



UNIVERSITÀ DEGLI STUDI DI FIRENZE
Facoltà di Ingegneria

Corso di Laurea in
INGEGNERIA ELETTRONICA

**Analisi del trasporto in
dispositivi ad effetto tunnel
intrabanda mediante
l'applicazione del formalismo di
Wigner**

19 Settembre 2002

Tesi di Laurea di

Omar Morandi

Relatori:

Prof. Giovanni Frosali

Prof. Gianfranco Manes

Anno Accademico 2001/2002

*Qui ci va
la paginata di ringraziamenti*

Indice

1	Formalismo di Wigner	6
1.1	introduzione	6
1.2	Traiettoria classica	7
1.3	Ensamble di particelle	8
1.4	Gerarchia BBGKY	9
1.5	Teoria semiclassica	11
1.5.1	Equazione di Liouville	12
1.5.2	Equazione di Boltzmann	13
1.6	Dinamica quantistica	14
1.7	Matrice di densità	15
1.8	Funzione di Wigner	16
1.8.1	Equivalenza Schrödinger-Wigner	19
1.8.2	f_w è reale	20
1.8.3	Casi particolari di moto	20
1.8.4	Operatore pseudo-differenziale	21
1.8.5	Momenti della f_w	21
1.8.6	Calcolo del “limite classico”	23
1.9	Scalatura equazione	23
1.10	Traiettorie classiche e quantistiche	26

1.11	Traiettoria quantistica	26
1.12	Equazione di Boltzmann quantistica	27
2	Modelli numerici	30
2.1	Analisi generale dei sistemi irreversibili	30
2.2	Sistemi “aperti”	31
2.2.1	Reservoirs	31
2.2.2	Condizioni al contorno di tipo inflow	33
2.3	Modelli basati sull’equazione di Schrödinger	34
2.3.1	Stati “misti” e stati “puri”	34
2.3.2	Descrizione degli urti nella formulazione di Schrödinger	36
2.4	Modello di Wigner	38
2.4.1	Apertura del sistema	38
2.4.2	Collisioni	41
3	Modello discreto	43
3.1	Introduzione	43
3.2	Equazione di Wigner stazionaria	45
3.3	Calcolo dei momenti	51
3.4	Soluzione non stazionaria	56
3.5	Equazione di Poisson	61
3.6	Traiettorie	71
3.6.1	Valutazione del tempo di tunnel	74
4	Risoluzione numerica del sistema Wigner-Poisson	75
4.1	RTD: tecnologie realizzative e principio di funzionamento . . .	75
4.1.1	L’eterogiunzione	78
4.1.2	Materiali impiegati nell’eterogiunzione	82
4.2	Applicazione dell’algoritmo numerico alla simulazione di RTD	86

4.2.1	Determinazione della funzione di equilibrio	87
4.3	Soluzione stazionaria	89
4.3.1	Calcolo del diagramma $I(V)$ in assenza di collisioni . .	89
4.3.2	Presenza di collisioni	93
4.4	Soluzione evolutiva	98
4.4.1	Effetti delle collisioni nei tempi di switch	99
4.5	Soluzione stazionaria auto-consistente	100
4.5.1	Criterio di convergenza	104
4.6	Soluzione evolutiva auto-consistente	106
4.6.1	Criterio di convergenza	111
4.6.2	Efficienza computazionale	112
4.6.3	Origine delle oscillazioni intrinseche dell'RTD	113
4.7	Calcolo delle quasi-traiettorie	115
4.7.1	Traiettorie risonanti	117
4.7.2	Conclusioni e sviluppi futuri	121
5	Appendici	124
5.1	Sviluppo in serie di potenze dell'operatore pseudo-differenziale	124
5.2	Operatore collisionale e valutazione di τ	126
5.2.1	Valutazione numerica del tempo di rilassamento	129

Elenco delle figure

2.1	Possibili condizioni di inflow per la funzione di Wigner	34
2.2	Circuito R C	39
2.3	Autovalori di $\mathbf{L}$	41
3.1	Griglia di discretizzazione del piano delle fasi	48
3.2	Matrice di Liouville	51
3.3	Direzione del flusso delle particelle	53
3.4	Stabilità degli schemi numerici	62
3.5	Diagramma a bande dell'RTD	63
3.6	Densità di carica e diagramma a bande risultanti dai primi quattro step dell'algoritmo di Gummel applicato ad un RTD .	65
3.7	Incrementi della f_w	73
3.8	Correzione della stima (x', p')	73
4.1	Principio di funzionamento di un RTD	76
4.2	Corrente misurata di un RTD alla temperatura di 77 K [17] .	77
4.3	Eterogiunzione all'equilibrio termico	80
4.4	Diagrammi a bande di un eterogiunzione non ideale	81
4.5	Diagrammi di energia per i vari tipi di discontinuità di banda .	82
4.6	Discontinuità delle bande di energia in una eterogiunzione $Al_xGa_{1-x}As - GaAs$	84

4.7	Energia di gap in funzione della costante reticolare.	85
4.8	Diagramma a bande e profilo del drogante nell'RTD simulato .	87
4.9	f_0 alla temperatura di 77 K	89
4.10	f_0 alla temperatura di 300 K	90
4.11	Diagramma I(V) balistico	91
4.12	Funzione di Wigner in corrispondenza del picco di corrente del diagramma I(V)	92
4.13	Funzione di Wigner nel punto di valle del diagramma I(V) . .	93
4.14	Legame fra tempo di rilassamento e temperatura	94
4.15	Effetto dell'inclusione degli urti nel diagramma I(V) dell'RTD	95
4.16	Rapporto picco-valle in funzione della temperatura	96
4.17	Densità di carica nel pozzo quantistico al variare della tensione	96
4.18	Densità di carica nel dispositivo al variare della temperatura .	97
4.19	Variazione temporale della corrente causata da uno switch da 0.13 a 0.22 Volt ad una temperatura di 77 K	98
4.20	Variazione temporale della corrente causata da uno switch da 0.22 a 0.13 Volt ad una temperatura di 77 K	99
4.21	Variazione temporale della corrente causata da uno switch da 0.13 a 0.22 Volt ad una temperatura di 300 K	100
4.22	Diagramma I(V) autoconsistente con collisioni alla temperatu- ra di 77 K e 300 K	101
4.23	Potenziale e densità autoconsistenti per l'RTD non polarizzato	102
4.24	Densità e potenziale autoconsistenti	103
4.25	Potenziale e densità autoconsistenti per $V_{bias} = 0.28$ eV . . .	105
4.26	Switch in tensione fra 0.26 V e 0.27 V	108
4.27	Switch in tensione fra 0.23 V e 0.24 V	109
4.28	Switch in tensione fra 0.24 V e 0.23 V	110

4.29 Oscillazioni intrinseche innescate dal solo rumore numerico, alla tensione di 0.24 V	110
4.30 Giustificazione delle oscillazioni intrinseche	114
4.31 Traiettorie f_w per RTD non polarizzato	115
4.32 Traiettorie f_w per RTD con $V_{Bias} = 0.03$ Volt	117
4.33 Analisi delle traiettorie a 77 K	119
4.34 Analisi delle traiettorie a 300 K	120
4.35 Tempi di attraversamento del dispositivo	121
4.36 Forza quantistica	122

Elenco delle tabelle

1.1	Grandezze tipiche nella modellizzazione di dispositivi meso- scopici	25
3.1	Coefficienti dello sviluppo Joyce-Dixon	70
4.1	Costi computazionali degli algoritmi risolutivi	112
5.1	Tempi di rilassamento caratteristici	130
5.2	Tempi di rilassamento totali	130

Capitolo 1

Formalismo di Wigner

1.1 introduzione

L'analisi del trasporto di carica riveste un ruolo fondamentale sia nell'ambito generico della fisica dello stato solido, sia nell'ambito più pratico e tecnologico dell'ingegneria elettronica delle alte frequenze, in quanto la progettazione e la realizzazione di dispositivi elettronici non può prescindere dalla comprensione fisica dei fenomeni che ne governano il comportamento. Da un punto di vista microscopico, a causa dell'elevata complessità dei fenomeni fisici coinvolti, il trasporto di carica non risulta un facile oggetto d'indagine. Grandezze come ad esempio la mobilità, non sono in generale ben conosciute, infatti quando se ne effettua una misura, quest'ultima dipende da un gran numero di proprietà caratteristiche del materiale in esame quali ad esempio la struttura a bande, il contenuto di impurezze e di difetti reticolari. Gli esperimenti che riguardano l'indagine del trasporto di carica quindi non sono solitamente in grado di fornire informazioni sufficienti alla stesura di un congruo modello matematico. D'altra parte l'elevato progresso delle tecnologie integrate dell'ultimo ventennio, ha portato allo sviluppo e alla produzione di

dispositivi quantistici che hanno trovato un sempre più largo impiego nelle applicazioni alle microonde. Si comprende quindi come sia particolarmente utile avere la possibilità di sviluppare attendibili codici di simulazione dei fenomeni di trasporto quantistico, prima di avviare costosi processi produttivi. Nel presente lavoro dimostreremo come rivesta un ruolo fondamentale la possibilità di disporre di una “veste cinetica” delle equazioni di moto quantistiche, per la simulazione dei dispositivi risonanti ad effetto tunnel (RTD). Le equazioni cinetiche da una parte, (assieme a modelli da esse derivate come le equazioni idrodinamiche), e le equazioni che derivano dalla formulazione di Schrödinger dall'altra, costituiscono lo stato dell'arte per la realizzazione di programmi di simulazione del trasporto in dispositivi submicrometrici.

1.2 Traiettoria classica

Come apparirà nel seguito della trattazione, si può stabilire una forte analogia formale fra le equazioni classiche di moto e la formulazione Wigneriana della meccanica quantistica. Tale analogia, oltre a fornire una visione intuitiva dei processi quantistici, aiuta a risolvere e completare il quadro descrittivo dei sistemi reali: in altre parole permette di includere fenomeni complessi come lo scattering elettrone-fonone e la modellizzazione di sistemi “aperti” tramite approssimazioni “fisicamente ragionevoli” delle equazioni di moto. In tale ottica rivedremo rapidamente l'applicazione della meccanica classica e semiclassica all'analisi del trasporto di carica.

1.3 Ensamble di particelle

Da un punto di vista modellistico, un dispositivo a semiconduttore è configurabile come un problema di moto di N elettroni interagenti, che si muovono in una regione delimitata da un volume V , sotto l'azione di un campo elettrico applicato. Un tale sistema rientra nell'ambito di studio della teoria statistica degli insiemi di Gibbs; ne descriveremo brevemente le tecniche basilari d'indagine.

Dato un sistema di N particelle si definisce “spazio Γ ” uno spazio $6N$ dimensionale con coordinate $(\mathbf{X}, \mathbf{P}) = (\mathbf{x}_1, \dots, \mathbf{x}_N, \mathbf{p}_1, \dots, \mathbf{p}_N)$ in cui $\mathbf{x}_i$ e $\mathbf{p}_i$ rappresentano rispettivamente la posizione e l'impulso dell' i -esimo elettrone. Un punto nello spazio Γ rappresenta quindi un intero sistema di particelle e calcolarne il moto significa determinare l'evoluzione temporale di ogni particella che compone l'insieme stesso. Si noti che una tale visione del problema esige però la conoscenza di $3N$ variabili di posizione $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_{3N}$ e $3N$ momenti $\mathbf{p}_1, \mathbf{p}_2, \dots, \mathbf{p}_{3N}$, ovvero la conoscenza del “microstato” del sistema in un qualche istante di tempo. Ogni punto di Γ , ovvero ogni configurazione del sistema sotto esame, prende il nome di punto rappresentativo del sistema. In ambito classico le coordinate di tale punto devono, nel tempo, risolvere le equazioni di Hamilton.

$$\begin{cases} \dot{x}_i = \frac{\partial H(x_i, p_i)}{\partial p_i} \\ \dot{p}_i = -\frac{\partial H(x_i, p_i)}{\partial x_i} \end{cases} \quad i = 1, 2, \dots, 3N. \quad (1.1)$$

in cui H è l'Hamiltoniana totale del sistema. Quello che riveste un'importanza rilevante, è come riuscire a estrarre, da una teoria totalmente deterministica quale la precedente, le poche basilari grandezze macroscopiche oggetto d'interesse nell'elettronica applicata. Queste grandezze macroscopiche, quali ad esempio corrente e densità, sono individuate da poche variabili ter-

modinamiche sperimentalmente osservabili (note come “momenti della distribuzione d'ensemble” che definiremo). Esse devono essere ricavate quindi dall'analisi dello stato microscopico (posizione e impulso di ogni particella) mediante opportune operazioni di media. Si può osservare che ad uno stesso valore di una variabile macroscopica (come la densità di portatori) corrispondano in generale più punti nello spazio Γ : infatti, ad esempio, scambiando due particelle fra loro, lo stato macroscopico non cambia, mentre si individuano due stati microscopici diversi, quindi due distinti punti dello spazio Γ . Nella teoria di Gibbs [1] si associa ad un sistema macroscopico l'insieme di tutti i punti nello spazio Γ che lo “descrivono” ovvero cui siano associati valori delle grandezze macroscopiche pari a quelli del sistema indagato; tale dominio prende il nome di macrostato. In altre parole se si considerano infinite copie dello stesso sistema, esse popolano lo spazio Γ con una costellazione di punti, ognuno dei quali si muove lungo la traiettoria relativa al proprio microstato, soluzione del sistema (1.1). Il moto di quest'insieme di punti viene quindi analizzato come fosse il moto di un fluido nello spazio delle fasi, caratterizzato da una distribuzione di densità $f(\mathbf{X}, \mathbf{P}) \geq 0$ che prende il nome di funzione di distribuzione d'ensemble e che fornisce il rapporto fra il numero dei punti rappresentativi contenuti in una regione “sufficientemente piccola” contenente il punto $(\mathbf{X}, \mathbf{P})$ e il volume della regione stessa [2].

1.4 Gerarchia BBGKY

Usualmente nello studio di sistemi a molti corpi, si definisce una teoria generica di dinamica rigorosa, e se ne deduce una formulazione ristretta da ipotesi semplificative al problema in esame, e tale che fenomeni complessi intrattabili teoricamente, possano essere tenuti in conto. L'inapplicabilità diretta di

una formulazione rigorosa di un problema di N corpi, è causata ovviamente dall'elevato numero di gradi di libertà del sistema stesso, in quanto, come si è visto, si deve ambientarne lo studio in uno spazio $6N$ dimensionale. In una formulazione cinetica al problema del trasporto, le variabili indipendenti sono posizione e impulso di ogni particella, cui si dovrebbero aggiungere, per la trattazione dei fenomeni di urto, addirittura tutte le “variabili fononiche”. Infine il sistema considerato è composto di particelle (elettroni) interagenti, quindi la velleitaria richiesta di una soluzione esatta di un problema di N corpi richiederebbe la soluzione di N (dell'ordine del numero di Avogadro) equazioni, che in più siano accoppiate fra loro; ovvero in cui ognuna delle quali non è risolubile, indipendentemente dalle altre. In generale in un tale processo di riduzione, tutti i gradi di libertà del sistema che non trovano esplicito interesse nell'ambito della simulazione, sono implicitamente inclusi attraverso operatori collisionali o potenziali efficaci di interazione. Parte dell'informazione sullo stato del sistema viene inevitabilmente persa ma per mezzo di tale semplificazione si riesce infine ad approssimare il problema originario con una formulazione di moto di singola particella: tale è l'idea alla base della procedura che va sotto il nome di “gerarchia BBGKY” e che verrà qui sotto analizzata.

Il metodo BBGKY, applicabile sia al caso classico della funzione di distribuzione di ensemble, sia, con congrue modifiche, al caso quantistico della pseudo-distribuzione di Wigner, segue l'iter seguente. Si deriva il sistema di equazioni scritto per il sotto-ensemble costituito da n particelle con n che varia fra 1 e N . Assumendo particolari forme d'interazione (interazioni deboli fra due particelle), al variare di n , si riesce a creare una catena di equazioni nella quale l'equazione di moto di singola particella coinvolge la densità del sistema composto da due particelle, la quale coinvolge a sua volta quella di tre

e così via. La gerarchia viene “chiusa” mandando formalmente N all’infinito e approssimando il campo elettrico visto dagli elettroni col “campo medio” da essi prodotto: si assume quindi che le particelle di un piccolo sotto-ensemble, si muovano indipendentemente l’una dall’altra [3]. Il sistema risultante ha la forma di un’equazione di Liouville (o nel caso quantistico di Wigner) di singola particella, accoppiata con l’equazione di Poisson: si origina il così detto sistema autoconsistente che sarà la base della nostra analisi e che verrà analizzato nel dettaglio in seguito.

1.5 Teoria semiclassica

L’approccio semiclassico al problema del trasporto si basa sull’assunzione che ai portatori di carica all’interno del solido cristallino possano essere associati contemporaneamente un valore della coordinata di posizione $\mathbf{x}$ e uno di quella dell’impulso $\mathbf{p} = \hbar\mathbf{k}$, dove $\mathbf{k}$ è il numero d’onda associato all’elettrone secondo la legge di Planck, e che essi si muovano, in risposta ad un campo applicato seguendo la legge

$$\dot{\mathbf{p}} = q\mathbf{E} \quad (1.2)$$

dove q rappresenta la carica dell’elettrone [11]. Nella presente trattazione verranno trascurati i gradi di libertà elettronici di spin in quanto, non essendo presenti rilevanti campi magnetici, essi sono del tutto influenti sul moto degli elettroni. Come discusso nel paragrafo precedente, la quantità fondamentale in questo tipo di trattazione è la funzione di distribuzione dei portatori $f(\mathbf{x}, \mathbf{p}, t)$. Essa è nota a meno di una costante moltiplicativa che è scelta in modo che la f risulti una distribuzione normalizzata ovvero

$$\int_V \int f(\mathbf{x}, \mathbf{p}, t) d\mathbf{x} d\mathbf{p} = N \quad (1.3)$$

dove N è al solito il numero di particelle che compongono il sistema. Nel presente lavoro, quando non vengano indicati esplicitamente, gli estremi di integrazione si intendono $-\infty$ e $+\infty$. L'equazione che descrive l'evoluzione della $f(\mathbf{x}, \mathbf{p}, t)$ è l'equazione del trasporto di Boltzmann [11]

$$\frac{\partial f}{\partial t} + v(\mathbf{p}/\hbar) \nabla_{\mathbf{x}} f + q \mathbf{E} \nabla_{\mathbf{p}} f = \left. \frac{\partial f}{\partial t} \right|_{coll} \quad (1.4)$$

in cui $v(\mathbf{p}/\hbar)$ è la legge di dispersione all'interno del solido cristallino. Se consideriamo nullo il termine $\left. \frac{\partial f}{\partial t} \right|_{coll}$, ovvero trascuriamo ogni fenomeno collisionale, l'equazione di Boltzmann si riduce alla più semplice equazione di Liouville.

1.5.1 Equazione di Liouville

Il teorema di Liouville afferma che il valore della funzione di distribuzione che risolve l'equazione di Liouville, resta costante su ogni traiettoria nello spazio delle fasi; in altre parole un osservatore che si muovesse lungo le traiettorie, vedrebbe una densità di particelle costante per ogni tempo [2]. Il teorema di Liouville trova immediata applicazione in meccanica classica in quanto da esso discende che il valore della funzione di distribuzione in un punto qualsiasi dello spazio delle fasi è ricavabile dalla determinazione della traiettoria che passa per quel punto. Infatti nel caso della simulazione di sistemi reali ed aperti, poiché il dominio di simulazione, a causa della dimensioni fisiche del sistema, è un sottoinsieme limitato di $\mathbb{R}^6$, per determinare il valore della f in un punto qualsiasi si determina la traiettoria che lo intercetta e la si percorre “all'indietro” (ovvero invertendo l'asse dei tempi) sino ad incontrare un punto appartenente al bordo del dominio di simulazione. Poiché in molti casi, per ragioni fisiche, sui punti del bordo è noto il valore della f , tale valore, in accordo col teorema di Liouville, sarà proprio quello cercato. Nel

caso quantistico della teoria di Wigner, le cose sono più complesse come descriveremo.

Una volta determinata la funzione di distribuzione in ogni punto del dominio, da essa si ricavano le quantità fisiche di interesse quali la velocità di deriva, l'energia cinetica media, ecc. Più precisamente il valor medio di una generica grandezza definita sullo spazio delle fasi è:

$$\langle A_{cl} \rangle = \int d\mathbf{x} \int d\mathbf{p} A(\mathbf{x}, \mathbf{p}) f(\mathbf{x}, \mathbf{p}) \quad (1.5)$$

dove $A(\mathbf{x}, \mathbf{p})$ è la funzione che rappresenta la grandezza di interesse. Casi di specifico interesse sono quelli in cui $A = cost.$, $A \sim \mathbf{p}$, $A \sim |\mathbf{p}|^2$; che sono noti come momenti di ordine zero, primo e secondo della distribuzione e tramite i quali si calcolano la densità di particelle

$$\int f(\mathbf{x}, \mathbf{p}) d\mathbf{p} = n(\mathbf{x}) \quad (1.6)$$

la corrente totale

$$\frac{1}{m} \int f(\mathbf{x}, \mathbf{p}) \mathbf{p} d\mathbf{p} = \mathbf{J}(\mathbf{x}) \quad (1.7)$$

e l'energia media

$$\int f(\mathbf{x}, \mathbf{p}) |\mathbf{p}|^2 d\mathbf{p} = E(\mathbf{x}) \quad (1.8)$$

1.5.2 Equazione di Boltzmann

Analizziamo come cambia la trattazione precedente qualora si consideri l'effetto delle collisioni nel moto elettronico, ovvero si includa il termine a secondo membro della (1.4) che tiene conto delle collisioni, noto come termine collisionale di Boltzmann: esso indica l'aumento o la diminuzione della densità di elettroni in un generico punto $(\mathbf{x}, \mathbf{p})$, dovuto alla presenza di eventi di urto (scattering) all'interno del semiconduttore. I fenomeni di urto si dividono in fenomeni del primo e del secondo ordine e sono rispettivamente: l'urto

fra elettrone e ione fisso appartenente al reticolo cristallino (urto elettrone-fonone) e l'urto fra due elettroni (scattering elettrone-elettrone). La precedente classificazione rimarca il fatto che i primi originano termini lineari nell'equazione di Boltzmann mentre i secondi, non lineari. Poiché si verifica che i fenomeni di urto che influenzano maggiormente il moto degli elettroni sono quelli elettrone-fonone, quest'ultimi saranno le uniche interazioni a corto raggio che considereremo nel presente lavoro. Sotto tale assunzione la forma esplicita del termine collisionale è la seguente [11],

$$\left. \frac{\partial f}{\partial t} \right|_{coll} = \int_{-\infty}^{\infty} d\mathbf{p}' [W_{\mathbf{p}' \rightarrow \mathbf{p}} f(\mathbf{p}, \mathbf{p}', t) - W_{\mathbf{p} \rightarrow \mathbf{p}'} f(\mathbf{x}, \mathbf{p}, t)] \quad (1.9)$$

in cui W è noto come nucleo di transizione, e verrà analizzata in appendice 5.1. L'equazione di Boltzmann ha una complessa forma integro-differenziale e non se ne conoscono soluzioni analitiche se non per pochi casi estremamente semplificati e non utilizzabili nella descrizione di un semiconduttore reale.

Inoltre la teoria semiclassica si fonda su ipotesi che non risultano neppure in via approssimativa, verificate nel caso reale dei semiconduttori oggetto della nostra indagine, essa non riesce perciò a giustificare minimamente fenomeni di cruciale interesse come il tunnel quantistico ed appare chiaramente inadeguata alle nostre esigenze. Vedremo però come i risultati di tale teoria aiutino nella descrizione quantistica del trasporto di carica.

1.6 Dinamica quantistica

Viste le limitazioni del modello semiclassico, passiamo alla definizione dei concetti basilari su cui si fonda la meccanica quantistica. Nell'ambito di tale teoria, il moto di un elettrone è descritto dall'equazione di Schrödinger :

$$i\hbar \frac{\partial \Psi}{\partial t} = H\psi \quad (1.10)$$

in cui H è l'operatore hamiltoniano quantistico di singola particella derivato da un arbitrario campo elettrico :

$$q\mathbf{E}(\mathbf{x}, t) = -\vec{\nabla}_{\mathbf{x}}U(\mathbf{x}, t) \quad (1.11)$$

(in cui U è espresso in eV) ovvero

$$H = -\frac{\hbar^2}{2m}\nabla_{\mathbf{x}}^2 - U(\mathbf{x}, t) \quad (1.12)$$

Una soluzione $\psi = \psi(\mathbf{x}, t)$ della (1.10) è detta funzione d'onda e il cui modulo quadro ha il seguente significato fisico:

$$\int_A |\psi(\mathbf{x}, t)|^2 d\mathbf{x} \quad (1.13)$$

è la probabilità di trovare l'elettrone in un sottoinsieme A dello spazio delle posizioni, all'istante t .

1.7 Matrice di densità

Il moto di un ensamble di particelle costituito da M elettroni, è ancora descritto tramite l'equazione di Schrödinger (1.10) ma in cui H è l'Hamiltoniana di M corpi:

$$H = -\frac{\hbar^2}{2m} \sum_{i=1}^M \nabla_{\mathbf{x}_i}^2 - U(\mathbf{x}_1, \dots, \mathbf{x}_M, t) \quad (1.14)$$

in cui $\mathbf{x}_i \in \mathbb{R}_x^{3M}$ denota la posizione dell' i -esimo elettrone.

Definiamo matrice di densità ρ , corrispondente ad una funzione d'onda ψ dell'ensamble di M elettroni, la funzione seguente :

$$\rho(\mathbf{r}, \mathbf{s}, t) = \overline{\psi(\mathbf{r}, t)}\psi(\mathbf{s}, t) \quad \mathbf{r}, \mathbf{s} \in \mathbb{R}_x^{3M} \quad (1.15)$$

Si verifica immediatamente che gli elementi così detti “diagonali” ($\mathbf{r} = \mathbf{s} = \mathbf{x}$) della matrice di densità rappresentano la densità di carica quantistica nel punto $\mathbf{x}$

$$\rho(\mathbf{x}, \mathbf{x}, t) = n_{quant.}(\mathbf{x}, t) \quad (1.16)$$

Differenziando la (1.15) rispetto al tempo e sfruttando la (1.10) si ottiene l'equazione che regola il moto della matrice di densità ρ :

$$i\hbar \frac{\partial \rho}{\partial t} = (H_s - H_r)\rho \quad (1.17)$$

in cui con H_s e H_r si sono indicate le Hamiltoniane scritte nelle variabili s e r rispettivamente. Con l'Hamiltoniana nella forma (1.14), la (1.17) si scrive più esplicitamente

$$i\hbar \frac{\partial \rho}{\partial t} = -\frac{\hbar^2}{2m} \sum_{i=1}^M \sum_{j=1}^M (\nabla_{\mathbf{s}_j}^2 \rho - \nabla_{\mathbf{r}_i}^2 \rho) - [U(\mathbf{s}, t) - U(\mathbf{r}, t)]\rho \quad (1.18)$$

Questa equazione è nota come equazione di moto di Heisenberg [1].

