UNIVERSITAT POLITÈCNICA DE CATALUNYA

Departament d'Enginyeria Electrònica

SIMULACIÓN MONTE CARLO DE TRANSISTORES BIPOLARES DE HETEROUNIÓN ABRUPTA (HBT)

Autor: Pau Garcias Salvà
Director: Lluís Prat Viñas

2. Modelo de arrastre-difusión ampliado para HBTs abruptos

El objetivo de este capítulo es ofrecer una descripción suficientemente detallada del modelo utilizado para obtener la solución inicial del proceso iterativo en que se basa el simulador Monte Carlo (MC) de dispositivos. Asimismo, esta solución se utiliza como referencia para comparar y discutir los resultados obtenidos por el simulador MCHBT.

El simulador numérico HBTSIM, basado en el modelo de arrastre-difusión ampliado para heterouniones, ha sido desarrollado por el grupo de investigación de dispositivos semiconductores (GDS) del Departament d'Enginyeria Electrònica de la Universitat Politècnica de Catalunya (DEE-UPC). Este simulador ha servido de soporte de una serie de publicaciones sobre HBTs, que pueden ser una referencia útil para ampliar su descripción [LópezG,1996], [LópezG,1997], [LópezG,1998].

El contenido de este capítulo se estructura de la siguiente forma. En el apartado 2.1 se describe el modelo de transporte en las regiones graduales del dispositivo, es decir, aquellas regiones en las que las variaciones en los niveles de energía son suficientemente suaves o graduales. En el apartado 2.2 se describe el modelo de transporte en las interfaces donde el cambio es abrupto. El procedimiento numérico para resolver las ecuaciones diferenciales planteadas en estos modelos se resume en el

apartado 2.3 . Para una descripción más completa de los modelos anteriores puede consultarse [LópezG,1994].

En el apartado 2.4 se presenta un modelo analítico simplificado que permite una interpretación física de los resultados obtenidos con el modelo numérico HBTSIM.

Finalmente, en el apartado 2.5 se discuten, a partir de la ecuación de transporte de Boltzmann, los límites de validez de los modelos basados en los mecanismos de arrastre-difusión. Con ello se justifica la necesidad de desarrollar modelos más exactos para la simulación de estos dispositivos. El modelo de MC que se desarrolla en los capítulos posteriores es una de las posibles alternativas para satisfacer esta necesidad.

2.1 Transporte en regiones graduales

Se define una región gradual como aquella en la que las variaciones en los niveles de energía son suaves. En concreto, cuando se cumpla que:

$$\frac{\Delta E}{\Delta x} << \frac{k_B T}{\ell} \tag{2.1}$$

donde ΔE es la variación del nivel de energía E en el intervalo Δx, k_B la constante de Boltzmann, T la temperatura en Kelvin y ℓ el camino libre medio [Berz,1985]. Esta situación suele darse en las regiones volumétricas de emisor, base y colector del HBT, aunque la composición de los materiales pueda variar gradualmente. Los cambios abruptos en los niveles de energía suelen tener lugar en las interfaces emisor-base y/o base-colector, en cuyo caso este modelo no sería válido.

El modelo que se presenta en este apartado se basa en el trabajo de Lundstrom y Schuelker [Lundstrom, 1983] y recoge y unifica diversas aportaciones teóricas que permiten extender el modelo convencional de arrastre-difusión a materiales que presentan una estructura de bandas no uniforme [Sutherland,1977], [Fitchner,1983],

[Yokoyama,1984], [Pejcinovic,1989], [Gray,1985], [Tait,1989], [Marty,1977], [Marshak,1984].

Considérese la Figura 2. 1, que muestra el diagrama de bandas de energía de una región gradual. Se observa que la composición del material varía a lo largo del dispositivo puesto que ni la anchura de la banda prohibida $E_g(x)$ ni la afinidad electrónica $q\chi(x)$ se mantienen constantes al recorrer el dispositivo. En estos casos es habitual definir unos niveles o valores de referencia, que en el caso de la figura son: E_o para el nivel de vacío E_l, $q\chi_o$ para la afinidad electrónica y E_{go} para la banda prohibida. La variable $\phi(x)$ de la figura representa el potencial electrostático.

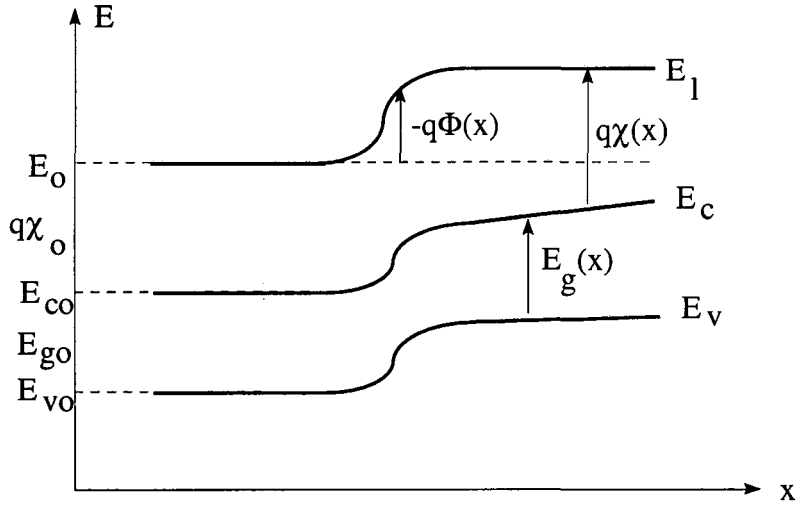

Figura 2. 1 Niveles de energía en un semiconductor gradual.

La modelización numérica de un dispositivo en régimen permanente requiere la resolución de tres ecuaciones básicas [Sze,1982]: la ecuación de Poisson,

$$\frac{d}{dx}\left[\varepsilon(x)\cdot\frac{d\phi(x)}{dx}\right]=-q\cdot\left[p(x)-n(x)+N_d^+(x)-N_a^-(x)\right] \tag{2. 2}$$

la ecuación de continuidad de electrones,

$$\frac{dJ_n(x)}{dx}=q\cdot R \tag{2. 3}$$

y la ecuación de continuidad de huecos,

$$\frac{dJ_p(x)}{dx} = -q \cdot R \tag{2.4}$$

donde $\phi(x)$ es el potencial macroscópico, ε la constante dieléctrica del medio, $n(x)$ y $p(x)$ las concentraciones de electrones y huecos, $N^+_d(x)$ y $N^-_a(x)$ las densidades de impurezas ionizadas donadoras y aceptadoras, R la velocidad neta de recombinación de portadores y, finalmente, $J_n(x)$ y $J_p(x)$ son las densidades de corriente para cada tipo de portador.

