was first used in 1957-Sah-Noyce-Shockley [33] as the high-level boundary condition in their investigation of the recombination current in the space-charge layer of p/n junctions. It was learned by Sah from Robert N. Hall's lecture on power rectifiers while attending, for doctoral credit, the historical first and only IRE-Device-Research-Conference Summer School in 1953 at the University of Illinois in Urbana-Champaign. This boundary condition has always been approximated by the transistor-device theorists and textbook authors [11,12,13,14] using the majority carrier solution, $PN=n_i^2\exp(U_{PN})$ (correct) and $P=P_{IM}$ (wrong) because $P-N=P_{IM}$, that is, the minority carriers cannot be neglected in the charge neutrality condition [33a], especially when U_{PN} is at a high forward bias. For p/n junctions, this approximation is valid at low injection levels, which is most of the cases except power devices at high currents, but for MOS devices, this approximation is not valid at the SiO_2/Si interface or bare surface near flatband. Flatband is in fact the equivalent high-level condition of the p/n junction, namely, the build-in or diffusion potential barrier height of the n+Source/p-Basewell-channel junction is reduced to zero by the gate voltage which flatbands the lower-doped p-basewell-channel (The V_{GB} or V_{GS} bends little the n+S surface.) so that the electric potential barrier is flattened or nearly flat from the n+Source-side to the p-Basewell-channel-side. Physically also, the flatband brings the remote charge neutrality boundary condition from $x=X_\infty$ to the SiO_2/Si interface at x=0. Thus, by not neglecting the minority carriers, the remote charge-neutrality boundary condition ($x=X_\infty$) is given by the **1957-SNS charge neutrality condition** [33,33a] in the MOST notation. It reduces to the commonly used approximation when U_{PN} is small or negative or P_{IM} is immensely larger than n_i.

$$\rho(x=\infty,y) = 0 = q[P(x=\infty,y) - N(x=\infty,y) - P_{IM}] \tag{21.18}$$

$$P(x=\infty,y) \times N(x=\infty,y) = n_i^2 \exp[U_{PN}(x=\infty,y)]$$

$$= n_i^2 \exp[U_{PN}(y)] \equiv n_i^2 \exp(U_{PNy}) \tag{21.19}$$

$$P(x=\infty,y) = n_i\{[(P_{IM}/2n_i)^2+\exp[U_{PNy}]]^{1/2} + (P_{IM}/2n_i)\} \tag{21.20}$$

$$= n_i\exp[+U_P(x=\infty,y)] = n_i\exp[+U_P(y)] \equiv n_i\exp(+U_{Py}) \tag{21.21}$$

$$N(x=\infty,y) = n_i\{[(P_{IM}/2n_i)^2+\exp(U_{PNy})]^{1/2} - (P_{IM}/2n_i)\} \tag{21.22}$$

$$= n_i\exp[-U_N(x=\infty,y)] = n_i\exp[-U_N(y)] \equiv n_i\exp(-U_{Ny}) \tag{21.23}$$

The spatially constant quasi-Fermi potentials (that was assumed in order to enable the x-integration of the Poisson Equation to give the two of the three solutions: $E_S(y) \triangleq E_X(x=0,y)$ in (21.12) and the Gauss Law in (21.9)) requires that $U_P(x,y) = U_P(y) = U_P(x=0,y) = U_{P0}(y) = U_{P\infty}$ and $U_N(x,y) = U_N(y) = U_N(x=0,y) = U_{N0}(y) = U_{N\infty}$. Then, F_{SI} in (21.17) reduces to the form given by Sah in his 1964 MOSC report [29] and his first MOST article [34], and it was described in details in 1965-Sah-Pao [35] (which was then used by the 1966 Pao-Sah Drift-Diffusion Model [36]). This reduced form gave the analytical solution of the MOST using the **three-layer** bulk-charge and **inversion-charge** model in which $\xi \triangleq -U_{PN}$ was defined. The negative sign was used for ξ because the drain and source p/n junctions were always reverse biased for MOST operation while the forward-biased p/n junction applications for monitoring and studying interface traps, investigated in 1961 [37,38], were not an MOST modeling objective at that time. The forward-bias mode was investigated substantially since 1995 by Sah and his group at Florida [39] to monitor the creation and annealing of the SiO_2/Si interface traps and the degradation of MOSTs by hot carriers.

However, it was just discovered by Sah while writing this keynote manuscript that this 1965-Sah-Pao solution [35] was one of two approximations among the three possible nonequilibrium or DC steady-state solutions. These two approximations are not consistent with the constant quasi-Fermi-potential assumption, which is the cause of the imaginary electric field or negative E_S^2 in a very small range of U_S near flatband {As pointed out to Sah by Zhou [31].}, while the third and self-consistent 'exact' solution does not have this physical unreality which we shall now show. (Note added during proof: I had already decided not to use the word "exact" and informed my ten co-authors [55] of the better term 'self-consistent', which was also rejected later by a reviewer of one of our manuscripts in which the reviewer stated "Nothing is exact in compact modeling.". Indeed this was an issue also raised rigorously by Brews in 1977 [32d] before the 1978-Brews analyses [40] and ably answered by El-Mansy and Boothroyd concerning one part of El-Mansy's PhD thesis on a new compact MOST model.) Using (21.17) for F_{SI}^2, in the following three sub-sections, 211, 212, and 213, the three solutions are obtained for the x-independent quasi Fermi potentials, written as $U_P(x,y)=U_P(y)$ and $U_N(x,y)=U_N(y)$.

211 Self-Consistent Solution (2005-Sah=1957-Sah-Noyce-Shockley)

This is the mathematically correct solution for the x-independent quasi-Fermi potentials giving by $U_{P0} = U_P(x=0,y) = U_{P\infty} = U_P(x=\infty,y) = U_P(y)$, $U_{N0} = U_N(x=0,y) = U_{N\infty} = U_N(x=\infty,y) = U_N(y)$, and $\xi \triangleq +U_{NP}(x=\text{any value}, y) = U_N(y) - U_P(y) = -U_{PN0} = -U_{PN\infty}$. The 'exact' or better **self-consistent** solution is

$$(F_{SI})^2 = +\,[\exp(-U_S) + (+U_S - 1)]\exp(+U_{P0}) + [\exp(+U_S) + (-U_S - 1)]\exp(-U_{N0}) \qquad (21.17)$$

$$(F_{SI})^2 = \{+\,[\exp(-U_S) + (+U_S - 1)]\exp(+U_{P0}) + [\exp(+U_S) + (-U_S - 1)]\exp(-U_{P0}-\xi)\} \qquad (21.24)$$

$$= \{+\,[\exp(-U_S) + (+U_S - 1)] + [\exp(+U_S) + (-U_S - 1)]\exp(-2U_{P0}-\xi)\}\exp(+U_{P0}) \qquad (21.25)$$

$$\lim(U_S\to 0)(F_{SI})^2 \to \{+\,[1 - U_S + U_S^2/2 - U_S^3/6 + (+U_S - 1)] + [1+U_S+(U_S^2/2)+(U_S^3/6)+(-U_S-1)]\exp(-2U_{P0}-\xi)\}\exp(+U_{P0})$$

$$= \{+[(U_S^2/2)(1-U_S/3)]+[U_S^2/2)(1+U_S/3)]\exp(-2U_{P0}-\xi)\}\exp(+U_{P0})$$

$$= (U_S^2/2)\{+(1-U_S/3)+(1+U_S/3)\exp(-2U_{P0}-\xi)\}\exp(+U_{P0}) \geq 0 \quad \textbf{QED} \qquad (21.26)$$

The asymptotic expansion to third order near flatband, $U_S \to 0$, given by (21.26), shows that $(F_{SI})^2$ is positive and real to third order (hence to all orders of U_S by just including the higher order terms in the Taylor series expansion of $\exp(\pm U_S)$, confirming that the remote charge-neutrality boundary condition and the mathematics are self-consistent with no error. However, there is another very-common error source that would make $(F_{SI})^2$ negative near flatband ($U_S \to 0$). This was just discussed that led to (21.18) to (21.23). It has not been recognized by most programmers and their device theorists, which is: in numerical computations, U_{P0} (and U_{N0}) must be computed using the exact remote charge-neutrality condition, given by (21.18) to (21.23), which do not drop the minority carriers.

212 Approximate Solution – 1 (1965-Sah-Pao [35] and 1966-Pao-Sah [36])

The earlier, if not the historical first, solution was given in 1965-Sah-Pao [35], which is not self-consistent, and which was used in 1966-Pao-Sah [36], 1978-Brews [40], 1996-Sah [41], 2001-2004-Gildenblat [32a,32b], 2002-McAndrew [32c], Jie-Sah [42,43], 2005-Zhou [44], and others. It was also used earlier in partial form in 1953-Brown-Shockley [18] and 1956-Garrett [21]. This solution was obtained by letting $U_{P\infty}=U_{N\infty}$ and $U_{PN\infty}=U_{P\infty}-U_{N\infty}=0$ at the distant boundary with charge neutrality, while retaining their interface value, $U_{PN0}=U_{P0}-U_{N0}=-\xi$ in regions near the SiO_2/Si interface at x=0. It violates the assumption that U_P and U_N are independent of x, and it was based on the device-physics intuition that in any real device, the far boundary would be at equilibrium or $U_P=U_N=U_F$, while at the SiO_2/Si (x-0) interface, they would have their nonequilibrium values due to the voltage applied between the drain and base and source and base terminals. Then from (21.15A), only two of three terms due to the minority carriers are multiplied by $\exp(-\xi)$ giving

$$F_{SX}^2 = +\left[\exp(+U_{P0}-U_S)-\exp(+U_P-U)+(+U_S-U)\exp(+U_{P\infty}-U_\infty)\right] + \left[\exp(+U_S-U_{N0})-\exp(+U-U_N)-(+U_S-U)\exp(+U_\infty-U_{N\infty})\right] \quad (21.15A)$$

$$F_{S\infty}^2 \approx +\left[\exp(-U_S)+(-1+U_S)\exp(+U_{P\infty}-U_{P0})\right]\exp(+U_{P0}) + \left[\exp(+U_S-\xi)-\exp(-U_{NX})-U_S\bullet\exp(U_{P0}-U_{N\infty})\right]\exp(-U_{P0}) \quad (21.27)$$

$$\approx \left\{+\left[\exp(-U_S)+(-1+U_S)\right] + \left[\exp(+U_S-\xi)-\exp(-\xi)-U_S\right]\exp(-2U_{P0})\right\}\exp(+U_{P0}) \quad (21.28)$$

The choice of U_{NX} in (21.15A) is crucial. If $U_{NX=\infty}=U_{N\infty}=0$, then the electron concentration term $\exp(-U_{NX}) = 1$ is independent of the drain and source voltage, ξ. Then, we have only one term that is modulated by the voltages applied to the drain and source which is the next or third solution or the second approximate solution given in section 213. In the 1965 Sah-Pao [35] and 1966 Pao-Sah [36] articles, the position X in F_{SX}^2 given by (21.15A) was chosen such that the three boundary conditions [$U(x=\infty,y)=0$, $\partial U(x=\infty,y)/\partial x=0$ and $\partial^2 U(x=\infty,y)/\partial x^2=0$] are nearly satisfied but the quasi-Fermi potential difference, $\xi(x=X)=U_N(x=X,y)-U_P(x=X,y)$, has not decreased significantly towards zero at the remote (or nearby) boundary $x = X \to \infty$. Due to the presence of the terms with the $\exp(-\xi)$, it is obvious that $F_{S\infty}^2$ given by (21.28) will become negative. The range of U_S for this physical reality violation can be easily obtained from an analytical expansion of (21.28), just like (21.26) for the self-consistent solution. Later investigators all used this model from 1965-Sah-Pao [35] and 1966-Pao-Sah [36], or textbooks such as 1981-Sze's and 1999-Tsividis's [13], and were burdened by this physical unreality trouble [31]. A second source of trouble came from dropping the minority carrier contribution, giving $\exp(U_F) = P_{IM}/n_i$ instead of

$\exp(U_F) - \exp(-U_F) = P_{IM}/n_i$ to compute the equilibrium Fermi potential U_F, and the nonequilibrium Fermi potentials, $U_P(x=X_\infty,y)$ and $U_N(X_\infty,y)$ using (21.18) to (21.23).

213 Approximate Solution – 2 (New)

A second alternative approximation which is also not self-consistent is to set the distant boundary condition to equilibrium or zero bias, $\xi_\infty = -U_{PN\infty} = U_{N\infty} - U_{P\infty} = 0$, while at the interface, still to have $\xi_0 = -U_{PN0} = U_{N0} - U_{P0} \neq 0$ so that it can be set to the voltage applied to the Drain/Base and Source/Base junctions at y=L and y=0 respectively. Then there is only one $\exp(-\xi)$ term.

$$(F_{SI})^2 = +\,[\exp(-U_S) + (+U_S - 1)\exp(-U_{P0}+U_{P\infty})]\exp(+U_{P0}) + [\exp(+U_S) + (-U_S - 1)\exp(+U_{N0}-U_{N\infty})]\exp(-U_{N0}) \qquad (21.17)$$

$$(F_{SI})^2 = +\,[\exp(-U_S) + (+U_S - 1)\exp(-U_{P0}+U_{P\infty})]\exp(+U_{P0}) + [\exp(+U_S-U_{P0}-\xi_0) - (U_S+1)\exp(-U_{P\infty})] \qquad (21.29)$$

$$= +\,[\exp(-U_S) + (+U_S - 1)\exp(-U_{P0}+U_{P\infty})]\exp(+U_{P0}) + [\exp(+U_S-\xi_0) - (+U_S + 1)\exp(+U_{P0}-U_{P\infty})]\exp(-U_{P0}) \qquad (21.30)$$

$$= +\{[\exp(-U_S) + (+U_S - 1)] + [\exp(+U_S-\xi_0) - (+U_S + 1)]\exp(-2U_{P0})\}\exp(+U_{P0}) \qquad (21.31)$$

The anticipated consequences on compact modeling from using these three boundary conditions, whether self-consistent or not, to design a compact transistor model for circuit simulation, should be immediately clear, namely, it would impact the solutions only near the flatband. In DCIV, seldom is the current computed down to flatband, $U_S=0$, which would be three mega-decades or 10^{-18} below the strong inversion current, since the practical standby or off range, a concern for battery operated equipment, is at about $U_S=U_F$ (the intrinsic surface at $N_S=P_S$) or two mega-decades lower. {See Fig.682.1(b) on p.655 of 1991-Sah [8] for the seldom if ever illustrated $U_S=0$, U_F and $2U_F$ markers on the semilog I_D-V_{GS}.} On digital and analog switching and RF-analog characteristics, the discontinuity or not-self-consistency could give a more serious theoretical design problem, such as waveform distortion limiting the accuracy of rise-fall-delay time estimates in synchronized logic circuits using linear piece-wise charge-control analysis, and charge-storage capacitance and conductance calculations for harmonic distortions, RF interferences, and voltage-controlled capacitors (varactors) operating in the depletion-accummulation ranges.

214 Fast Convergent Exact Iteration Formulas

Fast convergent exact iteration formulas were derived by Sah in 1993 {See Equations (411.16D) to (411.16U) on p. 129 of 1993-Sah [9].} for the equilibrium case. One formula was obtained for each of the three surface potential ranges (depletion, inversion and accumulation). They were soon extended and later reported in 1996-Sah [41] for the nonequilibrium 1965-Sah-Pao and 1966-Pao-Sah approximation (21.28). These exact fast-convergent formulas (for the not-self-consistent 1965-Sah-Pao x-equation) were employed to compute the results recently presented at the October 2004 ICSICT in Beijing by Jie and Sah [42,43]. The fast-convergent formulas for the self-consistent ('exact') solution (21.25), obtained by Jie and Sah in 2004 after an inquiry from Zhou [31], are given in the poster paper at the May-2005 WCM [45]. This self-consistent solution is used in Zhou May-2005 WCM [44], without the benefit of the fast convergent exact iteration formulas since his unified regional model does not need the iterative computation of the surface potentials. Additional tests on the mathematical approximations and

computation speeds of these three boundary conditions are anticipated. The 1996 formulas [41,42,43] are listed below, in sub-section 2141, which were extended to give the 2005 self-consistent formulas [45], also listed below, in sub-section 2142.