1.8 Funzione di Wigner

L'introduzione della funzione di Wigner ci consentirà di riformulare le equazioni quantistiche di moto in una forma cinetica. A tal fine introduciamo il seguente cambio di coordinate

$$\mathbf{r} = \mathbf{x} + \frac{\hbar}{2}\boldsymbol{\eta} \quad \mathbf{s} = \mathbf{x} - \frac{\hbar}{2}\boldsymbol{\eta} \quad (1.19)$$

nella matrice di densità, ovvero poniamo

$$u(\mathbf{x}, \boldsymbol{\eta}, t) = \rho\left(\mathbf{x} + \frac{\hbar}{2}\boldsymbol{\eta}, \mathbf{x} - \frac{\hbar}{2}\boldsymbol{\eta}, t\right) \quad (1.20)$$

Indicheremo con $(\mathcal{F}g)(\boldsymbol{\eta}) = \hat{g}(\boldsymbol{\eta})$ la trasformata di Fourier di $g(\mathbf{p})$ e con $(\mathcal{F}^{-1}h)(\mathbf{p})$ la sua inversa ovvero :

$$\hat{g}(\boldsymbol{\eta}) = (\mathcal{F}g)(\boldsymbol{\eta}) \doteq \int_{\mathbb{R}_p^{3M}} g(\mathbf{p}) e^{-i\mathbf{p}\boldsymbol{\eta}} d\mathbf{p} \quad (1.21)$$

$$(\mathcal{F}^{-1}h)(\mathbf{p}) \doteq \frac{1}{(2\pi)^{3M}} \int_{\mathbb{R}_{\boldsymbol{\eta}}^{3M}} h(\boldsymbol{\eta}) e^{i\mathbf{p}\boldsymbol{\eta}} d\boldsymbol{\eta} \quad (1.22)$$

La funzione di Wigner f_w relativa ad una funzione d'onda ψ (o equivalentemente ad una matrice di densità ρ) è definita come la trasformata inversa di Fourier di $u(\mathbf{x}, \boldsymbol{\eta}, t)$ rispetto ad $\boldsymbol{\eta}$

$$f_w(\mathbf{x}, \mathbf{p}, t) \doteq (\mathcal{F}^{-1}u)(\mathbf{p}) \quad (1.23)$$

ovvero

$$f_w(\mathbf{x}, \mathbf{p}, t) \doteq \frac{1}{(2\pi)^{3M}} \int_{\mathbb{R}_{\boldsymbol{\eta}}^{3M}} \rho\left(\mathbf{x} + \frac{\hbar}{2}\boldsymbol{\eta}, \mathbf{x} - \frac{\hbar}{2}\boldsymbol{\eta}, t\right) \exp^{i\mathbf{p}\boldsymbol{\eta}} d\boldsymbol{\eta} \quad (1.24)$$

L'equazione di moto della f_w è ottenuta trasformando l'equazione di Heisenberg (1.18). Notando che

$$\nabla_{\mathbf{r}}^2 - \nabla_{\mathbf{s}}^2 = \frac{2}{\hbar} \nabla_{\mathbf{x}} \nabla_{\boldsymbol{\eta}} \quad (1.25)$$

la (1.18) diventa nelle coordinate $(\mathbf{x}, \boldsymbol{\eta})$

$$\frac{\partial u}{\partial t} + \frac{i}{m} \operatorname{div}_{\boldsymbol{\eta}} (\operatorname{grad}_{\mathbf{x}} u) + i \frac{U(\mathbf{x} + \frac{\hbar}{2}\boldsymbol{\eta}, t) - U(\mathbf{x} - \frac{\hbar}{2}\boldsymbol{\eta}, t)}{\hbar} u = 0 \quad (1.26)$$

Inoltre ricordando che

$$\left(\mathcal{F}^{-1} \frac{\partial \hat{f}}{\partial \eta} \right) (p) = -ip (\mathcal{F}^{-1} \hat{f})(p) = -ip f \quad (1.27)$$

possiamo anti-trasformare secondo Fourier l'equazione (1.26)

$$\frac{\partial f_w}{\partial t} + \frac{\mathbf{p}}{m} \operatorname{grad}_{\mathbf{x}} f_w + \theta_{\hbar}[U] f_w = 0 \quad (1.28)$$

Dove $\theta_{\hbar}[U]$ è chiamato operatore pseudo-differenziale ed è tale che:

$$(\mathcal{F} \theta_{\hbar}[U] f_w)(\mathbf{x}, \boldsymbol{\eta}, t) = i \frac{U(\mathbf{x} + \frac{\hbar}{2}\boldsymbol{\eta}, t) - U(\mathbf{x} - \frac{\hbar}{2}\boldsymbol{\eta}, t)}{\hbar} (\mathcal{F} f_w)(\mathbf{x}, \boldsymbol{\eta}, t) \quad (1.29)$$

ovvero esplicitamente

$$(\theta_{\hbar}[U] f_w)(\mathbf{x}, \boldsymbol{\eta}, t) = \quad (1.30)$$

$$\begin{aligned} & \frac{i}{(2\pi)^{3M}} \int_{\mathbb{R}_\eta^{3M}} \int_{\mathbb{R}_{\mathbf{p}'}^{3M}} \frac{U(\mathbf{x} + \frac{\hbar}{2}\boldsymbol{\eta}, t) - U(\mathbf{x} - \frac{\hbar}{2}\boldsymbol{\eta}, t)}{\hbar} f_w(\mathbf{x}, \mathbf{p}', t) e^{i(\mathbf{p}-\mathbf{p}')\boldsymbol{\eta}} d\mathbf{p}' d\boldsymbol{\eta} \\ &= \frac{i}{(2\pi)^{3M}} \int_{\mathbb{R}_\eta^{3M}} \int_{\mathbb{R}_{\mathbf{p}'}^{3M}} \delta_\hbar U f_w(\mathbf{x}, \mathbf{p}', t) e^{i(\mathbf{p}-\mathbf{p}')\boldsymbol{\eta}} d\mathbf{p}' d\boldsymbol{\eta} \end{aligned}$$

avendo definito il simbolo

$$\delta_\hbar U(\mathbf{x}, \boldsymbol{\eta}) = \frac{U(\mathbf{x} + \frac{\hbar}{2}\boldsymbol{\eta}, t) - U(\mathbf{x} - \frac{\hbar}{2}\boldsymbol{\eta}, t)}{\hbar} . \quad (1.31)$$

La (1.28) è nota come equazione di Wigner del trasporto di carica [3].

Fin qui, nell'analisi della dinamica quantistica si è scelto, come fatto per la meccanica classica, di rimanere nella generica formulazione propria di un sistema a N corpi; come già accennato, la riduzione del numero di variabili della (1.28) seguirà dalla tecnica della gerarchia BBGKY, applicabile al presente caso, grazie alla formulazione di tipo cinetico data da Wigner alla meccanica quantistica. Il risultato finale del procedimento BBGKY è la nascita di una funzione di sette variabili $(\mathbf{p}, \mathbf{q}, t)$ che dovrà risolvere sia l'equazione di Wigner (scritta per singola particella) sia l'equazione di Poisson. Il sistema che deriva è noto come sistema di Wigner-Poisson che qui scriviamo nella forma 1-dimensionale

$$\frac{\partial f_w}{\partial t} + \frac{p}{m} \frac{\partial f_w}{\partial x} + \theta_\hbar[U] f_w = 0 \quad (1.32)$$

$$\epsilon \frac{d^2 U(x)}{dx^2} = q\rho(x) = q^2[C(x) - n(x)] \quad (1.33)$$

in cui $C(x)$ è la densità di ioni fissi droganti e $n(x)$ la densità di carica libera e che è legata al momento primo della f_w come vedremo tra breve.

Anche se si applica la stessa procedura sia al caso classico sia a quello quantistico i risultati finali della tecnica BBGKY sono concettualmente molto differenti: infatti la f_w , a differenza del caso classico, può assumere valori negativi (anche partendo da distribuzioni definite positive) e quindi non è

interpretabile come densità di probabilità. Tale differenza deriva dal fatto che nei sistemi quantistici, a causa del principio di indeterminazione, non si possono più introdurre densità di probabilità che specifichino contemporaneamente posizione e momento della particella, ma solo quantità formalmente analoghe, come appunto la funzione di Wigner, la quale pur non essendo una densità di probabilità può essere sfruttata per ottenere i valori di aspettazione delle grandezze di interesse, in maniera formalmente identica a quella usata con le distribuzioni classiche. E' per tale analogia che la funzione di Wigner in letteratura è nota come funzione di distribuzione di quasi-probabilità. La funzione di Wigner costituisce quindi il legame più diretto fra la funzione d'onda ψ e l'usuale descrizione in termini di distribuzione classica; ma rispetto a quest'ultima contiene informazioni aggiuntive concernenti la fase quanto-meccanica del sistema, che danno origine alle regioni dello spazio in cui la f_w assume valori negativi.

Analizzeremo nei restanti paragrafi le caratteristiche della f_w e la sua connessione col mondo classico.

1.8.1 Equivalenza Schrödinger-Wigner

Dal procedimento che ha permesso di giungere alla definizione della f_w , risulta chiaro come l'usuale formulazione della meccanica quantistica dovuta a Schrödinger, debba essere del tutto equivalente alla formulazione in termini della funzione di Wigner f_w [3]. Da un'indagine di tipo matematico della f_w si evince però che lo spazio delle soluzioni della (1.28) non comprende solo stati che corrispondono ad un'effettiva evoluzione quanto-meccanica, ma è più ampio e contiene anche soluzioni non fisiche, ovvero cui non corrisponde nessuno stato fisico. Solo con la scelta di "buone" condizioni iniziali, ovvero fisicamente coerenti, il precedente limite viene rimosso, facendo sì che l'equazione

di Wigner diventi in tutto e per tutto equivalente a quella di Schrodinger. Tatarskii [5] ha dimostrato che la condizione necessaria e sufficiente affinché la funzione di Wigner descriva uno stato quantistico è

$$\frac{\partial^2}{\partial r_1 \partial r_2} \ln \left[\int dp e^{ip(r_1 - r_2)/\hbar} f_w\left(\frac{r_1 + r_2}{2}, p/\hbar\right) \right] = 0 \quad (1.34)$$

1.8.2 f_w è reale

E' immediatamente verificabile dalla definizione della f_w come antitrasformata di u . Infatti dalla (1.20) segue che per la $u(x, \eta, t)$ vale $u(-\eta) = \overline{u(\eta)}$

1.8.3 Casi particolari di moto

La (1.28) è stata scritta in modo che appaia evidente l'analogia formale che essa ha con la (1.4). Le due equazioni coincidono sia nel caso di moto libero (assenza di potenziale)(e si riducono all'usuale equazione delle onde), sia nel caso dell'oscillatore armonico (potenziale quadratico); in generale si può verificare che le differenze nelle soluzioni fra le due equazioni saranno tanto più marcate quanto più rapidamente variabile sarà il potenziale applicato. Il confronto fra le due equazioni porta a considerare $\theta_\hbar[U]$ come una sorta di "generalizzazione" del concetto di forza classica: la differenza fra il caso quantistico e quello classico è che in quest'ultimo la forza è locale, ovvero la particella "sente" solo il valore della forza sul punto in cui si trova, mentre nel caso più complesso della meccanica quantistica, a determinare la funzione di distribuzione in un punto, concorre il valore del potenziale in ogni punto dello spazio. Tale visione verrà ripresa nel paragrafo 1.11

1.8.4 Operatore pseudo-differenziale

Analizziamo l'azione dell'operatore $\theta_{\hbar}[U]$ sulla f_w . Dalla (1.29) appare chiaramente che $\theta_{\hbar}[U]$ applicato a f_w , genera una funzione la cui trasformata di Fourier è pari al prodotto fra $\hat{f}(\eta)$ (la trasformata della f_w) ed il simbolo $\delta_{\hbar}U$: l'azione di $\theta_{\hbar}[U]$ su f_w è quindi quella che avrebbe un sistema lineare con funzione di trasferimento $\delta_{\hbar}U$ su un generico segnale $h(p)$. Quindi, poiché $\theta_{\hbar}[U]$ agisce sulla f_w come un prodotto di convoluzione esso è manifestamente non locale (come già accennato).

Adesso riscriviamo l'operatore pseudo-differenziale in una forma più utile, che si dimostra essere la “veste” migliore per risolvere numericamente l'equazione di Wigner. Riscriviamo la (1.30)

$$\begin{aligned}\theta_{\hbar}[U]f_w &= \\ &= \frac{i}{2\pi} \int_{\mathbb{R}_{p'}} f_w(x, p', t) dp' \int_{\mathbb{R}_{\eta}} \delta_{\hbar}U e^{i(p-p')\eta} d\eta \quad .\end{aligned}$$

Notando che, dalla definizione (1.31), $\delta_{\hbar}U$ è dispari rispetto a η , il secondo integrale lo si scrive

$$\int_{-\infty}^{\infty} \delta_{\hbar}U(\eta) e^{i(p-p')\eta} d\eta = \frac{2i}{\hbar} \int_0^{+\infty} \sin \left[\frac{(p-p')r}{\hbar} \right] \delta_{\hbar}U(r/\hbar) dr$$

avendo posto $r = \eta\hbar$ Infine

$$(\theta_{\hbar}[U])f_w = -\frac{1}{\hbar\pi} \int_{-\infty}^{\infty} dp' \int_0^{+\infty} dr \sin \left[\frac{(p-p')r}{\hbar} \right] \delta_{\hbar}U(r/\hbar) f_w(x, p', t)$$

1.8.5 Momenti della f_w

Quale significato fisico attribuire alla f_w non è ancora del tutto chiaro; ciò non deve stupire se si pensa che anche nella nota formulazione di Schrödinger, alla fase della funzione ψ non si può dare né un'interpretazione rigorosa diretta, né intuitiva in quanto tutta l'informazione fisica della funzione d'onda

è contenuta nel suo modulo; il quale ovviamente per essere valutato necessita l'esecuzione di un'operazione algebrica sulla ψ . In modo analogo, la f_w mostra il suo significato fisico nel calcolo dei suoi momenti, coerentemente con la sua struttura di quasi-distribuzione. Valutiamo i momenti di ordine zero e del primo ordine.

Momento di ordine zero: Sfruttando la nota proprietà della trasformata di Fourier, ovvero che l'integrale di una funzione è pari al valore della sua trasformata nell'origine

$$\int f(p)dp = \hat{f}(0) \quad (1.35)$$

e dalla (1.16) si ricava

$$\begin{aligned} n_{quant.}(x, t) &= u(x, \eta = 0, t) = \\ &= (\mathcal{F}f_w)(x, \eta = 0, t) = \int f_w(x, p, t)dp \end{aligned} \quad (1.36)$$

Ovvero il momento di ordine zero della f_w calcolato rispetto a p , non è altro che il valore della densità dei portatori $n(x)$.

Momento del primo ordine: Si può dimostrare [3] che dalla formula del calcolo della corrente quantistica scritta per la funzione d'onda ψ segue:

$$-\frac{q}{m} \int f(p, x, t)pdp = J_{quant.}(x) \quad (1.37)$$

Così, finché si considerano i momenti di ordine zero e primo, la funzione di Wigner si comporta in maniera analoga ad una distribuzione classica di particelle. A differenza di quest'ultima però, la f_w , in generale può assumere valori negativi e quindi non può essere interpretata come una distribuzione di probabilità.

1.8.6 Calcolo del “limite classico”

L'indagine che rafforza ulteriormente la necessità di vedere nella f_w il trait d'union fra la meccanica statistica classica ed il mondo quantistico è il calcolo del limite classico. Operando formalmente il limite per $\hbar$ tendente a zero nell'equazione di Wigner e notando che

$$\delta_{\hbar} U \xrightarrow{\hbar \rightarrow 0} \eta \text{grad}_x U \quad (1.38)$$

l'equazione di Wigner si modifica nella:

$$\frac{\partial f_w}{\partial t} + \frac{p}{m} \text{grad}_x f_w + \text{grad}_x U \text{grad}_p f_w = 0 \quad (1.39)$$

che non è altro che l'equazione classica di Liouville. Quindi l'equazione di Wigner che descrive i fenomeni quantistici, diventa la classica equazione di Liouville quando $\hbar$, opportunamente adimensionalizzata, può essere considerata una grandezza piccola rispetto agli altri parametri in gioco.

1.9 Scalatura equazione

Come sin qui messo in evidenza, la funzione di Wigner ha uno stretto legame con la classica equazione di moto di Liouville, ed in particolari casi, si riduce ad essa senza alcun tipo di approssimazione (oscillatore armonico e particella libera). In meccanica quantistica la costante di Planck fa da “spartiacque” fra il mondo classico e quello quantistico ovvero distingue fra i fenomeni la cui spiegazione si esaurisce con l'applicazione delle sole leggi di Newton (fisica classica), ed altri che richiedono il più complesso impianto numerico e concettuale della meccanica quantistica. Si verifica in pratica che il comportamento di un sistema diventa quantistico quando le grandezze in gioco diventano “in qualche modo” comparabili con la costante $\hbar$ (ovviamente si

confronteranno quantità omologhe ad $\hbar$, che ha le dimensioni di un'azione). Per esempio si può dimostrare che l'effetto tunnel “va come” $\hbar^2$, ovvero diventa significativo ogni qual volta non è più possibile trascurare $\hbar^2$ rispetto alle altre grandezze. In generale in tutte le soluzioni peculiarmente quantistiche è presente $\hbar$. Quanto detto apparirà in particolare evidenza nella formulazione cinetica Wigneriana della meccanica quantistica. Verificheremo adesso che, scalando opportunamente l'equazione (1.28), si possono introdurre coefficienti adimensionali il cui ordine di grandezza è indice della differenza fra le soluzioni dei due sistemi classico e quantistico, ovvero della necessità di una descrizione quantistica del moto [6]. A tale scopo assumiamo come metro di riferimento le grandezze seguenti: la dimensione della regione attiva L del dispositivo e la velocità di Fermi v_T . Riscriviamo l'equazione di Wigner in funzione delle grandezze scalate adimensionali $x' = \frac{x}{L}$, $t' = \frac{tv_T}{L}$, $p' = \frac{p}{mv_T}$. In tal modo la coordinata x' ad esempio, sarà misurata in unità di grandezza di L , e quindi con un range di variazione fra 0 e 1.

$$\begin{aligned} \frac{\partial f}{\partial t} + \frac{p}{mL} \frac{\partial f}{\partial x'} + \theta_{\hbar}[U]f &= 0 \\ \frac{v_T}{L} \left[\frac{\partial f}{\partial t'} + p' \frac{\partial f}{\partial x'} + \frac{L}{v_T} \theta_{\hbar}[U]f \right] &= 0 \end{aligned}$$

Il termine con l'operatore pseudo-differenziale può essere scritto nella forma

$$\begin{aligned} \frac{L}{v_T} \theta_{\hbar}[U]f &= \frac{iL}{2\pi v_T} \int d\eta \delta_{\hbar} U(x, \eta) \hat{f}(x, \eta) e^{ip\eta} = \\ \frac{LU_0}{v_T} \frac{i}{2\pi \hbar} \int d\eta' \delta_{\hbar} \mathcal{U} \left(x', \hbar \frac{\eta'}{Lmv_T} \right) \hat{f}(x', \eta') e^{ip'\eta'} &= \\ \frac{iU'_0}{2\pi \epsilon} \int d\eta' \delta_{\hbar} \mathcal{U}(x', \epsilon\eta') \hat{f}(x', \eta') e^{ip'\eta'} &= \\ \frac{iU'_0}{2\pi} \int d\eta' \frac{\mathcal{U}(x' + \epsilon\eta'/2) - \mathcal{U}(x' - \epsilon\eta'/2)}{\epsilon} \hat{f}(x', \eta') e^{ip'\eta'} &= U'_0 \theta'[\mathcal{U}]f \end{aligned}$$

dove

$$U'_0 \theta'[\mathcal{U}]f \equiv \frac{1}{2\pi} \int d\eta' \frac{\mathcal{U}(x' + \epsilon\eta'/2) - \mathcal{U}(x' - \epsilon\eta'/2)}{\epsilon} \hat{f}(x', \eta') e^{ip'\eta'} \quad (1.40)$$

$$(\delta_{\hbar}\mathcal{U})(x', \eta') \equiv \frac{\mathcal{U}(x' + \epsilon\eta'/2) - \mathcal{U}(x' - \epsilon\eta'/2)}{\epsilon} \quad (1.41)$$

Abbiamo sfruttato: $\eta' = mv_T\eta$, $U(x) = U_0\mathcal{U}(x')$, $\hat{f}(x', \eta') = \hat{f}(x, \eta)/(mv_T)$, $U'_0 = U_0/(mv_T^2)$ e $\epsilon = \hbar/(mLv_T)$. Concludendo, l'equazione di Wigner adimensionalizzata, si scrive

$$\frac{\partial f}{\partial t'} + p' \frac{\partial f}{\partial x'} + U'_0 \theta'[\mathcal{U}]f = 0 \quad (1.42)$$

Si nota dalla (1.40) che è proprio il termine “ ϵ ” a tener conto della non località dell' operatore pseudo-differenziale in quanto a parità di spettro della f_w (nota che sta nell'integrale a moltiplicare), la soluzione dell'equazione richiederà valori del potenziale tanto più lontani dal valore della coordinata x su cui la si sta calcolando, quanto più è grande il coefficiente ϵ . Quindi per grandi valori di ϵ viene esaltata la caratteristica di non località di $\theta_{\hbar}[U]$. Viceversa se ϵ tende a zero, si vede bene che, come nel caso del limite classico (ϵ ha $\hbar$ al numeratore), l'operatore $\delta_{\hbar}\mathcal{U}(\eta')$ diventa una derivata di U in dx e la f_w ridiventa la classica equazione di Liouville. Nei casi esaminati i valori tipici delle grandezze in gioco sono espressi in tab. 1.1: essi permettono, come risultato finale, di stimare che la non località dell'operatore si arresta su distanze di poco inferiori alla regione attiva dei dispositivi risonanti ad effetto tunnel (di circa 100 Å)

L (Å)	v_T (m/sec)	massa efficace m^* (kg)	U_0 (eV)	ϵ
550	10^6	$0.0667 \times 9.110^{-31}$	0.38	0.17

Tabella 1.1: Grandezze tipiche nella modellizzazione di dispositivi mesoscopici

1.10 Traiettorie classiche e quantistiche

Un'analisi di dispositivi elettronici che non prescindano dalla natura quantistica degli elettroni, deve abbandonare il concetto classico di traiettoria (ovvero del legame che sussiste fra il tempo e la posizione nello spazio di una particella), o più in generale, la localizzazione del punto rappresentativo del sistema nello spazio delle fasi. Ciò nonostante per la f_w si può continuare a parlare di quasi-traiettorie.

1.11 Traiettoria quantistica

Allo scopo di definire per la funzione di Wigner, l'analogo delle traiettorie classiche, conviene riscrivere la (1.28) in forma identica alla classica equazione di Liouville; vi si riesce senza introdurre nessuna approssimazione, definendo la “forza quantistica” $\tilde{F}$ che sostituisca la forza classica corrispondente ad un potenziale U . La forza classica dipende solo dal gradiente del potenziale ovvero

$$F = -\nabla U(r) \quad (1.43)$$

Nel caso quantistico $\tilde{F}$ invece sarà, in generale, funzione di tutte le derivate successive di U [7]; più precisamente posto

$$\tilde{F} = - \sum_{\lambda \text{ disp.}} \frac{1}{\lambda!} \left(\frac{\hbar}{2i} \frac{\partial}{\partial p} \right)^{\lambda-1} \frac{d^\lambda U(x)}{dx^\lambda} \quad (1.44)$$

e scrivendo l'operatore $\theta_\hbar$ in forma differenziale (vedi appendice 5.1) la funzione di Wigner acquista la forma

$$\frac{\partial}{\partial t} f_w = -\frac{p}{m} \nabla_r f_w - \tilde{F} \nabla_p f_w \quad (1.45)$$

In questa visione tutto funziona come se stessimo trattando particelle (quasi particelle) classiche (ovvero con posizione e momento ben definiti), ed in

cui l'effetto tunnel entra attraverso la “forza quantistica” che tali particelle sperimentano. Si può dimostrare che il teorema di Liouville vale ancora per l'equazione di Wigner scritta col termine di forza quantistica, ma solo finché non si annulli il gradiente della f_w rispetto a p . Nei punti in cui ciò si verifica, le curve che descrivono le traiettorie del sistema diventano singolari; si generano allora due casi: le traiettorie scompaiono (pozzi) oppure nascono (sorgenti). Una riprova immediata di ciò è che la f_w in alcuni casi assume valori negativi, anche se la funzione iniziale e le condizioni al contorno sono positive. Formalizziamo meglio il legame fra traiettorie classiche e quasi-traiettorie quantistiche: l'operatore $\tilde{F}$, quando applicato alla f_w , può essere scritto invertendo la (1.45), come

$$\tilde{F} = -\frac{\frac{\partial}{\partial t}f_w + p/m\nabla_x f_w}{\nabla_p f_w} \quad (1.46)$$

Quindi si possono definire le seguenti equazioni di Hamilton modificate [8]:

$$\begin{cases} \frac{dx}{dt} = \frac{p}{m} \\ \frac{dx}{dt} = \tilde{F} \end{cases} \quad (1.47)$$

La conservazione della f_w lungo tali curve, diretta conseguenza del teorema di Liouville, può essere verificata valutandone la derivata totale

$$\frac{df_w}{dt} = \frac{\partial f_w}{\partial t} + \nabla_r f_w \dot{p} + \nabla_r f_w \dot{x}$$

e sostituendovi la (1.46) e la (1.47), si ottiene ovviamente

$$\frac{df_w}{dt} = 0$$

1.12 Equazione di Boltzmann quantistica

In ultima analisi descriveremo come nell'equazione di Wigner si possa includere anche la descrizione dell'urto elettrone-fonone. Poiché non esiste una

teoria rigorosa che dia a questo proposito risultati semplici applicabili all'analisi del trasporto nei semiconduttori, si è costretti a ricorrere ad ipotesi intuitive e fisicamente ragionevoli. Poiché, come si è visto in sezione 1.8.6, il limite classico trasforma l'equazione di Wigner in quella di Liouville (che non è altro che un caso particolare dell'equazione di Boltzmann in assenza di collisioni), quando vengano considerati i fenomeni d'interazione a corto raggio, tale limite è ragionevole che dia proprio l'equazione di Boltzmann: è immediato verificare che ciò si ottiene quando si aggiunga al secondo membro della (1.28) il termine $\left. \frac{\partial f}{\partial t} \right|_{coll.}$ di formula (1.9). L'equazione che però si genera è praticamente irrisolvibile: si rende necessario quindi approssimare l'operatore collisionale di Boltzmann, ricorrendo alla sua approssimazione a “tempo di rilassamento” trovando:

$$\left. \frac{\partial f_w}{\partial t} \right|_{coll} = \frac{1}{\tau} \left[\frac{f_0(x, p)}{\int dp f_0(x, p)} \int dp f(x, p) - f(x, p) \right] .$$

Come descritto in appendice 5.2, quest'ultima deriva strettamente dall'approssimazione a tempo di rilassamento standard:

$$\left. \frac{\partial f}{\partial t} \right|_{coll.} \simeq \frac{f - f_0}{\tau} \quad (1.48)$$

L'analisi di quest'ultimo risultato segue dalla considerazione seguente:

i fenomeni collisionali sono fenomeni irreversibili che tendono a modificare la funzione di distribuzione e a portarla verso una distribuzione di equilibrio ed in questo senso possono essere visti come fenomeni di “rilassamento” del sistema. Tali fenomeni sono caratterizzati da una propria scala temporale ovvero hanno un caratteristico ordine di grandezza del tempo che rispecchia la “velocità” con cui il sistema “rilassa” alla funzione di equilibrio. Nell'approssimazione a tempo di rilassamento si introduce proprio un termine che forza la funzione di distribuzione a decadere esponenzialmente verso quella di equilibrio: lo si deve chiaramente dall'equazione (1.48) la cui soluzione è

proprio

$$f(t) = f_0 \exp^{-\frac{t}{\tau}} \quad (1.49)$$

Nei capitoli seguenti vedremo come questo risultato si inquadra nella visione più ampia di irreversibilità di moto (cap. 2) tradotta nel formalismo della matrice di Liouville (cap. 3) che segue dalla descrizione wigneriana del trasporto di carica.

Capitolo 2

Modelli numerici

2.1 Analisi generale dei sistemi irreversibili

Ogni sistema fisico reale complesso è ovviamente classificabile come processo irreversibile: infatti quella dell'irreversibilità è caratteristica propria di sistemi che contengano un gran numero di gradi di libertà interni (quella che in ambito modellistico si chiama una “dinamica nascosta”) e/o meccanismi di rilassamento (decadimento) che coinvolgano qualche (una o tutte) variabile termodinamica. Nel presente capitolo dimostreremo come l'irreversibilità sperimentale nei fenomeni di trasporto di un dispositivo reale, debba imputarsi alle seguenti cause:

- Il dispositivo è “aperto” nel senso che scambia particelle (e^-) con i circuiti con cui è contattato (che lo delimitano)
- La presenza di urti anelastici (ovvia causa dei fenomeni di decadimento)

La bontà di un modello numerico si giudicherà quindi dalla sua capacità di inglobare l'irreversibilità del sistema sotto questo duplice aspetto.

2.2 Sistemi “aperti”

Per “sistema aperto” in meccanica statistica si intende un sistema che globalmente non rispetta la legge di conservazione delle particelle. Il nostro interesse si limiterà all’analisi di sistemi aperti occupanti una “piccola” regione dello spazio all’interno della quale la materia è conservata, e connessi con altri di taglia molto più grande, che chiameremo ‘reservoirs’ (serbatoi) con i quali scambiano liberamente gli elettroni [1] [9]. Un circuito elettronico ha proprio tale particolarità: la sua interazione col mondo esterno avviene tramite due o più linee di adduzione (bus) che hanno il compito di rifornire il circuito di elettroni e di “drenare via” quelli in uscita del circuito stesso. Ogni bus connesso ai terminali del circuito è quindi una riserva di elettroni cui il dispositivo deve attingere per ricavarne la materia e l’energia necessarie al suo funzionamento.

2.2.1 Reservoirs

Nell’ambito dell’analisi cinetica dei circuiti submicrometrici, l’uso dei reservoirs si pone come una naturale necessità in quanto rappresenta la traduzione statistica del concetto di pila ideale, cui spesso si ricorre nella teoria elementare dei circuiti. Una pila diventa ideale quando la tensione ai suoi capi non dipende dalla corrente erogata; analogamente, un serbatoio di carica diverrà un reservoir ideale se il suo potenziale chimico (livello di Fermi) non varia a causa di una perdita (o accumulo) di elettroni. Nell’ottica statistica del presente lavoro la “pila ideale” quindi si comporta come un sistema “grande” (o meglio a molte particelle), che “interagisce debolmente” con l’esterno (come un grande serbatoio in cui il pelo libero dell’acqua varia talmente poco da potersi considerare a livello costante) [1]. Avendo messo in

risalto che i buses di alimentazione di un circuito non sono altro che serbatoi di carica caratterizzati dal proprio livello di Fermi, definiamo quali caratteristiche debbano essere loro attribuite nel ristretto ambito della simulazione circuitale. Un “bus” (reservoir) deve avere proprietà analoghe a quelle di un corpo nero: la distribuzione di elettroni che emette nel dispositivo è una distribuzione di equilibrio termico, e tutti gli elettroni che giungono al reservoir dal dispositivo, vengono assorbiti senza riflessione. Per un’implementazione pratica di tale caratteristica quindi, si deve essere in grado di poter discriminare il segno della velocità degli elettroni e la posizione dei contorni del dispositivo. Così la funzione di Wigner, grazie alla sua formulazione cinetica, sarà il contesto ideale per la rappresentazione dei sistemi aperti.