La velocidad neta de recombinación de portadores se puede descomponer en términos que reflejan los diferentes mecanismos físicos que contribuyen en el proceso de generación-recombinación:

$$R = R_{SRH} + R_A + R_{BB} \tag{2.5}$$

donde R_{SRH} representa la recombinación debida al mecanismo de recombinación por centros de impureza, también denominados recombinación Shockley-Read-Hall,

$$R_{SRH} = \frac{np - n_{ie}^2}{\tau_n(p + p_T) + \tau_p(n + n_T)} \tag{2.6}$$

R_A representa la recombinación Auger,

$$R_A = (np - n_{ie}^2) \cdot (C_{An} n + C_{Ap} p) \tag{2.7}$$

y R_B representa la recombinación radiativa o banda-banda:

$$R_{BB} = C_{BB} \cdot (np - n_{ie}^2) \tag{2.8}$$

En estas ecuaciones n_{ie} representa la concentración intrínseca efectiva, τ_n (τ_p) el tiempo de vida medio de los electrones (huecos) en la recombinación SRH, C_{An} y C_{Ap} son los coeficientes Auger para cada tipo de portador y C_{BB} el coeficiente de recombinación radiativa. Por su parte, n_T representa la concentración de electrones que habría en la banda de conducción (o, análogamente, huecos en la de valencia en el caso de p_T) si el nivel de Fermi coincidiese con el nivel de energía del centro de recombinación. De

forma similar se pueden añadir otras tasas de recombinación neta para incluir mecanismos adicionales, como por ejemplo la recombinación superficial y la interfacial.

En el modelo de arrastre y difusión para regiones graduales, las densidades de corriente de electrones y huecos, $J_n(x)$ y $J_p(x)$, de las ecuaciones (2. 3) y (2. 4) , se expresan como:

$$J_n(x) = q \cdot \mu_n \cdot n \cdot \frac{d(E_c - \Theta_n)/q}{dx} + k_B \cdot T \cdot \mu_n \cdot \frac{dn(x)}{dx} \tag{2. 9}$$

$$J_p(x) = q \cdot \mu_p \cdot p \cdot \frac{d(E_v + \Theta_p)/q}{dx} - k_B \cdot T \cdot \mu_p \cdot \frac{dp(x)}{dx} \tag{2. 10}$$

Estas ecuaciones muestran que las corrientes están constituidas por un término de arrastre y otro de difusión (primer y segundo término, respectivamente, del segundo miembro de las igualdades anteriores).

Obsérvese que el campo efectivo que arrastra a los electrones, dado por:

$$\frac{d(E_c - \Theta_n)/q}{dx} \tag{2. 11}$$

es distinto, en principio, del que arrastra a los huecos:

$$\frac{d(E_v + \Theta_p)/q}{dx} \tag{2. 12}$$

La existencia de un campo efectivo distinto para cada tipo de portador es un ejemplo ilustrativo de la *ingeniería del band gap.*

En las ecuaciones anteriores, Θ_n y Θ_p son términos que incorporan los efectos de las variaciones espaciales de masa eficaz ($N_c(x)/N_{co}$), la estadística de Fermi-Dirac ($F_{1/2}(\eta_c)/e^{(\eta c)}$) y la no parabolicidad de las bandas(α; -véase el capítulo siguiente-). La expresión para Θ_n es:

$$\Theta_n = k_B \cdot T \cdot \ln\left(\frac{N_c}{N_{co}}\right) + k_B \cdot T \cdot \ln\left[\frac{F_{1/2}(\eta_c)}{e^{\eta_c}} + \frac{3}{2} k_B \cdot T \cdot \alpha \frac{F_{3/2}(\eta_c)}{e^{\eta_c}}\right] \tag{2. 13}$$

donde $F_m()$ son las funciones de Fermi de orden m, α es el coeficiente de no-parabolicidad de las bandas, y η_c se calcula a partir del cuasinivel de Fermi para los electrones, E_{fn}, como:

$$\eta_c = \left(\frac{E_{fn} - E_c}{k_B \cdot T}\right) \tag{2. 14}$$

Cuando $N_c=N_{co}$, las bandas son parabólicas ($\alpha=0$) y la aproximación de Boltzmann es válida ($F_{1/2}(\eta_c)=e^{(\eta c)}$), entonces resulta que $\Theta_n=0$. En este caso, el campo eléctrico efectivo para los electrones se reduce a:

$$\frac{1}{q} \cdot \frac{dE_c}{dx} \tag{2. 15}$$

Pero, tal como se muestra en la Figura 2. 1:

$$E_c(x) = E_l(x) - q\chi(x) = E_0 - q\phi(x) - q\chi(x) \tag{2. 16}$$

por lo que resulta que si la afinidad electrónica es constante el campo eléctrico efectivo para los electrones se reduce a la expresión convencional:

$$\frac{1}{q} \cdot \frac{dE_c}{dx} = -\frac{d\phi}{dx} \tag{2. 17}$$

Análogamente ocurre para el campo eléctrico efectivo para los huecos.

Resolver el dispositivo significa determinar las tres variables fundamentales: $n(x)$, $p(x)$ y $\phi(x)$. Estas variables podrían ser determinadas a partir de las ecuaciones (2. 2), (2. 3) y (2. 4). Sin embargo, la aparición del cuasinivel de Fermi en (2. 14) hace más conveniente determinar previamente los potenciales $\phi_n(x)$, $\phi_p(x)$ y $\phi(x)$ y, a partir de estos, hallar finalmente las concentraciones de portadores $n(x)$ y $p(x)$.