2141 1965 Fast-Convergent Exact Iteration Formula

(1965-Sah-Pao [35] Model used in 1966-Pao-Sah [36]; 1996-Sah Formulas [41])

Accumulation Range $U_S \le 0$ (21.32)

$$U_S = -\log_e\{[(U_{GX} - U_{FB} - U_S)^2/4U_{AA}] + \Delta_A\}$$

Depletion Range $0 < U_S \le 2U_F + \xi$ (21.33)

$$U_S = U_{GX} - U_{FB} - 2U_{AA}\{[1+(U_{GX}-U_{FB}-\Delta_D)/U_{AA}]^{1/2} - 1\} \quad 0 \le U_S \le U_{GX}-U_{FB}$$
$$U_S = U_{GX} - U_{FB} + 2U_{AA}\{[1+(U_{GX}-U_{FB}-\Delta_D)/U_{AA}]^{1/2} + 1\} \quad U_{GX}-U_{FB} \le U_S \le 2U_F+\xi$$

Inversion Range $U_S \ge 2U_F + \xi$ (21.34)

$$U_S = 2U_F + \xi + \log_e\{[(U_{GX} - U_{FB} - U_S)^2]/(4U_{AA}\Delta_I)\}$$

Exact Iteration Correction Terms

$$\Delta_A = 1 - U_S - \{[\exp(+U_S) - 1]e^{-\xi} - U_S\}\exp(-2U_F)\} \approx 1 - U_S \quad (21.35)$$

$$\Delta_D = 1 - \exp(-U_S) - \{[\exp(+U_S) - 1]e^{-\xi} - U_S\}\exp(-2U_F) \le 1 - \exp(-2U_F) \quad (21.36)$$

$$\Delta_I = 1 - \{e^{-\xi} + U_S - [U_S - 1 + \exp(-U_S)]\exp(+2U_F)\}\exp[-(U_S-\xi)] \approx 1 + U_S\exp[-(U_S - 2U_F - \xi)] \quad (21.37)$$

$$U_{AA}=qV_{AA}/kT=(q/kT)(\varepsilon_S qP_{IM}/2C_O^2)=(P_{IM}/10^{17})(X_O/10nm)^2\times 69.59285mV/(kT/q)$$
$$kT/q=25.5564mV \text{ at } T=296.57K \text{ and } n_i=1.000\times10^{10}cm^{-3} \quad (21.38)$$

2142 2005-Fast Convergent Exact Iteration Formulas

(2004-Sah-Self-Consistent Model [31]; 2005-Jie-Sah formulas [45])

Accumulation Range $U_S \le 0$ (21.39)

$$U_S = -\log_e\{(U_{GB} - U_{FB} - U_S)^2[1-\exp(-2U_{P0}-\xi_0)]/4U_{AA} + \Delta_A\}$$

Depletion Range $+0 < U_S \le U_{P0} + U_{N0} = 2U_{P0} + \xi_0$ (21.40)

$$U_S = U_{GX} - U_{FB} - 2U_{AA}\{[1+(U_{GX}-U_{FB}-\Delta_D)[1-\exp(-2U_{P0}-\xi_0)]/U_{AA}]^{1/2} - 1\} \div[1-\exp(-2U_{P0}-\xi_0)] \quad 0\le U_S\le U_{GX}-U_{FB}$$

$$U_S = U_{GX} - U_{FB} + 2U_{AA}\{[1+(U_{GX}-U_{FB}-\Delta_D)[1-\exp(-2U_{P0}-\xi_0)]/U_{AA}]^{1/2} + 1\} \div[1-\exp(-2U_{P0}-\xi_0)] \quad U_{GX}-U_{FB}\le U_S\le 2U_{P0}+\xi_0$$

Inversion Range $U_S > U_{P0} + U_{N0} = 2U_{P0} + \xi_0$ (21.41)

$$U_S = 2U_{P0} + \xi_0 + \log_e\{(U_{GB}-U_{FB}-U_S)^2[1-\exp(-2U_{P0}-\xi_0)]/(4U_{II}\Delta_I)\}$$

Exact iteration correction terms

$$\Delta_A = 1 - U_S - [\exp(+U_S) - 1 - U_S]\exp(-2U_{P0}-\xi_0) \quad (21.42)$$

$$\approx 1 - U_S$$

$$\Delta_D = 1 - \exp(-U_S) - [\exp(+U_S) - 1 - U_S]\exp(-2U_{P0}-\xi_0)\} \quad (21.43)$$

$$\leq 1 - \exp(-2U_{P0})$$

$$\Delta_I = 1-\{(1+U_S)\exp(-\xi_0)-[U_S-1+\exp(-U_S)]\exp(+2U_{P0})\}\exp[-(U_S-\xi_0)] \quad (21.44)$$

$$\approx 1 + U_S \exp[-(U_S -2U_{P0}-\xi_0)]$$

$$U_{AA}=qV_{AA}/kT=(q/kT)(\varepsilon_S qP_{IM}/2C_O^2)=(P_{IM}/10^{17})(X_O/10nm)^2\times 69.59285mV/(kT/q)$$

$$kT/q=25.5564mV \text{ at } T=296.57K \text{ and } n_i=1.000\times 10^{10}cm^{-3}$$

215 Connections with the Inversion Charge and Threshold Voltage Models

An important result, just recognized by Sah during this keynote manuscript preparation, is the fact that these fast convergent exact iteration formulas can provide the rigorous derivations of the two approximate models (threshold voltage and inversion charge models) because the first term of these formulas gives the linear relationship sought in the expansion to relate the gate voltage to the surface potential. In addition, the dependencies of the quasi-Fermi potentials or the terminal voltages at the drain and source are also given in these exact fast-convergent iteration formulas. Thus, the nonlinear terms of these exact formulas lead naturally to the higher order corrections that could provide more accuracy to the threshold voltage and the inversion charge models while retaining their computation speeds.

22 The MOSR and MOST Theory (The Current Equation)

The derivation of the DCIV (Direct Current Current-Voltage) equation of the MOS transistor is continued in this section. The mathematical analysis of the 2-D (assuming wide transistor with no variation in the width-direction or z-direction so the 3-D is reduced to 2-D) MOS field-effect transistor was decomposed into two 1-D problems. The derivation of the 1-D MOSC Voltage-versus-Surface-Potential equation, $V_{GB}(U_S)$, was traced historically in the previous section. The 1-D MOS Resistor (MOSR) or conductor (conflict of acronym to MOS Capacitor) is a modulation of the non-uniform (y-dependent) resistance by the applied gate voltage via changing the electrical conductance or carrier concentration at the silicon semiconductor surface of the SiO_2/Si interface, the latter represented by the surface potential, U_S. Adding up the elemental resistances along the length of the surface channel between the source and drain regions gives the resistance of the MOS transistor $R_{CH}(U_S)$, or the terminal Current-versus-Surface Potential equation, $I_D(U_S)$. The history of the derivation of this current equation is described in this section. The two solutions, $V_{GB}(U_S)$ from section 21, and $I_D(U_S)$, are then combined to give the DC Current-Voltage (DCIV) equation by eliminating the surface-potential variable, U_S, via a variety of approximations in order to invert the implicit voltage equation $V_{GB}(U_S)$ to the explicit $U_S(V_{GB})$. The DCIV or current-voltage equation, $I_D(V_{GB},V_{DB},V_{SB})$, consists of four terminals and three terminal voltages as the independent variables with the base terminal as the reference. In the previous section, we learned that the historical route followed increasing complexity in the derivation of the voltage equation, $U_{GB}(U_S)$, for the MOSC, later used in MOST (to be discussed in this section). However, the exact formulation and rather detailed approximations of the MOSC or voltage equation had already been given in the analyses described by Garrett in 1955 [21] at

Bell Telephone Laboratories (BTL) to explain and match-and-fit the surface field-effect conductance experiments in order to give a measure of the surface states or electronic traps at the semiconductor surface. That was five years before the report on the first demonstration of the modern MOS transistor on silicon with stable thermally grown oxide [7] given by Kahng and Atalla in 1959-1960 at BTL [46,47] who also filed two patents on three different modes of operation of the MOST structure: by Kahng [48] on forward-biased mode and reverse-biased conventional MOST mode with doped p-channel and by Atalla [49] on space-charge-limited punch-through mode with intrinsic-base n-channel.

220 General Theory

(Origins and the History-Physics-Preferred Model Names)

In contrast to the derivations given in the history-bibliography articles of lesser mathematical rigor and fewer device-physics details, in this tutorial exposition of the MOS transistor compact modeling the derivation of the MOST output or drain current equation will be given some mathematical rigor, in the context of the Shockley Equations, in order to succinctly state and show by algebra the assumptions made in the 40+ years of analytical developments by the many transistor-theorists. These compact models have been based on the analyses of the MOS transistor characteristics given by Sah and Pao at the start of the 1960's in the three models described in the three articles they published. For ease of description, these were referred to at the end of the Voltage Equation subsection-21 as **1964-Sah** [34], **1965-Sah-Pao** [35] and **1966-Pao-Sah** [36]. They were for long channels. Some short channel effects in the long channel transistors of the 1960's era were recognized, such as the longitudinal electric-field gradient, and they were mathematically analyzed and correlated to experiments by Sah and his postdoc and predoc collaborators in two other articles, in 1964-Reddi-Sah and 1968-Chiu-Sah. These three long channel models were then approximated by Brews in **1978-Brews** [40], described tutorially in **1981-Brews** [50] and summarized in 1993-Arora [11], 1999-Tsividis [13] and 1981-Sze [51]. Some of Brews formulas were derived previously by Barron in 1972 [52] and simultaneously by Baccarani-Rudan-Spadini in 1978 [53]. Most of Brews' results were rigorously derived by Van de Wiele in 1979 [54] which has been referenced by the ten-author paper presented in this workshop [55] as the proofs of the approximations made by Brews [40]. Some of the 1978-Brews approximations were empirical-intuitive and most were device-physics based.

The following are the mathematical progressions and the assumptions made at each step. They began from the general 3-D time-dependent current equation of electrons (20.3) and holes (20.4). These are simplified to (20.9) and (20.10) for the DC steady-state. The two-mechanism and two-term (drift and diffusion) current equation is compacted into one with just one term, the **gradient** of the **electrochemical potential** or **quasi-Fermi-potential** [16,17] by the use of the **exponential transformation** or the **Boltzmann Representation** [8,9]. This one-term form was the basis of the 1966-Pao-Sah equation [36]. It still contains both the drift and the diffusion currents of electrons and holes since it is the gradient of the electrochemical potential, and the latter contains both the concentration and the concentration gradient of the charged particles or charge carriers. The model was further simplified to just one charge carrier type (electrons) in the surface channel, hence the term "unipolar" transistor, but in fact the presence of the other charge carrier type (holes) is responsible for many important but mostly undesirable electrical characteristics such as leakage current (increases standby power), generation of interface traps (noise) and

generation-charging of oxide traps (threshold voltage shifts) which all degrade the transistor performance.

We shall consider electron channel on the p-type basewell of the n-MOST. The two-term and one-term electron current equations, (20.9), were the origin of the two treatments and models given by 1964-Sah [34] and 1965-Sah-Pao [35] (the threshold voltage model), and 1966-Pao-Sah [36] (the surface potential model). From these, the inversion-charge model can also be derived.

The 1966-Pao-Sah Equation

The 1966-Pao-Sah equation [36] can be obtained by integrating (20.9) in all three coordinates, first the cross-sectional area, dxdz, assuming no variation in z, then dy along the length of the channel between the source and the drain, y=0 to y=L. The dxdz integrations give

$$-I_S = +\ I_D = -\int_z\int_x J_{NY}(x,y)\,dxdz \qquad (220.1)$$

$$= +\ Z\int_x q\mu_n(x,y)N(x,y)[\partial V_N(x,y)/\partial y]dx \qquad (220.2)$$

$$\cong +\ Z[\partial V_N(y)/\partial y]\int_x q\mu_n(x,y)N(x,y)dx \qquad (220.3)$$

$$\cong +\ Z[\partial V_N(y)/\partial y]\mu_{n\text{-}eff}(y)\int_x qN(x,y)dx \qquad (220.4)$$

In (220.1), I_D and I_S are the external currents flowing into drain and source terminals, where the negative sign of the double integral indicates that the electrons in the electron surface channel are flowing from the n+Source at y=0 to the n+Drain at y=L. The dz integration is taken over the width of the channel, Z, which gives the multiplier Z in (220.2) since we assume no z-dependences. The dx integration is from the SiO_2/Si interface $x=0_+$ to the remote boundary $x=X_\infty\rightarrow\infty$. The y-component of the electron current density, perpendicular to the cross-sectional area dxdz, given by (20.9), is substituted into (220.1) to give (220.2).

The approximation from (220.2) to give (220.3) uses $V_N(x,y)\cong V_N(y)$, the x-independent electron quasi-Fermi potential. This was used to integrate the Poisson Equation described in the preceding section to give the MOSC Voltage Equation. It is based on the gradual channel or long channel model first employed by Shockley for the junction-gate field-effect transistor in 1952 [19,20]. For the MOSFET, this long or graduate channel model assumes that the electron current is in parallel to the interface in the y-direction $J_{NY}>>J_{NX}\approx 0$. Then the x-component of the electron current density vector (20.9) gives

$$J_{NX}(x,y) = -q\mu_n N(x,y)\partial V_N(x,y)/\partial x \approx 0 \qquad (220.5)$$

$$\text{or}\quad U_N(x,y) = qV_N(x,y)/kT \approx qV_N(y)/kT = U_N(y) \quad \text{independent of x} \qquad (220.6)$$

An effective mobility is used to give (220.4) from (220.3). This was defined in 1964-Sah {Eq.(5) on p.330 of [34]} and given by (220.7) below. It is the conductivity mobility averaged over the electron charge distribution in the x-direction perpendicular to the SiO_2/Si interface from all the scattering mechanisms that randomize the electron velocity. It is measurable as a function of the gate voltage at zero-voltage between the source and the drain, which was known as the **field-effect mobility**, but truly the **conductivity mobility**, from $g_d=\partial I_D/\partial V_{DS}$, in contrast to other mobilities, such as the Hall-Effect mobility

and the transconductance mobility, from $g_m=\partial I_D/\partial V_{GS}$, which was originally called the **effective mobility** that also mixes in the effect of charged interface traps on the distortion of the transfer characteristics, I_D-V_{GB}. In compact model for device design, it is measured in a sample of transistors with a range of channel lengths in order to experimentally extract the parameters that are used to model the short channel effects.

$$\mu_{n\text{-}eff}(y) \triangleq \int_x \mu_n(x,y)N(x,y)dx / \int_x N(x,y)dx \triangleq \mu_n \qquad (220.7)$$

In the last definition above, we drop the –eff in the subscript of the field-effect mobility to simplify the notation with the x-averaging understood. This average takes into account of the random scattering of the electrons distributed in the electron channel by surface and bulk charges and phonons, and by surface roughness first analyzed by Schrieffer in 1955 [22,56] (using random square potential wells as a crude representation of the higher order point charge distributions, that give dipole and multipoles from randomly displayed host ion cores at the interface, i.e. the Si^{+4} and O^{+2} screened by the core electrons, and by the valence electrons (dielectric screening), further screened by the inversion channel electrons (Debye screening)) which was later applied by Pierret and Sah to MOS transistors in 1968 [57]. The experimental surface-channel mobilities under strong inversion were first measured by Leistiko-Grove-Sah on MOS transistors in 1965 [58] and by Tschopp and Ning in great detail which was then analyzed by Ning to separate surface-phonon and oxide-charge scatterings [59]. One of the frequently cited by the compact model developers was the 1980-Sun-Plummer surface electron mobility measurements [60]. In general, the mobility is used as a fitting parameter to match the theory to experimental measurements, usually at zero source-drain voltage, that is, the conductivity mobility, not the effective mobility which includes distortion from charged interface traps whose trapped charge density depends on band-bending or surface potential, U_S, that masks a measurement of the true surface mobility or the surface channel mobility alone.