Vediamo, come anticipato, quale legame c’è fra reservoir, “dinamica nascosta”, e irreversibilità del sistema fisico. Che sistema aperto sia sinonimo di sistema irreversibile lo si intuisce facilmente se si pensa che ogni qual volta una particella lascia il sistema, con essa si perde anche parte dell’informazione globale sullo stato dell’intero sistema. La connessione fra sistema aperto e dinamica delle variabili “nascoste” invece discende dall’analogia col “corpo nero” attribuita al reservoir: da quanto detto prima, la distribuzione che ne esce (e che quindi entra nel dispositivo) è indipendente da quella che vi entra. Questo è come dire che i gradi di libertà all’interno del reservoir sono così numerosi che si perde ogni informazione su un’arbitraria particella che vi entri. In altre parole, che le particelle al suo interno sono talmente tante che il loro moto è praticamente indipendente dalle poche particelle che vi giungono dal dispositivo. La connessione circuito-reservoir è un processo analogo alla dispersione termica che avviene al contatto col terreno di un grave in caduta: la dinamica deterministica della traiettoria del grave, al suo impatto col suolo si disperde in un’infinità di moti caotici di tipo termico impossibili da

seguirsi e quindi, nel loro complesso, tipicamente irreversibili.

2.2.2 Condizioni al contorno di tipo inflow

Vediamo come, grazie alla formulazione cinetica di Wigner, si riesce a modellizzare l'apertura del sistema, tramite naturali condizioni al contorno, note come “condizioni di inflow”. Per implementare la visione descritta precedentemente, ovvero che la particelle entranti nel sistema dipendono solo dallo stato del reservoir, e che quelle uscenti dipendano solo dallo stato del dispositivo, dobbiamo applicare all'equazione di Wigner le condizioni al contorno illustrate in fig(2.1). Matematicamente:

$$f(0, p)|_{p>0} = f_0^{sinistra}(p)$$

$$f(L, p)|_{p<0} = f_0^{destra}(p)$$

in cui la f_0 è (in entrambi i casi) la distribuzione di equilibrio termico di Fermi integrata rispetto le variabili trasverse ovvero [1]:

$$f_0(p) = \frac{K_B T m^*}{\pi \hbar^2} \ln \left\{ 1 + e^{-[p^2/(2m) - \mu_{d,s}]/K_B T} \right\} \quad (2.1)$$

con $\mu_{d,s}$ abbiamo indicato il potenziale chimico nei due lati del semiconduttore (rispettivamente destro e sinistro), il cui valore è individuabile imponendo la condizione di neutralità della carica nei contatti di anodo e di catodo.

Come ci aspettavamo dalla precedente discussione, tali condizioni al contorno non sono invarianti per inversione temporale. Una tale operazione richiederebbe infatti di cambiare il problema di fig(2.1/A) con quello di fig(2.1/B). Come vedremo l'irreversibilità temporale delle condizioni al contorno è di cruciale importanza perché garantisce stabilità al sistema numerico.

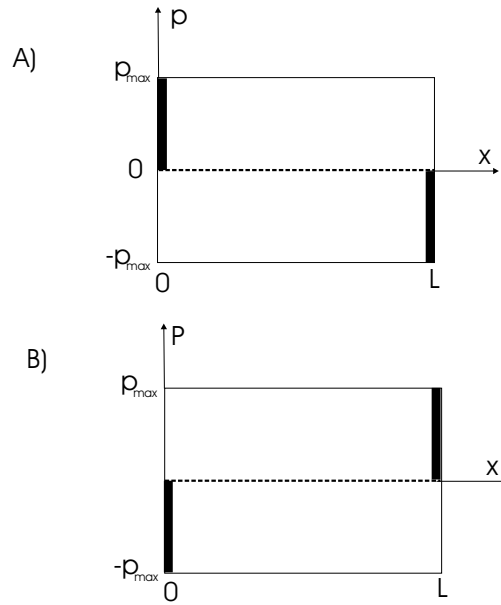


Figura 2.1: Possibili condizioni di inflow per la funzione di Wigner: i contorni segnati sono le zone in cui la f_0 è fissata; in fig. A), riguardo alle particelle in $x = 0$ ad esempio, tali condizioni permettono a quelle con velocità negativa, di uscire liberamente dal dispositivo per disperdersi nei reservoirs; mentre impone alle sole particelle entranti di seguire la statistica dei reservoirs

2.3 Modelli basati sull'equazione di Schrödinger

Prima di passare alla descrizione del modello cinetico di Wigner, analizzeremo in che modo i modelli standard basati sull'equazione di Schrödinger riescano a descrivere i fenomeni irreversibili, quali problemi ed errori ne derivino e come essi siano superati nella trattazione cinetica di Wigner. E' d'uopo premettere una breve introduzione sul concetto di stati “misti” e di stati “puri”.

2.3.1 Stati “misti” e stati “puri”

Lo stato di un sistema quantistico è definito “puro” se può essere ben descritto tramite una sola funzione d'onda ψ , che risolva l'equazione di Schrödinger

per un particolare autovalore E . Questo si verifica sempre quando si analizza il moto di una singola particella o di un insieme di particelle “preparate in maniera identica”, (ovvero che presentino le stesse caratteristiche energetiche) [4]. Quando, come nel presente caso, tale limite non è rispettato, si dimostra che in generale la descrizione dello stato del sistema richiede una congrua combinazione lineare delle funzioni d’onda, ognuna relativa ad un particolare autovalore. Tali stati vengono definiti “misti” e la matrice di densità ad essi associata si scrive

$$\rho(r, s) = \sum_{n,m} A_{n,m} \psi_n(r) \overline{\psi_m(s)} \quad (2.2)$$

Quando $A_{n,m} = \delta_{n,m}$ si dice che la matrice di densità è in forma diagonale. Ricordiamo il legame che sussiste fra funzione di Wigner e funzione d’onda di Schrödinger per uno stato puro

$$f_w(x, v, t) = \frac{1}{(2\pi)^{3M}} \int_{\mathbb{R}_\eta^{3M}} \overline{\psi\left(x + \frac{\hbar}{2m}\eta, t\right)} \psi\left(x - \frac{\hbar}{2m}\eta, t\right) e^{iv\eta} d\eta \quad .$$

Se si passa da uno stato puro ad uno misto la formula precedente si modifica [3]:

$$f_w(x, v, t) = \frac{1}{(2\pi)^{3M}} \sum_{l,j \in \mathbb{N}} \rho_{l,j} \int_{\mathbb{R}_\eta^{3M}} \overline{\psi^l\left(x + \frac{\hbar}{2m}\eta, t\right)} \psi^j\left(x - \frac{\hbar}{2m}\eta, t\right) e^{iv\eta} d\eta \quad .$$

Indipendentemente dal fatto che si usi una o l’altra formula, f_w deve in entrambi i casi risolvere la stessa equazione di Wigner: da ciò segue che nella trattazione di Wigner gli stati puri e quelli misti vengono descritti alla stessa maniera. Con la presente precisazione si è voluto mettere in evidenza che mentre l’equazione di Schrödinger riesce a descrivere solo stati puri (e l’analisi degli stati misti necessita di una somma di soluzioni indipendenti), l’equazione di Wigner descrive senza problemi (in maniera semplice) la dinamica degli stati misti, trattandoli allo stesso modo degli stati puri: una

soluzione dell'equazione di Wigner non “sa” se sta descrivendo uno stato puro o uno misto. Questa maggiore versatilità della descrizione Wigneriana rispetto alla tradizionale di Schrödinger fa sì che essa sia immune dagli errori e da alcuni limiti che affliggono una risoluzione pratica dell'equazione di Schrödinger, che sotto analizzeremo.

2.3.2 Descrizione degli urti nella formulazione di Schrödinger

Vediamo subito come in generale nei sistemi “tipo Schrödinger” il generico concetto di irreversibilità si coniughi necessariamente con “non conservazione della carica”. Ricordiamo che un sistema diventa irreversibile quando sono presenti termini che tendono nel tempo a far decadere, tipicamente in maniera esponenziale, alcune grandezze fisiche a particolari valori detti di equilibrio. Poiché in meccanica quantistica solo $\exp^{i\frac{Et}{\hbar}}$ dipende dal tempo, quando si descrive un sistema irreversibile bisogna far sì che l'autovalore E acquisti una componente immaginaria positiva. Ma poiché l'esponenziale precedente modula l'evoluzione temporale della funzione d'onda, conseguentemente anche la densità ($|\psi|^2$) subirà uno “smorzamento” nel tempo e si verrà a crearsi un deficit nel calcolo della carica totale del sistema. Il problema non si pone nella formulazione wigneriana poiché nella descrizione di moto della densità di quasi-probabilità, manca ogni esplicita indicazione degli autovalori dell'hamiltoniana del sistema [10].

Vediamo in maniera più sistematica come si approssimano i fenomeni di urto secondo il modello di Schrödinger. A causa della mancanza di una trattazione rigorosa percorribile, si ricorre all'approssimazione della “Master Pauli equation” [4]. L'assunzione che sta alla base della Master Pauli equation è che gli elettroni occupino solo autovalori dell'hamiltoniana, non sovrapposizioni di tali stati; ovvero che ρ sia diagonale secondo la definizione data nella

sez.(2.3.1)

$$\rho(r, s; t) = \sum_i P_i(t) \psi_i(r) \overline{\psi_i(s)} \quad (2.3)$$

in cui $P_i(t)$ è la probabilità che il sistema sia nello stato i . Nei processi di urto si assume che la probabilità di transizione fra due differenti autostati segua la regola aurea di Fermi [4] [11]. Vediamo come tale approssimazione porti alla violazione della legge di continuità della carica. Si considerino due autostati “legati”, ovvero localizzati in due punti differenti dello spazio che chiameremo x_0 e x_1 . Le funzioni d'onda dei due autostati quindi decadranno rapidamente a zero al di fuori dell'intorno rispettivamente di x_0 e x_1 ; ovvero si comporteranno, in via approssimativa, come $\psi_0(x) \sim \delta(x - x_0)$; $\psi_1(x) \sim \delta(x - x_1)$. Se la particella è localizzata in x_0 al tempo t_0 (ovvero descritta da ψ_0), a causa dello scattering anelastico introdotto nel modello, dopo un tempo $t_1 > t_0$, in accordo con la regola aurea, nascerà una probabilità non nulla che sia avvenuta una transizione fra i due stati; il che si traduce con una diminuzione del modulo di ψ_0 ed un aumento di quello di ψ_1 . Questo è un chiaro fenomeno di trasporto di carica, a causa del quale la funzione di distribuzione varierà nel tempo

$$\left. \frac{\partial n(x)}{\partial t} \right|_{x_0} < 0$$

$$\left. \frac{\partial n(x)}{\partial t} \right|_{x_1} > 0$$

Si dimostra però che in generale la densità di corrente che scorre fra due autostati di H deve essere uniforme ($\nabla J = 0$) [4]; da ciò segue immediatamente che la legge di conservazione non è più soddisfatta

$$\nabla J - \frac{\partial n(x)}{\partial t} \neq 0$$

L'errore cui si è giunti è diretta conseguenza dell'iniziale assunzione che siano significativi i soli termini diagonali della matrice di densità. Come puntual-

izzato in sez.(2.3.1) la formulazione di Wigner elimina il problema alla radice in quanto non pone nessun vincolo alla formula (2.2): riprova di ciò è che nel nostro modello cinetico, l'equazione di continuità sarà soddisfatta in maniera esatta (nei limiti degli errori macchina del calcolo) anche in presenza di urti.

2.4 Modello di Wigner

2.4.1 Apertura del sistema

Nella presente sezione vedremo, da un punto di vista matematico, come le condizioni al contorno proprie di un sistema aperto, introducano nella soluzione dell'equazione di Wigner termini che decadono esponenzialmente ad una soluzione stazionaria, ovvero un comportamento di tipo irreversibile della soluzione stessa. Apparirà quindi chiaro come la classificazione “sistema aperto” sia strettamente connessa a quella di sistema irreversibile come già messo in luce precedentemente.

Nel capitolo 3 dimostreremo che la soluzione dell'equazione di Wigner può essere riportata alla soluzione del seguente sistema lineare

$$\frac{\partial f}{\partial t} = \mathbf{L}f + b \quad (2.4)$$

in cui f e b sono vettori colonna ed $\mathbf{L}$ una matrice a coefficienti reali chiamata matrice di Louville. Dalla teoria dei sistemi di equazioni lineari sappiamo che condizione necessaria e sufficiente affinché il sistema precedente ammetta una soluzione $f(t)$ limitata, è che gli autovalori di $\mathbf{L}$ siano a parte reale negativa o nulla e quelli a parte reale nulla abbiano molteplicità algebrica pari a quella geometrica.

Gli autovalori di $\mathbf{L}$ a parte reale positiva corrispondono infatti, nel tempo, a soluzioni che divergono in maniera esponenziale e quindi ovviamente non

fisiche. Viceversa autovalori parte reale negativa sono legati a soluzioni modulate da un esponenziale decrescente: tali soluzioni descrivono stati transitori evanescenti, per cui la soluzione asintotica della (2.4) per t che tende all'infinito, dipenderà solo dal termine forzante b . Un esempio esplicito è il caso del semplice circuito di figura (2.2), che obbedisce ad un'equazione formalmente identica alla (2.4).

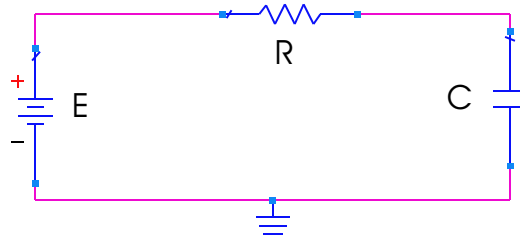


Figura 2.2: Circuito R C

L'informazione iniziale relativa allo stato del sistema (che in tal caso sarà ovviamente la tensione del condensatore v_{c_0}) viene persa e sostituita da quella imposta dal generatore secondo la formula

$$v_c(t) = (1 - e^{-t/\tau})(E - v_{c_0}) + v_{c_0}$$

Quindi se la parte reale degli autovalori di $\mathbf{L}$ è minore di zero, come nell'esempio del condensatore, il nostro modello descriverà un tipico comportamento di sistema irreversibile: verifichiamo qui di seguito che per il solo fatto che il sistema è aperto, gli autovalori di $\mathbf{L}$ hanno parte reale non nulla. La stabilità del modello numerico dipenderà quindi da una congrua scelta delle condizioni al contorno, in quanto è proprio tramite esse che si descrive l'apertura del sistema fisico ed inoltre si determina il segno della parte reale degli autovalori.

Sfrutteremo la seguente proprietà della $\mathbf{L}$: si può dimostrare che $i\mathbf{L}$ è hermitiana se e solo se lo è anche l'Hamiltoniana del sistema [10]. Dalle note

proprietà delle matrici hermitiane segue che se $i\mathbf{L}$ è hermitiana, gli autovalori di $\mathbf{L}$ saranno immaginari puri: la soluzione della (2.4) resta limitata e perciò fisicamente ragionevole. Quella dell'hermitianità di H si pone quindi come una preziosa proprietà di stabilità del sistema; si noti però che essa origina soluzioni oscillanti non smorzate quindi che non rilassano a valori di equilibrio come ci aspettiamo avvenga in un sistema fisico reale. Vedremo di seguito che un sistema "aperto" raramente possiederà un'hamiltoniana hermitiana. Si può notare, con l'applicazione del teorema di Green, che l'Hamiltoniana di un generico sistema rimane hermitiana se

$$\int_S J ds = 0 \quad (2.5)$$

dove J è il flusso di corrente e S è il dominio delle variabili cinetiche del sistema; è chiaro che se il sistema è "chiuso" oppure S è tutto lo spazio $\mathbb{R}^6$, H è hermitiana (infatti si assume sempre che la corrente all'infinito tenda rapidamente a zero). Nel caso di dominio limitato invece H rimane hermitiana solo se il flusso di particelle uscenti dal sistema è uguale a quello entrante. Poiché il numero totale di particelle di un sistema aperto può cambiare in risposta a condizioni esternamente imposte, tale richiesta appare troppo restrittiva. Un esempio pratico è lo stato risonante dell'RTD, che mostra un picco di densità di carica nel pozzo quantistico; perciò durante il processo di costruzione di tale stato, deve essere avvenuto un moto netto di carica entrante nel dispositivo, in aperta violazione della (2.5)

Concludendo: l'hermitianità di $i\mathbf{L}$ segue direttamente dal fatto che, in un sistema chiuso, lo è H . Un tale sistema (in assenza di urti), può mostrare solo un comportamento oscillatorio. Se si cambiano le condizioni al contorno, ovvero si "apre" il sistema, così da dividere fra le particelle che entrano e quelle che escono dal sistema, si viola l'hermitianità della matrice di Liouville. Ciò introduce una parte reale per alcuni autovalori di $\mathbf{L}$, originando

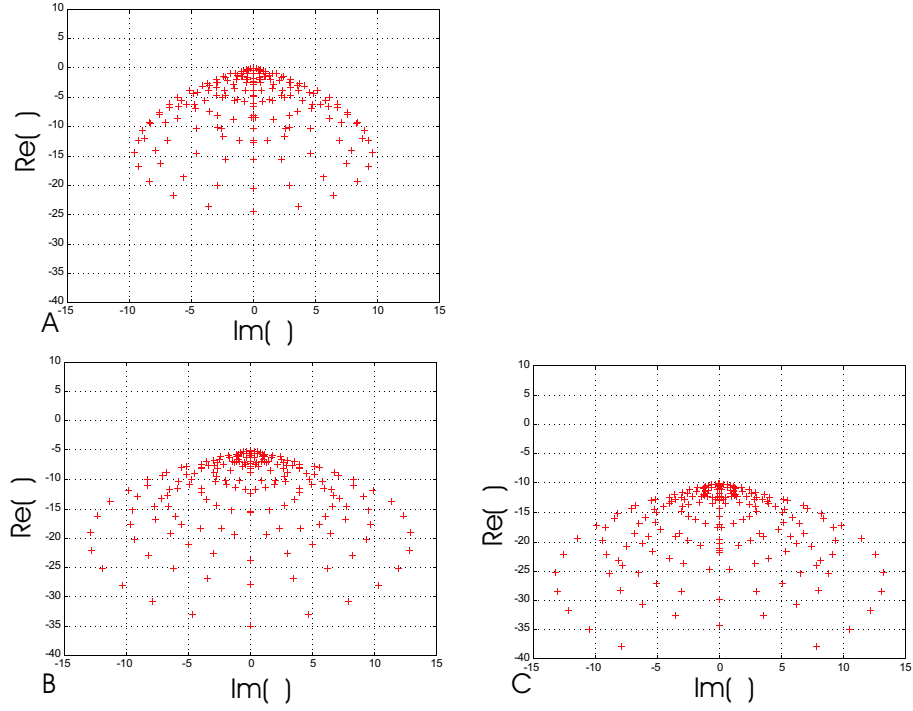


Figura 2.3: Autovalori di $\mathbf{L}$ per $N_p = 7$, $N_x = 7$. Fig.A $\tau = \infty$; fig.B $\tau = 1ps$; fig.C $\tau = 0.1 ps$

un comportamento temporale della soluzione di tipo esponenziale. Si dimostra che condizioni al contorno reversibili originano soluzioni divergenti, mentre quelle irreversibili (come quelle di inflow) portano a soluzioni stabili e decrescenti nel tempo [12].

2.4.2 Collisioni

Come già rimarcato nell'introduzione del presente capitolo, ogni sistema fisico, a causa dei processi dissipativi contenuti al suo interno, è “naturalmente” indotto ad uno stato di equilibrio, tanto che si può parlare di questa come di una semplice e fondamentale legge di natura. Sarà ragionevole aspettarsi dal nostro modello che l'inclusione delle collisioni elettrone-fonone descritte dal

termine $\frac{\partial f}{\partial t} \Big|_{col}$ tenda ad accentuare la proprietà di irreversibilità del sistema, diminuendo la parte reale degli autovalori di $\mathbf{L}$. La verifica di ciò è riportata in fig.2.3 in cui si illustrano le posizioni nel piano complesso degli autovalori al variare del tempo medio fra due collisioni (τ). Come ci aspettiamo, al diminuire di τ , il processo che porta il sistema all'equilibrio si accentua, diminuendo i tempi di rilassamento.

Capitolo 3

Modello discreto

3.1 Introduzione

La complessità dei problemi connessi con la simulazione del trasporto di carica nei semiconduttori, è tale che non è possibile sfruttare casi di moto analiticamente risolubili, per fare previsioni attendibili sul comportamento finale dei dispositivi reali. I risultati del modello Wigner-Poisson devono essere quindi valutati numericamente, ed il primo passo per farlo è quello di sostituire al dominio continuo delle variabili cinetiche una griglia di punti discreta. Riportiamo, per comodità del lettore, il sistema Wigner-Poisson nella sua formulazione unidimensionale continua (in forma integrale)

$$\frac{\partial f_w}{\partial t} = -\frac{p}{m} \frac{\partial f_w}{\partial x} - \frac{1}{\hbar} \int_{-\infty}^{+\infty} \frac{dp'}{2\pi\hbar} V(x, p-p') f_w(x, p') + \frac{\partial f_w}{\partial t} \Big|_{coll} \quad (3.1)$$

$$\epsilon \frac{d^2 U(x)}{dx^2} = q^2 [C(x) - n(x)] \quad (3.2)$$

in cui ϵ è la permeabilità dielettrica del semiconduttore, $C(x)$ la densità di atomi droganti e $n(x)$ la densità di elettroni liberi data da

$$n(x) = \int_{-\infty}^{+\infty} f_w(x, p) dp \quad (3.3)$$

La corrente è data da

$$J(x) = -\frac{q}{m} \int_{-\infty}^{+\infty} f_w(x, p) p dp \quad (3.4)$$

$V(x, p)$ nella (3.1) è dato da

$$V(x, p) = 2 \int_0^{+\infty} \sin\left(\frac{pr}{\hbar}\right) \left[U\left(x + \frac{r}{2}\right) - U\left(x - \frac{r}{2}\right) \right] dr \quad (3.5)$$

e $\left. \frac{\partial f_w}{\partial t} \right|_{coll}$

$$\left. \frac{\partial f_w}{\partial t} \right|_{coll} = \frac{1}{\tau} \left[\frac{f_0(x, p)}{\int dp f_0(x, p)} \int dp f(x, p) - f(x, p) \right] \quad (3.6)$$

Descriviamo velocemente quale sia la maggiore complessità nella risoluzione del sistema non lineare W-P, rispetto al caso originariamente lineare della singola equazione di Wigner. Nel caso lineare, dato un profilo di potenziale tempo invariante e attribuita una qualsiasi condizione iniziale alla f_w , l'equazione evolutiva di Wigner ammette soluzioni che tendono asintoticamente alla soluzione dell'equazione stazionaria risolta rispetto al medesimo profilo di potenziale. Come rimarcato ampiamente nel precedente capitolo, ogni soluzione stazionaria è quindi una soluzione (punto di equilibrio) stabile dell'equazione di Wigner. Ciò verrà sfruttato nel capitolo 4 per determinare i tempi di transizione fra due diversi punti di polarizzazione dell'RTD (ovvero il tempo di switch): la procedura sarà la seguente

- Si calcola la soluzione stazionaria per una fissata tensione di polarizzazione V_0 .
- Si fa evolvere la funzione precedente cambiando la tensione di polarizzazione da V_0 a V_1 .
- Per t tendente all'infinito il risultato finale sarà la soluzione stazionaria calcolata con una ddp ai capi dell'RTD di V_1 .

L'accoppiamento dell'equazione Wigner con quella di Poisson, rendendo il sistema non lineare, limita notevolmente la possibilità di fare previsioni sul comportamento della soluzione, come avveniva nel caso precedente; infatti in generale un sistema non lineare potrà presentare cicli limite e molteplici punti di equilibrio stabili o instabili: tutte situazioni di difficile analisi. Lo strumento di calcolo che ne deriva, più sofisticato e complesso, ci darà la possibilità di descrivere fenomeni non banali, che trovano riscontri sperimentali proprio nell'analisi delle proprietà degli RTD. Le caratteristiche dei sistemi non lineari che rendono più aderente alla realtà il nostro modello sono le seguenti:

Esistenza di molteplici punti di equilibrio: Poiché nella simulazione dell'RTD verificheremo la presenza di due punti di equilibrio stabili (due distinte soluzioni del sistema W-P con ognuna il proprio dominio di attrazione), saranno proprio le CI della f_w a decidere lo stato finale del dispositivo; in altre parole, per una stessa tensione applicata, saranno ammissibili due diversi valori di corrente: il diagramma $I(V)$ presenterà quindi un ciclo d'isteresi.

Cicli limite: La presenza di un ciclo limite si tradurrà invece con la nascita di una soluzione oscillante, non convergente a nessuna funzione stazionaria: tale fenomeno è la spiegazione numerica delle oscillazioni (cosiddette “intrinseche”), presenti ogni qual volta si cerchi di polarizzare il diodo nella sua zona di instabilità.

3.2 Equazione di Wigner stazionaria

Nel presente paragrafo illustreremo il metodo numerico che permette di risolvere la (3.1) nel caso stazionario, ovvero quando f_w non dipenda esplici-

tamente dal tempo. Poniamo quindi a primo membro $\frac{\partial f_w}{\partial t} = 0$. La presente procedura di discretizzazione, deriva da quella proposta da Frensley [10]: le variabili di posizione x e momento p saranno discretizzate in N_x e N_p punti rispettivamente

$$\begin{aligned} x_i &= i\Delta_x & \text{con } i &= 1, 2, \dots, N_x \\ p_j &= \frac{\pi\hbar}{\Delta_x} \left[\frac{(j-1/2)}{N_p} - \frac{1}{2} \right] & \text{con } j &= 1, 2, \dots, N_p \end{aligned}$$

Come appare dalla formula precedente, per quanto riguarda l'asse p , si è scelto un passo di campionamento pari a $\Delta_p = \frac{\pi\hbar}{N_p\Delta_x}$. Tale scelta è condizionata dalla richiesta che i punti siano simmetrici rispetto all'asse $p = 0$ e soddisfino l'equazione di “completezza di Fourier” che verrà discussa nel paragrafo (3.3) [10].

Il secondo membro della (3.1) è il risultato della somma di tre contributi, ognuno dei quali può essere interpretato come il risultato dell'azione di un operatore che agisca sulla f_w . Se chiamiamo tali operatori rispettivamente: operatore traslazione $\mathcal{T}$, operatore campo $\mathcal{V}$ e operatore collisionale $\mathcal{C}$, possiamo riscrivere la (3.1) nella forma più compatta

$$\mathcal{T}f_w + \mathcal{V}f_w + \mathcal{C}f_w = 0 \quad (3.7)$$

Analizziamo separatamente ogni operatore:

Operatore campo

$$\mathcal{V}f_w \doteq -\frac{1}{\hbar} \int_{-\infty}^{+\infty} \frac{dp'}{2\pi\hbar} V(x, p - p') f_w(x, p')$$

La versione discreta dell'operatore di campo è facilmente definibile, una volta approssimati gli integrali tramite somme finite, trovando:

$$(\mathcal{V}f)_{i,j} = -\frac{1}{\hbar} \sum_{j'=1}^{N_p} V_{i,j-j'} f_{i,j'} \quad (3.8)$$

in cui l'operatore discreto $V_{i,j}$ deriva dalla (3.5) e si scrive

$$V_{i,j} = \frac{2}{N_p} \sum_{i'=1}^{N_x/2} \sin\left(\frac{2j\Delta_p i' \Delta_x}{\hbar}\right) (U_{i+i'} - U_{i-i'}) \quad (3.9)$$

Operatore di traslazione

$$\mathcal{T}f_w \doteq -\frac{p}{m} \frac{\partial f_w}{\partial x}$$

La discretizzazione di $\mathcal{T}$ richiede la delicata operazione della stima del gradiente della f_w , per mezzo di tecniche numeriche. La più semplice approssimazione è la seguente (approssimazione di Eulero) [13]

$$\frac{\partial f_w}{\partial x} \approx \Delta_{UDS}^{\pm} f(i) = \pm[f(i) + f(i \pm 1)]/\Delta_x$$

in cui UDS è l'acronimo di “Schema differenziale up-downwind”. L'ambiguità del segno in formula precedente, può essere rimossa se si considera il ruolo svolto dalle condizioni al contorno: immaginiamo che quest'ultime siano rappresentate tramite punti extra della griglia di discretizzazione, posti appena al di fuori del dominio di f_w , come illustrato in fig (3.1). I valori della f_w lungo i contorni segnati sono fissati, e rappresentano “il numero di particelle entranti nel dispositivo” per unità di tempo; tali valori devono essere accoppiati con ogni punto all'interno del dominio di simulazione e ciò è reso possibile se si utilizza Δ^- per $p > 0$ e Δ^+ per $p < 0$; così quando $i = 1$ e $p > 0$, Δ^- richiederà i valori di $f_{i,j}$ per $x < 0$, ovvero proprio i punti extra della griglia fissati dalle condizioni al contorno. Analogamente per $p < 0$, Δ^+ coinvolgerà i valori di $f_{i,j}$ nei punti $x > L$. Lo schema UDS rivela un'accuratezza dell'ordine $O(\epsilon)$, in cui ϵ è proporzionale al tasso di discretizzazione; pur preservando una discreta rapidità di calcolo, non riesce a espletare un'accurata modellizzazione del diodo RTD specie nella zona di risonanza. Si introduce quindi uno schema differenziale del secondo ordine

(SDS) costruito su tre punti della funzione anziché due, accurato all'ordine $O(\epsilon^2)$ [14].