El cuasipotencial de Fermi para electrones, $\phi_n(x)$, se define como:

$$\phi_n(x) = -\left(\frac{E_{fn} - E_{fio}}{q}\right) \tag{2. 18}$$

y el de huecos como:

$$\phi_p(x) = -\left(\frac{E_{fp} - E_{fio}}{q}\right) \tag{2. 19}$$

donde E_{fi0} es el nivel de Fermi intrínseco de referencia y las variables E_{fn} y E_{fp} son los cuasiniveles de Fermi para electrones y huecos, respectivamente. Las concentraciones y corrientes vistas en las ecuaciones anteriores, pueden reexpresarse en función de los cuasipotenciales y del potencial térmico $V_T=k_BT/q$ de la siguiente forma:

$$n = n_{ien} \cdot e^{\left(\frac{\phi - \phi_n}{V_T}\right)} \tag{2. 20}$$

$$p = n_{iep} \cdot e^{\left(\frac{\phi_p - \phi}{V_T}\right)} \tag{2. 21}$$

$$J_n = -q \cdot \mu_n \cdot n \cdot \frac{d\phi_n}{dx} \tag{2. 22}$$

$$J_p = -q \cdot \mu_p \cdot p \cdot \frac{d\phi_p}{dx} \tag{2. 23}$$

En esta nueva formulación, n_{ien} y n_{iep} son las concentraciones intrínsecas efectivas para electrones y huecos, respectivamente. Su valor viene dado por:

$$n_{ien} = n_{i0} \cdot e^{\left(\frac{\Theta_n + q(\chi - \chi_0)}{k_B T}\right)} \tag{2. 24}$$

$$n_{iep} = n_{i0} \cdot e^{\left(\frac{\Theta_p - q(\chi - \chi_0) - (E_g - E_{g0})}{k_B T}\right)}$$

(2. 25)

donde n_{i0} es la concentración intrínseca de referencia, definida como:

$$n_{i0} = N_{c0} \cdot e^{\left(\frac{E_{fi0} - E_{co}}{k_B T}\right)} = N_{v0} \cdot e^{\left(\frac{E_{v0} - E_{fio}}{k_B T}\right)}$$

(2. 26)

Obsérvese que, a partir de (2. 20) y (2. 21) se tiene que:

$$n \cdot p = n_{ien} \cdot n_{iep} \cdot e^{\left(\frac{\phi_p - \phi_n}{V_T}\right)} = n_{ie}^2 \cdot e^{\left(\frac{\phi_p - \phi_n}{V_T}\right)}$$

(2. 27)

por lo que en equilibrio térmico, cuando $\phi_n(x)=\phi_p(x)$, la concentración intrínseca efectiva es:

$$n_{ie}^2 = n_{ien} \cdot n_{iep} = n_{i0}^2 \cdot e^{\left(\frac{\Theta_p + \Theta_n}{K_B T}\right)} \cdot e^{\left(\frac{E_g - E_{g0}}{K_B T}\right)}$$

(2. 28)

Cuando el dopado toma valores muy altos se produce una reducción de E_g. Esta reducción es debida a un desplazamiento simultáneo de los niveles E_c y E_v, que provocan un incremento de la afinidad $q\chi$ y una reducción de E_g: [LópezG,1997]:

$$\Delta(q\chi) = -\Delta E_c^{BGN}$$

(2. 29)

$$\Delta E_g^{BGN} = \Delta E_c^{BGN} + \Delta E_v^{BGN}$$

(2. 30)

Estas variaciones de los niveles E_c y E_v debidas a los efectos de alto dopado (efectos conocidos como *band gap narrowing, BGN*) se pueden evaluar a través del modelo de Jain-Roulston [Jain,1991], [LópezG,1997]:

$$\Delta E_c^{BGN} = C_1 \cdot \left(\frac{N}{10^{18}}\right)^{1/\alpha} + C_2 \cdot \left(\frac{N}{10^{18}}\right)^{1/2} \tag{2. 31}$$

$$\Delta E_v^{BGN} = C_3 \cdot \left(\frac{N}{10^{18}}\right)^{1/\beta} + C_4 \cdot \left(\frac{N}{10^{18}}\right)^{1/2} \tag{2. 32}$$

Nótese que este fenómeno afecta a las concentraciones intrínsecas de las ecuaciones (2. 24), (2. 25) y (2. 28):

$$n_{ien} = n_{i0} \cdot e^{\left(\frac{\Theta_n + \Delta E_c^{BGN} + q(\chi_l - \chi_0)}{k_B T}\right)} \tag{2. 33}$$

$$n_{iep} = n_{i0} \cdot e^{\left(\frac{\Theta_p - q(\chi_l - \chi_0) - (E_{gl} - E_{g0}) + \Delta E_v^{BGN}}{k_B T}\right)} \tag{2. 34}$$

donde el subíndice *l* indica bajo nivel de dopado y se ha ignorado la no-parabolicidad de las bandas ($\alpha=0$). Los parámetros Θ_n y Θ_p contienen la influencia de la estadística de Fermi y de la densidad efectiva de estados. Substituyéndolos por sus ecuaciones (2. 13) se obtiene:

$$n_{ie}^2 = n_{i0l}^2 \cdot \frac{N_c}{N_{c0}} \cdot \frac{F_{1/2}(\eta_c)}{e^{\eta_c}} \cdot e^{\left(\frac{\Delta E_g^{BGN}}{K_B T}\right)} \tag{2. 35}$$

La concentración intrínseca efectiva, por tanto, se ve afectada por la variación de la masa efectiva (N_c/N_{c0}), la influencia de la estadística de Fermi y el estrechamiento efectivo de la banda prohibida.

2.2 Transporte en interfaces de heterounión abruptas

En contraste con las regiones graduales tratadas a lo largo del apartado anterior, se considera que una región de semiconductor es abrupta cuando la variación de un nivel de energía ΔE cumple que:

$$\frac{\Delta E}{\Delta x} > \frac{k_B T}{\ell} \tag{2. 36}$$

En este caso, el modelo de arratre-difusión desarrollado para regiones graduales no es válido [Berz,1985].