The differential equation (220.4) can be 'solved' by simply integrating in the y-direction from the source at $y=Y_S=0$ to $y=Y_D=L$, since the terminal currents I_S and I_D are independent of y. Thus,

$$-\int I_S dy = -I_S L = +\int I_D dy = +I_D L$$

$$= -\int_y \int_z \int_x J_{NY}(x,y)\,dx dz dy \qquad (220.8)$$

$$= +Z\int_y \mu_n(y)[\partial V_N(y)/\partial y]dy \int_x qN(x,y)dx \qquad (220.9)$$

The y-dependence of the effective or conductivity mobility comes from the hot carrier effect. That is, the reduction of the mobility from the increase of the scattering rate of the electrons in the channel is due to energy loss to the optical phonons during each scattering of the electron by an optical phonon resulting in the emission or creation of an optical phonon whose energy is supplied by the electron kinetic energy gained from the accelerating electric field, or the voltage applied between the source and the drain terminals, while traveling along the y-direction from the source y=0 to the drain, y=L. The kinetic energy transferred to optical phonon is dissipated by the vibration of silicon cores in the lattice as heat transferred to the surrounding or heat-sink by acoustic and optical phonons as heat diffusion. The increase of the inelastic collision rate at higher electric field can be computed for the DC steady-state condition by the power balance equation, namely the power loss to optical phonon emission (Optical Phonon Energy divided by Optical-Phonon-Electron scattering time) $\hbar\omega_o/\tau_o$ is balanced by the power gained by the electron from acceleration in the electric field, which is the electron drift velocity ($\mu_n E_Y$) times the electric

force ($\mu_n E_Y \times qE_Y$). The maximum electron kinetic energy that can be gained is when it is equal to the optical phonon energy and immediately emits an optical phonon, $KE=(m_e/2)(\mu_n E_Y)^2=\hbar\omega_o$, giving a velocity and hence drain current saturation at $\Theta_{drift}=\mu_n E_Y = (2\hbar\omega_o/m_e)^{1/2} \approx 10^7$cm/s. {See sections 310 to 314 on pp. 233 to 254 of 1991-Sah [8] for an elementary and pictorial description of drift mobility without the use of the Boltzmann transport equation; see also Chapters 8 and 11 of 1950-Shockley [16].} So, in the velocity saturation range at high electric field, the mobility decreases inversely with the increasing electric field. This is steady-state power-balance, not the commonly-erroneously called energy balance. It is the electron kinetic energy threshold to generate an optical phonon with energy $\hbar\omega_o$. It is the rate of energy gain balanced by the rate of energy loss at steady state, or steady-state **power balance**. With this in mind, the solution given by (220.9) can be written as follows where μ_n is the effective mobility that takes into account of both the electron charge density variation in the x-direction perpendicular to the SiO_2/Si interface and the hot electron reduction of the mobility or the hot electron effect in the high longitudinal electric field in the channel along the y direction parallel to the SiO_2/Si interface.

$$-I_S = +I_D = +(Z/L)\mu_n \int_y [\partial V_N(y)/\partial y]\,dy \int_x qN(x,y)\,dx \qquad (220.10)$$

Since the electron concentration was represented by the electric and quasi-Fermi potentials in the Boltzmann or Maxwellian-Boltzmann representation (for hot electrons with temperature T_e), given by $N=n_i \exp[q(V-V_N)/kT] = n_i \exp(U-U_N)$, and we have shown the conditions under which $V_N(x,y)=V_N(y)$, so the 2-D integration in the (x,y) plane can be mapped to the $[V_N(y),V(x,y)]$ or (U_N,U) plane, given by

$$-I_S = +I_D = +(Z/L)\mu_n \int_y \partial V_N \int_x qN(U,U_N)\,\partial U/(\partial U/\partial x) \qquad (220.11)$$

$$= +(Z/L)qD_n n_i \int_y \exp(-U_N)\,\partial U_N \int_x \exp(U)\,\partial U/(F_x/L_{Di}) \quad \text{1966-Pao-Sah}$$

$$= +(Z/L)qD_n n_i L_{Di} \int_y \exp(-U_N)\,\partial U_N \int_x \exp(U)\,\partial U/F_x(U,U_N) \qquad (220.12)$$

where the normalized electric field $F_x(U,U_N)$ is given by (21.15). Its value at the interface, (x=0,y), is given in a similar form, (21.25), where $U(x=0,y)=U_S(y)$ and $U_N(x,y)=U_N(y)=U_{N0}(y)$. This result is the 1966-Pao-Sah equation [36], with which they had carried out the 2-D integration numerically using $U_N(y)-U_P(y) \triangleq \xi(y)$ to give the DCIV and C-V characteristics of the MOS transistor.

The Charge-Control Equation

The charge-control equation can be written down directly from the 1966 Pao-Sah just derived, but we shall take an earlier step of the Pao-Sah derivation to obtain the Charge-Control differential equation, or the **Inversion Charge Equation** for the inversion charge density, defined by Q_N or Q_I and q_i as used by some investigators. Instead of the final 1966-Pao-Sah equation for the drain or source terminal current given by (220.12), we stop and divert at the end of the dxdz integration over the cross-sectional area before starting the integration dy along the channel from the source to the drain. This was given by (220.4) which can be written in terms of the inversion charge density as follows, a differential equation in y or in the quasi-Fermi potential, $qV_N(y)/kT = U_N(y)$ and $\xi(y)=U_N(y)-U_P(y)$

$$-I_S = +I_D \cong +Z[\partial V_N(y)/\partial y]\mu_{n\text{-}eff}(y) \int_x qN(x,y)\,dx \qquad (220.4)$$

$$= - Z[\partial V_N(y)/\partial y]\mu_n(y)Q_N(y) \qquad (220.13)$$

where the electron charge or **inversion charge density** (areal charge density in Coulomb/cm^2) is

$$Q_N(y) \triangleq - q\int_x N(x,y)\,dx \qquad (220.14)$$

and the integration is carried from $x=X_\infty=+\infty$ to the interface x=0. Using an average mobility, the integration in dy recovers or gives the 1966-Pao-Sah (220.12) with the x-integration hidden in $Q_N(V_N,V_S)$ where $V_S=V_S(V_N)$ from the Voltage Equation solution, such as (21.16) and (21.17).

$$-I_S = + I_D = -(Z/L)\mu_n\int Q_N(V_N,V_S)\,dV_N \qquad (220.15)$$

The result given in (220.13) is in the form of the **Charge Control Equation** namely, Current = Charge divided by Characteristic Time, $I_k = Q_k/T_{Qk}$ at the k-th node. The Charge Control approach for electron device modeling is well-known for more than five decades [61-63]. In this MOST case, I_k is the drain or source terminal current or channel current, the charge is the charge density times the area, given by $Q_N(y)Z\Delta y$, so the local characteristic time for a differential length of the channel Δy in (220.13) is

$$\Delta T_Q(y) = \Delta y/|[\partial V_N(y)/\partial y]\mu_n(y)| \qquad (220.16)$$

which has the form of a length divided by a velocity, $(L/\mu E)$, or it is the transit time at the position y through the channel length Δy. For the entire channel, the total charge is $Q=\Sigma Q_N(y)Z\Delta y$, and using the current given by (220.15), the characteristic time would be defined by $I_D=Q/T_Q$ so that $T_Q\triangleq Q/I_D$ given below as the square of the distance travelled (drifted and diffused) divided by an averaged diffusivity or mobility, which is different from just adding up the differential time given by (220.16) through the channel length travelled y=0 to y=L. For two other calculations giving the identical expression for the transit time and charge storage time of the carriers through the length of the channel, see 1991-Sah. {See Equations (661.3) and (661.4) on page 579 of [8].}

$$T_Q \triangleq Q/I_D = L^2/\langle D_n\rangle = L^2[\int Q_N Z dy/\int \mu_n dV_N Q_N Z L] \qquad (220.17)$$

The charge control method has been used in electron device analysis for more than 50 years. It employs the general relationships of $i_k = q_k/t_k$, $C_{kj}=\partial q_k/\partial v_j$, $G_{kj}=\partial i_{kj}/\partial v_j$ among the terminal and internal nodes of a device (denoted by j and k for the jth and kth nodes). It was first introduced in 1957 to analyze bipolar junction transistors by Sparkes and Beaufoy of the British Telecommunication Research Ltd. In Taplow, Berkes, England. [61]. However, it was employed for decades previously to characterize photoconductors by Albert Rose and colleagues at the RCA Laboratories as stated in a comprehensive article by E. O. Johnson and Albert Rose of the RCA Laboratories in 1959 [62]. This RCA article included the analyses of unipolar and analog transistors [19,64-67], spacistor tetrodes (1957 Statz and colleagues at Raytheon), vacuum tubes, beam deflection tubes, and bipolar transistors [16,17]. An in-depth tutorial analysis of charge transport and charge-controlled (differential) small-signal equivalent circuit elements was given by Middlebrook in 1959 [63] for vacuum devices (diodes, triodes, tetrodes and pentodes; see Johnson-Rose for beam deflection or cathode ray tubes) and for semiconductor devices (m/n Schottky barrier diode, p/n junction diode, and n/p/n bipolar junction transistor [16,17]). Thus, the 3-word compound name "**Charge Control Model**" is the history and physics most-appropriate name for these MOS transistor

compact model developers, in addition to its match to the 3-word name "**Surface Potential Model**" employed by the second group of MOS transistor compact model developers. However, "**Inversion Charge Model**" is probably more suitable for the MOS transistors in view of its special focus on the minority carrier charges in the surface layer or channel because these minority carriers are the very ones which would invert the conductivity type of the surface layer to the type opposite to that of the semiconductor bulk, if the minority carrier concentration at the surface is higher than that of the majority carriers, $N_S \triangleq N(x=0,y) \geq P_S \triangleq P(x=0,y)$ for the p-type basewell or body of the nMOS transistor. Below this inversion condition, the surface is not inverted, but the current is still carried by the trying-to-invert minority carriers, electrons, since they are "injected" or "emitted" by the n+Source, and collected by the p-basewell-channel/n+Drain junction collector, just like that of the bipolar junction transistor [16,17].

The Two-Term (Drift and Diffusion) Four-Component Equation

We shall next derive the second form of the fundamental current equation of the MOST in which the drift and diffusion currents appear as separate terms. This was recently called the four-component model in 1996-Sah [41] which provided a rigorous 2-D analysis to include both the 2-D terms and also the nonlinear 1-D terms, such as the longitudinal field gradient (commonly call the lateral field gradient, definitely a misnomer). This 2-term solution was well-known since the first analysis of the diffusion current in the subthreshold range of MOST reported in 1972-Barron [52]. This two-term approach is also the very one that led to the differential equation of the inversion charge model for the MOST.

From the electron current density vector of the Shockley Equation (20.9), repeated below, the y-component of the electron current in the channel is given by (220.18). $[\mu_n(x,y), D_n(x,y), V(x,y), N(x,y)]$

$$\mathbf{J}_N = + q\mu_n N\mathbf{E} + qD_n\nabla N = - q\mu_N N\nabla V_N = - qD_N N\nabla U_N \qquad (20.9)$$

$$J_{Ny} = - q\mu_n N(x,y)\partial V(x,y)/\partial y + qD_n\partial N(x,y)/\partial y \qquad (220.18)$$

To give the total or terminal current, this is integrated over its entire cross-sectional area, $\int_z\int_x dxdz$:

$$-I_S = +I_D = -\int_z\int_x J_{NY}(x,y)\,dxdz \qquad (220.19)$$

$$= +\int_z\int_x q\mu_n N(x,y)[\partial V(x,y)/\partial y]dxdz - \int_z\int_x qD_n[\partial N(x,y)/\partial y]dxdz \qquad (220.20)$$

$$= +Z\int_x q\mu_n N(x,y)[\partial V(x,y)/\partial y]dx - Z\int_x qD_n[\partial N(x,y)/\partial y]dx \qquad (220.21)$$

$$= +Z\int_x q\mu_n N(x,y)[\partial V(x,y)/\partial y]dx + ZD_{n\text{-}eff}(y)[\partial Q_N(y)/\partial y] \qquad (220.22)$$

where the effective electron diffusivity is similarly defined as the effective mobility given by (220.7), i.e. averaged over the x-distribution of the electron concentration. Therefore, if the Maxwellian energy-distribution is used with an electron temperature, $\exp(-E_k/kT_e)$, the Einstein relationship could be used, $D_{n\text{-}eff}/\mu_{n\text{-}eff}=(kT_e/q)$. Thus far, the result is exact in the context of the assumptions made to obtain (220.18). However, it seems that further simplification is not possible since we cannot assume an x-independent electric potential gradient or y-component of the electric field in the drift current term, i.e., $\partial V(x,y)/\partial y \neq \partial V(y)/\partial y$. Nevertheless, as shown by 1996-Sah [41], an exact 2-D analytical 2-component solution of the drift current term can be obtained by partial integration and by using the solution of the Poisson equation

given by (21.7) to (21.17). This exact 2-component drift-current solution was bench-marked against the original 1966-Pao-Sah double-integral solution [36] and also tested against two compact-models derived by Jie and Sah recently [42,43,45] with reasonable accuracy (~1% up to 6% deviations).

The common practice of the compact model community has been the use of Brews charge-sheet approximation [40], with an intuitive justification or a premature assumption on the dominance of the inversion electron charge or depleted hole charge [52-54] which could be improved by a slightly more rigorous derivation as follows. Assuming that the electron charge density is a delta function at the SiO_2/Si interface, $-qN(x,y) = Q_N(x,y)\delta(x)$ or a very thin **charge sheet**, then the x-integration in the drift current term of (220.22) can be carried out to give the following results. However, there is no need to impose the Brews charge-sheet or delta-function approximation for the diffusion current in the second term of (220.22) although it is also a common practice. This gives the theoretical basis for the experimental determination of the mobility and diffusivity separately: mobility in the slightly above threshold range and diffusivity in the subthreshold range. The results listed below show that the effective mobility is also defined differently than (220.7), i.e. it is the value at x=0 rather than averaged over N(x,y). Therefore the Einstein relationship between thermal equilibrium mobility and diffusivity at any location (x,y) can no longer be used for the effective mobility and diffusivity defined below. It could be more convenient for compact modeling and for effective mobility and diffusivity extraction from experimental data to use the delta function for the diffusion current term also, so that $D_{n\text{-}eff} = (kT/q)\mu_{n\text{-}eff}$ and the approximate thin charge-sheet equation, (220.26) given below, is obtained.

$$I_D = +Z\int_x q\mu_n(x,y)N(x,y)[\partial V(x,y)/\partial y]dx + ZD_{n\text{-}eff}(y)[\partial Q_N(y)/\partial y] \quad (220.23)$$

$$= -Z\int_x \mu_n(x,y)Q_N(x,y)\delta(x)[\partial V(x,y)/\partial y]dx + ZD_{n\text{-}eff}(y)[\partial Q_N(y)/\partial y] \quad (220.24)$$

$$= -Z\mu_{n\text{-}eff}(y)[\partial V(0,y)/\partial y]Q_N(0,y) + ZD_{n\text{-}eff}(y)[\partial Q_N(y)/\partial y] \quad (220.25)$$

$$= -Z\mu_{n\text{-}eff}(y)[\partial V(0,y)/\partial y]Q_N(y) + Z(kT/q)\mu_{n\text{-}eff}(y)[\partial Q_N(y)/\partial y] \quad (220.26)$$

$$= -Z\mu_{n\text{-}eff}(y)\{[\partial V(0,y)/\partial y]Q_N(y) + (kT/q)[\partial Q_N(y)/\partial y]\} \quad (220.27)$$

This is a first order linear differential equation of the form (dy/dx) + P(x)y = Q(x) that can be solved using the integrating factor, exp(∫P(x)dx. The textbook solution is {H. B. Dwight, Tables of Integrals and Other Mathematical Data, MacMillan, NY, 1947, p.205, Section 891.1.}

$$I_D/ZD_{n\text{-}eff}(y) = -(\partial U_S/\partial y)Q_N(y) + \partial Q_N(y)/\partial y \quad (220.28)$$

$$I_D/[ZD_{n\text{-}eff}(\partial U_S/\partial y)] = -Q_N + \partial Q_N/\partial U_S \quad (220.29)$$

$$Q_{NL}\exp(-U_{SL}) - Q_{N0}\exp(-U_{S0}) = (I_D/Z)\int dU_S\{\exp(-U_S)/[D_{n\text{-}eff}(dU_S/dy)]\} \quad (220.30)$$

The last form is a general **charge control** solution of the charge sheet model without any additional approximations. The integral enables the inclusion of the hot carrier or high electric field effects on the mobility and diffusivity using the Maxwellian distribution with a hot carrier temperature that would extend the Einstein relationship to hot carriers. Several approximations to the surface electric field $dU_S/dy=(dUs/dU_{NP})\times(dU_{NP}/dy)$, derived in the previous section on MOSC, have been used by earlier [40,52-54] and recent [55] authors to evaluate the integral between the source, $U_S(y=0) = U_{S0}$, and drain,

$U_S(y=L) = U_{SL}$.