Se si considerano tre punti di campionamento di una generica funzione continua (indicati con f_i), quest'ultima potrà essere “fittata” con un polinomio del secondo ordine che passi per i tre punti dati. Ovvero verrà approssimata con

$$f(x) \approx \eta(\eta - 1)f_{i-1}/2 + (\eta^2 - 1)f_i + \eta(\eta + 1)f_{i+1}/2 \quad (3.10)$$

in cui $x = x_i + \eta\Delta_x$ e η varia fra -1 e 1. La derivata di f rispetto ad x si scrive

$$\frac{df(x)}{dx} \approx \left[\left(\eta - \frac{1}{2} \right) f_{i-1} + 2\eta f_i + \left(\eta + \frac{1}{2} \right) f_{i+1} \right] \frac{1}{\Delta_x} \quad (3.11)$$

se si pone $\eta = 0$ si ottiene lo schema standard alle differenze centrali, che si dimostra inadeguato alle nostre esigenze in quanto valido solo per potenziali simmetrici (ovvero per la simulazione del diodo non polarizzato) [13]. Perchè i vantaggi relativi alla possibilità di legare in maniera automatica i valori delle C.C. con i corrispondenti punti interni al dominio di simulazione, che ha lo schema up-downwind, siano

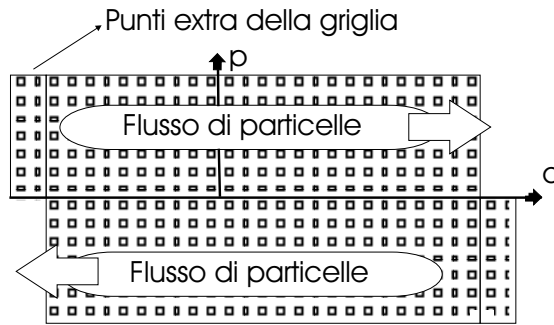


Figura 3.1: Griglia di discretizzazione del piano delle fasi (i quadratini rappresentano i punti di campionamento della f_0): i punti extra della griglia contengono i valori della distribuzione “noti a priori”

mantenuti anche nell'analogo schema al secondo ordine, si dovrà porre $\eta = \pm 1$. Esplicitando, la $\Delta^\pm$ diventa

$$\Delta_{DS}^\pm f_i = \pm \frac{1}{2} [-3f_i + 4f_{i\pm 1} - f_{i\pm 2}] \quad (3.12)$$

Quindi $\mathcal{T}$ si scrive

$$(\mathcal{T}f_w)_{i,j} = -\frac{p_j}{m\Delta_x} \times \begin{cases} +\frac{1}{2} [-3f_i + 4f_{i+1} - f_{i+2}] & \text{per } p_j < 0 \\ -\frac{1}{2} [-3f_i + 4f_{i-1} - f_{i-2}] & \text{per } p_j > 0 \end{cases}$$

Operatore collisionale

$$\mathcal{C}f_w \doteq \frac{1}{\tau} \left[\frac{f_0(x, p)}{\int dp f_0(x, p)} \int dp f(x, p) - f(x, p) \right]$$

In questo caso la discretizzazione è immediata e porta a

$$(\mathcal{C}f_w)_{i,j} = \frac{1}{\tau} \left[\frac{f_{i,j}^0 \Delta_p}{\rho_{i,j}^0} \sum_{j'=1}^{N_p} f_{i,j'} - f_{i,j} \right] \quad (3.13)$$

Per risolvere il sistema Wigner-Poisson tramite calcolatore, è necessario riformularlo in termini di equazione matriciale. A questo fine, definiamo un vettore colonna f tale che la sua $[(i-1)N_p + j]$ -esima componente, sia proprio $f_{i,j}$. Con tale assunzione ogni operatore della (3.7) rappresenterà l'azione di un'opportuna matrice sul vettore f . Conseguentemente introducendo gli operatori matriciali $\mathbf{T}$, $\mathbf{V}$ e $\mathbf{C}$ corrispondenti agli operatori $\mathcal{T}$, $\mathcal{V}$ e $\mathcal{C}$, la (3.7) potrà essere riscritta tramite il seguente sistema lineare

$$(\mathbf{T} + \mathbf{V} + \mathbf{C})f = \mathbf{L}f = b \quad (3.14)$$

Poiché le condizioni di inflow richiedono di fissare il valore della f_w sui contorni, tali vincoli si inseriscono facilmente nel calcolo definendo un vettore b in cui gli unici elementi non nulli siano quelli relativi ai confini del domino di simulazione e che abbiano proprio il valore della distribuzione di equilibrio

termico dei reservoirs. La matrice $\mathbf{L}$ risultante è detta “matrice di Liouville” e vedremo che la sua forma sarà quella di una matrice a $(N_x \times N_x)$ blocchi di dimensione $(N_p \times N_p)$ ciascuno. Di seguito indicheremo tali matrici con quattro indici, i primi si riferiranno al blocco considerato, i restanti allo scostamento relativo all’interno del blocco stesso:

$$\mathbf{L}_{i,i';j,j'} \quad \text{in cui} \quad \begin{cases} i, i' = 1, \dots, N_x \\ j, j' = 1, \dots, N_p \end{cases}$$

Definiamo matrice $\mathbf{V}$ dell’operatore di potenziale, la seguente

$$\mathbf{V}_{i,i';j,j'} = \delta_{i,i'} V_{i,|j-j'|} \quad (3.15)$$

con $V_{i,j}$ definito nella (3.9). Rimarchiamo che gli unici blocchi non nulli di $\mathbf{V}$ sono quelli della diagonale principale. Come si può facilmente valutare, la matrice derivante dalla formula (3.8), mappata su di un vettore con la regola descritta per la f_w , può essere scritta come il risultato dell’usuale prodotto fra la matrice $\mathbf{V}$ e il vettore relativo alla soluzione $f_{i,j}$. Analogamente l’operatore traslazione si scriverà:

$$\mathbf{T}_{i,i';j,j'} = -\frac{p_j}{2m\Delta_x} \delta_{j,j'} \times \begin{cases} -3\delta_{i,i'} + 4\delta_{i+1,i'} - \delta_{i+2,i'} & \text{per } p_j < 0 \\ +3\delta_{i,i'} - 4\delta_{i-1,i'} + \delta_{i-2,i'} & \text{per } p_j > 0 \end{cases}$$

Infine per l’operatore collisionale

$$\mathbf{C}_{i,i';j,j'} = \frac{\delta_{i,i'}}{\tau} \left[\frac{f_{i,j}^0 \Delta_p}{\rho_{i,j}^0} f_{i,j'} - f_{i,j} \right] \quad (3.16)$$

Il numero finale degli elementi da immagazzinare in $\mathbf{L}$ è $(4N_p + 1)N_p N_x$, e la sua topologia è illustrata in fig. 3.2, in cui è chiaramente evidente la forma pentadiagonale a blocchi di $\mathbf{L}$, che ha il pregio non trascurabile di ridurre enormemente le risorse necessarie a risolvere il sistema lineare rispetto al caso generale di matrice sparsa. L’algoritmo sfruttato per il calcolo del vettore f_w ,

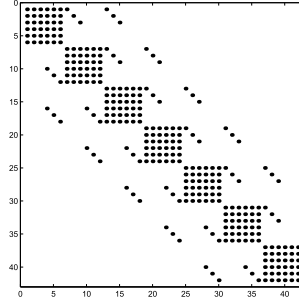


Figura 3.2: Matrice di Liouville per il metodo risolutivo matriciale della funzione di Wigner, per $N_p = 6$, $N_x = 7$

grazie alla forma peculiare di $\mathbf{L}$, limita il costo computazionale ad un ordine di $N_x \times N_p^3$ moltiplicazioni. Riportiamo la forma finale del sistema lineare

$$\begin{bmatrix} (\mathbf{T} + \mathbf{V} + \mathbf{C})_{1,1} & \mathbf{T}_{1,2} & 0 & 0 \\ \mathbf{T}_{2,1} & (\mathbf{T} + \mathbf{V} + \mathbf{C})_{2,2} & \mathbf{T}_{2,3} & 0 \\ 0 & \ddots & \ddots & \ddots \\ 0 & 0 & \mathbf{T}_{N_x, N_x-1} & (\mathbf{T} + \mathbf{V} + \mathbf{C})_{N_x, N_x} \end{bmatrix} \begin{bmatrix} f_1 \\ f_2 \\ \vdots \\ f_{N_x} \end{bmatrix} = \begin{bmatrix} b_1 \\ 0 \\ \vdots \\ b_{N_x} \end{bmatrix}$$

Come si vede dalla figura e dalle formule, i blocchi al di fuori della diagonale principale sono associati alle parti dell'operatore di traslazione corrispondenti ai termini $f(i \pm 1, j)$ e $f(i \pm 2, j)$.

3.3 Calcolo dei momenti

Riportiamo il calcolo dei momenti di ordine zero e primo sia della funzione di Wigner (ovvero il calcolo della densità di carica e di corrente) sia dell'omonima equazione (da cui si originano le equazioni di continuità e di bilancio del momento primo).

Densità di carica:

$$n(x, t) = \int f(x, p, t) dp$$

Coerentemente con l'approssimazione degli integrali tramite somme finite impiegata nella sez.(3.2) il calcolo della densità di carica si converte semplicemente in:

$$n(x_i, t) = \Delta_p \sum_{j=1}^{N_p} f_{i,j}(t) \quad (3.17)$$

Densità di corrente

$$-\frac{q}{m} \int f(x, p, t) p dp = J(x)$$

Il calcolo del flusso di corrente che scorre nel dispositivo, non è banalmente implementabile come nel caso precedente ma richiede una certa cautela. Come già rimarcato, grazie alla sua ambientazione nel piano delle fasi, lo studio dell'evoluzione temporale della f_w é riportabile all'analisi delle traiettorie di pseudo-particelle, il cui senso di percorrenza é indicato in figura (3.3). Il modo migliore per valutarne la corrente prodotta, si dimostra quello di calcolarla su di una superficie posta *nel mezzo* a due successivi punti di campionamento dell'asse x (che in figura è stata indicata con S): così facendo la divergenza di J sarà funzione della differenza fra flussi di corrente relativi intervalli adiacenti ad ogni punto di campionamento e quindi associata al loro punto comune. Il risultato è che ∇J rimarrà definita sui punti di campionamento stessi. Inoltre come vedremo sotto, tale scelta è dettata dalla necessità di calcolare $J(x)$ in un modo che permetta eliminare gli errori di discretizzazione presenti nel calcolo dell'equazione di continuità. Indicando con

Figura 3.3: Direzione del flusso delle particelle

$J_{i+\frac{1}{2}}$ la corrente che scorre nell'intervallo tra x_i e x_{i+1} , definiamo:

$$J_{i+\frac{1}{2}} = \Delta_p \left(\sum_{j|p_j < 0} \frac{p_j}{m} (3f_{i+1,j} - f_{i+2,j}) + \sum_{j|p_j > 0} \frac{p_j}{m} (3f_{i,j} - f_{i-1,j}) \right) \quad (3.18)$$

Si noti che la formula precedente richiede i valori di due file adiacenti di punti paralleli all'asse p , vedremo che questa è necessaria conseguenza dell'aver adottato lo schema di calcolo SDS (che è uno schema del secondo ordine) e che è indispensabile per la seguente verifica.

Equazione di continuità

$$\nabla J - \frac{\partial n(x)}{\partial t} = 0$$

Analizziamo il momento di ordine zero dell'equazione di Wigner che, nella formulazione continua, non è altro che l'equazione di continuità. Verifichiamo quali siano le condizioni richieste dal modello discreto affinché, in assenza di collisioni, la formula precedente sia anche un'i-

dentità numerica. Integriamo la (3.7) scritta nella forma più generale tempo-dipendente

$$\int \frac{\partial f_w}{\partial t}(x, p, t) dp = \int \mathcal{T} f_w(x, p, t) dp + \int \mathcal{V} f_w(x, p, t) dp$$

dove si è tenuto conto che l'integrale del termine collisionale è zero. Discretizzando e sfruttando la similitudine fra la formula del calcolo dell'operatore traslazione e quella precedentemente definita di $J_{i+\frac{1}{2}}$ si ottiene

$$\frac{\partial n_i}{\partial t} = -\frac{1}{\Delta_x} \left(J_{i+\frac{1}{2}} - J_{i-\frac{1}{2}} - \Delta_p \sum_{j=1}^{N_p} \sum_{j'=1}^{N_p} V_{i;j-j'} f_{i,j'} \right) \quad (3.19)$$

Perché tale equazione si riduca a quella di continuità, il contributo al secondo membro legato all'operatore potenziale, dovrà annullarsi. Allo scopo si consideri la somma sui j ; ricordando l'espressione di V (3.9) si ottiene

$$\sum_{j=1}^{N_p} V_{i;j-j'} = \sum_{j=1}^{N_p} V_{i,j} = \frac{2}{N_p} \sum_{i'=1}^{\frac{N_x}{2}} \sum_{j=1}^{N_p} \sin \left(\frac{2j \Delta_p i' \Delta_x}{\hbar} \right) (U_{i+i'} - U_{i-i'}) \quad (3.20)$$

Questa sommatoria si annulla se

$$\sum_{j=1}^{N_p} \sin \left(\frac{2j \Delta_p i' \Delta_x}{\hbar} \right) (U_{i+i'} - U_{i-i'}) = 0 \quad (3.21)$$

ovvero quando Δ_p soddisfa la seguente relazione, che prende il nome di relazione di completezza di Fourier, già anticipata in sez. 3.2.

$$\Delta_p = \frac{\pi \hbar}{N_p \Delta_x}$$

Grazie a questo vincolo, il modello numerico soddisfa la seguente equazione di continuità in forma discreta

$$\frac{\partial n_i}{\partial t} = -\frac{1}{\Delta_x} (J_{i+\frac{1}{2}} - J_{i-\frac{1}{2}}) \quad (3.22)$$

Equazione di bilancio del momento primo

$$\frac{\partial J}{\partial t} = -\frac{n(x)}{m} \frac{\partial U(x)}{\partial x} - \frac{1}{\hbar} \int_{-\infty}^{+\infty} \frac{dp}{2\pi\hbar} \frac{\partial f_w}{\partial x}$$

Il calcolo del momento primo dell'equazione di Wigner, noto appunto come bilancio dei momenti, mostra i limiti del modello numerico adottato. L'importanza dell'equazione di bilancio del momento primo sta nella descrizione degli stati di equilibrio: essa assicura infatti che la corrente che scorre in un dispositivo non polarizzato è identicamente nulla indipendentemente dal profilo del potenziale dell'eterostruttura (e lo si dimostra con l'unica ipotesi che le funzioni di distribuzione nei due reservoir siano identiche funzioni dell'energia, ovvero in equilibrio termico). Nel nostro modello in cui, a differenza dell'equazione di continuità, quella di bilancio del momento primo non può essere soddisfatta in maniera esatta, si verifica infatti che strutture dal profilo di potenziale asimmetrico producono una corrente non nulla che scorre nel dispositivo anche in assenza di tensione applicata, e la cui ampiezza va a zero come $O(\Delta_x)$. Mains e Haddad [19], tramite calcoli che non riportiamo, hanno dimostrato che l'equazione di bilancio può essere soddisfatta se si “finestra” la trasformata di Fourier (rispetto ad x) del profilo di potenziale, con una maschera rettangolare, ovvero se si sostituisce al potenziale dato, la sua convoluzione con un $\text{sinc}(x)$ opportunamente scelto. Tale operazione ha l'effetto di “addolcirlo” eliminando ogni cambio brusco di tensione all'interno del semiconduttore. Modificare il profilo di potenziale rischia però di generare errori nella valutazione degli effetti quantistici. Inoltre si è verificato che l'impiego di tale tecnica è conveniente solo nella discretizzazione USD mentre non mostra vantaggi apprezzabili nell'implementazione SDS seguita nel presente lavoro, in cui quindi non è stata inserita.

3.4 Soluzione non stazionaria

Nel paragrafo 3.2 abbiamo mostrato come l'equazione continua (3.1), nel caso stazionario, possa essere approssimata col sistema lineare (3.14). Illustreremo qui di seguito il metodo iterativo che ci permetterà di estendere l'impiego di tale sistema alla risoluzione l'equazione (3.1) nella sua forma completa tempo-dipendente. Formalmente è sufficiente aggiungere alla (3.14) il primo membro della (3.1), ed inserire l'eventuale dipendenza dal tempo del potenziale, per ottenere

$$\frac{\partial f_w}{\partial t} = \mathbf{L}f_w - b \quad (3.23)$$

che non è altro che l'equazione differenziale (2.4) discussa nel capitolo precedente. Come la presenza di b non incide sulla stabilità del sistema appena scritto, così anche per suoi algoritmi risolutivi, sarà solo la matrice $\mathbf{L}$ a contenere informazioni necessarie all'analisi della loro stabilità numerica. Ci limiteremo quindi allo studio del più semplice sistema

$$\frac{\partial f_w}{\partial t} = \mathbf{L}f_w \quad . \quad (3.24)$$

Rimarchiamo che nella (3.24) il tempo è ancora una variabile continua: ciò è stato determinante nella valutazione delle sue caratteristiche di convergenza. Una soluzione numerica necessita quindi dell'ulteriore discretizzazione dell'asse dei tempi, ed il modo con cui viene effettuata incide sulla stabilità numerica dell'algoritmo; in generale si implementano tecniche di calcolo iterative (più o meno stabili) del tipo [10]:

$$f(t + \triangle_t) = \mathbf{L}_d f(t) \quad (3.25)$$

E' noto che il sistema iterativo precedente è stabile se e solo se gli autovalori di $\mathbf{L}_d$ hanno modulo minore dell'unità. Poiché abbiamo dimostrato che il sistema continuo è stabile, dobbiamo verificare che la sua discretizzazione si

mantenga tale: in pratica quindi l'algoritmo di calcolo ottimale sarà quello in cui si possa dimostrare il seguente

Teorema 1. Gli autovalori di $\mathbf{L}_d$ hanno modulo minore di uno se e solo se gli autovalori di $\mathbf{L}$ hanno la parte reale negativa

La condizione precedente è anche necessaria poiché non ha senso aspettarsi che un sistema originariamente instabile acquisti stabilità quando venga discretizzato. Per illustrare come in generale non si possa stabilire una connessione così diretta fra sistema continuo e discreto, analizzeremo due tecniche di calcolo note come “formula esplicita” e “formula implicita” di Eulero in cui la proposizione precedente è verificata solo nel secondo [10] caso.

La soluzione della (3.24) si scrive formalmente

$$f(t) = \left[\exp \left(\int_{t_0}^t \mathbf{L}(t') dt' \right) \right] f(t_0) \quad (3.26)$$

Schema esplicito. E' anche l'approccio numerico più semplice e consiste nell'approssimare l'integrale dentro l'esponenziale con una sommatoria finita, per quindi ottenere un'espansione in prodotti dell'operatore esponenziale:

$$\exp \left(\int_{t_0}^t \mathbf{L}(t') dt' \right) \approx \exp \left(\sum_{t'=t_0}^t \mathbf{L}(t') \Delta_t \right) \approx \prod_{t'=t_0}^t \left[1 + \mathbf{L}(t') \Delta_t \right] \quad (3.27)$$

dove $t' \in \{t_0, t_0 + \Delta_t, t_0 + 2\Delta_t, \dots, t\}$.

Si verifica immediatamente che la (3.27) è equivalente al seguente schema iterativo

$$f(t + \Delta_t) = f(t) + \Delta_t \mathbf{L}(t) f(t) = (1 + \Delta_t \mathbf{L}(t)) f(t) \quad (3.28)$$

che è equivalente alla (3.25) con

$$\mathbf{L}_d = 1 + \Delta_t \mathbf{L}(t)$$

Indicando con λ gli autovalori della $\mathbf{L}$ e con λ_d quelli di $\mathbf{L}_d$, vale ovviamente

$$\lambda_d = 1 + \lambda \Delta_t \quad .$$

La validità di tale schema è limitata all'esistenza di autovalori di $\mathbf{L}$ con una grande componente reale negativa. Il modulo di λ_d può essere reso minore dell'unità a costo però di scegliere un Δ_t sufficientemente piccolo come mostrato in fig. 3.4/A. Poiché il valore risultante implica spesso un numero di iterazioni (e quindi tempi di calcolo) eccessivamente elevato, tale metodo di risoluzione è quasi sempre impraticabile.

Schema implicito. E' possibile costruire un processo di integrazione sempre stabile, espandendo l'operatore esponenziale nella seguente maniera:

$$\exp \left(\int_{t_0}^t L(t') dt' \right) \approx \prod_{t'=t_0}^t [1 - L(t') \Delta_t]^{-1} \quad (3.29)$$

In cui si è semplicemente riscritto e^L come $\frac{1}{e^{-L}}$ prima di effettuarne formalmente lo sviluppo in serie. Alla (3.29) è associata la seguente equazione iterativa

$$(1 - \Delta_t \mathbf{L}) f(t + \Delta_t) = f(t) \quad (3.30)$$

che riscritta nella notazione di formula (3.25) diviene

$$\mathbf{L}_d^{-1} f(t + \Delta_t) = f(t) \quad \text{con} \quad \mathbf{L}_d^{-1} = 1 - \Delta_t \mathbf{L}$$

il legame fra gli autovalori vale:

$$\lambda_d^{-1} = 1 - \lambda \Delta_t \quad .$$

Verifichiamo che la stabilità del sistema continuo implica quella del sistema discreto ovvero che $\text{Re}(\lambda) < 0 \Rightarrow |\lambda_d|^{-1} > 1$. Posto $\lambda = x + iy$

$$|\lambda_d|^{-2} = (x \Delta_t - 1)^2 + y^2 \Delta_t^2 \geq (x \Delta_t - 1)^2 > 1 \quad \text{se} \quad x < 0$$

come richiesto dal teorema 1. Dalle precedenti disuguaglianze si ha anche che la $\mathbf{L}_d$ è non singolare. Concludendo anche lo schema implicito può essere scritto

$$f(t + \Delta_t) = \mathbf{L}_d f(t)$$

ma in questo caso lo schema numerico è certamente stabile.

La maggiore stabilità dello schema implicito si paga però in termini di un maggior di costo computazionale richiesto, rispetto allo schema esplicito, per ogni iterazione temporale: lo schema implicito richiede infatti, per ogni step temporale, l'inversione della matrice $\mathbf{L}_d$, operazione che ha la stessa complessità computazionale del calcolo di una soluzione stazionaria il cui costo, come detto, è dell'ordine di $N_x \times N_p^3$ prodotti; per contro un'iterazione dello schema esplicito richiede un solo prodotto matrice per vettore del costo di $N_x \times N_p^2$ prodotti. Nonostante ciò, lo schema implicito rimane il più conveniente tra i due grazie alla libertà offerta di poter scegliere un Δ_t più grande. Com'è ovvio, però, all'aumento di Δ_t i risultati dello schema implicito saranno sempre meno accurati. Analizziamo proprio quest'ultimo aspetto: l'errore commesso ad ogni iterazione deriva dal fatto che si cerca di approssimare $e^{\lambda t}$ con $(1 - \lambda \Delta_t)^{-1}$ che perde di validità quando $|\lambda \Delta_t| > 1$. Il sistema provvede comunque ad una sorta di compensazione di tali errori in quanto, fissato il valore di Δ_t gli errori maggiori affliggeranno la valutazione degli stati associati con grandi autovalori, ovvero negli stati ad elevata energia, che però sono quelli meno popolati.

Schema di Cayley. Presentiamo adesso lo schema di Cayley, che si configura come una “via di mezzo” fra i due precedenti algoritmi, ma che si rivela maggiormente accurato. Possiede le stesse caratteristiche di stabilità e onere di calcolo dello schema implicito ma assicura la maggior precisione nel calcolo della (3.26). Lo schema Cayley viene usato

spesso in meccanica quantistica per seguire l'evoluzione temporale della funzione d'onda di Schrödinger, in cui è necessario risolvere la [10]

$$e^{i\mathbf{H}t/\hbar} \quad (3.31)$$

che ha una certa analogia con la nostra (3.26). La differenza maggiore è che, se $\mathbf{H}$ è hermitiana (come spesso accade), la (3.31) risulta essere un operatore unitario. Questa importante proprietà dell'operatore di evoluzione temporale viene preservata approssimando la (3.31), per un tempo infinitesimo Δ_t , con

$$e^{i\mathbf{H}\Delta_t/\hbar} \approx \frac{1 + i\mathbf{H}\Delta_t/2\hbar}{1 - i\mathbf{H}\Delta_t/2\hbar} \quad (3.32)$$

Nel nostro caso ovviamente la (3.26), a causa della parte reale non nulla degli autovalori di $\mathbf{L}$, non è unitaria e quindi lo sviluppo precedente non sembrerebbe necessario. Nonostante ciò risulta comunque utile in quanto possiede un'accuratezza nello sviluppo dell'esponenziale rispetto a Δ_t del second'ordine, contro quella del primo ottenibile con gli algoritmi di Eulero. Possiamo verificarlo immediatamente. Riprendiamo innanzitutto la (3.26) e, come sempre, approssimiamo l'integrale con una produttoria

$$\exp\left(\int_{t_0}^t \mathbf{L}(t')dt'\right) \approx \exp\left(\sum_{t'=t_0}^t \mathbf{L}(t')\Delta_t\right) = \prod_{t'=t_0}^t \exp\left[\mathbf{L}(t')\Delta_t\right] \quad (3.33)$$

Posto

$$\exp(\mathbf{L}(t')\Delta_t) \approx \frac{1 + \mathbf{L}(t')\Delta_t/2}{1 - \mathbf{L}(t')\Delta_t/2} \quad (3.34)$$

sviluppando il denominatore rispetto a $\mathbf{L}\Delta_t$ si ottiene

$$\frac{1 + \mathbf{L}\Delta_t/2}{1 - \mathbf{L}\Delta_t/2} \approx 1 + \frac{\mathbf{L}\Delta_t}{2} + \frac{\mathbf{L}^2\Delta_t^2}{4}$$

che è lo sviluppo al second'ordine dell'esponenziale (3.34). L'algoritmo di Cayley infine si scrive

$$(1 - \Delta_t \mathbf{L}) f(t + \Delta_t) = (1 + \Delta_t \mathbf{L}) f(t) \quad (3.35)$$

che, come si nota, ingloba sia la (3.28) che la (3.30). Per quanto riguarda la stabilità ci riportiamo come al solito alla (3.25), in cui adesso gli autovalori della $\mathbf{L}_d$ risultano ovviamente

$$\lambda_d = \frac{1 + \lambda \Delta_t}{1 - \lambda \Delta_t} \quad .$$

E' immediato verificare che

$$|\lambda_d| < 1 \quad \Longleftrightarrow \quad \text{Re}(\lambda) < 0$$

Detto $\lambda = x + iy$

$$|\lambda_d| = \frac{(x+1)^2 + y^2}{(x-1)^2 + y^2}$$

la tesi segue da $(x+1)^2 < (x-1)^2$ se $x < 0$. Quando dimostato è illustrato in fig.3.4/B in cui risulta chiaramente che l'autovalore λ_d ha modulo minore di uno se e solo se la parte reale dei λ (gli autovalori della matrice $\mathbf{L}$) è minore di zero. Anche in questo caso quindi la stabilità è garantita.

3.5 Equazione di Poisson

Nel caso unidimensionale, e nell'ipotesi di uniformità della permeabilità dielettrica ϵ lungo tutto il semiconduttore, l'equazione di Poisson (PE) si scrive

$$\epsilon \frac{d^2 u(x)}{dx^2} = q\rho(x) = q^2[C(x) - n(x)] \quad (3.36)$$

dove $u(x)$ è il potenziale, q la carica dell'elettrone, $n(x)$ la densità di carica elettronica libera, e $C(x)$ la densità di ioni fissi positivi del drogante.