En las regiones abruptas el modelo de transporte se basa en la emisión termoiónica ampliado con la transmisión túnel [Crowell,1969], [Stratton,1962], [Chang,1970], [Grinberg,1984]. Considérese la Figura 2. 2:

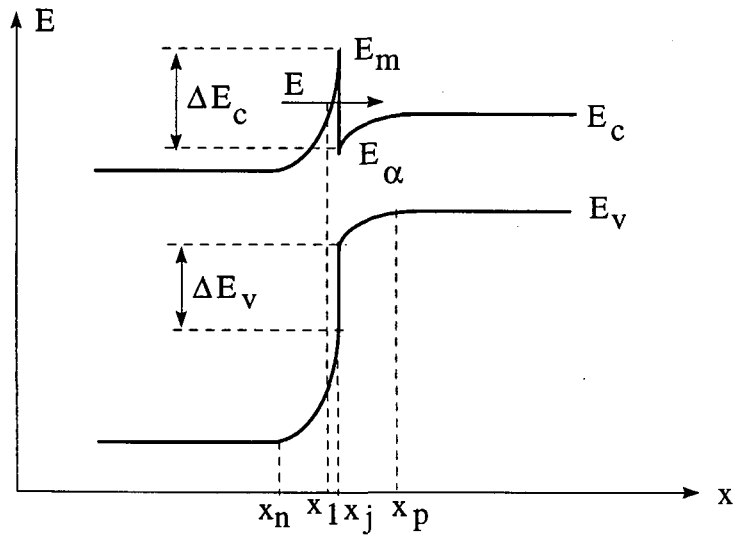

Figura 2. 2 Niveles de energía en una interfaz abrupta.

La corriente de electrones a través de esta interfaz es:

$$J_n(x_j) = q \cdot \sqrt{\frac{K_B T}{2\pi m^*}} \cdot \frac{N_c}{k_B T} \int_{E_\alpha}^{\infty} T(E) \cdot \ln\left[1 + e^{-\left(\frac{E + E_c - E_F}{k_B T}\right)}\right] \cdot dE \tag{2. 37}$$

donde m^* es la masa efectiva del electrón y $T(E)$ es el coeficiente de transmisión cuántico, que debe hallarse resolviendo la ecuación de Shrödinger. Usando la aproximación WKB [Chang,1970], [Das,1988], el coeficiente de transmisión viene dado por:

$$T(E) = 1 \qquad si \quad E \geq E_m \tag{2. 38}$$

$$T(E) = e^{-\frac{2}{\hbar}\sqrt{2m^*}\int_{x_1}^{x_2}\sqrt{qV(x)-E}\cdot dx} \qquad si \quad E < E_m \tag{2. 39}$$

donde $qV(x)$ es la energía potencial (E_c) y x_1, x_2 son las intersecciones del nivel de energía E del electrón con $E_c(x)$; en concreto, $x_2 = x_j$. Tal como se observa en la primera ecuación, la aproximación WKB ignora la reflexión cuántica para energías del electrón superiores a la de la barrera de potencial.

Suponiendo que la barrera sea parabólica [Crowell,1969] la ecuación (2. 39) se puede resolver analíticamente y se convierte en:

$$T(E) = e^{-\Gamma(E)} \qquad si \quad E < E_m \tag{2. 40}$$

con:

$$\Gamma(E) = \frac{E_m}{E_{oo}}\left[\sqrt{1-\alpha} - \alpha\cdot\ln\left(\frac{1+\sqrt{1-\alpha}}{\sqrt{\alpha}}\right)\right] \tag{2. 41}$$

$$\alpha = \frac{E}{E_m} \tag{2. 42}$$

$$E_{oo} = \frac{q\hbar}{2}\sqrt{\frac{N}{\varepsilon\cdot m^*}} \tag{2. 43}$$

A partir de (2. 37) y usando esta aproximación, Grinberg [Grinberg,1984] demostró que la corriente de electrones a través de la interfaz puede expresarse como:

$$J_n(x_j) = q \cdot v_{neff} \cdot \left[\frac{N_c(x_j^+)}{N_c(x_j^-)} \cdot n(x_j^+) - n(x_j^-) \cdot e^{-\Delta E_c / k_B T} \right] \tag{2.44}$$

donde v_{neff} es la velocidad efectiva de los electrones, dada por:

$$v_{neff} = \sqrt{\frac{k_B T}{2 \cdot m^*}} \cdot (1 + p_t) \tag{2.45}$$

siendo p_t el factor túnel,

$$p_t = \frac{1}{K_B T} \cdot e^{\left(\frac{E_m}{K_B T} \right)} \cdot \int_{E_\alpha}^{E_m} T(E) \cdot e^{-\left(\frac{E}{K_B T} \right)} \cdot dE \tag{2.46}$$

$n(x_j^+)$ y $n(x_j^-)$ son las concentraciones de electrones a ambos lados de la interfaz x_j y el término $N_c(x_j^+)/ N_c(x_j^-)$ toma en consideración las diferencias de masa efectiva entre las dos regiones.

De forma similar, la corriente de huecos a través de la interfaz se expresa por:

$$J_p(x_j) = q \cdot v_{peff} \cdot \left[p(x_j^-) - \frac{N_v(x_j^+)}{N_v(x_j^-)} \cdot p(x_j^+) \cdot e^{-\Delta E_v / k_B T} \right] \tag{2.47}$$

si bien en este caso el factor túnel p_t es nulo.

2.3 Simulador numérico HBTSIM

Las ecuaciones descritas en los apartados 2.1 y 2.2 son resueltas de forma numérica por el simulador HBTSIM. Los detalles de esta resolución numérica se exponen en la referencia [LópezG,1994]. A continuación se expone un resumen de las principales características del simulador.

Las tres ecuaciones diferenciales que describen el dispositivo (ecuación de Poisson (2. 2), ecuación de continuidad de electrones (2. 3), y ecuación de continuidad de huecos (2. 4)) junto con las ecuaciones complementarias (concentraciones de electrones y huecos (2. 20) y (2. 21), corrientes de electrones y huecos (2. 22), (2. 23), (2. 44) y (2. 47), y ecuaciones de recombinación (2. 5) a (2. 8)) forman un sistema de tres ecuaciones diferenciales de las variables $\phi(x)$, $\phi_n(x)$ y $\phi_p(x)$ (potencial electrostático y cuasipotenciales de Fermi para electrones y para huecos).