The three MOST compact models, **Threshold Voltage Model**, **Charge Control Model (or Inversion Charge Model)**, and **Surface Potential Model**, all originated from these two **general solutions**, (220.15) without and (220.30) with the charge-sheet approximation. However, some of the earliest derivations of the drift current that gave the first threshold voltage models had used shortcuts instead of the general approach given above. We shall now discuss these earlier and later models.

221 First Generation Threshold-Voltage and Surface Potential Models

(Drift-Current-Only Charge-Control Model with Constant Bulk Charge)
(Parabolic DCIV Equation)

The analysis to describe the DC current-voltage (DCIV) of the MOS transistor began using the electron-channel or n-channel on a p-Si (p-type Silicon substrate or body or basewell) by ignoring the hole current (20.10) and solving the electron current density equation (20.9)

$$\mathbf{J}_N = +q\mu_n N\mathbf{E} + qD_n\nabla N = -q\mu_N N\nabla V_N = -qD_N N\nabla U_N \qquad (20.9)$$

in one dimension along the length (y) of the surface channel without the diffusion current

$$J_{NY} = -q\mu_n N(x,y)\partial V(x,y)/\partial y \qquad (221.1)$$

The drain terminal current is then given by

$$I_D = -\int\int J_{NY}dxdz = -Z\int J_{NY}dx = +Z\int q\mu_n N dx(\partial V/\partial y) \qquad (221.2)$$

$$\simeq +Z\mu_{nS}\int qNdx(\partial V/\partial y) \qquad (221.3)$$

$$\simeq +Z\mu_{nS}(\partial V_S/\partial y)\int qNdx \equiv -Z\mu_{nS}(\partial V_S/\partial y)Q_N(y) \qquad (221.4)$$

$$= +Z\mu_{nS}(\partial V_S/\partial y)C_O(V_{GB} - V_S - V_{FB}) \qquad (221.5)$$

where the inversion charge, $Q_N(y)$ in (221.4), is obtained by the 1-D Gauss Law

$$0 = Q_{OX} + Q_{OT} + Q_{IT} + Q_S \qquad (221.6)$$

$$= Q_{OX} + Q_{OT} + Q_{IT} + q\int(P - N - P_{IM})dx \qquad (221.7)$$

$$= Q_{OX} + Q_{OT} + Q_{IT} + q\int(P - P_{IM})dx - q\int Ndx \qquad (221.8A)$$

$$= Q_{OX} + Q_{OT} + Q_{IT} + Q_B + Q_N \qquad (221.8B)$$

$$\equiv C_o(V_{GB} - V_{FB} - V_S) + Q_N \qquad (221.9)$$

$$V_{FB} = -(Q_{OX} + Q_{OT} + Q_{IT} + Q_B)/C_O = \text{constant} \neq f(V_S,V) \qquad (221.10)$$

The MOST current differential equation, (221.5), can be solved by integration along y from the source, y=0 and $V_S(y=0)=V_{S0}$, to the drain, y=L and $V_S(y=L)=V_{SL}$, giving the parabolic DCIV equation of the MOST for the surface potential model (SP) and the threshold voltage model (TV) respectively as follows.

$$I_D = (Z/L)(\mu_{nS}C_O/2)\left[(V_{GB} - V_{FB0} - V_{S0})^2 - (V_{GB} - V_{FBL} - V_{SL})^2\right] \quad \mathbf{SP} \qquad (221.11)$$

$$I_D \approx (Z/L)(\mu_{nS}C_O/2)\left[(V_{GB} - V_{TS} - V_{SB})^2 - (V_{GB} - V_{TD} - V_{DB})^2\right] \quad \mathbf{TV} \qquad (221.12)$$

To get the threshold voltage equation, (221.12), the approximations of $V_{SL}=V_{DB}-V_{Dbuilt\text{-}in}$ and $V_{S0}=V_{SB}-V_{Sbuilt\text{-}in}$ are made in (221.11) since the diffusion current is not considered.

The surface potential (SP) model equation of the drain current, (221.11), is a new form, obtained from the known general solution including bulk-charge and diffusion current. It is included here just for this historical review in order to generalize the presentation. It shows all the desired features sought by the compact model developers during the last decade, for examples, (i) it is completely symmetrical with respect to the source and drain, regardless of the physical asymmetry such as different impurity concentrations between the drain and source boundaries, and (ii) it does not have the difficulties associated with the threshold voltage model, (221.12), which are described below. However, the SP model equation given by (221.11) is not an explicit equation relating the current to the terminal voltages, but rather to the surface potentials at the drain and source boundaries, which must be computed from another implicit equation between the applied gate, drain and source voltages and the surface potential at the source and the drain boundaries which was derived and given in the MOSC section by inverting the implicit relationship, $V_{GB}(U_S)$, to give $U_S(V_{GB})$, such as (21.16) and (21.17).

There have been three troubles in the result given by the threshold voltage model (221.12). **(1)** It implies a negative differential resistance or decreasing I_D when $V_{SB}=0$ and $V_{DB} > V_{DBsat} = V_{GB} - V_{TD}$ (or the source output bias configuration $V_{DB}=0$ and $V_{SB} > V_{SBsat} = V_{GB} - V_{TS}$) after I_D reaches the peak I_{Dsat} at $V_{DB} = V_{DBsat} = V_{GB}-V_{TD}$. Actually, it does not decrease and it does not give a negative differential resistance (first noted by Shockley during his invention analysis of the JGFET [19], cited by him as an entry in his Bell Laboratory patent notebook). Instead, the drain or channel current stays at the maximum value, independent of further increase of V_{DB}. The underlying physics is that the minority or channel carriers are swept to the drain by the drain electric field $\mathbf{E}_Y$ from the applied drain voltage V_{DB}, and this region becomes depleted of carriers, $Q_N(y \geq Y_D)=0$, and hence can no longer influence the minority carrier concentration, which is now completed controlled by the voltage applied between the gate and the source terminal which lowers the source/base junction's diffusion potential barrier. **(2)** A second trouble occurs if the source and drain are not identical (in dopant impurity concentration and oxide thickness) so that the local threshold voltages are not equal or $V_{TD} \neq V_{TS}$. Then $I_D \neq 0$ when $V_{DB}=0$ and $V_{SB}=0$ or $V_{DB}=V_{SB}=0$ or $V_{DS}=0$. This again comes from ignoring the diffusion current which produces a diffusion barrier or build-in potential difference in the source and drain p/n junctions. Including this diffusion-barrier difference between the source and drain p/n junctions in their threshold voltages referred to the drain and source, the threshold voltage difference is then eliminated, giving $V_{TD}=V_{TS}$, and $I_D=0$ when $V_{DB}=V_{SB}$. This is known as the Gummel Symmetry condition. [See also the 1991-Sah [8] p.586, Equations (663.1), (663.2A) and (663.2B) for switching transient calculations of MOST inverter circuits.] The correct answer, after the fact, is $V_{TD} = V_{TS} = V_{FBL} + V_{Dbuild\text{-}in} = V_{FB0} + V_{Sbuild\text{-}in}$, which may seem contradictory to well-known experimental fact for physically asymmetrical transistors, but indeed correct for this formula which applies to strong inversion because the channel current in strong inversion approaches independence of dopant impurity concentration or bulk charge when the gate-voltage induced channel charge dominates, $|-Q_N| >> |-Q_B|$ and we have ignored both the channel drift current in the weak inversion range that is strongly affected (depressed) by the bulk charge as shown by 1965-Sah-Pao [35] and the channel diffusion current in the subthreshold range which was analyzed by later authors using the 1966-Pao-Sah double integral formulation [36] and the 1996-Sah exact formulation [41].

We now give the historical developments of the theory. The initial descriptions of the experimental measurements and mathematical derivation of the DCIV characteristics of the MOST was

given by Dawon Kahng at the June 1960 IRE-AIEE Solid-State Device Research Conference in Pittsburgh. These results were covered in a Bell Telephone Laboratories (BTL) Memorandum for file written by Kahng [47] dated January 16, 1961. The analysis was limited to the drift current, (221.1), while his MOSC analysis was more advanced, following the formulation given by 1955-Garrett (of BTL) [21]. However, Kahng deleted the bulk-charge term Q_B in (221.8B) by assuming strong inversion so that the electron density is much higher than the bulk charge density, $|Q_N| \equiv |-q\int N dx| >> |Q_B| = |q\int (P - P_{IM}) dx|$. The high concentrations of electrons are assumed to reside in a very thin sheet at the SiO_2/Si interface, $qN(x,y) = Q_N(x,y)\delta x$ which later, in 1978, was called by Brews, also a member of the BTL, the charge-sheet model [40], which has been known as Brews Charge-Sheet model by all subsequent investigators.

There were two additional simplifications of this charge-sheet assumption: (i) the electron mobility $\mu_n(x,y)$ is taken as its value the interface $\mu_n(x=0,y) \equiv \mu_{nS}(y)$, giving (221.3) and (ii) only the electric field $E_y = -\partial V(x=0,y)/\partial y$ in the y-direction at the SiO_2/Si, x=0, matters, and it can be taken out of the x-integration as a delta-function at x=0, as illustrated by (221.4) and assumes the value at the interface, x=0 labeled by subscript S, show in (221.4). These assumptions gave the simple differential equation (221.5) which can be integrated along the length of the channel to give the parabolic DCIV equation of the MOST shown in (221.11) as a function of the surface potential at the source y=0, V_{S0}, and at the drain y=L, V_{SL}. A drastic but good approximation was made by assuming a proportionality between the surface electric potential and the surface electron and hole quasi-Fermi potentials or the voltages applied to the drain and source terminals relative to the body or basewell terminal. This reduces (221.11) to (221.12). The approximation (221.12) fails when the drain voltage is higher than the gate voltage, $V_{DB} > V_{GB} - V_{TD}$. The additional or excess drain voltage drains out all the electrons and creates a depleted carrier space-charge region surrounding the n+Drain/p-body p/n junction, where the proportionality of the electron quasi-Fermi potential and the electric potential is no longer maintained, as they are in the channel region between the source and the drain junctions.

Kahng's analysis was followed by a much simplified and elementary distributed resistance-capacitor analysis given by Hofstein and Heiman of the RCA Laboratories which was reported in a September 1963 article [68]. This included also only the drift current without the bulk charge and the experiments were made on Silicon. The description gave the induced and doped channel transistors, defined the depletion mode and enhancement mode operations, and also the current saturation phenomena as the drain voltage is increased to above the gate voltage causing the decrease and depletion of the electrons near the drain. Then, Ihantola and Moll gave a analyses made by Ihantola in his Stanford PhD thesis dated September 1961 [69] and published three years later [70], including only drift current. The theory did not include the diffusion current but the bulk charge was included, to be analyzed in the next section. Only the input and output capacitances were computed based on the differential capacitance model of the total change of charge or total charge flow into a terminal, Δq (measurable by an integrating current-electrometer), when a small voltage change ΔV (such as a voltage step) is applied between the terminal and a reference terminal, $C = \Delta q/\Delta V$.

A detailed description was given by Sah in 1964 [34] for the drift current and strong inversion range. Some thirty concepts of the MOST analysis and operation were discussed or proposed, including picking the MOS transistor name (MOST without the FET), the terms source, gate and drain borrowed from the Junction Gate field effect transistors (JGFET) described by Shockley [19,64-66], the spatial

variation of the quasi-Fermi potentials first described by Shockley in 1949 for p/n junctions and bipolar junction transistors [16,17] and computed by Sah in 1966 for symmetrical abrupt and linear graded p/n junctions [71], and 1-D energy band versus x-direction. The three-dimensional constant energy surfaces was given also in the 1964 Sah description [34], following Sah's 1961-1962 studies on the effects of surface recombination and channel on p/n junction and transistor characteristics [37,38] which was modeled after Brown's 1953 description of electron potential energy surface in a n/p/n transistor with a surface channel [18]. These 3-D constant energy surfaces [34] were particularly helpful in describing pictorially the geometric features of the MOST operation, including the pinch-off of the electrical channel thickness and the shortening of the channel after pinch-off or drain current saturation, the use of the term threshold voltage and its definition, the analyses of the bulk charge which was then assumed constant or voltage independent to give the simple parabolic current voltage characteristics. The 1964-Sah paper [34] also computed and illustrated in figures the voltage dependences of the intrinsic small-signal capacitances. This parabolic current-voltage and small-signal charge control parameter model was used in the initial version of the Berkeley SPICE by Nagel and Pederson in 1973 [72,73]. (See also the excellent review by Narain Arora in his 1993 book [11].)

222 Second Generation Threshold Voltage and Surface Potential Models

(Drift-Current-Only Charge-Control Model with Spatially Varying Bulk Charge)
(Un-parabolic DCIV Equation)

The bulk impurity ions in the surface space-charge layer reduce the inversion-layer charge induced by the voltage applied to the gate, hence lower the drain current. The common assumption $-Q_N \equiv q\int N dx >> -Q_B \equiv q\int(P_{IM} - P)dx$ is not made, and all terms in Q_S are retained. From (221.6) and (221.7) repeated below,

$$0 = Q_{OX} + Q_{OT} + Q_{IT} + Q_S \qquad (221.6)$$

$$= Q_{OX} + Q_{OT} + Q_{IT} + q\int(P - P_{IM} - N)dx \qquad (221.7)$$

$$\triangleq Q_{OX} + Q_{OT} + Q_{IT} + Q_B + Q_N$$

we get

$$I_D = -\int\int J_{NY}dxdz = -Z\int J_{NY}dx = +Z\int q\mu_n(\partial_y V/\partial y)Ndx \qquad (222.1)$$

$$\triangleq +Z\mu_{nS}\int(\partial_y V/\partial y)qNdx \quad \{\text{Next use Brews charge-sheet } Q_N(x)\delta(x).\} \qquad (222.1A1)$$

$$\approx +Z\mu_{nS}\int(\partial_y V/\partial y)[-Q_N(x)\delta(x)]dx \equiv -Z\mu_{nS}(\partial V_S/\partial y)Q_N(y) \qquad (222.1A2)$$

$$= +Z\mu_{nS}(\partial V_S/\partial y)\ [C_O(V_{GB} - V_{FB} - V_S) + Q_B] \qquad (222.1A3)$$

$$\triangleq +Z\mu_{nS}(\partial V_{NP}/\partial y)[C_O(V_{GB} - V_T - V_{NP}) + Q_B] \qquad (222.1A4)$$

where the bulk charge term can be approximated by the three-layer approximation first used for MOSC by 1964-Sah [29] and then by 1965-Sah-Pao for Si MOSTs [35]. Using (21.14) and (21.15) for the x-component of the electric field, $E_X = -\partial V/\partial x = -(kT/q)(\partial U/\partial x)$, the three-layer approximation gave

$$Q_B = -q\int(P_{IM} - P)dx = q\int P_{IM}[1-\exp(-U)]\partial_x U/(\partial_x U/\partial x) \qquad (222.2)$$

$$\triangleq -qP_{IM}(x_1 + x_2 + x_3) \qquad (222.3)$$

$$= -(2\varepsilon_S kTP_{IM})^{1/2}\times\left[e^{-1}(2U_F+1+U_{NP})^{-1/2}+\sqrt{2}\,(2U_F-1+U_{NP})^{-1/2}+(2U_F-3+U_{NP})^{1/2}\right] \qquad (222.4)$$

where x_1 is the inversion layer thickness, x_2 is the transition layer thickness between the inversion and

depletion layer, and x_3 is the depletion layer thickness. This approximation is obviously not good near flatband U_S=0 and when the n+Drain/p-Body or n+Source/p-Body junction is forward-biased U_{NP}<0. The 1-layer surface potential charge-sheet approximation given by 1978-Brews [40] empirically and its compact model simplification are given by:

$$Q_B = -q\int (P_{IM} - P)dx = q\int P_{IM}[1-\exp(-U)]\partial_x U/(\partial_x U/\partial x) \quad (222.5)$$

$$\approx -(2\varepsilon_S kTP_{IM})^{1/2}(U_S-1)^{1/2} \text{ 1978-Brews Approximation [40]} \quad (222.6)$$

$$\approx -(2\varepsilon_S kTP_{IM})^{1/2}(U_S-0)^{1/2} \text{ Compact Model Simplification} \quad (222.7)$$

A origin of the "−1" term guessed by 1978-Brews can be deduced a little more rigorously by noting that this is equivalent to replacing the three terms in last square-root of (222.4) by two terms from the main surface space-charger layer given in the square-root of (222.6): $(U_S - 1) \rightarrow (2U_F - 3 + U_{NP})$ or $U_S = (2U_F - 2) - U_{PN}$. This is a familiar expression of the p/n junction theory, showing a build-in or diffusion barrier height of $2U_F - 2$ at both the drain and source p/n junctions. This simple result does not seem to take into account of the dopant impurity profile differences between the source and the drain junction and the variation of the basewell-channel impurity concentration from the source to the drain, but indeed it could as shown below. Furthermore, the finite thicknesses of the inversion layer, x_1, and the boundary between the inversion and depletion layers, x_2, are not evident. A more bothersome point is that the effects of the applied voltages to the drain and source junctions are no longer explicitly evident but are imbedded in the surface potentials at the drain and source boundaries as indicated below.