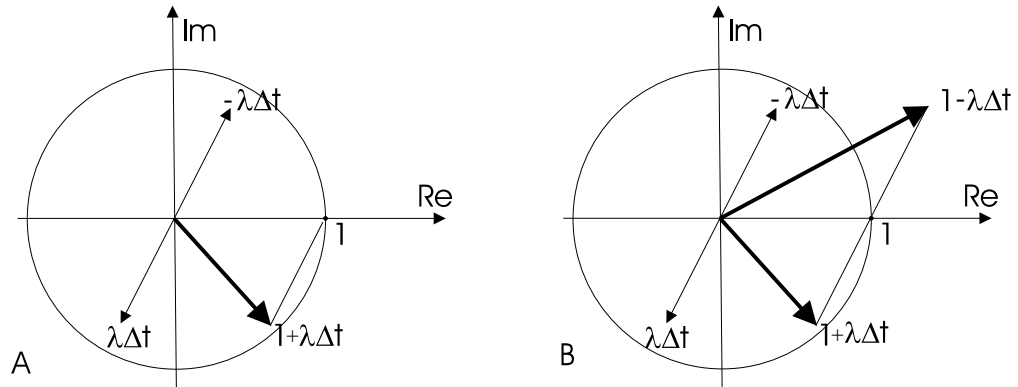


Figura 3.4: A) Stabilità nello schema esplicito; B) Stabilità nello schema Cayley

La complessità computazionale aggiunta dall'accoppiamento dell'equazione di Poisson con quella di Wigner, ovvero del calcolo autoconsistente rispetto al caso precedente, è che il sistema, diventando non lineare, necessita di procedimenti iterativi per il calcolo della sua soluzione. Oltre ad un aumento dei tempi di calcolo, nascono problemi di instabilità numerica, cui si deve ovviare sfruttando particolari algoritmi che verranno qui di seguito descritti. Nella presente trattazione seguiremo il metodo iterativo noto come metodo di Newton: esso verrà applicato sia nel calcolo della soluzione stazionaria del sistema W-P che, in una sua versione semplificata (nota come metodo Gummel), nel calcolo non stazionario del problema evolutivo. Per soluzione “autoconsistente” del sistema Wigner-Poisson, intendiamo la ricerca di un profilo di potenziale $U(x)$ che sia coerente col profilo della densità di carica ρ del dispositivo, nel senso che l'equazione di Poisson (il cui dato di ingresso è la ρ e la cui uscita è la $U(x)$) e l'equazione di Wigner (WF) (in cui i dati di ingresso e uscita sono invertiti rispetto all'equazione di Poisson) siano simultaneamente risolte: la PE sfrutta la densità di carica per il calcolo del profilo del diagramma a bande del circuito, mentre la WF usa tale risultato per de-

terminare (fra l'altro) la distribuzione di carica ρ . Il minimo della banda di conduzione ($U(x)$) sarà calcolato sommando al potenziale autoconsistente, l'offset $\delta U(x)$ del profilo della banda di conduzione dell'eterostruttura.

$$U(x) = u(x) + \delta U(x) \quad . \quad (3.37)$$

Poichè l'unico dato immesso dall'esterno sarà la differenza di potenziale fra le due facce del dispositivo, le condizioni al contorno per l'equazione PE si scrivono semplicemente

$$u(0) = 0 \quad (3.38)$$

$$u(L) = [\delta U(0) - \delta U(L)] - qV_{BIAS} \quad (3.39)$$

in cui il potenziale è misurato rispetto (vedi fig (3.5)) al livello di Fermi in $x = 0$ cosicchè $\delta U(0) = -\mu_0$; nel caso pratico di struttura simmetrica, il primo termine a secondo membro della (3.39) è ovviamente zero. Poichè la PE non

Figura 3.5: Diagramma a bande dell'RTD analizzato nel cap.4

pone grossi problemi di discretizzazione, ricorreremo alla sua approssimazione secondo il classico schema delle differenze finite [15]

$$u_{i+1} - 2u_i + u_{i-1} = \frac{q^2 \Delta_x^2}{\epsilon} [C_i - n_i] \quad (3.40)$$

Allo scopo di costruire un procedimento iterativo che porti alla soluzione del sistema W-P, indicheremo con l'indice n l'ennesima distribuzione ($n^n(x)$ o $u^n(x)$) soluzione dell'omonimo passo iterativo dei seguenti algoritmi di calcolo: metodo di Gummel e metodo di Newton [19]

Metodo Gummel:

Questo è il metodo concettualmente più semplice per risolvere il sistema Wigner-Poisson in quanto, data $n^n(x)$, si provvede ad integrarla tramite la (3.36) per trovare il nuovo profilo di potenziale u^{n+1} da inserire nella WE; questa a sua volta fornirà la nuova n^{n+1} . Nell'algoritmo di Gummel si ripetono iterativamente i passi precedenti finchè il metodo (sperabilmente) non converga: poiché non si prevede nessun meccanismo di controllo, si verifica in pratica che nel calcolo di soluzioni stazionarie, un tale algoritmo spesso non porta a nessuna soluzione, ovvero non esiste una funzione cui tende la f_w , che oscilla indefinitamente.

Un esempio pratico può spiegarne il motivo: si consideri l'invasione balistica degli elettroni in un RTD uniformemente drogato. Il potenziale "visto" dagli e^- al passo zero, sarà semplicemente l'offset dell'eterostruttura; gli elettroni in questa prima iterazione si muoveranno quindi senza risentire del potenziale di mutua repulsione, e si distribuiranno come indicato in figura (3.6/A). Se adesso (siamo ancora all'iterazione zero) integriamo l'equazione di Poisson rispetto alla $n^0(x)$ appena trovata, determineremo un potenziale netto repulsivo (che si sovrapporrà all'offset del diagramma a bande), sicuramente sovrastimato rispetto a quello realmente presente all'interno del dispositivo. La causa è la descrizione non fisica che si è precedentemente data agli elettroni, avendoli trattati in pratica come fossero portatori neutri, e che conseguentemente, ha portato ad un "sovraffollamento" di elettroni all'interno del semiconduttore il cui effetto è stato la nascita di un forte potenziale

repulsivo. Nel passo successivo (iterazione uno), la barriera di potenziale formatasi, tenderà a respingere gli elettroni al di fuori del dispositivo (vedi fig.(3.6/B)). Il risultato netto sarà che alla fine del passo uno il potenziale presente nel semiconduttore sarà stimato essere una buca di potenziale (fig.(3.6/C)). Conseguentemente al passo due si registrerà un'elevata concentrazione di portatori, il cui integrale è un potenziale di tipo repulsivo

Figura 3.6: Densità di carica e diagramma a bande risultanti dai primi quattro step dell'algoritmo di Gummel applicato ad un RTD

che ha intensità maggiore di quello calcolato nel passo zero (fig.(3.6/D)). In questo come in svariati altri casi, succede quindi che il sistema non converge a nessuna soluzione mostrando invece un comportamento oscillante che è puramente frutto dell'incapacità dell'algoritmo numerico di trovare la soluzione richiesta. Ricordiamo infatti, che quelli calcolati non sono passi temporali, ovvero non stiamo seguendo l'evoluzione del sistema nel tempo: quindi quelle ottenute sono oscillazioni puramente numeriche senza nessun significato fisico; non devono essere confuse con le oscillazioni intrinseche del diodo che verranno analizzate successivamente.

Analizziamo adesso il più stabile algoritmo di Newton

Metodo Newton:

Per poter sviluppare il metodo numerico di Newton, al posto dell'equazione di Poisson, conviene definire il seguente “funzionale di Poisson” [19]

$$\mathcal{P}\{u, n\}(x) \equiv \epsilon \frac{d^2 u(x)}{dx^2} - q^2 [C(x) - n(x)] \quad (3.41)$$

relativo ad un dato profilo di potenziale e ad un generico profilo di carica. Si noti che il funzionale di Poisson diventa identicamente nullo solo quando i suoi ingressi sono proprio il potenziale autoconsistente, e la carica che lo genera: in questo senso va visto come una sorta di funzionale di errore per la determinazione della coerenza fisica fra distribuzione di carica e potenziale. Per chiarezza espositiva consideriamo il primo step dell'algoritmo di Newton

Passo 0: Definizione del potenziale iniziale u^0 , risoluzione della WE rispetto ad u^0 e calcolo della densità $n^0(x)$ risultante.

Poiché il calcolo del funzionale di Poisson con gli ingressi u^0, n^0 , darà per risultato una funzione errore ovviamente non identicamente nulla, decidiamo di variare il potenziale u^0 con $u^1 = u^0 + \delta u^1$; nel criterio di scelta di u^1 risiede la caratteristica peculiare del metodo di Newton. Il vincolo sulla u^1 ,

non sarà quello di annullare il funzionale di Poisson scritto per n^0 e u^1 , ma di annullare la funzione errore che scriveremo fra u^1 e la distribuzione incognita n^1 , che deve essere ancora calcolata. In altre parole richiediamo che si annulli $\mathcal{P}(u^1, n^1)$ e non $\mathcal{P}(u^1, n^0)$. Quindi, poiché la n^1 dipenderà a sua volta da u^1 (ricordiamo che il loro legame è proprio la WE) dovremo cercare di valutare quale sarà l'entità del cambiamento di n^0 a n^1 ad opera del cambiamento di u da u^0 a u^1 . Descritto a parole ciò significa: determinare la variazione del numero di portatori liberi in un dato punto (da n^0 a n^1), ad opera di un'arbitraria variazione del potenziale nello stesso punto. Poiché l'entità di tale variazione sta scritta nella WE, che sappiamo risolvere solo numericamente, nessuna formula analitica può venirci in soccorso nel caso generale, e dovremo ricorrere a metodi iterativi, frutto comunque di ipotesi che semplificano il legame fra potenziale e carica all'interno del semiconduttore.

Vediamo più precisamente come si riesce a calcolare la δu^n ad ogni iterazione, che soddisfi la richiesta $\mathcal{P}(u^n, n^n) \equiv 0$: allo scopo calcoliamo la variazione del funzionale $\mathcal{P}$, in risposta di una variazione del potenziale in ingresso $u(x)$. Per esplicitarlo si noti che, fissato $x = \bar{x}$, $u(\bar{x})$ (e $n(\bar{x})$) diventa un numero e come tale può essere considerato una variabile qualsiasi. Quindi, per ogni x , ha senso collegare la variazione del funzionale di Poisson, con la derivata di $\mathcal{P}$ rispetto ad u , (essendo sparita la dipendenza dalla x). Qui di seguito indicheremo con P quando il funzionale $\mathcal{P}$ è pensato funzione della variabile u e non della funzione $u(x)$. Cerchiamo qual'è la variazione al primo ordine di $P(u)$ se si sostituisce ad u , $u + \delta u$, ovvero quello che avviene ad ogni passo iterativo. Formalmente, ricordando che

$$u^{n+1} = u^n + \delta u^{n+1} \quad (3.42)$$

e sfruttando come si è detto, il vincolo

$$\mathcal{P}(u^{n+1}, n^{n+1}) = \mathcal{P}^{n+1} = 0 \quad (3.43)$$

si ha

$$0 = \mathcal{P}(u^{n+1}, n^{n+1}) = \mathcal{P}(u^n, n^n) + \frac{\partial \mathcal{P}^n(u^n)}{\partial u^n} [\delta u^{(n+1)}] \quad (3.44)$$

Come vedremo la (3.44) acquista senso completo solo quando si sostituirà la versione discreta del funzionale di Poisson, al posto della sua formulazione continua. La versione discreta della (3.42) si scrive indicando come sempre tramite l'indice basso, l'i-esimo punto di campionamento (quindi $u(x)$ diventa u_i con $x = i\Delta_x$)

$$u_i^{n+1} = u_i^n + \delta u_i^{n+1} \quad (3.45)$$

Per quanto riguarda la (3.44), visto che stiamo cercando di calcolare l'incremento di $\mathcal{P}$ nel punto x_0 ad opera della variazione di $u(x)$ nel generico x , dobbiamo trattare la $\mathcal{P}$ come funzione delle n variabili indipendenti che sono i campioni di u (ovvero $P = P(u_0, u_1, \dots, u_N)$). La derivata totale della P calcolata in x_0 , in accordo con la teoria delle funzioni di più variabili, è pari alla somma delle derivate parziali della P rispetto alla generica variabile $\left(\frac{\partial P}{\partial u_i} \Big|_{x_0}\right)$, per l'incremento di quest'ultima (δu_i). La (3.44) si scrive:

$$\sum_{i'} \frac{\partial P_i^n}{\partial u_{i'}^n} [\delta u_{i'}^{(n+1)}] = -P_i^n \quad (3.46)$$

in cui

$$P_i^n = u_{i+1} - 2u_i + u_{i-1} - \frac{q^2 \Delta_x^2}{\epsilon} [C_i - n_i] \quad (3.47)$$

L'indice alto della P , come già spiegato, è relativo all'iterazione considerata ovvero $P^n = P(u^n, n^n)$. Infine

$$\frac{\partial P_i^n}{\partial u_{i'}^n} = \delta_{i+1, i'} - 2\delta_{i, i'} + \delta_{i-1, i'} + \frac{q^2 \Delta_x^2}{\epsilon} \left[\frac{\partial n_i^n}{\partial u_{i'}^n} \right] \quad (3.48)$$

dove $\delta_{i,i'}$ è la funzione delta di Kronecker. Dalla (3.46) e dalla (3.47) si può notare che pur risolvendo un'equazione diversa dalla semplice equazione di Poisson, quando δu tende a zero per ogni x , la (3.47) tende alla funzione di Poisson discreta (3.40): ovvero se il sistema converge ad una soluzione, questa è la soluzione voluta, in quanto il funzionale di Poisson (che è la nostra funzione errore) tende anch'esso a zero. Quindi se fra due iterazioni dell'algoritmo, la $u(x)$ rimane invariata, tale $u(x)$ è proprio la soluzione dell'intero sistema Wigner-Poisson come desiderato.

Il valore aggiunto della trattazione di Newton rispetto a quella di Gummel, sta nella presenza della derivata al secondo membro della (3.48). Infatti è di facile verifica come, se quest'ultima viene posta a zero, l'algoritmo di Newton, si riduca all'algoritmo Gummel.

Riassumendo: l'algoritmo Gummel sfruttando la risoluzione diretta dell'equazione di Poisson (3.36), porta al calcolo dell'esatto profilo di potenziale $u(x)$, per la data distribuzione di portatori liberi $n(x)$ valutata ad ogni step risolutivo. L'approccio di Newton, per contro, sfrutta il termine $\frac{\partial n}{\partial u}$ per tener conto dei cambiamenti futuri della $n(x)$, risultanti dalla soluzione parziale $u(x)$, cercando di predire quale sarà l'esatta distribuzione di potenziale finale. Il termine $\frac{\partial n}{\partial u}$, che provvede a cambiare l'ingresso del sistema in funzione dell'uscita futura, può quindi essere visto come un termine di feedback; come tale è in grado di rendere più robusto il metodo numerico adottato, in maniera analoga a quanto avviene nel controllo dei sistemi retroazionati. Come anticipato, la stima quantitativa della variazione della densità di carica $n(x)$ causata da una variazione del potenziale $u(x)$, sta scritta nell'equazione di Wigner, da cui però non deriva nessuna espressione analitica in forma chiusa che ci sia di aiuto. La valutazione di $\frac{\partial n}{\partial u}$ può esser ricavata solo se si approssima la distribuzione reale dei portatori all'interno del dispositivo, con quella

di equilibrio termodinamico di Fermi. Poichè in tal caso è noto che la densità dipende dal potenziale tramite l'integrale di Fermi $\frac{1}{2}$

$$\int_U^\infty \sqrt{E - U} f_F(E) dE = n(x)$$

dove U è il minimo della banda di conduzione di formula (3.37) e f_F la distribuzione di Fermi, si sfruttano sviluppi in serie noti come approssimazioni di Joyce-Dixon per ottenere il risultato cercato [16]. La procedura di calcolo, semplice ma lunga, viene in questa sede omessa, ma ne riportiamo i risultati

$$\frac{\partial n(u)}{\partial u}(x) = \frac{N_c}{k_B T} \left[\frac{N_c}{n(u(x))} + \sum_{m=1} \left(\frac{a_m}{m} \frac{N_C}{n(u(x))} \right)^{m-1} \right]^{-1} \quad (3.49)$$

in cui N_C è la densità di drogante sui contatti, e i coefficienti a_m sono riportati in tabella 3.1. Nel metodo Newton si dimostra critica la scelta del numero di

Coeficiente	Valore
a_1	1
a_2	3.53553 10^{-1}
a_3	5.75499 10^{-2}
a_4	5.76396 10^{-3}
a_5	4.01950 10^{-4}

Tabella 3.1: Coefficienti dello sviluppo Joyce-Dixon

termini in cui sviluppare la sommatoria; in quanto spingersi a considerarne un numero troppo elevato accelera la convergenza ma si rivela in qualche caso, a sua volta instabile; sceglierne un numero piccolo viceversa richiede lunghi tempi i convergenza: nel presente lavoro si abbiamo adottato uno sviluppo fino al quarto ordine che si è mostrato un buon compromesso.

3.6 Traiettorie

Come descritto in sez.1.10, grazie alla sua forma di tipo cinetico, anche per la funzione di Wigner possiamo parlare di traiettorie, o meglio di “quasi traiettorie”. Verificheremo che la loro analisi aiuta nella comprensione intuitiva dei fenomeni fisici alla base del funzionamento dei dispositivi ad effetto tunnel. Il calcolo delle traiettorie è un’operazione non banale e si presta a vari approcci: Jacoboni ad esempio [7] fonda lo studio e la risoluzione dell’equazione di Wigner proprio sul calcolo delle quasi-traiettorie tramite sviluppi perturbativi; nella presente trattazione in cui un metodo risolutivo dell’equazione di Wigner è già stato sviluppato, sfrutteremo le tecniche di calcolo proposte da Buot e Jensen [14]: un primo metodo grossolano per il calcolo delle traiettorie, lo si ricava semplicemente valutando la variazione della $f_w(x, p)$, causata da un incremento infinitesimo $\delta x, \delta p$:

$$f(x + \delta x, p + \delta p) \approx f(x, p) + \frac{\partial f(x, p)}{\partial x} \delta x + \frac{\partial f(x, p)}{\partial p} \delta p$$

Dividendo entrambi i membri per δt , nell’ipotesi di distribuzioni stazionarie si ottiene

$$0 \approx \frac{\partial f(x, p)}{\partial x} \frac{\delta x}{\delta t} + \frac{\partial f(x, p)}{\partial p} \frac{\delta p}{\delta t} \quad (3.50)$$

che non è altro che l’approssimazione del differenziale totale della f_w , calcolato lungo la spezzata di fig.3.7: le curve che risolvono la (3.50) sono quindi le curve di livello della f_w . Un approccio più accurato richiede invece la risoluzione delle equazioni di moto di Newton (1.47). Il nostro programma per la risoluzione di tale sistema, esegue una stima iniziale della posizione x' e della quantità di moto p' al tempo $t + \delta t$ usando

$$p' = p(t) + \dot{p} \delta t$$

$$x' = x(t) + [p' + p(t)] \frac{\delta t}{2}$$

in cui $\dot{p}$ è dato dalla (1.46), e correggendola successivamente, spostando il punto trovato lungo le direzioni a massima pendenza. Riportiamo per comodità l'espressione della forza quantistica

$$\dot{p} = \frac{\frac{1}{2\pi\hbar^2} \int_{-\infty}^{\infty} dp' V(x, p - p') f(x, p') - \left. \frac{\partial f}{\partial t} \right|_{coll}}{\partial f / \partial p} \quad (3.51)$$

la cui discretizzazione segue immediatamente dalle (3.8) e (3.13). Vediamo adesso in dettaglio la procedura di calcolo: innanzitutto premettiamo che dobbiamo conoscere la $f(x, p)$ e le sue derivate in ogni punto del piano e non solo su quelli della griglia di campionamento. Allo scopo, il metodo che abbiamo adottato, sfrutta polinomi interpolanti e si dimostra quindi adatto solo per funzioni lentamente variabili, in caso contrario i polinomi oscillano drammaticamente. Unico palliativo è l'infittimento della griglia quando la f_w diventa oscillante, come accade nei pressi della risonanza. Nei punti in cui comunque il metodo fallisce, possiamo coadiuvarlo con l'inseguimento delle curve di livello della f_w precedentemente descritto. Sappiamo che nel caso unidimensionale, esiste un unico polinomio di grado n che passi per n punti dati: tale risultato non è estendibile a più dimensioni. Si è scelto il seguente algoritmo per l'approssimazione della $f_w(x, p)$ e delle sue derivate: denominiamo con x_i e p_j i punti più vicini a x e p e tali che $z = (x - x_i)/\Delta_x$ e $s = (p - p_i)/\Delta_p$ siano positivi e minori dell'unità. Scegliendo la griglia 4×4 di fig.(3.7) approssimiamo la f_w con due successive discretizzazioni: prima per l'asse x

$$f(x, p(j)) = f(x, j) \approx [-z(z-1)(z-2)f(i-1, j) + 3(z^2-1)(z-2)f(i, j) - 3z(z+1)(z-2)f(i+1, j) + z(z^2-1)f(i+2, j)]/6$$

e successivamente rispetto all'asse p

$$f(x, p) \approx [-s(s-1)(s-2)f(x, j-1) + 3(s^2-1)(s-2)f(x, j) +$$

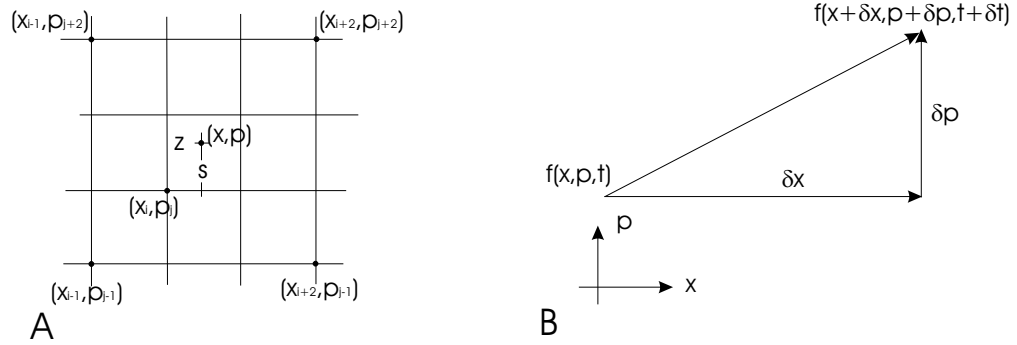


Figura 3.7: A) Differenziale esatto di f_w ; B) Griglia di discretizzazione per il calcolo di derivate e interpolazione di f_w

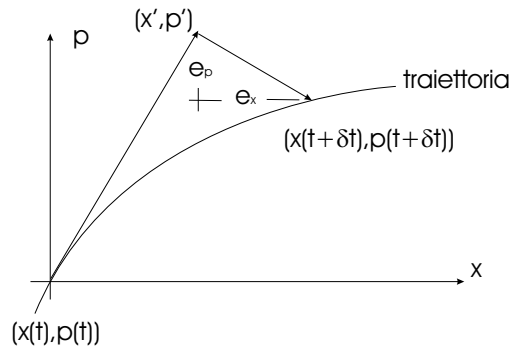


Figura 3.8: Correzione della stima (x', p')

$$-3s(s+1)(s-2)f(x, j+1) + s(s^2-1)f(x, j+2)]/6$$

in cui con $f(x, j)$ abbiamo indicato $f(x, p(j))$ e con $f(i, j)$ i valori della f_w in corrispondenza dei punti di campionamento. Facendone poi le derivate rispetto a $x = x_i + z\Delta_x$ e $p = p_i + s\Delta_p$, ricaviamo le approssimazioni di $\frac{\partial f}{\partial x}$ e $\frac{\partial f}{\partial p}$. La prima è sfruttata nella valutazione della forza quantistica (vedi form.(3.51)) ed entrambe per il calcolo delle correzioni a (x', p') . Se indichiamo con $(x(t+\delta t), p(t+\delta t))$ le coordinate esatte del punto all'istante $t+\delta t$,

una stima degli errori commessi e_x e e_p giunge scrivendo,

$$f(x(t + \delta t), p(t + \delta t)) = f(x', p') + e_x \frac{\partial f(x', p')}{\partial x} + e_p \frac{\partial f(x', p')}{\partial p}$$

e ricordando che vogliamo applicare il teorema di Liouville anche alla funzione di Wigner, ovvero che deve valere $f(x(t + \delta t), p(t + \delta t)) = f(x(t), p(t))$. Le correzioni e_x e e_p sono univocamente determinate quando si esegua la correzione della stima (x', p') lungo le linee a massima pendenza. Quest'ultima è perpendicolare alla direzione precedentemente calcolata, che era quella tangente alle traiettorie come illustrato in fig.(3.8): evitiamo così che le soluzioni “scappino” lungo la tangente.

3.6.1 Valutazione del tempo di tunnel

La ricerca del tempo di percorrenza delle particelle fra due punti di una traiettoria è immediata, in quanto richiede semplicemente l'integrazione dell'inverso della velocità, calcolata fra i punti di partenza e di arrivo

$$t(x) = \int_{x_0}^x \frac{m^*}{p} dx \quad .$$

Approssimando l'integrale con la regola dei trapezi si ottiene

$$t(x) = \sum_1^{n_x-1} m^* (p_{i+1}^{-1} + p_i^{-1}) (x_{i+1} - x_i)$$

La formula precedente sarà impiegata per il calcolo del tempo di attraversamento delle barriere risonanti di un RTD (tempo di tunnel).

Capitolo 4

Risoluzione numerica del sistema Wigner-Poisson

4.1 RTD: tecnologie realizzative e principio di funzionamento

Nell'ultimo ventennio si sono spese sempre maggiori risorse (e con sempre maggiori successi grazie allo sviluppo tecnologico dei processi realizzativi e di controllo nell'elettronica submicrometrica) nello studio e produzione di dispositivi strettamente quantistici e che si prefigurano come il futuro della tecnologia integrata. Ultimamente hanno trovato sempre un più largo impiego nelle applicazioni in dispositivi alle microonde e nella progettazione di circuiti oscillanti [9].

L'idea di poter sfruttare due barriere risonanti in dispositivi singola banda per creare dispositivi a resistenza differenziale negativa, è relativamente vecchia. La possibilità però di concretizzarla per mezzo della “band gap-engineering”, applicata a strutture a eterogiunzione, si è realizzata solo grazie allo sviluppo

tecnologico dell'ultimo ventennio ed i primi esempi realizzativi si devono a Esaky. Da allora i materiali utilizzati sono stati prevalentemente $AlGaAs$, In , e solo negli ultimi anni si è riusciti ad ottenere i primi risultati in tal senso, nella più standard (tradizionale) e quindi (soprattutto) economica tecnologia del Si , con l'impiego delle leghe $Si - Ge$. Il funzionamento dei dispositivi risonanti ad effetto tunnel (RTD) è dovuto alla formazione, fra due barriere di potenziale, di una scatola quantistica (che chiameremo pozzo quantistico), all'interno della quale si localizzano pochi livelli energetici permessi, ed il cui autovalore è funzione della differenza di potenziale applicata alle estremità del semiconduttore stesso. Come illustrato in fig.4.1/A,

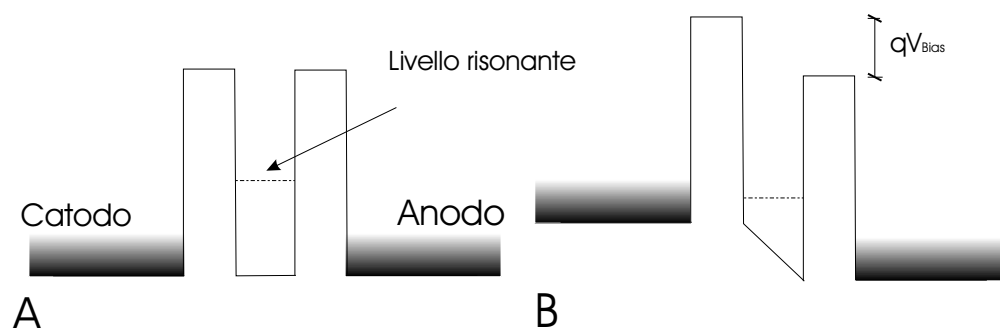


Figura 4.1: Principio di funzionamento di un RTD

in assenza di polarizzazione, la probabilità che un elettrone migri dall'anodo al catodo è praticamente nulla, in quanto le barriere sono sia troppo alte per permettere passaggio di corrente termoionica, sia troppo spesse per essere attraversate per effetto tunnel (la lunghezza di tunnel che è tipicamente dell'ordine di qualche lunghezza d'onda di de Broglie ovvero circa $60\text{\AA}$; per contro la struttura attiva si estende per $120\text{\AA}$). Quando si applica all'anodo una tensione positiva, il diagramma a bande del dispositivo si deforma e tale contatto è spinto ad energie inferiori (fig.4.1/B). Il livello energetico del

pozzo quantistico inizia a diventare degenere con gli stati occupati del lato catodico ed una corrente inizia a scorrere lungo la struttura. Quello che si è formato è infatti un livello risonante che fa da ponte per il passaggio degli elettroni, riducendo la distanza di tunnel allo spessore di una singola barriera. Se adesso la d.d.p. viene ulteriormente aumentata, dopo aver raggiunto il massimo accoppiamento, la corrente di tunnel tenderà a diminuire, annullandosi quando il livello discreto scende al di sotto del minimo della banda di conduzione del catodo. L'unica corrente che adesso ha la possibilità di scorrere nel dispositivo, è quella di emissione termoionica, che diventa perciò dominante nella parte alta della caratteristica $I(V)$ del diodo. Misure condotte su diodi praticamente identici a quello che verrà descritto nel dettaglio in sez.4.2 (differiscono solo per la densità di drogante e per lo spessore degli spacer) portano al diagramma tensione-corrente di fig.4.2, in cui è evidente

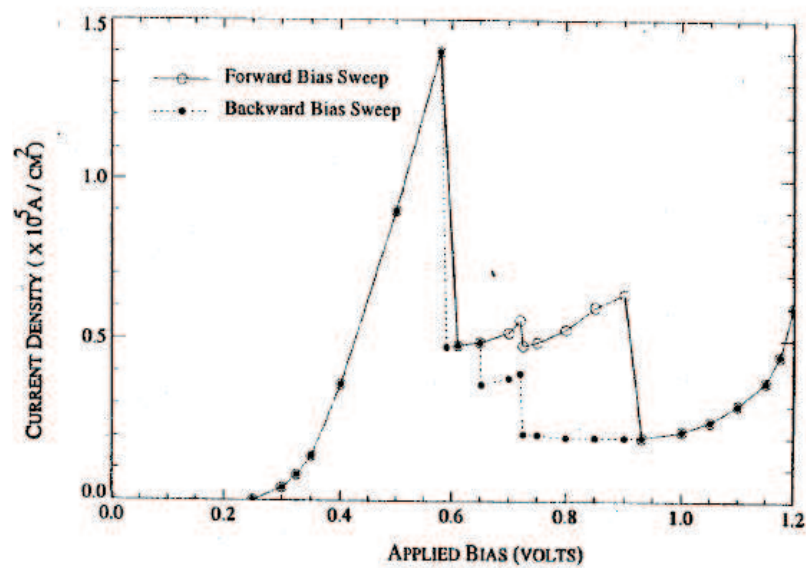


Figura 4.2: Corrente misurata di un RTD alla temperatura di 77 K [17]

il comportamento risonante dell'RTD. Vedremo nel corso dell'esposizione del presente lavoro, come i risultati delle simulazioni siano in ottimo accordo con tali misure. Come precedentemente accennato, la tecnologia che ha permesso la realizzazione dei dispositivi moderni e che si pone come frontiera d'indagine per i futuri sviluppi nel campo dell'elettronica micrometrica, è la tecnologia delle eterogiunzioni: ricapitoliamo brevemente la natura delle eterogiunzioni e i progressi fatti negli ultimi anni dall'elettronica dello stato solido per la loro applicazione a nuovi materiali.