La resolución numérica del dispositivo consiste en hallar las funciones incógnita en un conjunto de N puntos del dispositivo. Mediante un proceso de discretización basado en "cajas finitas", se transforma el sistema de ecuaciones diferenciales en un sistema de ecuaciones algebraicas de 3N incógnitas: los potenciales ϕ^i, $\phi_n{}^i$ y $\phi_p{}^i$ en cada uno de los puntos de la discretización. Este sistema algebraico de 3N ecuaciones en 3N incógnitas, que es no lineal, se linealiza por el método de Newton y se resuelve el sistema algebraico lineal resultante por técnicas de descomposición matricial.

Hay dos aspectos peculiares en el modelo que conviene resaltar debido a su significado físico: la modelización de los contactos y la modelización del comportamiento dinámico del HBT.

En cuanto a la modelización de los contactos de emisor y de colector, se supone que los dos fuerzan la neutralidad de carga en la región del contacto y que en ellos la velocidad de recombinación es infinita. A partir de estas premisas se llega a las siguientes condiciones de contorno:

$$\phi(x_t) = \phi_0(x_t) + V_a \tag{2. 48}$$

$$\phi_n(x_t) = V_a \tag{2. 49}$$

$$\phi_p(x_t) = V_a \tag{2. 50}$$

donde x_t es el punto del contacto (emisor o colector) y V_a es la tensión de polarización aplicada al contacto.

El contacto de base se modela según la condición de contorno de Gummel [Gummel,1964]: el cuasipotencial de Fermi de los portadores mayoritarios se supone fijado por la tensión de polarización aplicada. Si x_b es la posición del contacto de base y V_b es la tensión de polarización de base en un transistor *npn*, esta condición se convierte en:

$$\phi_p(x_b) = V_b \tag{2. 51}$$

El comportamiento dinámico del HBT se ha modelado siguiendo la aproximación cuasiestacionaria propuesta por Gummel [Gummel,1969]. La frecuencia de ganancia unidad f_T se calcula como:

$$f_T = \frac{\Delta J_c}{2\pi \cdot \Delta Q_T} \tag{2. 52}$$

donde las variaciones incrementales de carga total acumulada en el dispositivo, ΔQ_T, y de corriente de colector, ΔJ_c, se calculan a partir de las soluciones de dos estados estacionarios correspondientes a polarizaciones muy próximas: una polarización inicial (V_{BE}, V_{CE}) y una polarización final ($V_{BE}+\Delta V_{BE}$, V_{CE}), siendo ΔV_{BE} una variación comparativamente pequeña frente al valor nominal o inicial de la polarización. De forma análoga se obtienen otros parámetros dinámicos del HBT.

2.4 Modelo analítico aproximado del HBT

El simulador numérico presentado en el apartado anterior proporciona las características eléctricas del dispositivo a partir de sus características físicas (geometría, dopados, etc.). Se trata, en definitiva, de un proceso similar al que ocurre en la fabricación de un dispositivo real: a partir de unos parámetros físicos de proceso (temperatura, tiempos, etc.) se obtiene un dispositivo que presenta unas determinadas características eléctricas. Es importante en ambos casos saber *a priori* a qué parámetros obedecen los resultados obtenidos a fin de poder controlar las características deseadas del dispositivo.

Para lograr el objetivo de "entender" el funcionamiento del dispositivo, se hace casi imprescindible disponer de un modelo simple que resalte a grandes rasgos el funcionamiento del dispositivo, aunque sea a costa de tener poca precisión en los valores estimados. Esta funcionalidad de interpretación física es el objetivo de los modelos analíticos.

El modelo analítico que hemos desarrollado para los HBTs [LópezG,1997], [LópezG,1999], se basa en el modelo de arrastre-difusión en las regiones graduales del dispositivo, y en la transmisión termoiónica y por efecto túnel en la heterounión abrupta. Su punto de partida consiste en establecer que la densidad de corriente de electrones a través de la interfaz, $J_n(x_j)$, debe ser igual (por continuidad en régimen permanente) a la que entra en la región neutra de la base, $J_n(x_p)$, más la que se pierde por recombinación en la región de carga espacial situada entre la interfaz y la región neutra, J_{nRD} (Figura 2.2):

$$J_n(x_j) = J_n(x_p) + J_{nRD} \tag{2. 53}$$

En este modelo analítico suponemos que $J_{nRD} << J_n(x_p)$, por lo que se simplifica la ecuación anterior:

$$J_n(x_j) \approx J_n(x_p) \tag{2. 54}$$

La corriente de electrones a través de la base es fácil de evaluar si se supone condición de baja inyección (el dopado típico de base en HBTs es del orden de $10^{19}cm^{-3}$), un perfil de dopado uniforme, campo eléctrico nulo y que en el extremo de colector de la región neutra de la base la velocidad de electrones es la de saturación, v_{nsat}. También se supone que la base es muy estrecha, de forma que no hay pérdidas por recombinación y la corriente de electrones a través de la misma se mantiene constante. A partir de estas hipótesis puede plantearse que:

$$J_n(x_p) = q \cdot D_n \cdot \frac{dn(x_p)}{dx} \approx q \cdot D_n \cdot \frac{n(x_p) - n(x_p + w_B)}{w_B} \tag{2. 55}$$

donde w_B es el espesor de la zona neutra de la base y x_p es el punto fronterizo con la zona de carga espacial de emisor. En el extremo de colector de la región neutra de la base se tiene que:

$$J_n(x_p + w_B) \approx J_n(x_p) \approx q \cdot n(x_p + w_B) \cdot v_{nsat} \tag{2. 56}$$

A partir de (2. 55) y (2. 56) obtenemos:

$$n(x_p + w_B) = n(x_p) \cdot \frac{(D_n / w_B)}{(D_n / w_B) + v_{nsat}} \tag{2. 57}$$

y substituyendo (2. 57) en (2. 55):

$$J_n(x_p) \approx q \cdot v_B \cdot n(x_p) \tag{2. 58}$$

$$\frac{1}{v_B} = \frac{1}{v_{nsat}} + \frac{1}{D_n / w_B} \tag{2. 59}$$

Es decir, la corriente de difusión a través de la base depende de la concentración en la coordenada x_p y de una velocidad efectiva de difusión v_B. De acuerdo con (2. 59), esta velocidad será siempre inferior a la más pequeña de sus componentes, v_{nsat} o (D_n/w_B). Para bases muy anchas, $(D_n/w_B) << v_{nsat}$, por lo que $v_B \approx (D_n/w_B)$; pero si el espesor de la base w_B se reduce suficientemente, (D_n/w_B) puede ser comparable a v_{nsat}, y la velocidad efectiva v_B será menor que (D_n/w_B).