An important result in the three-layer charge-sheet approximation (222.6) given by 1965-Sah-Pao [35], is the voltage dependence of the inversion layer thickness, $x_1 \propto (2U_F+1+U_{NP})^{-1/2}$, which decreases with increasing reverse biased ($U_{NP}=U_{DB}\geq 0$) applied between the n+Drain and p-Basewell. If the inversion layer thickness is defined electrically as the point in space where the electron surface concentration is equal to the p-type dopant impurity concentration, $N(x=X_{INV}) = P_{IM}$, then the inversion or electron channel indeed thins down to zero ($X_{INV}=0$) at a given gate voltage, U_{GB}, when the reverse-biased drain voltage, U_{DB}, is increased to a value equal to or large than that given by $U_S = 2U_F+U_{NP} = 2U_F+U_{DB}$ or $U_{DB} \geq U_S - 2U_F$. The value of this surface potential, is determined by both the applied gate and drain voltages, as given by the general solution (21.16) and (21.17), and the self-consistent solution (21.24) and (21.25) repeated below, where the quasi-Fermi potentials are obtained from the exact remote charge neutrality condition.

$$U_{GB} - U_{FB} - U_S = U_{OX} = 2(U_{II})^{1/2} \times F_{SI}(U_S, U_{P0}, U_{P\infty}, U_{N0}, U_{N\infty}) \quad (21.16)$$

$$(F_{SI})^2 = +\left[\exp(-U_S)+(+U_S - 1)\exp(-U_{P0} + U_{P\infty})\right]\exp(+U_{P0}) + \left[\exp(+U_S)+(-U_S - 1)\exp(+U_{N0} - U_{N\infty})\right]\exp(-U_{N0}) \quad (21.17)$$

$$\rho(x=\infty,y) = 0 = q[P(x=\infty,y) - N(x=\infty,y) - P_{IM}] \quad (21.18)$$

$$P(x=\infty,y)\times N(x=\infty,y) = n_i^2\exp[U_{PN}(x=\infty,y)] \equiv n_i^2\exp(U_{PN\infty}) \equiv n_i^2\exp(-\xi) \quad (21.19)$$

$$P(x=\infty,y) = n_i\left\{\left[(P_{IM}/2n_i)^2 + \exp(U_{PN\infty})\right]^{1/2} + (P_{IM}/2n_i)\right\} \quad (21.20)$$

$$= n_i\exp[+U_P(x=\infty,y)] = n_i\exp[+U_P(y)] \equiv n_i\exp(+U_{P\infty}) \quad (21.21)$$

$$N(x=\infty,y) = n_i\left\{\left[(P_{IM}/2n_i)^2 + \exp(U_{PN\infty})\right]^{1/2} - (P_{IM}/2n_i)\right\} \quad (21.22)$$

$$= n_i\exp[-U_N(x=\infty,y)] = n_i\exp[-U_N(y)] \equiv n_i\exp(-U_{N\infty}) \quad (21.23)$$

$$(F_{SI})^2 = \left\{+\left[\exp(-U_S) + (+U_S - 1)\right]\exp(+U_{P\infty})\right.$$

$$+\ [\exp(+U_S) + (-U_S - 1)]\exp(-U_{P\infty}-\xi)\ \} \qquad (21.24)$$

$$= \{+\ [\exp(-U_S) + (+U_S - 1)]$$
$$+\ [\exp(+U_S) + (-U_S - 1)]\exp(-2U_{P\infty}-\xi)\}\exp(+U_{P\infty}) \qquad (21.25)$$

Integrating the current (222.1A4) along the y-axis from the source to the drain using (222.4) or (222.6) for Q_B, we then obtain respectively the bulk-charge depressed parabolic DCIV equation for the threshold voltage or the surface potential models listed below.

Threshold Voltage Model

$$I_D=(Z/L)\mu_{nS}(C_O/2)\ [(V_{GB} - V_T - V_{SB})^2 - (V_{GB} - V_T - V_{DB})^2]$$
$$-(Z/L)\mu_{nS}(2q\varepsilon_S P_{IM})^{1/2}(2/3)\times[(V_{DB}+2V_F-3kT/q)^{3/2} - (V_{SB}+2V_F-3kT/q)^{3/2}]$$
$$-(Z/L)\mu_{nS}(2q\varepsilon_S P_{IM})^{1/2}(kT/q)\times$$
$$\{+(2/e)[(V_{DB}+2V_F + kT/q)^{1/2} - (V_{SB}+2V_F + kT/q)^{1/2}]$$
$$+(2\sqrt{2})[(V_{DB}+2V_F - kT/q)^{1/2} - (V_{SB}+2V_F - kT/q)^{1/2}]\} \qquad (222.8)$$

$$\approx (Z/L)\mu_{nS}(C_O/2)\ [(V_{GB} - V_T - V_{SB})^2 - (V_{GB} - V_T - V_{DB})^2]$$
$$-(Z/L)\mu_{nS}(2q\varepsilon_S P_{IM})^{1/2}(2/3)\times[(V_{DB}+2V_F)^{3/2} - (V_{SB}+2V_F)^{3/2}]$$
$$-(Z/L)\mu_{nS}(2q\varepsilon_S P_{IM})^{1/2}[(2/e)+2\sqrt{2}]\times[(V_{DB}+2V_F)^{1/2} - (V_{SB}+2V_F)^{1/2}](kT/q) \qquad (222.8A)$$

Surface Potential Model

$$I_D = (Z/L)\mu_{nS}(C_O/2)\ [(V_{GB} - V_{FBL} - V_{SL})^2 - (V_{GB} - V_{FB0} - V_{S0})^2] \qquad (222.9)$$
$$-(Z/L)\mu_{nS}(2/3)(2q\varepsilon_S P_{IM})^{1/2}[(V_{SL} - kT/q)^{3/2} - (V_{S0} - kT/q)^{3/2}]$$

$$I_D \approx (Z/L)\mu_{nS}(C_O/2)\ [(V_{GB} - V_{FBL} - V_{SL})^2 - (V_{GB} - V_{FB0} - V_{S0})^2] \qquad (222.9A)$$
$$-(Z/L)\mu_{nS}(2/3)(2q\varepsilon_S P_{IM})^{1/2}[(V_{SL} - V_{\Theta L})^{3/2} - (V_{S0} - V_{\Theta 0})^{3/2}]$$

The threshold voltage equation, (222.8), has the same discontinuity difficulties of (221.12), the voltage-independent bulk-charge solution. The surface potential equation, (222.9), is free of these difficulties and also terminal- or circuit-symmetric for physically asymmetrical source and drain if Brews empirical choice of $V_{\Theta L} = V_{\Theta 0} = (kT/q)$ is replaced by 0.

A simpler bulk-charge effect in the threshold-model is given in (222.8A) in which the 3kT/q and ±kT/q terms are dropped. It shows that the two terms from the inversion and transition layers, X_1 and X_2, gives less than 16% contribution, $[(2/e)+2\sqrt{2}](kT/q) = 3.564(kT/q)$ compared with $(2/3)2V_F = (2/3)2\times(kT/q)\log_e(P_{IM}/n_i) = (4/3)7\log_e(10)(kT/q) = 21.490(kT/q)$ for $P_{IM}/n_i = 10^{17}/10^{10}$ or 19% for $10^{18}/10^{10}$. This is a substantial contribution and should not be dropped in comparison with that from the depletion layer, X_3.

Historically, the bulk-charge contribution was first analyzed by W. L. Brown at the Bell Telephone Laboratories under the direction of Shockley in 1953. {See acknowledgement at the end of the 1953-Brown article [18] in which the 1953-Brown-Shockley theory was described in detail.} This pioneering

1953-Brown-Shockley theoretical work was based on concepts introduced by Shockley including quasi-Fermi-potentials [16,17], carrier-depleted space-charge-layer [17] and depletion charge or the bulk-charge [17], the inversion layer and inversion charge, and a most relevant concept to the compact model developers, the pinch-off voltage or the applied drain voltage at which the carriers in the channel is depleted, with a quantitative definition given by Eq.(7) in [18], $U_S = U_{DB} - U_{P\infty} + \log_e(U_{DB\text{-}sat} - U_{DB} + 1)$, which is the basis of the inversion-charge model. In the 1953-Brown-Shockley analyses [18], the thickness of the inversion layer was also obtained which is the first layer of the three layer model for the bulk charge used by 1965-Sah [35]. The 1953-Brown-Shockley analyses were carried further and in great detail by Garrett in 1955 as described in 1955-Garrett [21] for several experimental field-effect measurement conditions and structures which were described in the previous section on the MOSC Voltage Equation. The application of this bulk charge analysis to the DCIV characteristics of the MOS transistor was first made by Kahng in 1961 [47] in which he attributed the derivation to Atalla, rather than the earlier result of 1953-Brown-Shockley [18] and 1956-Garratt [21]. However, when solving the current equation to give the DCIV characteristics, Kahng dropped the bulk-charge term for simplicity. The first inclusion of the bulk charge term was made by Ihantola in his 1961 Stanford PhD thesis under the direction of John Moll which was published in 1964 [69,70]. Ihantola-Moll gave a neat derivation of the bulk charge, $(V_{DS}+2V_F)^{3/2}-(2V_F)^{3/2}$, making it so easy to see without the multiple –(kT/q), but also, most important(ly) avoids two of the problems of the threshold voltage model (drain source asymmetry and imaginary value at low voltage in high resistivity substrate when $2V_F=2(kT/q)\log_e(P_{IM}/n_i) <$ multiple×(kT/q) and when V_{DS} is near zero. A comprehensive device-physics-based analysis of the bulk-charge term for the threshold voltage model was given by 1965-Sah-Pao [35] in which the softened and bulk-charge-shifted turn-off of the I_D-V_{GB} characteristics due to bulk-charge alone (no diffusion) was demonstrated and analyzed {Fig.3(b) in [35]} with a bulk-charge-shifted threshold-voltage defined for the first time which was immediately taken up in SPICE-1 to give a better-model SPICE-2 [11,12,72,73]. The softening was not recognized in the first compact model analysis of the MOST characteristics made in 1972-Barron [52] which had a rather small bulk-charge softening and large diffusion-current softening due to the very low substrate concentration ($P_{IM}=2.5\text{x}10^{15}\text{cm}^{-3}$ without ion implantation) and very thick gate oxide (1400Å) of the earliest technology (see Fig. 8 of [52]). The bulk charge depressment of the drift and diffusion currents have been included in all subsequent theoretical analyses for compact models.

223 Third Generation Threshold Voltage and Surface Potential Models

(Drift and Diffusion Currents with Spatially Varying Bulk Impurity Charge)

Diffusion current was first included in the 1966-Pao-Sah analysis [36] in search of a mathematical and device-physics formulation to account for the negative conductance beyond the peak drain current-voltage characteristics in the drift-current-only threshold-voltage models, given in the first-generation, (221.12), which did not include bulk charge, and in the second generation, (222.8) and (222.8A), which included the bulk charge. The theoretical negative conductance in the threshold-voltage models was indicated by the theoretical drain current reaching a peak and then decreasing as the drain voltage is increased beyond the gate voltage, $V_{DS}>V_{DS\text{-}sat} \equiv V_{GS}-V_{GT}$ where V_{GT} includes the bulk-charge contribution. This theoretical negative conductance was first discovered in the JFET theory worked out by Shockley in 1951 [19] which he entered in his Bell Labs notebook anticipating patent filing for negative resistance signal generation source applications. {Private conversation with Shockley in 1958 (See also [33].) and told

by him in several of his articles on transistor invention history as one of his two mistakes in transistor theory, the other, the Zener tunneling instead of interband-impact electron-hole-pair generation mechanisms for current divergence in p/n junction at breakdown voltage. (See [33], also section 536 on pp.441-450 of 1991-Sah's sophmore textbook [8].) As a student of Shockley, this author also made transistor physics errors at the early age (like teacher like student but one more than teacher at least to his knowledge as of this moment). One was the use of the high-level boundary condition to solve the p/n diode I-V characteristics; an algebraic error led to the incorrect asymptotic formula at low level and the wrong diffusivity given in the appendix of 1957-SNS [33] which was later pointed out by Jean Hoerni, the planar fame, [33b]. The second young-person confusion was the description of the drain saturation current in 1966-Pao-Sah [36], given by Sah not the poor-PhD-student Pao, not the physics, but the mathematics of not recognizing that U_S saturates as $V_{DS} >> V_{DS\text{-}sat} \approx V_{GB}-V_T$, which ($U_S$ saturation) were repeated pointed out by several recent compact modeling authors using it as the lead figure in their articles on "compacting" the models [55]. The third was the 20041130@1243 realization by Sah [34], told by Zhou to Sah, of the imaginary electric field of the 1965-Sah-Pao [35] formula near flatband, and its earlier 2002-GildenblacAndrew empirical correction [32c] and the 2004-Sah response with the physics-based self-consistent solution given by (21.18) to (21.23) to correct this mistake.} The 1964-1965 story was that the cause of the negative conductance in the MOST I_D-V_D curve beyond $V_{DB}>V_{DBsat}$ was known to be due to the reduction of the electron-number or electron-charge density to zero at the drain/channel boundary when V_{DS} reaches $V_{DS\text{-}sat}$ [34,35]. However, the inclusion of the diffusion current given by 1966-Pao-Sah [36], in such a neat-looking but hard-to-compute double integral (220.9)-(220.12), elucidated a transient euphoria that masked the realization that diffusion current was not the physics-answer of the negative conductance, but rather the use of the voltage equation $V_{GB}(U_S,V_{NP}\triangleq\xi,V_{DB},V_{SB})$ to compute the surface potential U_S needed in the $I_D(U_S,V_{NP}\triangleq\xi,V_{DB},V_{SB})$ double integration that eliminated the negative conductance. This surface-potential approach prevented the inversion charge particle-number at the drain/channel boundary, $y=Y_D<L$, from becoming negative, i.e., $-Q_N/q=\int Ndx<0$. If the voltage equation, $V_{GB}(U_S)$ given by (21.16) and (21.17), were used in the first-generation bulk-charge-less MOST model by 1964-Sah [34] and bulk-charge MOST model by 1965-Sah-Pao [35], then the negative conductance trouble would have been resolved in 1964-1965.