4.1.1 L'eterogiunzione

Quando due differenti semiconduttori vengono posti in intimo contatto si ottiene la cosiddetta eterogiunzione (in contrapposizione con l'omogiunzione in cui viene usato lo stesso semiconduttore). E' possibile realizzare sia giunzioni di tipo p-n che di tipo n-n o p-p: se la giunzione è di tipo p-n viene chiamata eterogiunzione anisotipa ed in questo caso la corrente di cariche è dovuta all'iniezione di portatori minoritari, mentre se è di tipo n-n o p-p è chiamata eterogiunzione isotipa ed in questo caso è un dispositivo a portatori maggioritari (ovviamente se usate come giunzioni raddrizzanti). Le moderne tecniche di crescita del cristallo quali MOCVD (Metal-Organic-Chemical-Vapour-Deposition) o MBE (Molecular-Beam-Epitaxial) hanno avuto un ruolo fondamentale nello sviluppo di tecnologie per realizzare eterogiunzione di elevatissime qualità, cosicché attualmente è possibile disporre di eterogiunzioni estremamente raffinate dal punto di vista delle proprietà elettriche e cristallografiche. Il primo modello di eterogiunzione ideale venne proposto da R.L.Anderson il quale fece riferimento all'eterogiunzione $Ge - GaAs$: in questo caso il "mismatch" esistente fra il passo reticolare dei due semiconduttori è inferiore all' 1 %, cosicché è possibile parlare di

struttura “lattice-matched” e la giunzione presenta caratteristiche di elevata qualità. In fig.4.3/a è riportato il diagramma a bande dei due semiconduttori isolati (prima della formazione della giunzione): il semiconduttore 1 (*Ge*) ha energia di gap E_{g1} minore ed è supposto drogato di tipo p, mentre il semiconduttore 2 ha energia di gap E_{g2} maggiore ed è supposto drogato di tipo n. Quando vengono posti in intimo contatto si verifica il flusso di elettroni-lacune attraverso la giunzione fino a che non è raggiunto l’equilibrio termico: si forma una regione svuotata di donatori ionizzati nel lato n e di accettori ionizzati nel lato p, allo stesso modo di quanto avviene in una omogiunzione p-n. Per disegnare il diagramma a bande nella situazione di equilibrio dobbiamo in primo luogo disegnare un comune livello di Fermi E_f e collocare rispetto questo, nelle regioni p e n neutre, le posizioni delle rispettive bande di valenza, conduzione e livello del vuoto. Nella regione svuotata si varia gradualmente il livello del vuoto dal semiconduttore-1 al semiconduttore-2 ed infine si traccia l’andamento dei margini di banda E_{c1} , E_{c2} , E_{v1} ed E_{v2} disegnandoli paralleli al livello del vuoto; si ottiene così l’andamento di fig.4.3/b. Di seguito sono riportate le grandezze che caratterizzano l’eterogiunzione. Il potenziale di built-in è dato da:

$$V_i = (V_{i1} + V_{i2}) = \Phi_1 - \Phi_2$$

Dal momento che E_{c1} ed E_{c2} sono paralleli al livello del vuoto nei rispettivi semiconduttori si vede che si forma una brusca discontinuità ΔE_c in banda di conduzione data da:

$$\Delta E_c = q(\chi_1 - \chi_2)$$

In modo simile abbiamo una discontinuità ΔE_v anche nella banda di valenza data dalla differenza delle affinità delle lacune:

$$\Delta E_v = (E_{g2} - E_{g1}) - q(\chi_1 - \chi_2)$$

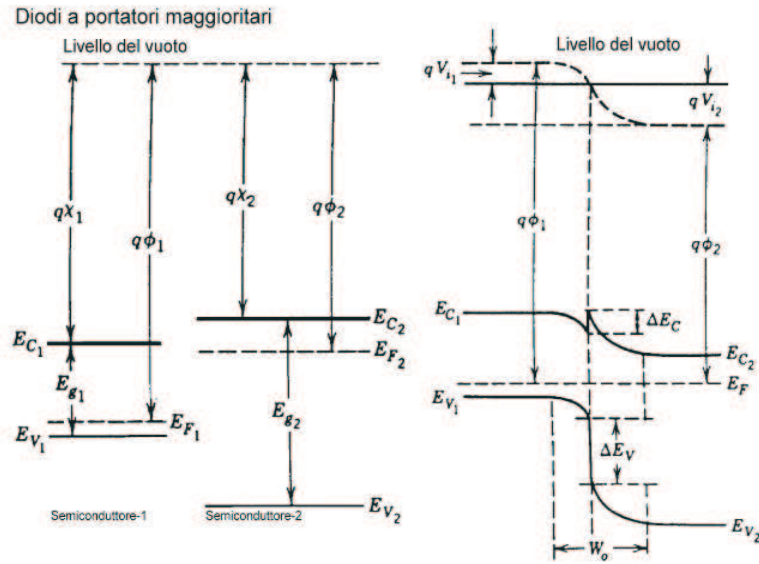


Figura 4.3: Diagrammi a bande di energia per (a) due semiconduttori neutri isolati e per (b) eterogiunzione all'equilibrio termico.

vale inoltre la seguente relazione

$$\Delta E_C + \Delta E_V = (E_{g2} - E_{g1})$$

Le precedenti considerazioni si riferiscono ad una eterogiunzione di tipo ideale; in un'eterogiunzione di tipo reale, a causa di diffusioni di materiale in direzioni indesiderate durante il processo di crescita, interazioni chimiche di varia natura, danni in corrispondenza della superficie di separazione, possono nascere fenomeni quali strati di materiale all'interfaccia che rendono la giunzione graduale e non più brusca, stati occupabili in banda proibita collocati in prossimità dell'interfaccia che possono contenere una carica netta positiva o negativa, strati all'interfaccia contenenti dipoli superficiali che concorrono a modificare la caduta di potenziale sulla giunzione. Le modifiche del diagramma a bande rispetto al caso ideale sono riportate in fig.4.4. Le applicazioni elettroniche necessitano di eterogiunzioni della più elevata qualità; a tal fine,

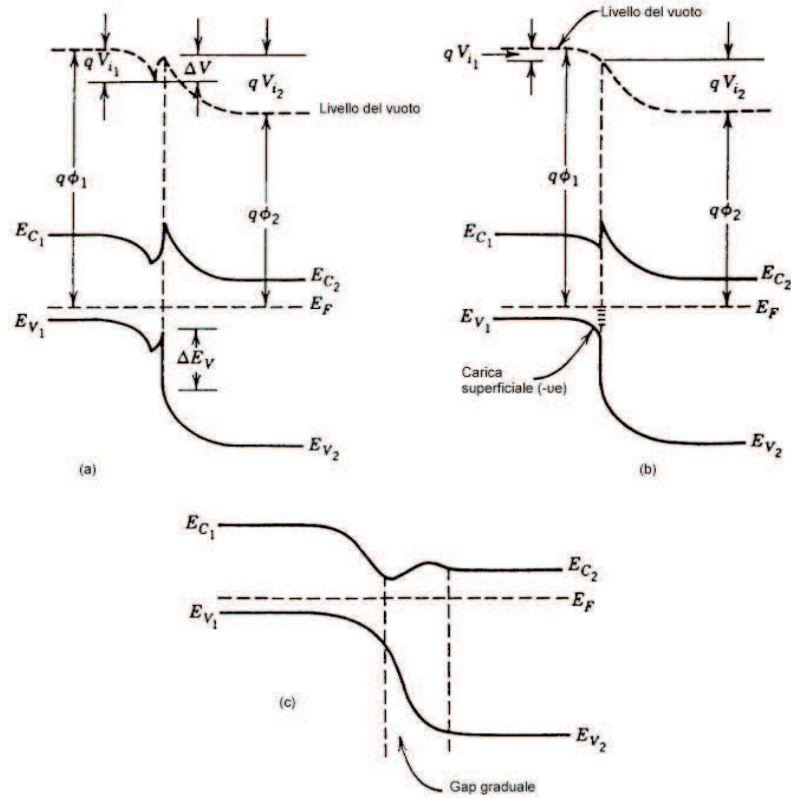


Figura 4.4: Diagrammi a bande di un eterogiunzione non ideale: (a) stati occupabili in banda proibita, collocati in prossimità dell'interfaccia, che possono contenere una carica netta, (b) strati indesiderati all'interfaccia contenuti dipoli superficiali, (c) giunzione graduale.

occorre che i due semiconduttori abbiano la stessa struttura cristallina e la medesima costante reticolare: quando ciò accade si dice che il sistema è "Lattice matched". Nei processi di crescita viene indicato con il termine "misfit" la differenza fra la costante reticolare dello strato accresciuto e quella del substrato; se la struttura non è "Lattice-matched" e quindi il "misfit" è elevato, si possono creare delle imperfezioni all'etero-interfaccia che possono addirittura rendere inutilizzabile la giunzione per certe applicazioni elettron-

iche. Esiste una considerevole libertà nella fabbricazione di eterogiunzioni: si possono scegliere diversi materiali per ottenere i valori desiderati per le energie di gap Eg_2 ed Eg_1 e per i cosiddetti “offset” di banda ΔE_c ed ΔE_v che, come abbiamo visto, sono i parametri più significativi dell’eterogiunzione; si possono verificare tre distinte situazioni quando accostiamo due differenti semiconduttori che danno origine ad altrettante configurazioni, riportate in fig.4.5.

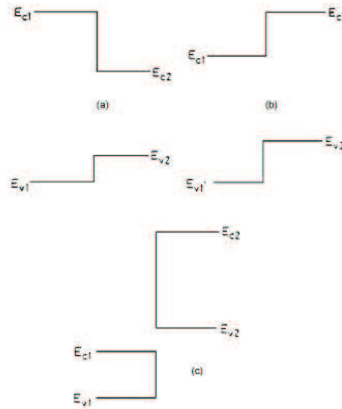


Figura 4.5: Diagrammi di energia per discontinuità di banda di tipo 1 (a), 2 (b) e 3 (c).

4.1.2 Materiali impiegati nell’eterogiunzione

Quando viene costituita un eterogiunzione ponendo a contatto due differenti semiconduttori, la struttura elettronica di ciascuno dei due rimane pressoché invariata eccetto che nella regione di interfaccia; siccome le energie di gap sono differenti, in corrispondenza dell’interfaccia sorgono delle discontinuità sia in banda di valenza che in banda di conduzione, i cui parametri di controllo sono gli offset di banda ΔE_c ed ΔE_v : tali parametri sono importantissimi per

le applicazioni elettroniche in quanto determinano l'altezza delle barriere per il trasporto delle cariche e le proprietà ottiche della struttura. Differenti valori per tali offset possono essere ottenuti scegliendo differenti combinazioni di semiconduttori. Le leghe conduttrici rappresentano un metodo naturale per variare tali parametri: in esse infatti l'ampiezza degli offset varia in maniera pressoché continua in funzione proprio della composizione della lega stessa. Per fare un esempio prendiamo in considerazione il materiale a gap diretto $GaAs$: se l'alluminio (Al) viene parzialmente sostituito al posto del gallio (Ga) durante il processo di crescita del cristallo, nasce una lega semiconduttrice indicata con $Al_xGa_{1-x}As$ che ha circa la stessa costante reticolare del $GaAs$; pertanto è possibile crescere uno strato di $Al_xGa_{1-x}As$ su substrato di $GaAs$ ed il sistema è perfettamente Lattice-matched. La quantità x è detta frazione molare ed è il parametro che controlla la composizione chimica della lega: può variare nel range $0 \leq x \leq 1$ i cui estremi corrispondono alle leghe binarie Arseniuro di Gallio ($GaAs$) per $x = 0$ e Arsenide di Alluminio ($AlAs$) per $x = 1$; la caratteristica veramente importante è che l'energia di gap varia in modo continuo ed in prima approssimazione lineare in funzione proprio della frazione molare passando da quella del $GaAs$ pari a 1.42 eV a 300 °K a quella del $AlAs$ pari a 2.17 eV a 300°K. Se si realizza una giunzione con i semiconduttori $GaAs$ e $Al_xGa_{1-x}As$ ci troviamo nella situazione descritta in fig.4.6, nella quale si vede che in corrispondenza dell'interfaccia si hanno discontinuità sia in banda di valenza che in banda di conduzione dovute alle differenti energie di gap e quantificabili con gli offset ΔE_c ed ΔE_v . Ci interessa principalmente il caso $x \leq 0.45$ in quanto, il $Al_xGa_{1-x}As$ esibisce un comportamento da materiale a gap diretto; in questa zona le discontinuità ΔE_c ed ΔE_v variano approssimativamente con legge

$$\Delta E_g \simeq 1.225x$$

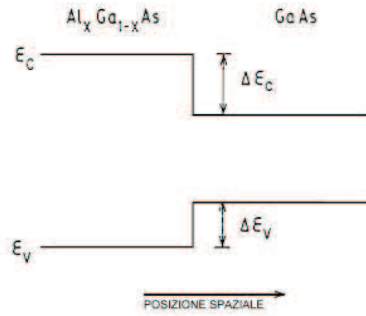


Figura 4.6: Discontinuità delle bande di energia in una eterogiunzione $Al_xGa_{1-x}As - GaAs$

siccome vale anche che:

$$\Delta E_c \simeq 0.62 \Delta E_g$$

allora

$$\Delta E_c \simeq 0.78x$$

quindi variando la frazione molare si possono controllare le caratteristiche dell'eterogiunzione. Non sempre però è possibile un perfetto adattamento reticolare, a tale proposito si può fare la seguente precisazione: nei processi di crescita cristallina in cui si deposita su un singolo cristallo uno strato di materiale diverso, tale crescita è detta epitassiale e con il termine misfit si intende la differenza fra le costanti reticolari del substrato e del film epitassiale. Se il misfit è sufficientemente basso (minore del 7%) e se lo spessore del film è al di sotto di un valore critico, tipico del materiale, si ha un rilassamento degli atomi del film in direzione perpendicolare all'interfaccia che riduce la forza elastica e mantiene gli atomi su piani paralleli all'interfaccia stessa: tale tipo di crescita epitassiale è detta pseudomorfica; il disadattamento reticolare dei materiali viene in questo modo assorbito elasticamente riducendo la difettosità della struttura che si riflette direttamente

sulle proprietà elettriche ed ottiche della giunzione. In ogni caso, per ottenere eterogiunzioni della migliore qualità possibile, occorre utilizzare combinazioni di materiali semiconduttori il più possibile lattice matched; a tal proposito è molto utile fare riferimento alla fig.4.7 dove vengono riportate le energie di gap e le costanti reticolari dei più comuni semiconduttori. Si vede benissimo come il *GaAs* e l'*AlAs* sono materiali che hanno all'incirca la stessa costante reticolare e lo stesso si può dire del composto ternario $Al_xGa_{1-x}As$ che si trova sul segmento che unisce i due materiali: spostandosi su tale segmento attraverso la frazione molare si può idealmente variare ampiamente l'energia di gap.

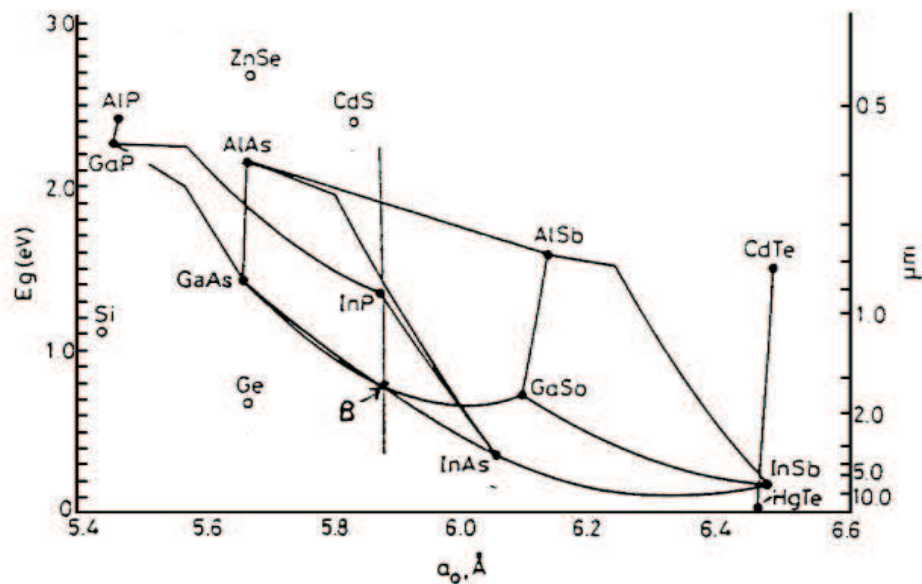


Figura 4.7: Energia di gap in funzione della costante reticolare.

4.2 Applicazione dell'algoritmo numerico alla simulazione di RTD

Per applicare il nostro algoritmo di calcolo ad un caso concreto e testarne l'efficacia, nel seguito del capitolo, simuleremo una struttura standard, spesso usata da altri autori e presa come test per il confronto dei risultati nei diversi metodi numerici: tale dispositivo consta di un substrato in *GaAs*, sopra il quale sono adagiati due strati di $Al_{0.3}Ga_{0.7}As$ dello spessore di 0.3 nm, separati da uno strato di *GaAs* di 0.5 nm che funge da pozzo quantistico [18] [13] [19]. Si creano così due barriere dell'altezza di 0.3 eV. I lati esterni delle barriere confinano con gli spacers, dello spessore ciascuno di 0.3 nm, non drogati, che hanno lo scopo di evitare la diffusione del drogante all'interno della regione attiva; la parte esterna del dispositivo è quindi drogata di tipo n con una concentrazione uniforme di donori pari a $2 \times 10^{18} \text{ cm}^{-3}$. Il diagramma a bande risultante è mostrato in fig. 4.8. Il reticolo cristallino entra nell'approssimazione di massa efficace, che risulta essere 0.0667 volte la massa dell'elettrone libero. La regione simulata è scelta essere di 5.5 nm, ed il numero di punti della griglia di discretizzazione pari a $N_x = 86$, $N_p = 72$. Consideriamo le differenti utilità d'impiego fra il calcolo stazionario e quello transitorio: di principio solo un metodo transitorio è utile nella simulazione di fenomeni tempo-dipendenti, come la risposta del dispositivo ad un brusco cambio di potenziale, o l'innescare di oscillazioni intrinseche come avviene nell'RTD. Per contro i metodi stazionari, più semplici e di più rapida soluzione, sono sfruttati per il calcolo della caratteristica $I(V)$ statica, anche se è controversa la questione di come un metodo stazionario possa essere in grado di calcolare la corrente nei punti di instabilità del dispositivo (cosa che verrà discussa in sez.4.6).



Figura 4.8: Diagramma a bande e profilo del drogante nell'RTD simulato

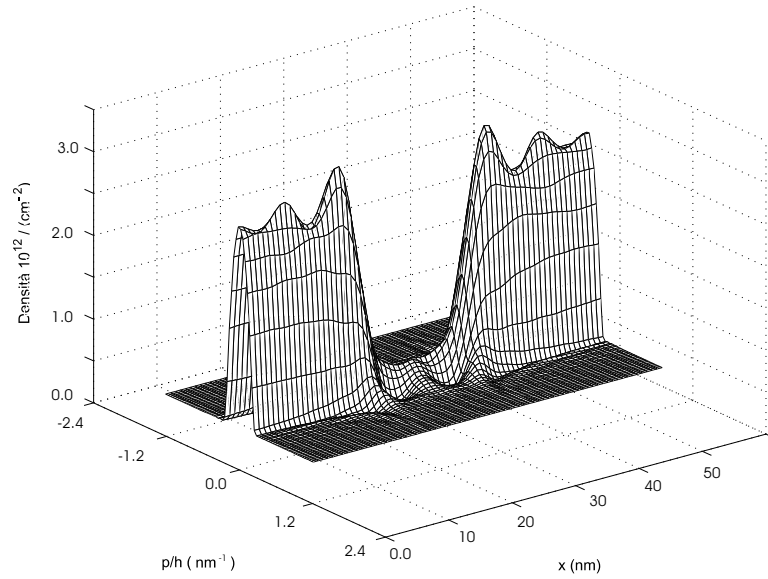
4.2.1 Determinazione della funzione di equilibrio

L'approssimazione dell'operatore collisionale di Boltzmann con un termine a tempo di rilassamento (vedi appendice (5.1)), richiede la conoscenza della funzione di equilibrio termodinamico cui “tende” la f_w : tale funzione deve rispecchiare il modo con cui gli elettroni si dispongono all'interno del semiconduttore per il solo effetto delle collisioni. L'unica funzione di equilibrio termodinamico disponibile dalla teoria della meccanica quantistica, ed utile ai nostri scopi, è la distribuzione di Fermi, applicabile però solo nel caso di semiconduttori *omogenei*. Nel caso reale di dispositivi ad effetto tunnel, tale risultato non è sufficiente, in quanto solo sui contatti è lecito pensare

che il semiconduttore sia praticamente omogeneo (e qui appunto sono localizzati i reservoir, che immettono nel dispositivo proprio la distribuzione di Fermi). Invece di identificare la funzione di equilibrio f_0 con quella di Fermi, si preferisce stimarla approssimandola con la distribuzione di carica presente nello specifico dispositivo, a zero bias (che indicheremo con f_w^{0V}) [20]; così facendo, la scelta ricade su di una funzione che rappresenta il miglior “adattamento” della distribuzione di Fermi, al particolare profilo di potenziale: essa infatti tiene conto che in realtà gli elettroni, in moto di agitazione termica sui contatti, diffondono all’interno del semiconduttore, per poi interagire e quindi essere riflessi dalle barriere. Le motivazioni quindi addotte a pro della scelta fatta di $f_0 = f_w^{0V}$ sono le seguenti:

- 1) La densità di elettroni e la loro distribuzione varia significativamente in prossimità delle barriere, cosa non prevista da una funzione di distribuzione che non dipenda dalla variabile x , come quella di Fermi, cioè una $f_0 = f_0(p)$
- 2) L’equazione di bilancio dettagliato è soddisfatta .

L’ultimo punto significa che con $f_0 = f_w^{0V}$, l’operatore collisionale è identicamente nullo ($\frac{\partial f}{\partial t} \big|_{coll}$) nelle simulazioni di dispositivi non polarizzati (ovvero nel caso di stati di equilibrio). Lo si dimostra semplicemente, notando che la f_w^{0V} è soluzione dell’equazione di Wigner e quindi annulla sia il primo membro della 1.28) che $\frac{\partial f}{\partial t} \big|_{coll}$ (vedi (3.6)). Numericamente comunque, abbiamo verificato che le differenze nel calcolo della f_w , con l’una o l’altra funzione di equilibrio, sono piccole e riflettono il fatto che l’estensione della regione di simulazione è dell’ordine del libero cammino medio degli elettroni, ovvero il trasporto è principalmente balistico.

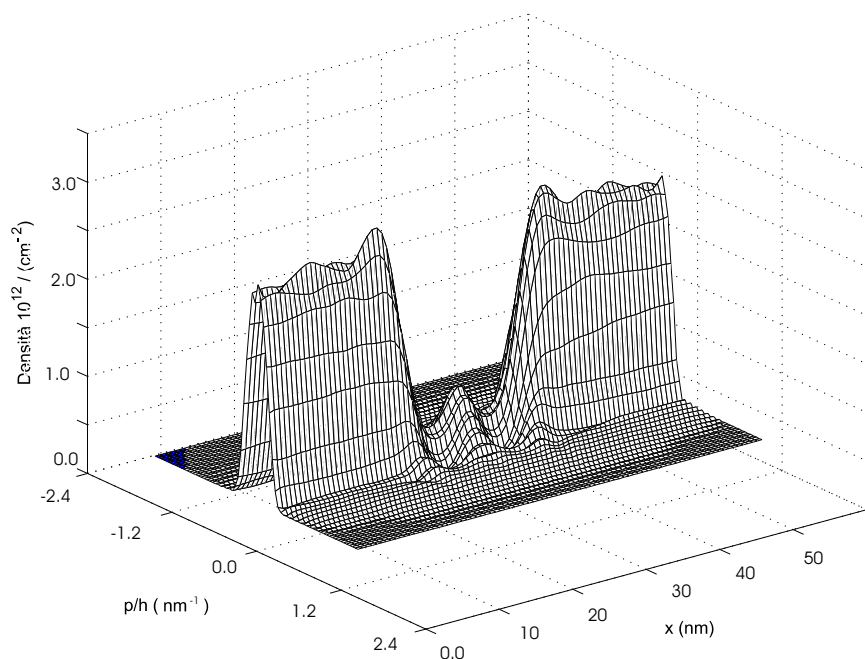
Figura 4.9: f_0 alla temperatura di 77 K

4.3 Soluzione stazionaria

Si è applicato l'algoritmo stazionario non autoconsistente discusso in sez.(3.2), al profilo di potenziale precedentemente descritto, in assenza di polarizzazione: le funzioni di Wigner risultanti rispettivamente per $T=77$ K e $T=300$ K sono mostrate in fig.4.9 e 4.10 e verranno discusse in dettaglio nella prossima sezione. Come anticipato, tali funzioni verranno prese come distribuzione di equilibrio f^0 nei calcoli successivi.

4.3.1 Calcolo del diagramma $I(V)$ in assenza di collisioni

Per il calcolo della relazione tensione-corrente del diodo, è necessario conoscere come il potenziale applicato fra anodo a e catodo “cade” all'interno del semi-

Figura 4.10: f_0 alla temperatura di 300 K

conduttore. Nel presente calcolo non autoconsistente adottiamo una approssimazione così detta a “iniezione di elettrone caldo” [21], nella quale si assume che i contatti siano elettricamente neutri, che la caduta sia lineare e che si localizzi sostanzialmente in prossimità della regione attiva. Il potenziale di polarizzazione sarà dato quindi da $V(x) = V_{Bias}(x - x_1)/(x_1 - x_2)$ e sovrapposto al profilo dell’eterostruttura; x_1 e x_2 si posizionano a $30 \text{ \AA}$ rispettivamente prima della prima barriera e dopo la seconda. Come verificheremo nel seguito, specialmente per valori alti di V_{Bias} , la precedente è una buona approssimazione del vero potenziale autoconsistente che calcoleremo in sez. 4.5. Il diagramma $I(V)$ risultante è riportato in fig. 4.11 in cui è evidenziata la regione a resistenza differenziale negativa (NDR). Notiamo inoltre come la curva a 300 K mostri, contrariamente all’esperienza, una corrente di picco

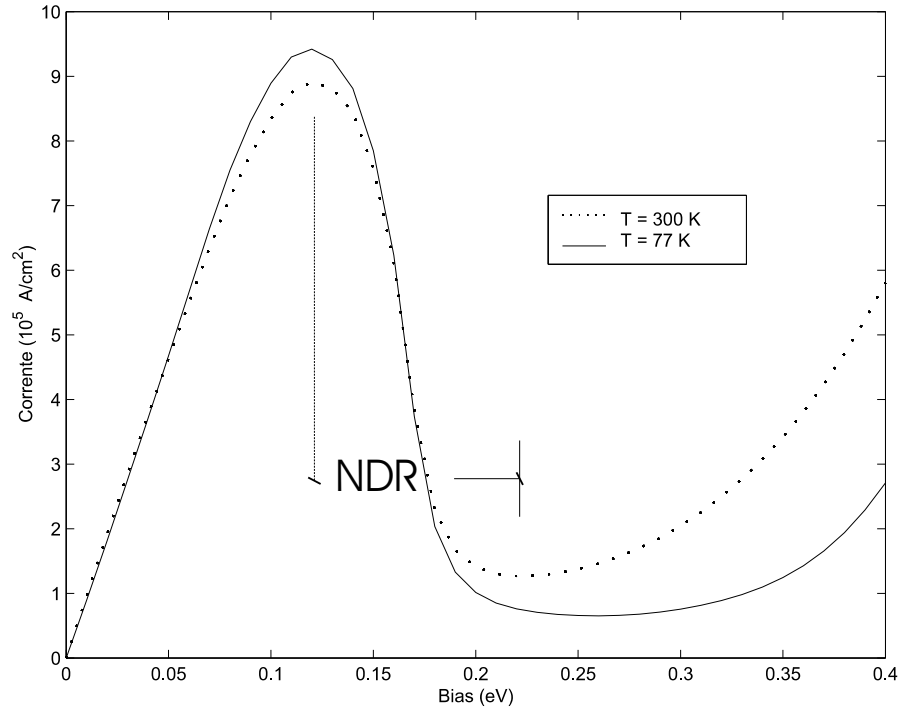


Figura 4.11: Diagramma $I(V)$ nel caso di assenza di fenomeni collisionali (balistico) per $T=77 \text{ K}$ e $T=300 \text{ K}$

praticamente uguale a quella di 77 K , invece di essere notevolmente ridotta. Tale incongruenza è causata dall'omissione delle collisioni che, a temperatura ambiente, sono un fenomeno determinante nel moto degli elettroni: la loro inclusione nel prossimo paragrafo, correggerà tale tendenza, producendo risultati in pieno accordo con le misure sperimentali. Commentiamo brevemente i risultati ottenuti nel corso della simulazione: in fig.4.9 è mostrata la funzione di equilibrio (zero-bias) in cui è evidente l'elevata concentrazione di elettroni nelle regioni degli elettrodi, e il loro brusco calo in prossimità delle barriere. Piccole oscillazioni compaiono in varie regioni del piano, riflettendo gli effetti di interferenza quantistica presenti all'interno del semiconduttore. La fig. 4.12 mostra invece la funzione stazionaria di Wigner per una ten-

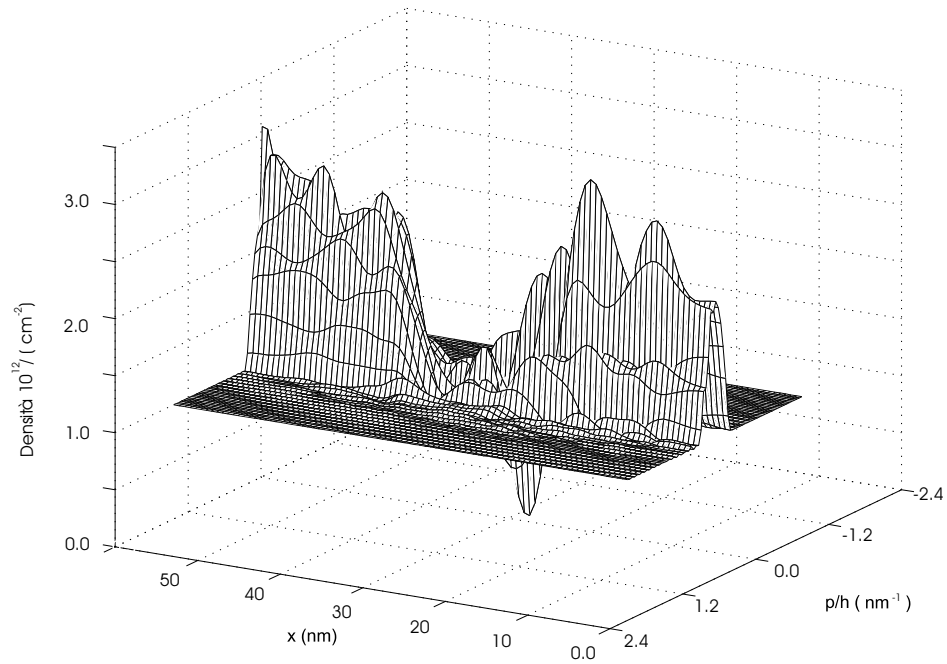


Figura 4.12: Funzione di Wigner in corrispondenza del picco di corrente del diagramma $I(V)$

sione di 0.11 V, che corrisponde al picco della curva $I(V)$. Le oscillazioni sono più pronunciate, fino alla formazione di un minimo *a valori negativi* della funzione, che non ha nessun corrispettivo classico e che caratterizza la forte natura quantistica dei fenomeni di trasporto in questa regione della caratteristica $I(V)$. Inoltre si vede come la zona di interferenza quantistica si estenda fino a giungere in prossimità dei contatti. All'opposto, la funzione di Wigner (mostrata in fig.4.13) per una maggiore tensione di 0.25 V e che corrisponde alla corrente di “valle” è molto più simile a quella di equilibrio.