La ecuación (2. 54) establece la igualdad entre $J_n(x_p)$ dada por (2. 58) y $J_n(x_j)$ dada por (2. 44). Ésta última es función de la concentración de electrones en x_j^+ y x_j^-. Si suponemos válida la aproximación de Boltzmann para los electrones en la zona de carga espacial:

$$q \cdot V_p = E_c(x_p) - E_c(x_j^+) = k_B T \cdot \ln \frac{n(x_j^+)}{n(x_p)} \tag{2. 60}$$

$$q \cdot V_n = E_c(x_j^-) - E_c(x_n) = k_B T \cdot \ln \frac{n(x_n)}{n(x_j^-)}$$

(2. 61)

donde V_p y V_n son las caídas de tensión en la zona de carga de espacio de la parte P (base) y de la parte N (emisor) de la unión BE del HBT. Evidentemente, $V_p+V_n=V_{bi}-V_{BE}$. Despejando $n(x_j^+)$ de (2. 60) y $n(x_j^-)$ de (2. 61), y substituyendo en (2. 44), se obtiene:

$$J_n(x_j) = q \cdot v_{neff} \cdot e^{V_p/V_T} \cdot e^{-\Delta E_c/k_B T} \cdot \left[n_0(x_p) \cdot e^{V_{BE}/V_T} - n(x_p) \right]$$

(2. 62)

Para llegar a esta expresión se ha hecho uso de que $J_n(x_j)=0$ en equilibrio térmico. La notación de (2. 62) puede simplificarse definiendo una velocidad efectiva de los electrones a través de la heterounión, v_s:

$$J_n(x_j) = q \cdot v_s \cdot \left[n_0(x_p) \cdot e^{V_{BE}/V_T} - n(x_p) \right]$$

(2. 63)

$$v_s = v_{neff} \cdot e^{(qV_p - \Delta E_c)/k_B T}$$

(2. 64)

Igualando (2. 63) con (2. 58) resulta que:

$$n(x_p) = \frac{n_0(x_p) \cdot e^{V_{BE}/V_T}}{1 + v_B/v_s}$$

(2. 65)

Esta última expresión contiene una interpretación física clave para explicar el comportamiento de los HBTs. Si la velocidad efectiva a través de la interfaz, v_s, es mucho mayor que v_B, la concentración de electrones en x_p toma el valor habitual de un BJT (el correspondiente al numerador de (2. 65)). Pero si $v_s<<v_B$, entonces resulta que la concentración en x_p es mucho menor, lo que significa que *el spike limita la inyección de electrones.*

La corriente de colector puede expresarse substituyendo (2. 65) en (2. 58):

$$J_c = J_n(x_p) = q \cdot \frac{v_B \cdot v_s}{v_B + v_s} \cdot n_0(x_p) \cdot e^{V_{BE}/V_T}$$

(2. 66)

Si v_s fuera muy superior a v_B, la corriente estaría limitada por la difusión de electrones a través de la base. Si denomináramos J_{nB} a esta corriente, resultaría que:

$$J_{nB} = q \cdot v_B \cdot n_0(x_p) \cdot e^{V_{BE}/V_T} \tag{2. 67}$$

Si por el contrario $v_B >> v_s$, la corriente estaría limitada por la transmisión a través del *spike*, J_{nS}:

$$J_{nS} = q \cdot v_S \cdot n_0(x_p) \cdot e^{V_{BE}/V_T} \tag{2. 68}$$

La corriente de colector (2. 66) puede, por tanto, expresarse en función de estos valores asintóticos:

$$\frac{1}{J_c} = \frac{1}{J_{nB}} + \frac{1}{J_{nS}} \tag{2. 69}$$

La ecuación (2. 69) pone de manifiesto que la corriente total de electrones a través del HBT es el resultado de *dos mecanismos restrictivos que actúan simultáneamente: la transmisión a través del spike y la difusión a través de la región neutra de la base.* El más restrictivo de los dos mecanismos es el que fija el valor final de la corriente de colector.

En la Figura 2. 3 se representa la corriente de colector en función de la tensión de polarización VBE para un HBT de InP/InGaAs. La curva ***a*** es el resultado obtenido por el simulador HBTSIM, mientras que la curva ***b*** corresponde al modelo analítico descrito por las ecuaciones (2. 68), (2. 66) y (2. 67). La aproximación analítica es buena para tensiones de polarización medias y bajas. Para tensiones superiores a 0.75V se observa una cierta discrepancia debida a que el modelo analítico se ha desarrollado sin considerar la caída de tensión en la región neutra de base. El campo eléctrico en esta región puede expresarse como [Prat,1990]:

$$E_{el} = \frac{J_p}{q \cdot \mu_p \cdot (N_B + n)} + \frac{k_B T}{q} \cdot \frac{1}{(N_B + n)} \cdot \frac{dn}{dx} \tag{2. 70}$$

que produce una caída de tensión ΔV_B en la región neutra de la base:

$$\Delta V_B = -\int_{x_p}^{x_p + w_B} E_{el} \cdot dx \tag{2. 71}$$

que debe ser sumada a V_{BE}. Con esta corrección se obtiene la curva ***c***, que se ajusta muy bien al resultado del simulador.

La Figura 2. 4 muestra la comparación entre J_{nB} , J_{nS} y la corriente de colector proporcionada por el modelo analítico. Como puede verse, para tensiones inferiores a 0.4V la corriente de colector queda determinada por J_{nB}, es decir, el mecanismo más restrictivo es la difusión a través de la base, mientras que para tensiones superiores a 0.55V el mecanismo que fija J_c es la transmisión a través del *spike*.