The transport through the depleted (or nearly depleted) space-charge layer at the drain/channel boundary is still due to drift, but in high field with drift velocity saturation at high fields or high excess drain voltages, $V_{DS} - V_{DS\text{-}sat}$, as discussed in both the 1965-Sah-Pao [35] and the 1966-Pao-Sah [36] articles, rather than diffusion that was thought to serve current continuity. Nevertheless, the 1966-Pao-Sah [36] formulation included diffusion and gave the double-integral solution for the drain current and has served as the benchmark for checking the accuracy of new compact models.

For this section in which we want to provide an analysis for the diffusion current, we will go to the 1-D form of the original 2-term electron current density equation (20.9), instead of manipulating the two general 1-D 1-term solutions, (220.25) and (220.30), to recover the diffusion current, as some authors have done [40,50-55] as the starting or subsidiary starting point. We shall follow the tutorial route of the previous sections, by analyzing the diffusion current term, for example, in (220.28) that was unnecessarily simplified and limited from the use of Brews' charge-sheet approximation, or (220.23) without the charge-sheet approximation. In either case, the charge-sheet only restricted the diffusivity, so the general definition of the effective diffusivity is that of electron-concentration-weighted average in x-direction without an assumed electron-concentration distribution. So, integrating the diffusion component of (220.23) along the length of the channel (y-direction), we get the following solution for the diffusion

current component.

$$I_{DIFFUSION} = (Z/L)D_{n\text{-}eff}[-Q_N(y=0) + Q_N(y=L)] \quad (223.1)$$

where $-Q_N(y) = q\int N(x,y)dx = q\int N(U,U_N)\partial_x U/(\partial_x U/\partial x)$

$$= qn_i L_{Di}\exp(-U_N)\int \exp(U)\partial_x U/F_X(U,U_N) \quad (223.1A)$$

and $Q_G = C_O V_{OX} = C_O(V_{GB}-V_{FB}-V_S) = -Q_S = +Q_N+Q_P+Q_{IM} = +Q_N+Q_B$ 1-D Gauss Law (223.1B)

$$= -Q_S = -\varepsilon_s E_S = \varepsilon_s \partial V_S/\partial x = \varepsilon_s(kT/q)(\partial_x U_S/\partial x) = \varepsilon_s(kT/q)L_{Di}F_{SI} \quad (223.1C)$$

where $(F_{SI})^2 = \{+ [\exp(-U_S) + (+U_S - 1)]$

$$+ [\exp(+U_S) + (-U_S - 1)]\exp(-2U_{P0}-\xi)\}\exp(+U_{P0}) \quad (223.1D)$$

where $\xi = U_N - U_P$ is assumed x-independent, i.e. $U_N(x,y) = U_N(y)$, $U_P(x,y) = U_P(y)$, and

$$P(x=\infty,y)\times N(x=\infty,y) = n_i^2\exp[U_{PN}(x=\infty,y)] \equiv n_i^2\exp(U_{PN\infty}) \equiv n_i^2\exp(-\xi) \quad (21.19)$$

$$P(x=\infty,y) = n_i\{[(P_{IM}/2n_i)^2 + \exp(-\xi)]^{1/2} + (P_{IM}/2n_i)\} \quad (21.20)$$

$$= n_i\exp[+U_P(x=\infty,y)] = n_i\exp[+U_P(y)] \equiv n_i\exp(+U_{P\infty}) \quad (21.21)$$

$$N(x=\infty,y) = n_i\{[(P_{IM}/2n_i)^2 + \exp(-\xi)]^{1/2} - (P_{IM}/2n_i)\} \quad (21.22)$$

$$= n_i\exp[-U_N(x=\infty,y)] = n_i\exp[-U_N(y)] \equiv n_i\exp(-U_{N\infty}) \quad (21.23)$$

In the above, we collected the previous results from the derivation of the MOSC or voltage equation in order to see, on one page, the algebraic steps and approximations used in the evaluation of the inversion charge density integral, Q_N. The charge densities (223.1A) and (223.1B) were defined in (21.10) and (21.11) while applying the 1-D Gauss Law or integrating the Poisson Equation from the gate-contact metal to the body-contact metal, where both charges and electric fields are zero, and F_{SI} in (223.1D) was derived and given by (21.17). The quasi-Fermi-potentials of holes, U_{P0}, and electrons $\xi = -U_{N0}+U_{P0}$, in (223.1D) were obtained from the remote charge-neutrality condition and the x-independence of the quasi-Fermi-potential approximation given in (21.18) to (21.23). Using these results, the inversion charges and hence the diffusion current can be written in two forms, one for the strong inversion range and one for the weak inversion or subthreshold range.

For the strong inversion range, we can use the 1-D Gauss Law to split the inversion charge into the two terms as indicated by (223.1B),

$$Q_N = Q_G - Q_B = C_O V_{OX} - Q_B = C_O(V_{GB} - V_{FB} - V_S) - Q_B$$

then using the three-layer and 1-layer approximations for the bulk charge, given previously by

$$Q_B = -q\int (P_{IM} - P)dx = q\int P_{IM}[1-\exp(-U)]\partial_x U/(\partial_x U/\partial x) \quad (222.2)$$

$$\triangleq -qP_{IM}(x_1 + x_2 + x_3) \quad (222.3)$$

$$= -(2\varepsilon_S kTP_{IM})^{1/2}\times[e^{-1}(2U_F+1+U_{NP})^{-1/2} + \sqrt{2}(2U_F-1+U_{NP})^{-1/2} + (2U_F-3+U_{NP})^{1/2}] \quad (222.4)$$

$$Q_B = -q\int(P_{IM} - P)dx = q\int P_{IM}[1-\exp(-U)]\partial_x U/(\partial_x U/\partial x) \quad (222.5)$$

$$\approx -(2\varepsilon_S kTP_{IM})^{1/2}(U_S-1)^{1/2} \text{ 1978-Brews Approximation [40]} \quad (222.6)$$

$$\approx -(2\varepsilon_S kTP_{IM})^{1/2}(U_S-0)^{1/2} \text{ Compact Model Simplification} \quad (222.7)$$

we have the following threshold-voltage and surface-potential solutions for the diffusion current.

Strong Inversion Range

$$I_{DIFFUSION} = (Z/L)D_{n\text{-}eff}[+Q_N(y=L) - Q_N(y=0)] \text{ General Solution} \quad (223.1)$$

Threshold-Voltage Model (223.2)

$$I_{DIFFUSION} = (Z/L)D_n\{C_o(kT/q)(U_{DB} - U_{SB}) + (2\varepsilon_S kTP_{IM})^{1/2} \times$$
$$[e^{-1}(2U_F+1+U_{DB})^{-1/2} + \sqrt{2}(2U_F-1+U_{DB})^{-1/2} + (2U_F-3+U_{DB})^{1/2}$$
$$-e^{-1}(2U_F+1+U_{SB})^{-1/2} - \sqrt{2}(2U_F-1+U_{SB})^{-1/2} - (2U_F-3+U_{SB})^{1/2}]\}$$
$$\approx (Z/L)D_n\{C_o(kT/q)(U_{DB} - U_{SB}) + (2\varepsilon_S kTP_{IM})^{1/2} \times$$
$$[(e^{-1}+\sqrt{2})(2U_F+U_{DB})^{-1/2} + (2U_F+U_{DB})^{1/2}$$
$$-(e^{-1}+\sqrt{2})(2U_F+U_{SB})^{-1/2} - (2U_F+U_{SB})^{1/2}]\} \quad (223.2A)$$

Surface-Potential Model

$$I_{DIFFUSION}$$
$$= (Z/L)D_n\{C_o(kT/q)(U_{SL} - U_{S0}) + (2\varepsilon_S kTP_{IM})^{1/2}[(U_{SL}-1)^{1/2} - (U_{S0}-1)^{1/2}]\} \quad (223.3)$$
$$= (Z/L)D_n\{C_o(kT/q)(U_{SL} - U_{S0}) + (2\varepsilon_S kTP_{IM})^{1/2}[(U_{SL}-D_L)^{1/2} - (U_{S0}-D_0)^{1/2}]\} \quad (223.3A)$$

Weak Inversion or Subthreshold Range

In this range, $|-Q_B| >> |-Q_N|$, so a direct evaluation of the Q_N integral is more desirable than to compute the difference between two nearly equal terms, C_oV_{OX} and Q_B. Then using

$$(F_I)^2 = \{+[\exp(-U) + (+U - 1)]\exp(+U_{P\infty})$$
$$+ [\exp(+U) + (-U - 1)]\exp(-U_{P\infty}-\xi)\} \quad (21.24)$$
$$= \{+[\exp(-U) + (+U - 1)]$$
$$+ [\exp(+U) + (-U - 1)]\exp(-2U_{P\infty}-\xi)\}\exp(+U_{P\infty})$$

which reduces to the following for the weak inversion subthreshold range $0<U<U_S<2U_{P\infty}+\xi$

$$\approx \{+[\cancel{\exp(-U)} + (+U - 1)]$$
$$+ [\cancel{\exp(+U)} + \cancel{(-U - 1)}]\exp(-2U_{P\infty}-\xi)\}\exp(+U_{P\infty}) \quad (223.4)$$

Using $\partial_x U/\partial x = (L_{Di})^{-1}F_I$ we then have

$$Q_N = -q\int Ndx = -q\int n_i \exp(U-U_N)\,\partial_x U/(\partial_x U/\partial x) \quad (223.5)$$

$$= -qn_i L_{Di}\int \exp(U-U_N)\,\partial_x U/F_I$$

$$\approx -qn_i L_{Di}\exp(-U_{N0}-U_{P\infty}/2)\int \exp(U)\,\partial_x U/\sqrt{U} \quad (223.5A)$$

$$\approx -2qn_i L_{Di}\exp(-U_{N0}-U_{P\infty}/2)[\exp(U_S)-1] \quad (223.5B)$$

$$\approx -2qn_i L_{Di}\exp(-\xi_0-U_{P0}-U_{P\infty}/2)[\exp(U_S)-1] \quad (223.5C)$$

$$= -2qn_i L_{Di}\exp[-\xi_0-(3U_{P0}/2)+U_S][1-\exp(-U_S)] \quad (223.5D)$$

$$\approx -2qn_i L_{Di}\exp[-\xi_0-(3U_{P0}/2)+U_S] \quad (223.5E)$$

In the above, we used the single term approximation for F_I given by (223.4) to give (223.5A) which is then an exponential integral of the form $2\int\exp(Z^2)dZ$ with $Z=U^{1/2}$ which is evaluated approximately by a delta function at the interface, x=0. This result can also be obtained using the depletion solution of the Poisson equation, $\varepsilon\partial E_X/\partial x=\rho=q(P-N-P_{IM})\approx -qP_{IM}$, which gives a depletion layer or surface space-charge layer thickness of $X_{SC}=[2\varepsilon_s(kT/q)U_S/qP_{IM}]^{1/2}$. Using the exact iteration formula to give the surface potential in the depletion range, $0 \le U_S \le 2U_{P0}+\xi_0$, given by (21.40), which can be approximated as follows since U_{AA} and Δ_D are both about 1, and $U_{GX}-U_{FB}>>1$ also $2U_{P0}+\xi_0>2U_{P0}\sim 35>>1$ for $P_{IM} \sim 10^{17}$ to 10^{18} cm^{-3}. So the surface potential is proportional to $V_{GB}+A\sqrt{V_{GB}}$ and the diffusion current is proportional to $\exp[U_{GB}-U_{FB}\pm A\sqrt{(U_{GB}-U_{FB})}]$ showing the exponential slope of the I_D vs V_{GB} decreases from about 60mV per decade due to higher basewell impurity concentration and thicker oxide in the characteristic voltage $U_{AA} \propto P_{IM}\times X_O^2$ defined by

$$U_{AA}=qV_{AA}/kT=(q/kT)(\varepsilon_S qP_{IM}/2C_O^2)=(P_{IM}/10^{17})(X_O/10\text{nm})^2\times 69.59285\text{mV}/(kT/q)$$

in the square root dependent term shown below: $2[U_{AA}(U_{GX}-U_{FB})]^{1/2}$

$$U_S = U_{GX} - U_{FB} - 2U_{AA}\{[1+(U_{GX}-U_{FB}-\Delta_D)[1-\exp(-2U_{P0}-\xi_0)]/U_{AA}]^{1/2} - 1\}$$
$$\div[1-\exp(-2U_{P0}-\xi_0)] \qquad 0\le U_S\le U_{GX}-U_{FB}$$
$$\approx U_{GX} - U_{FB} - 2\{[(U_{AA})^2 + U_{AA}(U_{GX}-U_{FB}-\Delta_D)]^{1/2} - U_{AA}\}$$
$$\approx U_{GX} - U_{FB} - 2[U_{AA}(U_{GX}-U_{FB})]^{1/2} \qquad 0\le U_S\le U_{GX}-U_{FB} \quad (223.6A)$$
$$U_S \approx U_{GX} - U_{FB} + 2[U_{AA}(U_{GX}-U_{FB})]^{1/2} \qquad U_{GX}-U_{FB}\le U_S\le 2U_{P0}+\xi_0 \quad (223.6B)$$

This reproduces a general and well-known result since 1978-Brews [40], namely, in the subthreshold or weak inversion range, the surface potential is nearly constant along the channel between the source and the drain. Using this approximate solution of the surface potential, then

$$Q_N = -q\int Ndx$$
$$\approx -2qn_i L_{Di}\exp[-\xi_0-(3U_{P0}/2)+U_S] \quad (223.7A)$$
$$\approx -2qn_i L_{Di}\exp[-\xi_0-(3U_{P0}/2)+U_{GX} - U_{FB} + 2[U_{AA}(U_{GX}-U_{FB})]^{1/2}] \quad (223.7B)$$

$$I_{DIFFUSION} = (Z/L)D_{n\text{-}eff}\{-Q_N(y=0) - [-Q_N(y=L)\} \quad \text{General Solution} \quad (223.1)$$

$$= (Z/L)D_{n\text{-}eff}2qn_i L_{Di}\exp[-(3U_{P0}/2)]$$
$$\times [\exp(U_{SL}-U_{SB}) - \exp(U_{S0}-U_{DB}] \quad \text{Surface Potential Model} \quad (223.8A)$$

$$= (Z/L)D_{n\text{-}eff}2qn_i L_{Di}\exp[U_{GX} - U_{FB} + 2[U_{AA}(U_{GX}-U_{FB})-(3U_{P0}/2)]$$
$$\times [\exp(-U_{SB}) - \exp(-U_{DB})] \quad \text{Threshold Voltage Model} \quad (223.8B)$$

Note again the nearly constant surface potential along the channel between the source and the drain although the explicit difference is shown in the surface potential model (223.8A) to give more accurate solution using the exact iteration formula for the surface potential given by (21.37) and (21.39) if needed. The second feature is that the diffusion current indeed reaches a saturation, independent of the reverse-biased n+D/p-Base voltage, $U_{DB}>>1$ or $V_{DB}>>(kT/q)$.