Affinché i risultati delle simulazioni siano attendibili si deve garantire che i limiti della regione di simulazione siano sufficientemente lontani dalle barriere, altrimenti, se gli effetti di interferenza quantistica si estendono fino ad

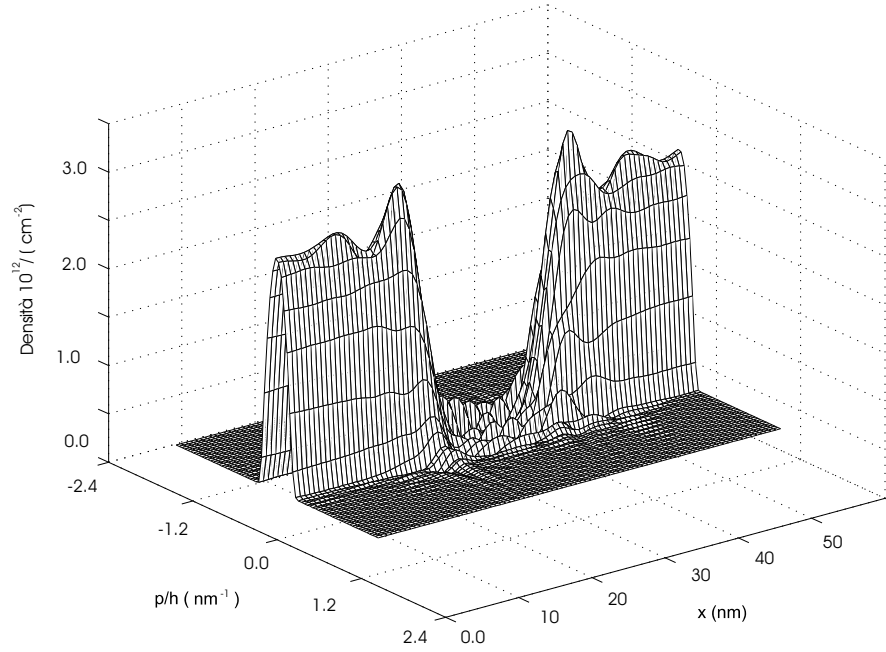


Figura 4.13: Funzione di Wigner nel punto di valle del diagramma I(V)

essi, l'assunzione che gli elettroni iniettati dai reservoir nel circuito siano in equilibrio termico diventa inaccurata e artificiosa. Per forzare l'annullamento degli effetti quantistici sui bordi, oltre all'ovvio aumento della regione di simulazione, si può scegliere un N_p sufficientemente più piccolo di N_x in modo da annullare $V(x, p)$ di formula (3.9) e conseguentemente della (3.5) sul contorno, diminuendo quindi artificialmente la non località del potenziale.

4.3.2 Presenza di collisioni

Vediamo come il diagramma I(V) precedente si modifica se si includono i fenomeni collisionali elettrone-fonone. Gli eventi di urto sono caratterizzati dall'avere una forte dipendenza dalla temperatura, essendo quest'ultima una misura indiretta proprio dell'interazione fra portatori e reticolo. Nel nostro

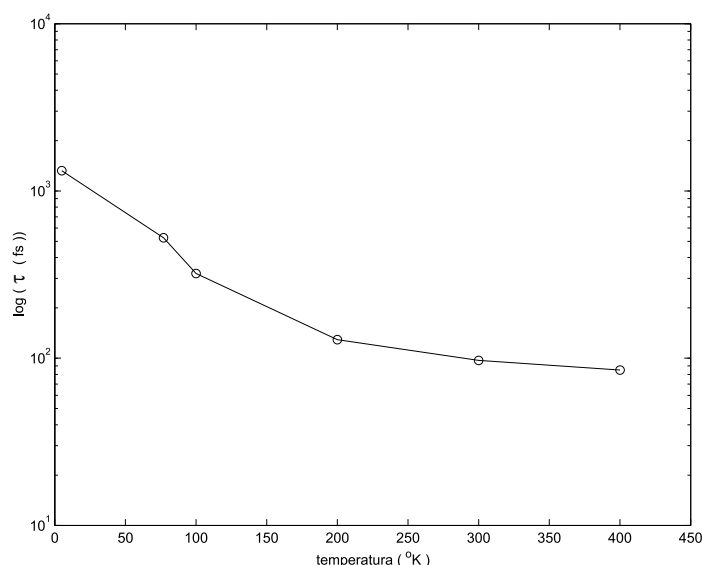


Figura 4.14: Legame fra tempo di rilassamento e temperatura [22]

modello la temperatura gioca un duplice ruolo: modifica il tempo τ di rilassamento, come mostrato in fig.4.14, e la varianza della distribuzione di Fermi nei reservoir. Un aumento di T provocherà perciò un maggior sparpagliamento degli elettroni, favorendo il trasporto di tipo termoionico e quindi con esso un generale aumento della corrente, ma anche un accoppiamento più stretto fra la f_w e la f^0 , che contrasterà con la tendenza precedente. Ci aspettiamo in particolare che la presenza di quest'ultimo meccanismo interferisca fortemente (e soprattutto) con i meccanismi di tunnel risonante, e quindi particolarmente visibile proprio nella regione a resistenza negativa. Nella NDR infatti la f_w è caratterizzata da una forma oscillante particolarmente irregolare: come ogni fenomeno risonante, anch'esso è poco robusto e criticamente sensibile ad ogni interferenza che tenda “sporcare” la forma della funzione, come l'inserimento della $\left. \frac{\partial f}{\partial t} \right|_{coll}$. Una discussione più approfondita di quest'aspetto è possibile attraverso l'analisi delle così dette “traiettorie

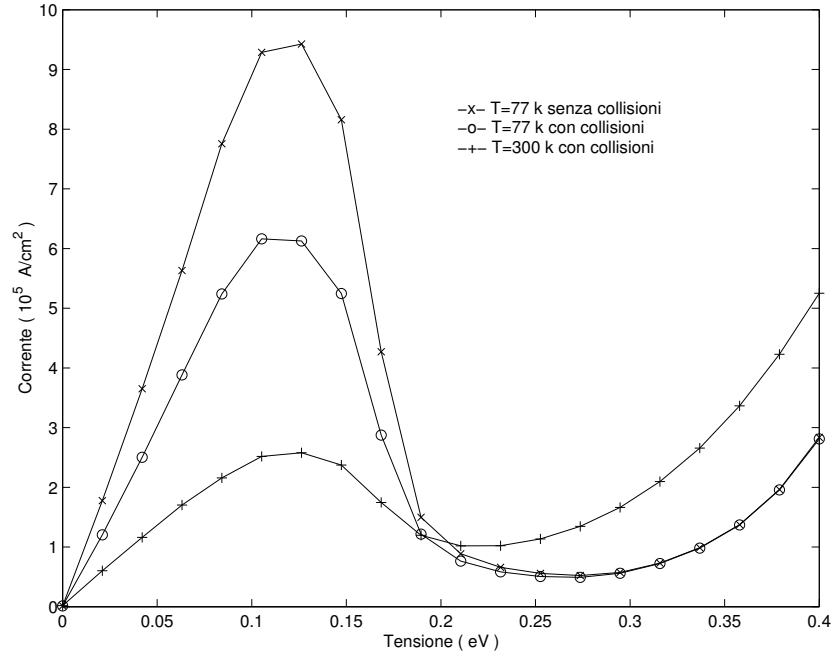


Figura 4.15: Effetto dell'inclusione degli urti nel diagramma I(V) dell'RTD

risonanti" di sez.4.7.

Abbiamo simulato il comportamento del diodo nel range di temperatura fra 50 e 300 K: in fig.4.15 ne sono riassunti i risultati salienti, che si dimostrano in accordo con le precedenti affermazioni. Una conferma della validità del nostro metodo numerico giunge dal calcolo della degenerazione del rapporto picco-valle a causa dell'aumento della temperatura: i risultati illustrati in fig.4.16 concordano infatti con la seguente formula sperimentale [23]

$$\frac{I_p}{I_v} = R \ln \frac{T_0}{T}$$

in cui $r = 5.35\Omega$ e $T_0 = 451.4$ K, entro una precisione del 5%. In fig.4.17 riportiamo il grafico della densità di carica nel dispositivo, ottenuta integrando la $f_w(x, p, t)$ rispetto a p , a quattro differenti tensioni. Inizialmente la carica intrappolata fra le barriere aumenta, ma dopo gli 0.11 V inizia a decrescere, cosicché tale densità, al termine della NDR, è significativamente più piccola

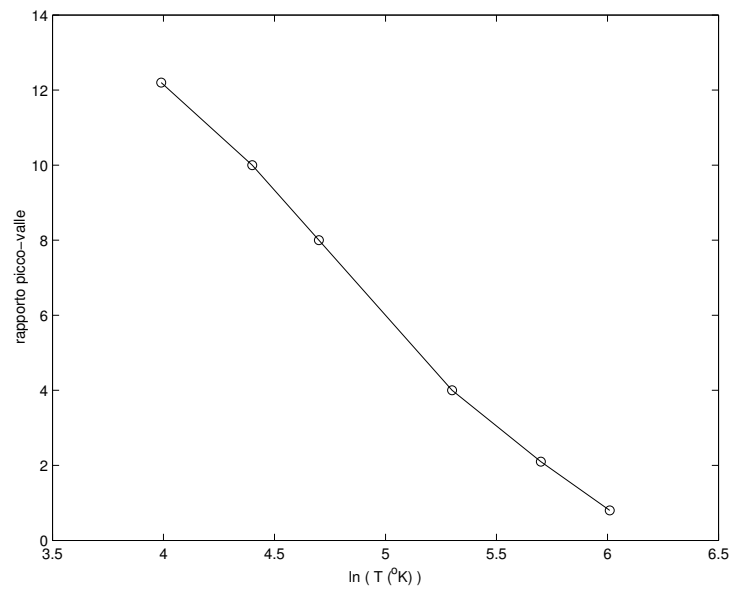
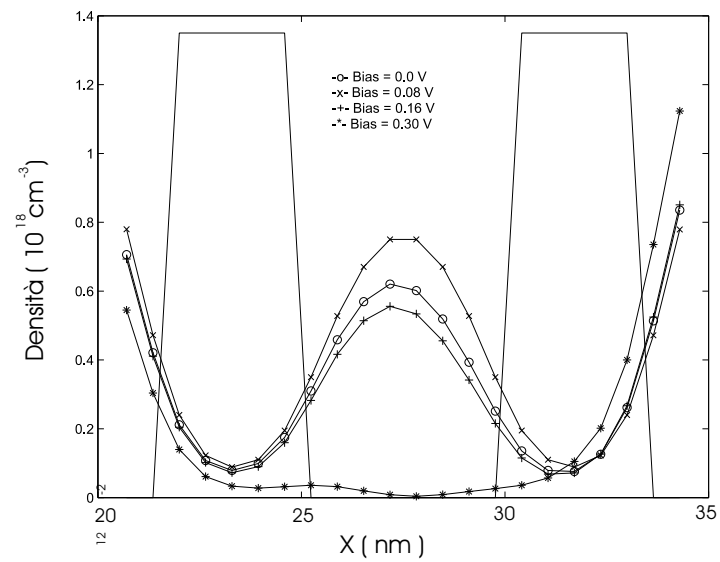


Figura 4.16: Rapporto picco-valle in funzione della temperatura


 Figura 4.17: Densità di carica nel pozzo quantistico al variare della tensione per $T=300 \text{ K}$

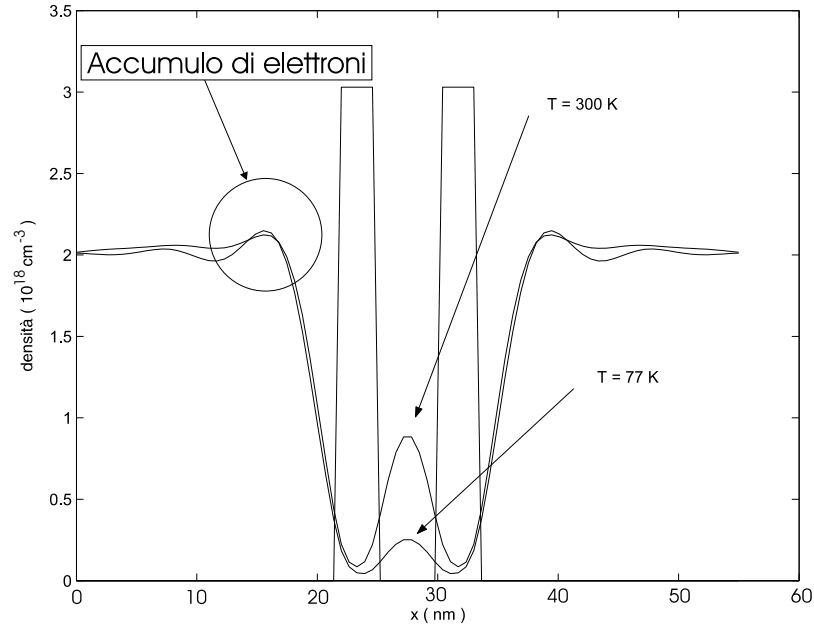


Figura 4.18: Densità di carica nel dispositivo al variare della temperatura

del caso zero-bias. Tale trend è confermato anche sperimentalmente [24]. Un ulteriore aumento della tensione ricrea un debole accumulo di carica nel pozzo quantistico.

Il leggero aumento della densità di carica in prossimità dei lati esterni delle barriere segnalato in fig.4.18, invece è un effetto tipicamente quantistico e riflette la natura non locale del potenziale. Infine confrontiamo il diverso profilo di carica al variare della temperatura: è evidente dalla fig.4.18 come l'accumulo di portatori nel pozzo centrale a 300 K sia molto più pronunciato di quello a 77 K.

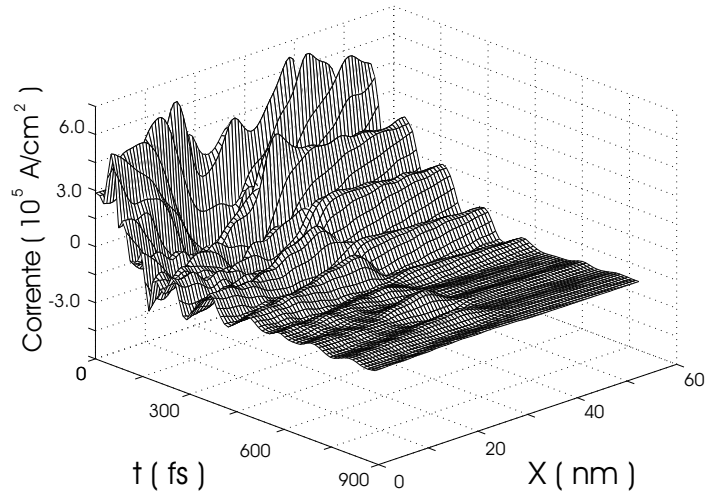


Figura 4.19: Variazione temporale della corrente causata da uno switch da 0.13 a 0.22 Volt ad una temperatura di 77 K

4.4 Soluzione evolutiva

Nella presente sezione studieremo il comportamento transitorio dell'RTD, in risposta a brusche variazioni del potenziale applicato: i risultati ottenuti verranno poi (in sez.4.6) confrontati con quelli più completi delle simulazioni auto-consistenti. Mentre uno stato stazionario, come diretta conseguenza della legge di continuità della carica, prevede solo distribuzioni di corrente costanti (tradotto in ambito circuitale, ciò significa che al dispositivo descritto è applicabile il secondo primo principio di Kirkoff, per contro in uno stato transitorio il gradiente della corrente non è più nullo. La fig. 4.19 mostra l'andamento della corrente all'interno dell'RTD a causa di una variazione istantanea di tensione da 0.13 a 0.22 V. Si nota come il tempo di assestamento del circuito (tempo di switch) sia dell'ordine dei 600 fs. Un'investigazione delle caratteristiche degli stati risonanti ha mostrato che i tempi fra una transizione e la sua inversa sono differenti fra loro: per il caso precedente

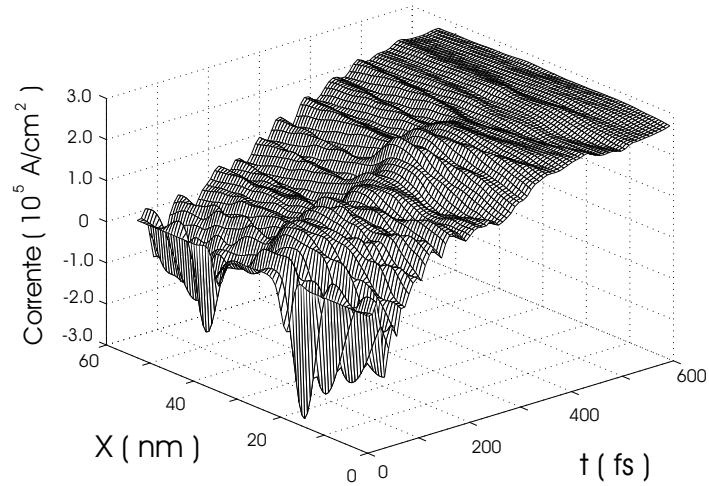


Figura 4.20: Variazione temporale della corrente causata da uno switch da 0.22 a 0.13 Volt ad una temperatura di 77 K

infatti uno switch da 0.22 a 0.13 V richiede un tempo 1.5 volte inferiore (come mostrato in fig.4.20). Tale osservazione, accoppiata con l'analisi del profilo di densità relativo agli stati di partenza e di arrivo, presentato prima, indica come il dispositivo richieda più tempo a svuotare la carica immagazzinata fra le due barriere, che ad accumularla.

4.4.1 Effetti delle collisioni nei tempi di switch

Per analizzare l'effetto delle collisioni nei tempi di switch, abbiamo calcolato le distribuzioni stazionarie in assenza dell'operatore di scattering: è emerso che la densità di carica nel pozzo aumenta. Poiché quello che pesa di più nel tempo di transizione fra due livelli è la quantità di carica che deve essere rimossa dal centro del pozzo, questo porta alla strana conclusione che le collisioni hanno il paradossale effetto di diminuire i tempi di switch: un confronto fra la fig.4.19 e la fig.4.21 conferma tale supposizione. Una giustificazione di tale fenomeno segue dall'analisi delle traiettorie eseguita in sez. 4.7.

4.5 Soluzione stazionaria auto-consistente

Mostriamo i risultati dell'applicazione del metodo di Newton, descritto in sez.3.5, al calcolo del diagramma $I(V)$ dell'RTD. La determinazione della soluzione auto-consistente del sistema W-P, richiede una stima iniziale della f_w , necessaria per poter avviare l'algoritmo iterativo di Newton: la scelta naturale è la soluzione stazionaria non autoconsistente dell'RTD, calcolata in assenza di polarizzazione (ovvero quella di fig.(4.9)). Al termine dell'algoritmo di calcolo si ottengono il profilo e la funzione di Wigner auto-consistenti dell'RTD non polarizzato; tali funzioni sono scelte come punto di partenza per la seguente procedura: si incrementa la tensione applicata al diodo, sovrapponendo al profilo di potenziale una funzione a rampa che crei un dislivello pari ad una tensione dV ; con questi nuovi dati si riapplica il metodo di Newton: la soluzione trovata fornirà il punto di partenza per un ulteriore aumento di tensione. Raggiunta la tensione massima, quest'ultima viene

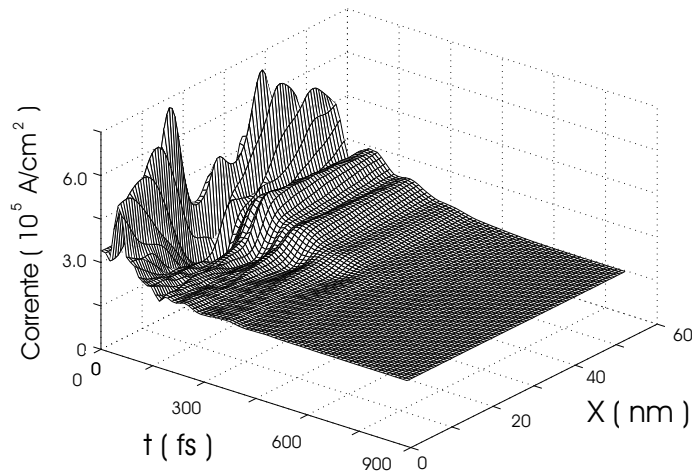


Figura 4.21: Variazione temporale della corrente causata da uno switch da 0.13 a 0.22 Volt ad una temperatura di 300 K: è evidente la diminuzione dello switch time rispetto ai casi precedenti

decrementata. Nel grafico di fig.4.22 si riportano i risultati relativi ad una tensione massima di 0.4 eV, suddivisa in 40 step. Le simulazioni sono state effettuate con $T=77$ K perché a tale temperatura il dispositivo mostra meglio le sue caratteristiche peculiari quali il ciclo di isteresi e le oscillazioni intrinseche. Se paragoniamo i risultati ottenuti con quelli precedenti riportati in fig.4.15 si nota che la corrente di picco si è mossa a valori più alti di tensione, come ci si potrebbe aspettare alla luce del profilo autoconsistente trovato in fig.4.24 nel quale il potenziale del pozzo non è così basso come approssimato nel modello di caduta di tensione lineare. Conseguentemente una tensione maggiore è richiesta per allineare il livello di Fermi con lo stato risonante. Inoltre si verifica anche un'avvicinamento fra la tensione di picco e quella di

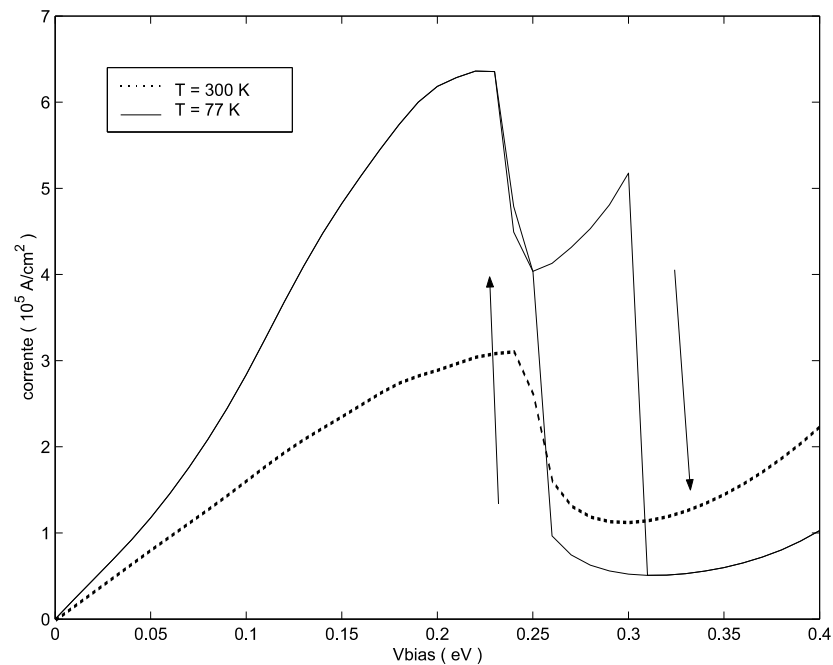


Figura 4.22: Diagramma $I(V)$ autoconsistente con collisioni alla temperatura di 77 K e 300 K; si nota come alla temperatura ambiente il ciclo d'isteresi sia praticamente scomparso

valle (più simile ai risultati sperimentali). Il dispositivo mostra un comportamento bistabile che, unito all'inerzia presentata nella redistribuzione della carica al variare della tensione, spiega la presenza del ciclo d'isteresi. Per una stessa tensione sono possibili due distinte soluzioni del sistema W-P, che si differenziano per la concentrazione di elettroni nella zona fra le barriere: in una ve ne sono molti, nell'altra praticamente nessuno. All'aumentare della tensione, l'iniziale quantità di carica nel pozzo tende ad essere mantenuta, e il potenziale elettrostatico che ne deriva, disegna il primo braccio del ciclo d'isteresi. Al termine della NDR comunque, la carica centrale (come già osservato e in questa sede confermato), viene quasi completamente rimossa. Quando però la V ridiminuisce, al passaggio del livello risonante nella regione proibita, non c'è più nessun meccanismo che rifornisca di carica il pozzo: gli elettroni dal lato catodico hanno un'energia troppo alta, mentre quelli dell'anodo non ne hanno a sufficienza. Rispetto al caso precedente,

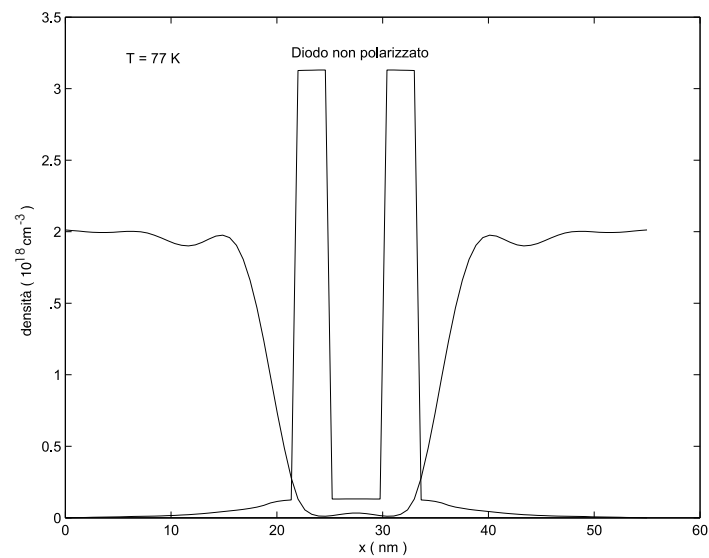


Figura 4.23: Potenziale e densità autoconsistenti per l'RTD non polarizzato

tale deficit di carica deforma il potenziale auto-consistente, innalzando il diagramma a bande dal lato catodico dell'eterostruttura. Un riallineamento

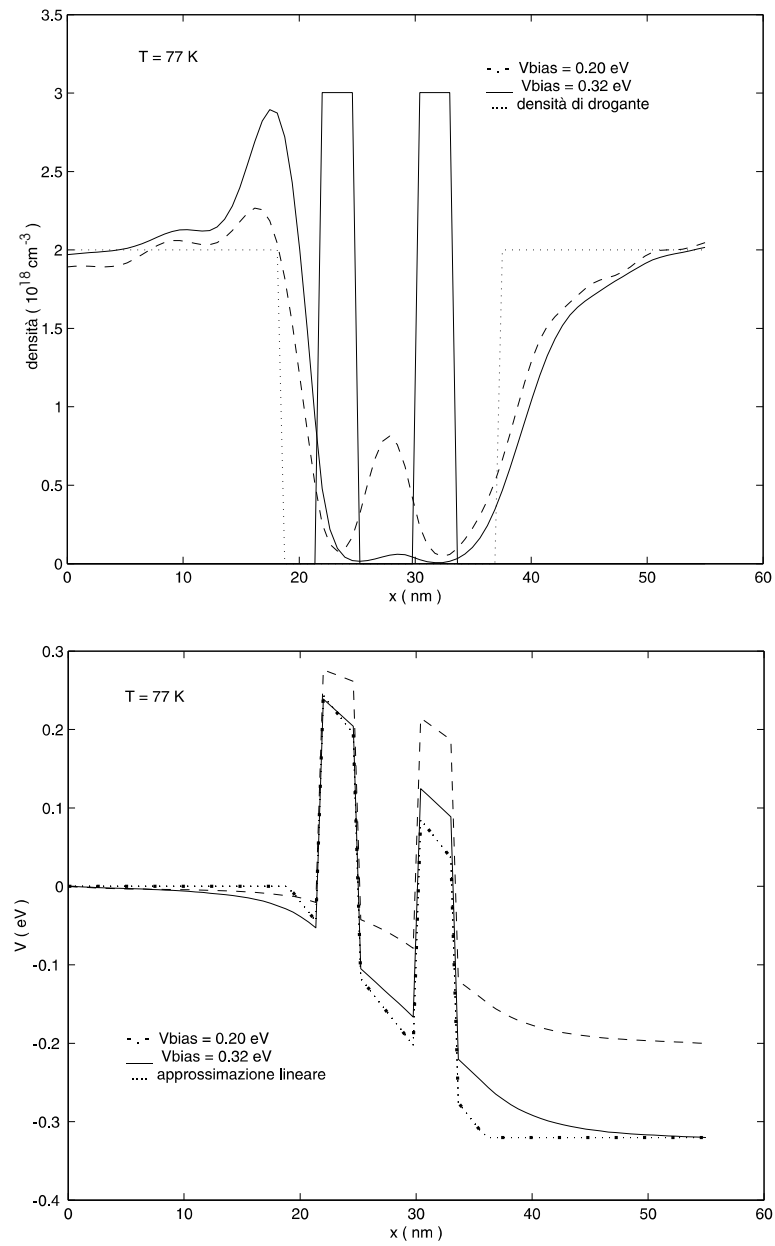


Figura 4.24: Densità e potenziale autoconsistenti

dei livelli energetici della scatola quantistica, con gli stati occupati dell'anodo (indispensabile al riaccumulo di carica) richiede quindi una tensione più bassa. Conseguentemente la corrente manterrà valori bassi di valle per un range maggiore di tensione, tracciando il secondo braccio del ciclo d'isteresi; in fig.4.25 mostriamo le due distinte distribuzioni di densità e potenziale che originano i due rami del diagramma $I(V)$.