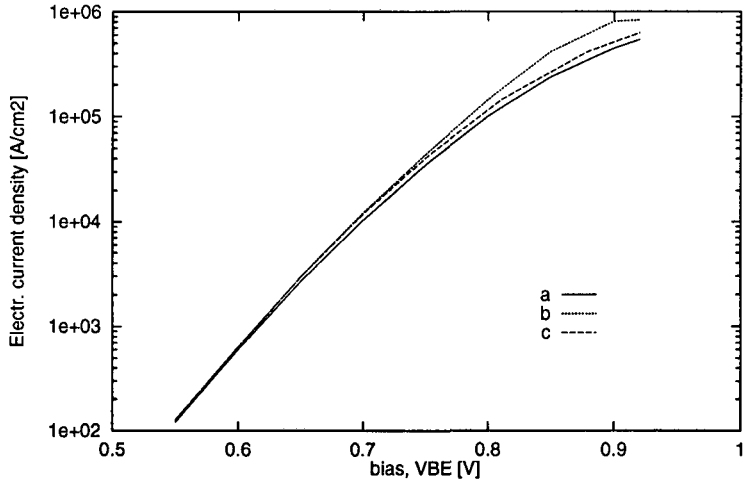

Figura 2. 3 Comparación entre el modelo numérico (curva *a*) y analítico (curva *b*; curva *c* : modelo analítico con la caída de tensión en la zona neutra de la base).

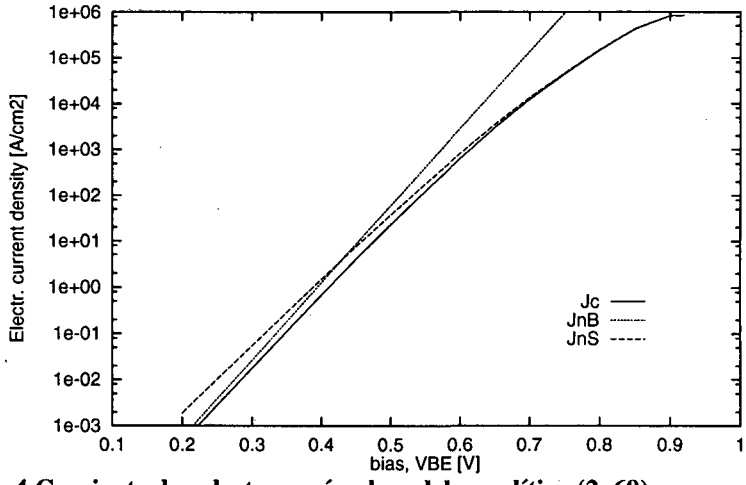

Figura 2. 4 Corriente de colector según el modelo analítico (2. 69) y sus componentes J_{nB} y J_{nS}.

En la Tabla 2. 1 se presentan las concentraciones de electrones en la región neutra de la base según datos obtenidos por tres modelos: el modelo numérico HBTSIM, el modelo analítico dado por las ecuaciones (2. 65) y (2. 57), y el modelo habitualmente aplicado para el BJT convencional (ecuaciones (2. 65) con $v_s \rightarrow \infty$ y (2. 57) con $v_{nsat} \rightarrow \infty$). Como puede observarse, el ajuste del modelo analítico con el numérico es muy bueno y ambos discrepan mucho del modelo convencional para el BJT. El efecto limitador del *spike* es muy acusado en la concentración en $n(x_p)$, y el efecto de la velocidad de saturación es también muy significativo en $n(x_p+w_B)$.

Modelo:	$n(x_p)$ [cm^{-3}]	$n(x_p+w_B)$ [cm^{-3}]
numérico	$2.82\ 10^{15}$	$1.29\ 10^{15}$
analítico	$2.63\ 10^{15}$	$1.35\ 10^{15}$
BJT convencional	$1.45\ 10^{18}$	≈ 0

Tabla 2. 1 Concentraciones de electrones en los extremos de la zona neutra de la base según diferentes modelos.

La corriente de base de los HBTs de InP/InGaAs está dominada por la recombinación en la región de la base [Seabury,1993], [Gee,1993] para tensiones de polarización V_{BE} medias-altas. Este hecho permite realizar un modelo analítico simple para la corriente de base [LópezG,1999]. En efecto,

$$J_B = q \cdot \int_{x_p}^{x_p + w_B} R \cdot dx$$

(2. 72)

Suponiendo que la distribución de electrones en la región neutra de base sea lineal, la ecuación anterior conduce a:

$$J_B = q \cdot \frac{w_B}{2} \cdot n_0(x_p) \cdot e^{V_{BE}/V_T} \cdot f_R \cdot \frac{1}{1+v_B/v_s} \cdot \left(\frac{v_{nsat} + 2D_n/w_B}{v_{nsat} + D_n/w_B} \right)$$

(2. 73)

$$f_R = C_{Ap} \cdot N_B^2 + (C_{SRH} + C_{BB}) \cdot N_B$$

(2. 74)

donde C_{Ap}, C_{SRH}, C_{BB}, son los coeficientes de recombinación debidos a los mecanismos Auger, Shockley-Read-Hall y banda-banda. La ecuación (2. 73) muestra que J_B también está afectada por la limitación en la inyección que representa el *spike* (factor v_B/v_s) y por el efecto de v_{nsat}. Si el *spike* no limita la inyección ($v_s >> v_B$) y $D_n/w_B << v_{nsat}$, la corriente de base toma el valor convencional de los BJTs con la salvedad que la recombinación es mucho mayor como consecuencia del alto dopado de base, que incrementa considerablemente la recombinación Auger.

La ganancia de corriente del HBT puede calcularse a partir de (2. 66) y (2. 73):

$$\beta = \frac{I_C}{I_B} = \frac{2 \cdot D_n}{w_B^2 \cdot f_R} \cdot \frac{1}{\left(1 + \frac{2D_n/w_B}{v_{nsat}}\right)} \tag{2. 75}$$

Esta expresión indica que β es independiente del efecto del spike (v_s) y que, si $D_n/w_B << v_{nsat}$, entonces β toma el valor de los BJT convencionales (es decir, sólo el primer factor de la ecuación).