23 General Surface Potential Models

(Drift and Diffusion Currents with Spatially Varying Bulk Impurity Charge)
(The Inversion Charge Model)

We shall just collect the results of the 2-term model for the drift current given by (222.9A) and diffusion current given by (223.3), (223.3A) and (223.8) in the previous two sections to give the 1-D surface model. The more general one including the longitudinal electric field gradient term, $\partial EY/\partial y$, and 2-D terms, derived in 1996-Sah [41] and analyzed in [42,43,45] will be just copied from these original articles for ease of referencing. In the two bulk-charge terms, the optimization parameter formulas [42,43] are also listed. In 1978-Brews model, these four parameters are chosen as 1 or kT/q. In threshold-voltage compact models, these optimization parameters are set to zero if theoretical results near $V_{DB}=V_{SB}=0$ in low P_{IM} region are need. In the surface-potential-based compact model applications, they are set to zero, but can be optimized [42,43] for table-lookup applications to extract experimental data.

$$I_{DRIFT} \approx (Z/L)\mu_{nS}(C_O/2)\left[(V_{GB}-V_{FBL}-V_{SL})^2-(V_{GB}-V_{FB0}-V_{S0})^2\right] \qquad (222.9A)$$
$$-(Z/L)\mu_{nS}(2/3)(2q\varepsilon_S P_{IM})^{1/2}\left[(V_{SL}-V_{\theta L})^{3/2}-(V_{S0}-V_{\theta 0})^{3/2}\right]$$

$$I_{DIFFUSION-INVERSION}$$
$$=(Z/L)D_n\{C_o(kT/q)(U_{SL}-U_{S0})+(2\varepsilon_S kTP_{IM})^{1/2}\left[(U_{SL}-1)^{1/2}-(U_{S0}-1)^{1/2}\right]\} \qquad (223.3)$$
$$=(Z/L)D_n\{C_o(kT/q)(U_{SL}-U_{S0})+(2\varepsilon_S kTP_{IM})^{1/2}\left[(U_{SL}-D_L)^{1/2}-(U_{S0}-D_0)^{1/2}\right]\} \qquad (223.3A)$$

$$I_{DIFFUSION-SUBTHRESHOULD}$$
$$=(Z/L)D_n(2qn_iL_{Di})\exp\left[-(3U_{P0}/2)\right]\times\left[\exp(U_{SL}-U_{SB})-\exp(U_{S0}-U_{DB})\right] \qquad (223.8A)$$

24 Four-Component 2-D Exact Solution

This was derived in 1996-Sah [41] and its 1-D form was tested in both the threshold voltage and surface-potential compact model forms with and without the four optimization parameters [42.,43,45] for the inversion range. As expected, excellent accuracy were observed. Testing for the subthreshold range is incomplete.

25 Latest Developments

Latest developments using the surface potential and inversion charge approaches are reported in the 10-author paper presented in this conference [55]. The aim of advanced compact modeling is to meet both the speed and the accuracy, the latter especially in analog applications when distortion and noise are of prominent importance in small-signal sinusoidal applications which demand accuracy of first, second, and third if not also higher derivatives. An approach using threshold voltage (the fastest), surface

potential (the most accurate), and regional charge (for derivatives) could be the best compromise to cover both digital and analog applications. One of the optimized approaches was recently proposed by Zhou [44,74] and some initial tests have been reported.

III. Summary

This keynote presentation has attempted to provide a history of the early developments of compact models of the MOS field-effect transistors for use in circuit simulators. The survey covers the start of the surface field-effect experiments on bare germanium surface in the earlier 1950's at Bell Telephone Laboratories under the direction of Shockley, Brattain and Bardeen, since the theoretical analyses and ideas of the surface field-effect experiments given by 1953-Brown-Shockley [18] and 1955-Garrett-Brattain [21] formed the bases from which the MOSFET compact model theory has evolved. The survey stopped at the end of the long channel or gradual channel theory, around 1980, since the short channel and 2-D effects have all been made as corrections and modifications of the long channel theory, while the long channel theory was and can be formulated with mathematical rigor with quantitative justification of the approximations. Furthermore, the early history covering this period (~1950 to ~1980) is one the current generations of semiconductor engineers are not familiar with due to lack of time, while doing the doctoral thesis and undertaking a career job, not only industrial but also academic, the latter rather alarming. It is with the hope of this review of the early history, based on the mathematical developments, however incomplete, that the latter could be helped to improve to some extent.

IV. Acknowledgement

It is impossible to write this historical review in about three months from nearly scratch without the helps from four active veteran investigators (Xing Zhou, Colin McAndrew, Gennady Gildenblat and Mitiko Miura-Mattausch). The first two, the author only got to know at the start of this project exactly four months ago when he accepted Xing Zhou's invitation to participate. It is an improbable task to survey 40+years of history in three months even the author had studied, became familiar with, and developed the device-physics underlying (rather skin-depth for a green Ph.D. educated in traveling-wave tubes) the three MOS transistor models (threshold voltage [34,35], inversion charge [35], and surface potential [36] forty some years ago, then dropped the subject after his first four doctoral students completed and left. The author thanks these four veteran colleagues deeply for bringing the author back into this exciting and most important integrated circuit engineering activity on MOS transistor compact modeling. It has been indeed a pleasure. Special thanks go to Xing Zhou for seeking me out, by email, and convincing me to give an invited talk on any subject that could help compact modeling developers, which turned into this first keynote of the Workshop on Compact Modeling he founded three years ago, and which became one on a history of MOS transistor compact modeling. I also thank Xing Zhou for helping me in nearly a hundred email exchanges (averaging one per day). The email communications with Dr. Colin McAndrew and Professors Gennady Gildenblat and Mitiko Miura-Mattausch and my studies of their numerous journal and conference articles on MOS transistor compact models have been indispensible in helping me to understand the bases of the compact models which they, our co-authors [55] and others have developed. I am also indebted to Bin B. Jie, without his presence as my postdoc, I would not have suggested the invited talk to him which he delivered on a 4-component-optimized MOST compact model at the ICSICT on October 18, 2004 in Beijing, where Xing Zhou noticed that I had not retired and decided to contact me to give an invited talk at the WCM. I would also like to mention the support of an Intel Research Grant during 1993-1996 on compact modeling, that was initiated by Dr.

Wallace W. L. Lin (a former PhD student of mine at Illinois) to get me familiar with this engineering subject. The Intel grant was monitored by Dr. Shiuh-Wuu Lee (Lin's manager of the Intel CAD group) after Wallace Lin transferred to another Intel technology group before the start of the grant. This inserted a 3-year activity into the 40-year gap, however, the engineering purposes and results were not evident during that 3-year pursuit. I would also like to thank Bin Jie for help to acquire the reprints and books cited in this manuscript since the copies I studied forty years ago were buried in file cabinets and I have not kept up with the later literature. I also thank Xing Zhou, Bin Jie and Fred Tsang for reading the drafts and making suggestions that clarified the descriptions in the earlier drafts, and Bin Jie for several thorough proof-readings of the sixth and final version. I also thank the WCM Proceedings editors and organizers (Matt, Bart, and Sarah) to give me repeatedly ~1-week delays to allow me to have the extra days to figure out how to present the essence of this vast literature of 40+year's of history on MOS transistor compact modeling. Still, I can only cover the early history before manpower-increase and overwhelming literature began in the early eighties.

References

[1] J. E. Lilienfeld, "Method and apparatus for controlling electric currents," U.S. Patent 1,745,175. Application filed October 8, 1926, granted January 18, 1930.

[2] J. E. Lilienfeld, "Device for controlling electric current," U.S. Patent 1,900,018. Application filed March 28, 1928, granted March 7, 1933.

[3] J. E. Lilienfeld, "Amplifier for electric currents," U.S. Patent 1,877,140. Application filed December 8, 1928, granted September 13, 1932.

[4] Oscar Heil, "Improvements in or relating to electrical amplifiers and other control arrangements and devices," British Patent 439,457, application filed March 4, 1935, granted December 6, 1935. Germany Convention Date, March 2, 1934.

[5] Chih-Tang Sah, "Evolution of the MOS Transistor - From conception of VLSI," Proceedings of the IEEE, 76(10), 1280-1326, October 1988.

[6] Chih-Tang Sah, et. al. "A history of MOS Transistor Compact Modeling (a tutorial exposition)" Planned for post-WCM-keynote distribution on internet and future posting in both English and Chinese.

[7] M. M. Atalla, M. Tannenbaum and E. J. Scheibner (Bell Telephone Laboratories), "Stabilization of silicon surface by thermally grown oxides," Bell System Technical Journal, 38(3), 123-140, May 1959.

[8] Chih-Tang Sah, ***Fundamentals of Solid-State Electronics***, 1001pp, 1991, World Scientific Publishing Company, Singapore. Reference to the page number for the relevant descriptions and figures are stated in the text. (**1991-Sah**)

[9] Chih-Tang Sah, ***Fundamentals of Solid-State Electronics-Study Guide***, 423pp, 1993, World Scientific Publishing Company, Singapore. Reference to the page number for the relevant descriptions and figures are stated in the text. (**1993-Sah**).

[10] Chih-Tang Sah, ***Fundamentals of Solid-State Electronics-Solution Manual***, 200pp, 1996, World Scientific Publishing Co., Singapore. Reference to the page number for the relevant descriptions and figures are stated in the text. Appendix "***Transistor Reliability***," pp.101-200. (**1996-Sah-FSSE-SM**).

[11] Narain Arora (Digital Equipment Corporation, Cadence), ***MOSFET Models for VLSI Circuit Simulation, Theory and Practice***, 605pp, Springer-Verlag, Wien, New York, 1993, Chapter 6, pp.230-324. (**1993-Arora**) Chapter 11 covers SPICE Diode Model and MOSFET Models Level 1 to 4.

[12] Daniel P. Foty, ***MOSFET Modeling with SPICE – Principles and Practice***, 653pp, Prentice Hall, Inc. Upper Saddle River, New Jersey, 1997. Chapters 5 to 12 cover Level 1-3, BSIM 1-3, HPSPICE28, and MOS9. Later models BSIM5, MOS11, HiSIM are covered in [55] presented at this conference.

[13] Yannis Tsividis (Columbia University), ***Operation and Modeling of the MOS Transistor***, 605pp, 1999. McGraw-Hill Book Company, New York. (**1999-Tsividis**) Specific SPICE transistor models are not discussed but criterias for compact models are given in Chapter 10.
[14] William Liu (Texas Instr.), ***MOSFET Models for Spice Simulation, Including BSIM3v3 and BSIM4***, 588pp, 2001, John Wiley & Sons, Inc. New York.
[15] Xing Zhou (Nanyang Technological University, Singapore.) A collection of references on compact modeling is continuously updated by Professor Xing Zhou and his staff at this website which was first set up after the 2002 Workshop on Compact Modeling. (Private communication, March 11, 2005.) http://www.ntu.edu.sg/home/exzhou/WCM/link.htm#Book (**2002-Zhou**)
[16] William Shockley (Bell Telephone Laboratories), ***Electrons and Holes in Semiconductors***, 558pp, 1950. D. Van Nostrand Co. New York. (**1950-Shockley**)
[17] W. Shockley (Bell Telephone Laboratories), "Theory of p-n junctions in semi-conductors and p-n junction transistors," Bell System Technical Journal, 28(7), 436-489, July 1949. (**1949-Shockley**)
[18] Walter L. Brown (Bell Telephone Laboratories, now with Lehigh University, Materials Science and Engineering Department), "n-type surface conductivity on p-type Ge," Physical Review 91(3), pp.518-527, August 1, 1953. Brown worked at the Bell Telephone Laboratories starting in 1950 while finishing Harvard's PhD requirements during 1945-1950 receiving the degree in 1951. As acknowledged by Brown at the end of this paper, this work was undertaken under the supervision of W. Shockley and contained many ideas and expressions originated from Shockley. This article will be referenced as **1953-Brown-Shockley**. This was the first article giving a three-dimensional view of the constant-potential-energy surface of electrons in an n/p/n transistor with a surface channel. See also [18a], [18b], [18c] listed below for the follow-up experimental results.
[18a] W. L. Brown, "Surface potential and surface charge distribution from semi-conductor field effect measurements," Physical Review 100, 590-591, October 15, 1955. This second paper by Brown showed that by matching to the theory, the experimental minimum and shape of the conductance versus applied gate voltage, the surface state density can be extracted. This was the first paper on obtaining the density of state of the surface and interface traps from experiments. Detailed results were reported in two later papers listed below [18b,18c].
[18b] H. C. Montgomery and W. L. Brown, "Field-induced conductivity changes in germanium," Physical Review 103, 865-870, August 15, 1956.
[18c] H. C. Montgomery, "Field effect in germanium at high frequencies," Physical Review 106(3), 441-445, May 1, 1957.
[19] W. Shockley (Bell Telephone Laboratories), "A unipolar field-effect-transistor," Proc. IRE, 40(11), 1365-1376, November, 1952. (**1952-Shockley-JFET**)
[20] R. C. Prim and W. Shockley, "Joining solutions at the pinch-off point in 'field-effect' transistor," Transactions of the IRE Professional Group on Electron Devices pp. 1-14, December 1953. This has no volume and issue number. The paper was printed in its original manuscript form without typesetting and submission and acceptance dates. It was the first paper of this 4th and last issue before the Transaction on Electron Devices was serialized in 1954. Very unfortunately it is not in the IEEE Electron Device Society Archival Collection 1954-2004 on DVD(8GB) sold at the IEDM-2004.
[21] C, G, B, Garrett and W. H. Brattain (Bell Telephone Laboratories) "Physical theory of semiconductor surfaces," Physical Review, vol. 99(2), 376-387, July 15, 1955. (**1955-Garrett**)
[22] R. H. Kingston (editor. MIT Lincoln Laboratory) ***Semiconductor Surface Physics***, 413pp, 1957, University of Pennsylvania Press, Philadelphia. Proceedings of the Conference on the Physics of Semiconductor Surfaces, Philadelphia, June 4-6, 1956.
[23] John L. Moll (Bell Labs then EE Professor, Stanford University), "Variable capacitance with large capacity change," IRE Wescon Convention Record, Part 3, pp.32-36, August 1959.

[24] W. G. Pfann and C. G. B. Garrett (BTL), "Semiconductor varactors using surface space-charge layers," Proc.IRE, 47(11), 2011-2012, November 1959.
[25] Daniel R. Frankl (General Telephone and Electronics, then Professor of Physics, Pennsylvania State University), "Some effects of material parameters on the design of surface space-charge-varactors," Solid-State Electronics, 2(1), 71-76, January 1961.
[25a] See also, ***Electrical Properties of Semiconductor Surfaces***, 310pp, 1967, Pergamon Press, Oxford, London, New York.
[26] Robert H. Kingston and Siegfried F. Neustadter (MIT Lincoln Laboratory), "Calculation of the Space Charge, Electric Field, and Free Carrier Concentration at the Surface of a Semiconductor," Journal of Applied Physics, 26(6), 718-720, June 1955. (**1955-Kingston-Neustadter**)
[27] R. Lindner (Bell Telephone Laboratories), "Semiconductor surface varactor," Bell System Technical Journal, XLI(3), 803-831,May 1962. Received October 19, 1961.
[28] Andrew S. Grove, Edward H. Snow, Bruce E. Deal and Chih-Tang Sah (Fairchild Semiconductor Research and Development Laboratory), "Simple physical model for the space-charge capacitance of metal-oxide-semiconductor structures," Journal of Applied Physics 35(8), 2458-2460, August 1964.
[29] Chih-Tang Sah (University of Illinois), ***Theory of the Metal Oxide Semiconductor Capacitor***, 139pp, appendix 21pp, Solid-State Electronics Laboratory Report No. 1, December 14, 1964, 4th Printing, February 1, 1974, University of Illinois, Urbana, Illinois. Copy may be obtained from the Engineering Library, University of Illinois at Urbana-Champaign or the author. More than 250 copies (100+50+50+50 for the four printing) of this 160-page report were printed and widely distributed in response to private and library requests and a mailing list has been kept.
[30] Lewis M. Terman (Stanford PhD thesis, IBM Research Laboratory), "An investigation of surface states at a silicon/silicon oxide interface employing metal-oxide-silicon diodes," Solid-State Electronics, 5(5), 285-299, September-October, 1962. Submitted 16 Oct. 1961, revised 8 Feb. 1962.
[31] Xing Zhou called Sah's attention in a November 30, 2004 email (copied below) to this negative number problem and pointed out to Sah the 2002 McAndrew description [32c]. See also references cited by Gildenblat [32a] and [32b], and McAndrew [32c].

From: Zhou Xing (Assoc Prof)
To: Chih-Tang Sah T41-UFL
Sent: Tuesday, November 30, 2004 12:43
Subject: Pao-Sah model near flat-band

Dear Professor Sah,

I'd like to consult with you on your original Pao-Sah (P-S) model (I think you're the best person to ask).

As you know, surface-potential (fs)-based model is now the trend, notably, fully iterative (HiSIM from Hiroshima, and MM11 from Philips) and fully explicit (SP from Penn State), which is derived after a few "analytical iterations" of the P-S implicit equation **[1]**. However, it is known to have convergence problem when Vgb is approaching Vfb (flat-band). This is well described by McAndrew's paper **[2]**, page 2, Eq. (2) which is your ***original*** model, and Eq. (6) which is setting "K=1" to avoid negative values in the sqrt() and its derivative for aiding Newton-Raphson iteration. Eq. (6) (with K=1) is now being cited as the basis of P-S and used in deriving fs (e.g., **[1]**). In the latest development (SP model) **[3]**, P-S model has even been "mathematically conditioned" to change the form.

My question to you is: why your original physically derived (drift-diffusion/Poisson) model had negative values inside a mathematical square-root function when fs is approaching +Vfb (i.e., very small +value of fs)? And what's your comment on these "conditionings" of the original P-S model?

We're pursuing unified ***regional*** charge-based approach, and already extended our model to full regions including polysilicon depletion, accumulation, inversion (the latter two have not been modeled in existing models), and forward-biased Vbs, as well as strained-Si MOSFETs.

Thanks,
Joe

For your convenience, I attach the three papers I referenced.
[1] 45sse01-chen.pdf
[2] 49ed01-mcandrew.pdf
[3] 51ed07-wu.pdf

[32a] T. L. Chen and G. Gildenblat (Pennsylvania State University), "Analytical approximation for the MOSFET surface potential," Solid-State Electronics 45(3), 335-339, March 2001.

[32b] Weimin Wu, Ten-Lon Chen, Gennady Gildenblat, and Colin C. McAndrew, "Physics-based mathematical conditioning of the MOSFET surface potential equation," IEEE Transaction on Electron Devices, 51(7), 1196-1200, July 2004.

[32c] Colin C. McAndrew and James J. Victory (Motorola, Freescale Semiconductor), "Accuracy of approximations in MOSFET charge models," IEEE TED-49(1), 72-81, January 2002.

[32d] J. R. Brews, "Comments on 'A new approach to the theory and modeling of IGFET's," IEEE Transactions on Electron Devices ED24(12), 1369-1370, December 1977.

[32e] Y. A. El-Mansy and A. R. Broothroyd, "Authors' reply to "Comments on 'A new approach to the theory and modeling of IGFET's." IEEE Trans. on Electron Devices ED25(3), 393-394, March 1978.

[33] Chih-Tang Sah, Robert N. Noyce, and William Shockley (Shockley Transistor Laboratory), "Carrier generation and recombination in p-n junctions and p-n junction characteristics," Proc. IRE, 45(9), 1228-1242, September 1957. Submitted March 23, 1957, revised May 13, 1957. Presented at the American Physical Society Winter Meeting, December 27, 1956 in Monterey, California. Abstracts i Bulletin American Physical Society, II, vol. 1, H9 and H10, p.382, December 27, 1956. (**1957-SNS**)

[33a] On the importance of minority carriers in the quasi-neutral boundary condition, see Sah's 1991-sophmore device-core-course textbook, FSSE [8], Equation (242.7) on page 189, and its 1993 Study Guide [9] Equations (242.7) and (242.8) to (242.9A) for n-type semiconductor and (242.10) to (242.11A) for p-type semiconductor. The 1957-SNS high-injection-level space-charge-neutral boundary condition given in the appendix of [33] was not analyzed in this 1991 first-edition of the sophmore device core-course textbook.

[33b] J. A. Hoerni, "Carrier mobilities at low injection level," Proc.IRE 46(2), 502, Feb.1958. There were also many mistakes made by later textbook and journal-article authors on using another incorrect asymptotic diffusivity at high levels of twice the low-level diffusivity. The correct high-level ambipolar diffusivity is $D_H=2D_nD_p/(D_n+D_p)$ and not $2D_p$ or $2D_n$ and lifetime $\tau_H=\tau_{n0}+\tau_{p0}$.

[34] Chih-Tang Sah (University of Illinois), "Characteristics of the Metal-Oxide-Semiconductor Transistor," IEEE Transaction on Electron Devices, ED-11(7), 324-345, July 1964. Received January 31, 1964. Accepted March 24, 1964. (**1964-Sah**). This and [35] had the most complete description of new concepts and phenomena in MOST based on device physics. It also included all the intrinsic small-signal capacitances. My count came up with some 30 odd concepts, most of which were not discussed in depth by authors of previous papers.

[35] Chih-Tang Sah and Henry Pao (University of Illinois), "The effects of fixed bulk charge on the characteristics of metal-oxide-semiconductor transistors," IEEE Transactions on Electron Devices, ED-13(4), 393-409, April 1966. Received Aug. 24, 1965. Accepted Nov. 1, 1965. (**1965-Sah-Pao**)

[36] Henry C. Pao and Chih-Tang Sah (University of Illinois. EE PhD thesis of Pao), "Effects of diffusion current on characteristics of metal-oxide (insulator)-semiconductor transistors," Solid-State Electronics, 9(10), 927-937, October, 1966. Received 4 April 1966. Accepted 10 May 1966. (**1966-Pao-Sah.**) The double integral formula has been known as the Pao-Sah formula and it served as the bench-mark for numerical accuracy of all subsequent MOS transistor compact models.

[37] Chih-Tang Sah (Fairchild Semiconductor), "A new semiconductor tetrode, the surface-potential controlled transistor," Proceedings of the IRE, 49(11), 1623-1634, November 1961. This and [38] were

the first two articles by Sah while he was learning MOS device physics.
[38] Chih-Tang Sah (Fairchild Semiconductor), "Effects of surface recombination and channel on p-n junction and transistor characteristics," IRE Transactions on Electron Devices, ED-9(1), 94-108, January 1962. This and [37] were the first two articles by Sah while learning MOS device physics.
[39] Chih-Tang Sah (University of Florida), "DCIV diagnosis for submicron MOS transistor design, process, reliability and manufacturing," Plenary Address. Proceedings of the 6th International Conference on Semiconductor and Integrated Circuit Technology, vol. 1, pp.1-15, October 22, 2001, Shanghai, China. IEEE Publication Catalog No. 01EX433.
[40] J. R. Brew (Bell Laboratories), "A charge-sheet model of the MOSFET," Solid-State Electronics 21(4), 345-355, April 1978. Submitted 7 May 1977, accepted 9 June 1977. (**1978-Brews**)
[41] Chih-Tang Sah (University of Florida), "Space Charge Theory of the MOS Transistor," Version 1.1, December 12. 1996; Version 2.1, 24pp, March 31, 1997. (**1996-Sah**)
[42] Bin B. Jie and Chih-Tang Sah (University of Florida), "Evaluation of a surface-potential-based bulk-charge model for MOS transistors," Submitted to and accepted by the IEEE Transaction on Electron Devices. (2003-Jie-Sah)
[43] Bin B. Jie and Chih-Tang Sah (University of Florida), "Physics-based exact analytical drain current equation and optimized compact model for long channel MOS transistors," Invited. Proceedings of the 7th International Conference on Solid-State and Integrated Circuits Technology, vol.2, 941-945, October, 2004, Beijing China. IEEE Publication Catalog No. 04EX863. (2003-Jie-Sah)
[44] Xing Zhou, Siau Ben Chiah, Karthik Chandrasekaran, Guan Huei See, Wangzuo Shangguan, Shesh Mani Pandey, Michael Cheng, Sanford Chu, and Liang-Choo Hsia, "Unified regional charge-based versus surface-potential-based compact modeling approaches," Proceedings of the 2005 Workshop on Compact Modeling, 2005 Nanotechnology Convention and Show, May 10, 2005. Publisher: NSTI.
[45] Bin B. Jie and Chih-Tang Sah (University of Florida), "Optimized compact MOS transistor model from the exact 4-compoent equation," Workshop on Compact Modeling, 2005 Nanotechnology Convention and Show, May 10, 2005, Anaheim, California. Publisher: NSTI. (2005-Jie-Sah)
[46] Dawon Kahng and M. M. Atalla (Bell Telephone Laboratories), "Silicon-Silicon dioxide field induced surface devices," IRE-AIEE Solid State Device Research Conference, June 1960.
[47] Dawon Kahng (Bell Telephone Laboratories), "Silicon-silicon dioxide surface device," Memorandum for file MH-2821DK--pg, 23pp, 10 figures. Bell Telephone Laboratories, Jan. 16, 1961. The MOSC theory is formulated with three layers from Garrett but did not quote the Garrett-Brattain 1955 paper [21]. However, when applied to the transistor, Kahng solved only the inversion I-V without bulk charge, giving the parabolic equation. He described the "pinch-off" range of I-V. According to a private communication from Kahng on February 17, 1988, Figures 1, 2, 3, 7, 8, 9 and Equations (23) and (24) were presented at the IRE-AIEE Solid State Device Research Conference in June 1960 [46]. See also D. Kahng, "A historical perspective on the development of MOS transistors and related devices," IEEE Transaction on Electron Devices, ED-23(7), 655-657, July 1976.
[48] Dawon Kahng (Bell Telephone Laboratories), "Electric field controlled semiconductor device," US Patent 3,102,230, Filed May 31, 1960. Issued August 27, 1963. Claims made include voltage regulator with forward-biased drain junction and current limiter with reverse-biased drain junction, in the conventional p/n/p p-channel MOST with doped n-type base.
[49] M. M. Atalla (Bell Telephone Laboratories), "Semiconductor triode," U. S. Patent 3,056,888, Filed August 17, 1960. Issued October 2, 1962. This patent claimed an n/π/n MOS transistor with intrinsic base region, operated under punch-through of the pi(π) base region by a large reverse drain voltage.
[50] John R. Brews, "Physics of the MOS transistor," in ***Silicon Integrated Circuits, Part A***, pp. 2-118, edited by Dawan Kahng, Academic Press, NY, 1981. (**1981-Brews**)
[51] Simon M. Sze, ***Physics of Semiconductor Devices***, 2nd Edition, Chapter 8, MOSFET, 868pp, John

Wiley & Sons, New York 1981. (**1981-Sze**)
[52] M. B. Barron (Stanford PhD thesis, GE Research Laboratory) "Low level currents in insulated gate field effect transistors," Solid-State Electronics 15(3), 293-302, March 1972. Received 24 March 1971, revised 11 August 1971. Starting from the 1966-Pao-Sah double integral, this author gave the first theory on subthreshold I-V characteristics including the diffusion current and he also gave correlation with experimental data.
[53] G. Baccarani, M. Rudan (Universita di Bologna, Italy) and G. Spadini (CRN, SGS-Ates, Milano, Italy), "Analytical i.g.f.e.t. model including drift and diffusion currents," IEE Journal on Solid-State and Electron Devices, 2(2), 62-68, March 1978. Received 25 October 1977, revised form 2 February 1978. Reference to 1978-Brews [40] as reference 13 on page 67.
[54] F. Van de Wiele (U. Catholique de Louvain, Belgium) "A long-channel MOSFET model," Solid-State Electronics 22(12), 991-997, December 1979. Received 9 April 1979, revised 20 June 1979.
[55] Josef Watts. Colin McAndrew, Christian Enz, Carlos Galup-Montoro, Gennady Gildenblat, Chenming Hu, Ronald van Langevelde, Mitiko Miura-Mattausch, Rafael Rios, and Chih-Tang Sah, "Advanced compact models for MOSFETs," Proceedings of the Workshop of Compact Modeling, Nanotechnology Conference and Trade Show, May 8-12, 2005, Anaheim, CA. Publisher: Nano Science and Technology Institute (NSTI).
[56] J. R. Schrieffer, "Effective carrier mobility in surface space-charge layers," Physical Review 97(3), 641-646, February 1, 1955.
[57] R. F. Pierret and C. T. Sah, "An MOS oriented investigation of the effective mobility theory," Solid-State Electronics 11(3), 279-290, March, 1968.
[58] O. Leistiko, A. S. Grove and C. T. Sah, "Electron and hole mobilities in the inversion layers of thermally oxidized silicon surfaces," IEEE Trans. on Electron Devices, ED12(5), 543-560, May 1965,
[59] T. H. Ning, L. L. Tschopp, and C. T. Sah, "The scattering of electrons by surface oxide charge and by lattice vibration at the silicon-silicon oxide interface," Surface Science 32(9), 561-575, Sept. 1972.
[60] S. C. Sun and James D. Plummer, "Electron mobility in inversion and accumulation layers on thermally oxidized silicon surfaces," IEEE Transactions on Electron Devices, ED-27(8), 1497-1508, August 1980.
[61] J. J. Sparkes and R. Beaufoy (British Telecom Research Ltd. Taplow, Berks, England), "The junction transistor as a charge controlled device," ATE J. pp.310-327, October 1957; Proceedings of IRE 45(12), 1740-1742, December 1957.
[62] E. O. Johnson and Albert Rose (RCA Laboratories), "Simple general analysis of amplifier devices with emitter, control, and collector functions," Proceedings of IRE 47(3), 407-418, March 1959.
[63] R. David Middlebrook (Stanford University, California Institute of Technology), "A modern approach to semiconductor and vacuum device theory," International Convention on Transistors and Associated Semiconductor Devices, 22nd May 1959; Written version received 28th July 1959; The Institution of Electrical Engineers, Paper No. 3180E, 887-902, March 1960.
[64] W. Shockley (Bell Telephone Laboratories), "Transistor electronics: imperfections, unipolar and analog transistors," Proc. IRE, 49(11), 1289-1313, November 1952.
[65] G. C. Dacey and I. M. Ross (Bell Telephone Laboratories), "Unipolar "field-effect" transistor," Proc. IRE 41(8), 970-979, August 1953
[66] G. C. Dacey and I. M. Ross BSJT, "The field effect transistor," Bell System Technical Journal XXXIV(6), 1149-1189, November 1955.
[67] W. Shockley and R. C. Prim, "Space-charge limited emission in semiconductors," Physical Review 90(5), 753-758, June 1, 1953,
[68] S. R. Hofstein and F. P. Heiman (RCA Laboratories), "The silicon insulated-gate field-effect transistor," Proc. IEEE, 51(9), 1190-1202, September, 1963.

Introduced the two types of MOSFETs: the induced channel type and the doped channel. Explained the induced channel only works in the enhance mode, while the doped channel can operate in both the depletion mode and the enhancement mode. The theory is distributed capacitance-resistance model; no bulk charge was included, resulting in only the simple parabolic IV characteristics. Give explanation of pinch off, but explanation based on drain fringe electric field given for the saturation resistance for thin-film transistor on insulator substrate, not correct and not applicable to inversion channel or enhancement mode transistor structure but applicable to depletion mode doped channel transistor that works like a JGFET.

[69] H. K. J. Ihantola (Stanford EE PhD thesis), "Design theory of a surface-field-effect transistor," Stanford Electronics Laboratory Report No. 1661-1, September 1961.

[70] H. K. J. Ihantola (Institute of Technology, Helsinki, Finland) and J. L. Moll (Stanford Electronics Laboratories), "Design Theory of a Surface Field-Effect Transistor," Solid-State Electronics, 7(4), 423-430, April 1, 1964. Submitted September 23, 1963. Accepted December 16, 1963. No subthreshold and no diffusion. Bulk charge term included. Input capacitance and output conductance were computed.

[71] Chih-Tang Sah (University of Illinois) "The spatial variation of the quasi-Fermi potentials in p-n junctions," IEEE Transaction on Electron Devices, ED-13, 839-846, December 1966.

[72] L. Nagel and D. O. Pederson (UC Berkeley), ***SPICE (Simulation Program with Integrated Circuit Emphasis)***, Memorandum ERL-M382, April 12, 1973, University of California, Electronics Research Laboratory, Berkeley, California

[73] L. W. Nagel, ***SPICE2 A Computer Program to Simulate Semiconductor Circuits***, Ph.D. dissertation, May 1975, University of California, Department of Electrical Engineering, Berkeley.

[74] S. B. Chiah, X. Zhou, K. Chandrasekaran, W. Shangguan, G. H. See, S. M. Pandey, "Single-piece polysilicon accumulation/depletion/inversion model with implicit/explicit surface-potential solutions," Submitted to Applied Physics Letters, 23 November 2004.