2.5 Limitaciones del modelo de arrastre-difusión

El modelo de arrastre-difusión presentado en este capítulo es una simplificación de la ecuación de transporte de Boltzmann, cuya validez está limitada a dispositivos que cumplen una serie de restricciones.

La ecuación de balance del momento, deducida a partir de la ecuación de Boltzmann (véanse los capítulos 3 y 5) establece [Lundstrom,1990]:

$$J_n + \tau_m \cdot \frac{\partial J_n}{\partial t} = \frac{q^2 n}{m^*} \cdot \tau_m \cdot E_{el} + \frac{2q}{m^*} \cdot \tau_m \cdot \nabla W \tag{2. 76}$$

donde τ_m es el valor medio del tiempo de relajación del momento y W es el tensor de energía cinética media de los electrones, cuyas componentes son:

$$W_{ij} = \frac{1}{2\Omega} \sum_{p} v_i \cdot p_j \cdot f$$

(2. 77)

con Ω un pequeño volumen de normalización alrededor del punto espacial considerado, f la función de distribución de electrones, p_j el momento del electrón en la dirección j, y v_i su velocidad en la dirección i.

Definiendo la movilidad como

$$\mu_n = \frac{q}{m^*} \cdot \tau_m$$

(2. 78)

resulta:

$$J_n + \tau_m \cdot \frac{\partial J_n}{\partial t} = q \cdot \mu_n \cdot n \cdot E_{el} + 2 \cdot \mu_n \cdot \nabla W$$

(2. 79)

Si la velocidad del portador se expresa como suma de una componente de agitación térmica c debida a colisiones en el cristal más otra componente de velocidad media de arrastre v_d, se demuestra que:

$$W_{ii} = W_{arrastre} + W_{térmica} = n \cdot \left[\frac{1}{2} m^* v_d^2 + \frac{1}{2} m^* < c^2 > \right]$$

(2. 80)

A partir del término $W_{térmica}$ se define la temperatura de los portadores, T_c, como:

$$\frac{3}{2} \cdot k_B \cdot T_c \cdot n = \frac{1}{2} \cdot m^* \cdot < c^2 > \cdot n = W_{térmica}$$

(2. 81)

La ecuación de balance del momento (2. 79) se simplifica a la de arrastre-difusión en las siguientes condiciones:

a) Durante un intervalo de tiempo τ_m la corriente no varía apreciablemente:

$$\tau_m \cdot \frac{\partial J_n}{\partial t} \to 0$$

(2. 82)

b) El tensor W es diagonal y la componente de arrastre es muy inferior a la térmica ($W_{arrastre} << W_{térmica}$). En este caso:

$$2 \cdot \mu_n \cdot \nabla W \rightarrow \frac{2}{3} \cdot \mu_n \cdot \nabla W_{térmica} = q \cdot D_n \cdot \nabla n + q \cdot S_n \cdot \nabla T_c \tag{2. 83}$$

$$D_n = \frac{k_B \cdot T_c}{q} \cdot \mu_n \tag{2. 84}$$

$$S_n = \mu_n \cdot \frac{k_B}{q} \cdot n \tag{2. 85}$$

c) La temperatura de los portadores, T_c, es igual a la del cristal, T_L:

$$\nabla T_c \rightarrow 0 \tag{2. 86}$$

Con todas estas condiciones la ecuación de balance del momento (2. 79) se simplifica a:

$$J_n = q \cdot \mu_n \cdot n \cdot E_{el} + q \cdot D_n \cdot \nabla n \tag{2. 87}$$

en la cual μ_n es función biunívoca del material y del campo eléctrico.

Estas condiciones pierden validez en HBTs rápidos y de pequeñas dimensiones. En efecto, la condición *a)* deja de cumplirse en la medida que J_n experimenta un incremento significativo en un tiempo τ_m (recuérdese que la frecuencia de transición f_T en estos transistores supera ampliamente los 200 GHz).

Las condiciones b) y c), $W_{arrastre} << W_{térmica}$ y $T_c = T_L$, también están en cuestión en los HBTs abruptos. En efecto, tal como muestra la Figura 2. 5 , los electrones inyectados por el emisor a la base tienen una energía cinética considerable.

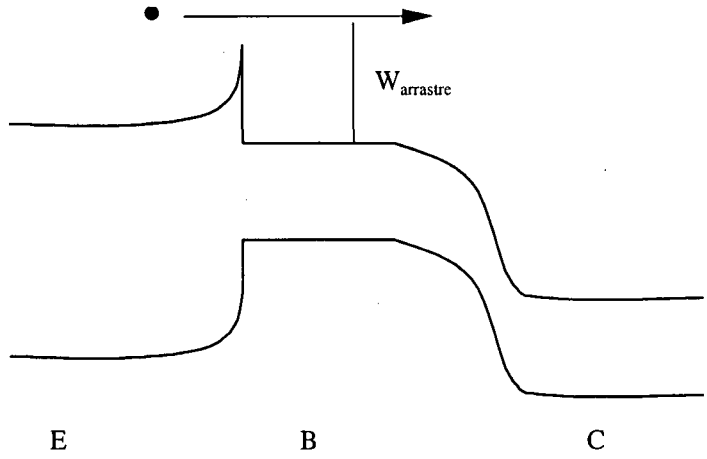

Figura 2. 5 Energía cinética de un electrón inyectado desde el emisor a la base en un HBT abrupto.

La dependencia habitual de μ_n y D_n *sólo* del material y del campo eléctrico local deja de cumplirse en dispositivos muy pequeños. Estos parámetros dependen también de la estructura del dispositivo y de la polarización aplicada. Esto se debe a que τ_m, el tiempo medio de relajación del momento, depende de la función de distribución *f(r,p,t)*:

$$\hat{\tau}_m = \frac{1}{<< 1/\tau_m(p) >>} = \frac{P_i(r,t) - P_i^0(r,t)}{\frac{1}{\Omega}\sum_p f(r,p,t) \cdot P_\tau / \tau_m} \tag{2. 88}$$

y *f(r,p,t)* depende de la estructura del dispositivo cuando sus dimensiones son muy pequeñas.

Por estas razones, el modelo de arrastre-difusión tiene una validez limitada en HBTs de alta velocidad.