I metodi stazionari autoconsistenti sono stati introdotti dalla necessità di analizzare le caratteristiche statiche dei dispositivi, senza ricorrere ad i più onerosi metodi tempo-dipendenti: una critica verso tale impiego deve essere mossa. Dimosteremo che nel range di tensione 0.24 - 0.26 Volt l'RDT oscilla; in questo caso come è possibile che un metodo stazionario sia in grado predire il comportamento di un sistema dinamico? La risposta è che, poiché il sistema oscilla attorno ad una soluzione di equilibrio instabile del sistema W-P, è proprio tale stato instabile che emerge dal calcolo stazionario: il valore di corrente statico precedentemente calcolato coincide infatti col valor medio della corrente oscillante della soluzione dinamica. A riprova abbiamo preso la soluzione stazionaria a 0.24 V, e ne abbiamo seguito l'evoluzione temporale: il rumore numerico è stato sufficiente affinché la corrente iniziasse ad oscillare con le stesse modalità che risultano da un calcolo interamente tempo-dipendente; ciò è illustrato in fig. 4.29. Poiché il metodo Newton è un metodo iterativo, riportiamo i criteri di convergenza adottati nelle presenti simulazioni.

4.5.1 Criterio di convergenza

Stabilire un adeguato criterio di convergenza pone di fronte alla necessità di prediligere la rapidità di calcolo (che nel nostro caso è un problema tutt'altro che trascurabile) o la precisione dei risultati. Purtroppo si è riscontrato che

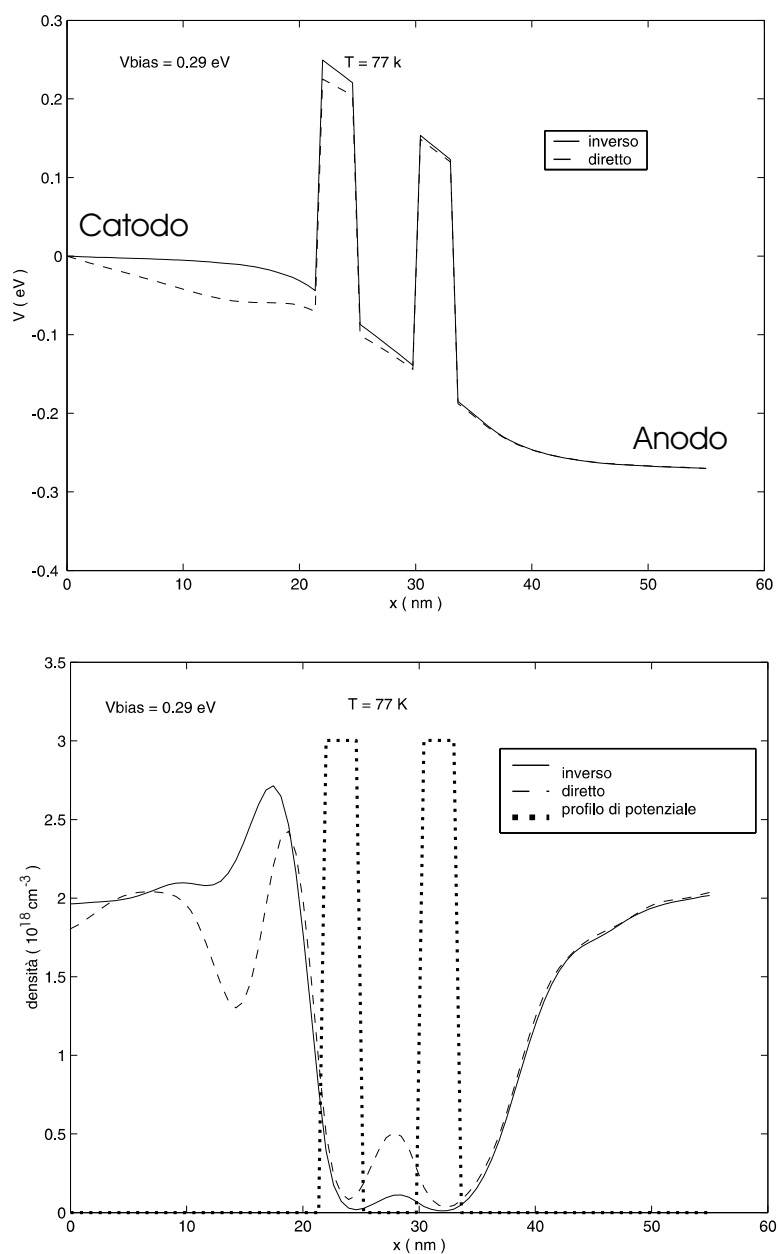


Figura 4.25: Potenziale e densità autoconsistenti per $V_{bias} = 0.28$ eV : con diretto intendiamo che la soluzione è trovata aumentando il potenziale di polarizzazione, con inverso l'opposto.

piccoli errori possono essere disastrosi, in qualche caso, per la valutazione dell'effettiva funzione di equilibrio; conseguentemente siamo stati costretti a sacrificare lunghi tempi di calcolo (la cui entità è riassunta in tab.4.1) in favore di criteri di convergenza particolarmente esigenti. Nell'algoritmo Newton, un ovvio criterio di convergenza, è la verifica che l'equazione di Poisson (3.41) sia soddisfatta con alto grado di precisione. Si è scelto

$$P_i < 10^{-8} \text{ eV con } (0 < i < N_x) \quad (4.1)$$

Tale criterio anche se necessario, non sempre è sufficiente: per assicurarsi che due successive soluzioni siano non oscillanti, dobbiamo richiedere che anche le correzioni al potenziale, fra due successive iterazioni, in ogni punto, siano praticamente nulle. Imponiamo

$$\delta u_i < 10^{-6} \text{ eV con } (0 < i < N_x) \quad . \quad (4.2)$$

4.6 Soluzione evolutiva auto-consistente

In contrasto col metodo stazionario, quello evolutivo cerca di seguire l'esatta evoluzione temporale del sistema (dispositivo); condizione necessaria affinché una soluzione tempo dipendente dell'equazione di moto (nella sua forma completa auto-consistente) sia plausibile, è che l'equazione di Poisson (in ogni istante) sia risolta in maniera esatta, e non per successive approssimazioni come avviene nella tecnica di Newton. La richiesta di consistenza fra la distribuzione di carica in un certo istante e il potenziale elettrostatico da essa prodotto, ci fa quindi prediligere l'algoritmo di Gummel, presentato in sez. 3.5, al posto di quello di Newton. L'algoritmo di Newton era stato introdotto per stabilizzare il metodo numerico *stazionario* e per accelerare la determinazione di una soluzione congiunta delle due equazioni (Wigner e Poisson).

Per un metodo di evoluzione, al contrario, se si sceglie una congrua condizione iniziale e uno step temporale sufficientemente piccolo, le modifiche della f_w , (e conseguentemente della distribuzione di carica), fra t e $t + dt$ saranno di modesta entità, permettendole così di adattarsi dinamicamente alla conseguente variazione del potenziale elettrostatico. Nel caso stazionario invece, fra due successivi step risolutivi, le funzioni di distribuzione non sono fortemente correlate fra loro: questo perché una soluzione stazionaria *non* è un problema ai valori iniziali, ma solo un problema ai valori al contorno. Quindi è necessario premunirsi affinché non si innescino oscillazioni che minino la stabilità del sistema numerico.

Ricapitoliamo brevemente l'algoritmo di Gummel: il punto di partenza di ogni iterazione è l'esatto profilo di potenziale ricavato integrando l'equazione di Poisson (3.36) rispetto all'attuale densità di portatori; si lascia quindi evolvere il sistema per un dt fissato risolvendo l'equazione transitoria di Wigner; infine il potenziale è corretto in accordo con la nuova densità trovata. Il procedimento prosegue finché, o si approda ad uno stato stazionario (soddisfacendo il criterio di convergenza sotto esposto), o si supera il numero massimo di iterazioni prefissato. Per condizione iniziale nel calcolo del primo punto del diagramma $I(V)$ è stata presa la soluzione stazionaria autoconsistente dell'RTD non polarizzato: con una procedura del tutto simile a quella di sezione precedente si è aumentato il potenziale fino a 0.4 eV, e si è decrementato fino ad annullarlo; come anticipato, il risultato è che si riproducono esattamente (anche se con un dispendio di calcoli notevolmente superiore) i valori di corrente di fig.4.22, con l'unica differenza che nei punti in cui l'RTD oscilla, quest'ultima è determinata come la media fra i valori massimo e minimo. Focalizziamo l'attenzione sulle soluzioni non stazionarie, che sono il valore aggiunto alla modellizzazione dell'RTD. La simulazione converge per

tutte le tensioni eccetto quelle nella regione $0.24 \leq V_{Bias} \leq 0.26$ Volt. Come mostrato nelle fig.4.27 e 4.26, ad ogni cambio di tensione, la corrente oscilla fino ad assestarsi o su un valore stazionario, o sulla propria frequenza naturale di oscillazione. I tempi di assestamento sono dell'ordine del ps, e

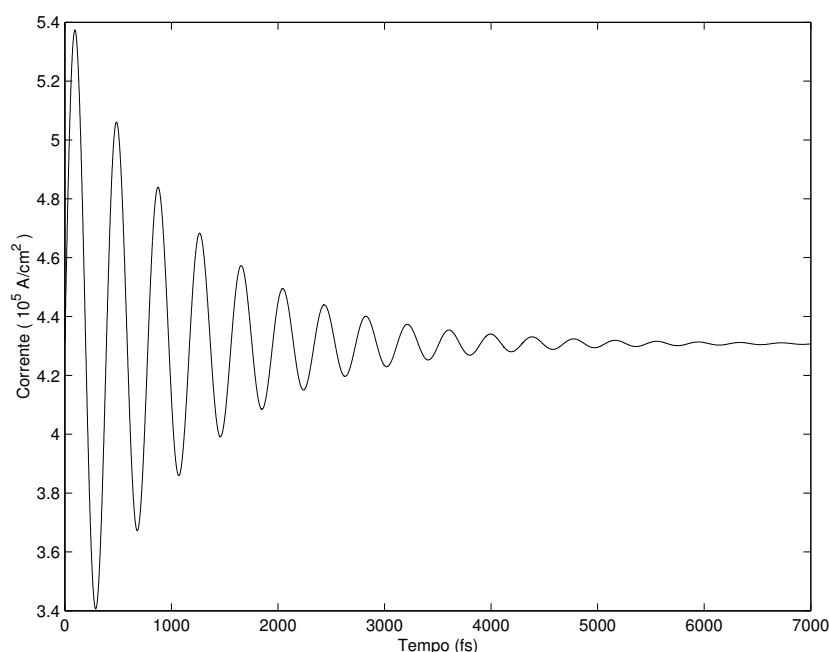


Figura 4.26: Switch in tensione fra 0.26 V e 0.27 V

raggiungono il loro massimo in vicinanza dei punti di instabilità: laddove infatti sono sufficienti 1500 iterazioni per raggiungere uno stato stazionario a 0.31 V, per contro ne occorrono almeno 5000 a 0.27 V. Analizziamo la zona di stabilità ($V_{Bias} < 0.24$ V e $V_{Bias} > 0.27$ V): per mostrare come le oscillazioni originate da un cambio di tensione, qui siano effettivamente smorzate, ovvero che il sistema tenda ad uno stato stazionario, riportiamo il grafico (fig. 4.26) dell'evoluzione temporale della corrente del catodo, a seguito di uno switch della tensione da (0.26 V a 0.27 V): le oscillazioni decrescono con regolarità e decadono con una costante di tempo pari a 0.2 ps^{-1} ; quindi detta A

l'ampiezza delle oscillazioni:

$$A(t) \approx 0.6 \cdot 10^{-5} e^{-t/5}$$

Viceversa, nella regione di instabilità, si innescano oscillazioni auto-sostenute, come illustrato in figura 4.27, che mostra il risultato della transizione da 0.23 V a 0.24 V. In fig.4.28 riportiamo invece la transizione inversa ovvero

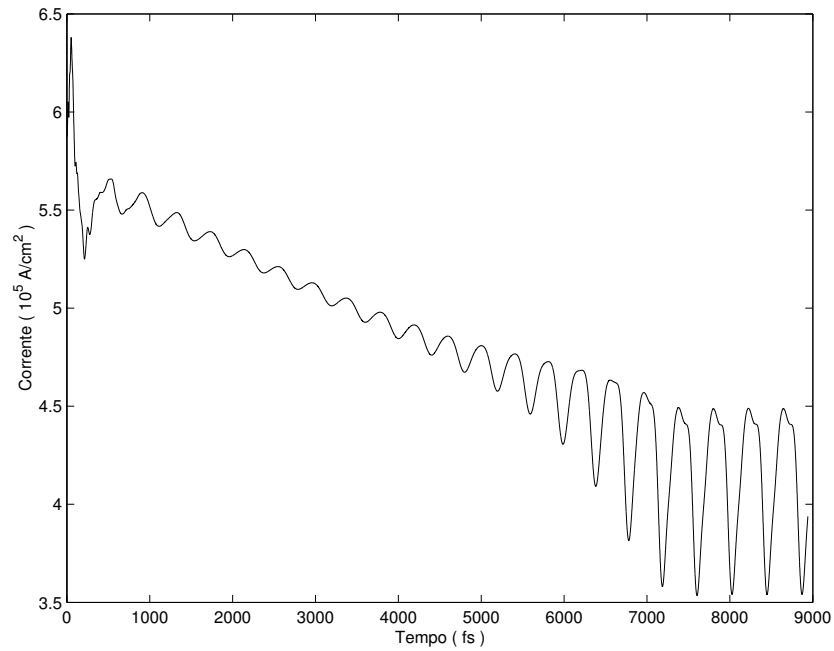


Figura 4.27: Switch in tensione fra 0.23 V e 0.24 V

fra 0.24 V (instabile) e 0.23 V (stabile). Come già anticipato in sez.4.5, per assicurarci che le oscillazioni intrinseche siano perpetue e non smorzate, abbiamo simulato l'evoluzione del sistema (vedi fig.4.29) prendendo come condizione iniziale la soluzione dell'algoritmo di Newton stazionario (risolta con alto grado di precisione), per la tensione di polarizzazione di 0.24 V: anche se molto lentamente si nota come si instaurino le stesse oscillazioni precedentemente mostrate in fig.4.27.

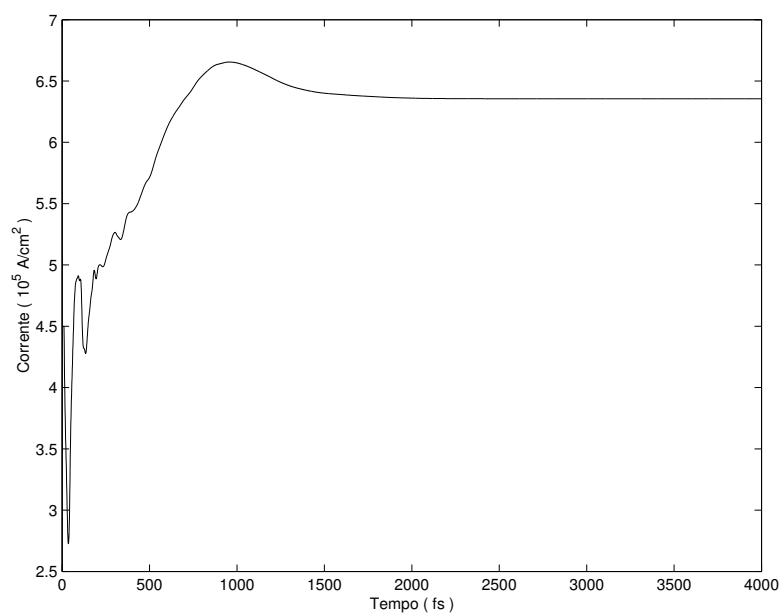


Figura 4.28: Switch in tensione fra 0.24 V e 0.23 V

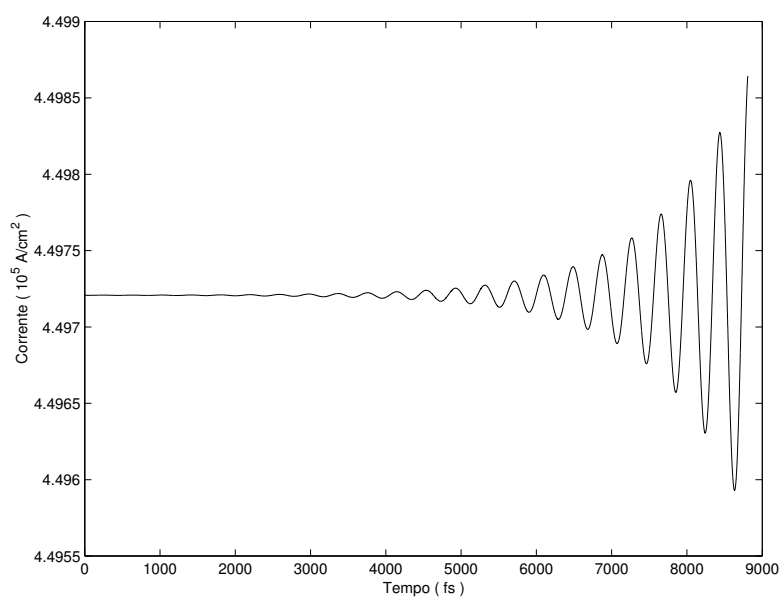


Figura 4.29: Oscillazioni intrinseche innescate dal solo rumore numerico, alla tensione di 0.24 V

Basandoci sulle nostre simulazioni concludiamo che la regione di isteresi è divisibile in due parti: una instabile (0.24-0.26 V) in cui l'RTD oscilla sempre, ed una stabile (0.26 - 0.31 V) in cui eventuali oscillazioni, originate da impulsi iniziali, si smorzano fino ad annullarsi.

4.6.1 Criterio di convergenza

Nel caso evolutivo, un criterio di convergenza si rende necessario quando si voglia analizzare il comportamento globale dell'RTD su di un'ampia scala di valori del potenziale, come ad esempio per il calcolo della caratteristica tensione-corrente; come già accennato nel paragrafo precedente, il diagramma $I(V)$ è identico sia che venga calcolato in maniera stazionaria sia evolutiva. Nonostante il considerevole aumento del numero di calcoli richiesto in quest'ultimo caso, il calcolo della caratteristica statica tramite simulazioni tempo dipendenti è un'indispensabile verifica per la dimostrazione tutt'altro che ovvia della validità del calcolo per stati stazionari. Dobbiamo quindi stabilire quando una simulazione tempo dipendente approdi ad uno stato stazionario. Dei due criteri dati in sez. 4.5.1 solo la (4.2) rimane valida e applicabile, in quanto la (4.1) perde di utilità poiché l'algoritmo Gummel, risolvendo esattamente l'equazione di Poisson ad ogni iterazione, annulla P_i durante tutto il corso della simulazione. Alla (4.2) però possiamo affiancare una nuova disuguaglianza che riguardi la densità di corrente, che adesso non è più uniforme lungo tutto il dispositivo. Diremo che una simulazione transitoria ha raggiunto uno stato stazionario quando la variazione δJ di corrente diventa sufficientemente piccola:

$$\delta J = (J_{max} - J_{min}) < 1000 \text{ A/cm}^2 \quad .$$

4.6.2 Efficienza computazionale

In questa sezione confronteremo i costi delle varie soluzioni adottate nell'analisi del comportamento dell'RTD: dalla più semplice soluzione stazionaria non autoconsistente, alla complessa e onerosa soluzione transitoria autoconsistente. Nelle considerazioni seguenti ci riferiremo a simulazioni protratte su un range di tensione che varia tra 0 e 0.4 V, suddiviso in 41 punti, con un incremento quindi di 0.01 V a punto. Il metro di misura sarà il numero di risoluzioni richieste del sistema Wigner-Poisson, per il calcolo di ogni singolo punto del diagramma. Nel caso stazionario non autoconsistente tale rapporto è ovviamente di 1:1; un calcolo stazionario autoconsistente (Newton) richiede in media dalle 50 alle 100 iterazioni; per contro il metodo non stazionario (Gummel) richiede ben 5000 iterazioni a punto (il tempo di simulazione è di 1 fs, mentre i tempi di assestamento del sistema sono dell'ordine di 5 ps). Come anticipato in sez. 3.5 il fattore che abbatte i tempi di calcolo nel metodo Newton stazionario, è il numero di termini cui si interrompe lo sviluppo Dixon (3.49): il numero fra parentesi di tabella (4.1) sarà relativo proprio all'indice superiore della sommatoria. L'ultima colonna riporta i tempi di calcolo, relativi ad un normale PC, che richiede circa 30 secondi per ogni iterazione di calcolo.

Tipo di simulazione	n. iterazioni richieste	CPU time (h)
Non auto-consistente	41	0.14
Stazionario autoconsistente (Newton) (1)	2200	7
Stazionario autoconsistente (Newton) (4)	800	2.5
Transitorio autoconsistente (Gummel)	20000	80

Tabella 4.1: Costi computazionali degli algoritmi risolutivi

4.6.3 Origine delle oscillazioni intrinseche dell'RTD

L'origine delle oscillazioni intrinseche negli RTD è stata molto dibattuta e tuttora non del tutto risolta: una prima spiegazione, proposta da Tsui e Cunningham [25], attribuiva tale comportamento alla bistabilità del dispositivo, ovvero la presenza del ciclo d'isteresi, congiunta con una sorta di meccanismo feedback a carico del potenziale elettrostatico. Anche altri autori, ed in particolare Buot e Jensen [20], ipotizzavano un'esistenza di tipo causa-effetto fra isteresi e oscillazioni auto-sostenute, confortati dall'osservazione che nella pratica sperimentale i due fenomeni quasi sempre coesistono. Viceversa, ricercatori come Sollner, per esempio, si pongono in una posizione di netto contrasto, affermando che al contrario le oscillazioni sono indotte nell'RTD dai circuiti esterni, comprese le probes di misura. Le recenti simulazioni circuitali riescono a gettare una luce maggiore sul problema, permettendo così di avanzare spiegazioni più esaustive e probabilmente fondate [20]. Dai risultati emersi nel corso delle nostre simulazioni, è possibile avanzare la seguente giustificazione circa il comportamento oscillante dell'RTD: postuliamo che le oscillazioni si originino a causa del passaggio del livello energetico discreto (localizzato nel pozzo quantistico e che indicheremo con QWL), nella regione proibita del catodo. Per regione proibita, non si deve intendere solo quella corrispondente al gap che separa banda di conduzione con quella di valenza: una regione proibita può anche crearsi *sopra* al minimo della banda di conduzione, in virtù della quantizzazione degli stati elettronici fra la regione del catodo e la prima barriera. Tale situazione è mostrata in fig.4.30/A nella quale si illustra come si possa formare un'ulteriore scatola quantistica, stavolta triangolare, tra catodo e zona centrale del pozzo quantistico (i cui livelli energetici discreti saranno indicati con EQWL). In ciò che segue spiegheremo come l'allineamento (e successivo disallineamento) fra QWL e gli EQWL,

sia la causa delle oscillazioni ad alta frequenza simulate (e sperimentalmente verificabili) dell'RTD. Una diminuzione del QWL, con conseguente passaggio nella regione proibita, provoca ovviamente una caduta della corrente che scorre nel dispositivo. Gli elettroni che successivamente giungono dal catodo, non riuscendo a oltrepassare la barriera, tenderanno a riflettersi indietro dando vita all'immane effetto di interferenza quantistica fra onda viaggiante e riflessa. E' noto che il risultato di tale fenomeno è la creazione di ripple nella densità di carica, con conseguente accumulo e deficit di elettroni. Questa nuova distribuzione degli elettroni tende, aumentando la regione su cui cade il potenziale, ad allargare la scatola triangolare, e quindi a formarvi una distribuzione continua di livelli accessibili al suo interno (fig.4.30/B). Il loro conseguente riallineamento col QWL, genera un nuovo aumento della corrente nel dispositivo, che svuotando la regione del catodo, e ci riporta nella situazione di partenza. Tali oscillazioni permangono quindi solo finché il QWL è al di sopra del minimo della banda di conduzione del catodo, il che si

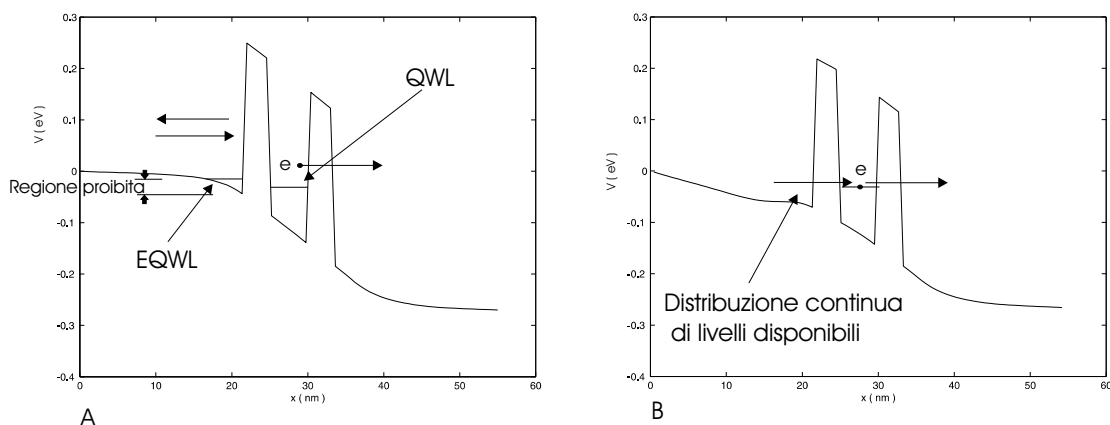


Figura 4.30: Giustificazione delle oscillazioni intrinseche. Nota: abbiamo amplificato la deformazione del potenziale, per motivi di esposizione, fra i punti di massimo e di minimo della corrente durante un ciclo d'oscillazione

verifica fino a una ddp 0.25 V, dopo di che, come già rimarcato, svaniscono.

4.7 Calcolo delle quasi-traiettorie

Per completare il quadro e l'analisi delle proprietà dell'RTD, riportiamo il calcolo delle quasi-traiettorie, eseguito in conformità a quanto descritto in sez. 3.6. L'utilità dei seguenti risultati è essenzialmente pratica e qualitativa: riusciremo a caratterizzare il moto degli elettroni tramite delle quasi-traiettorie, in maniera analoga a quanto avviene in ambito classico, rendendone così più intuitiva la comprensione. All'interno del semiconduttore verificheremo in-

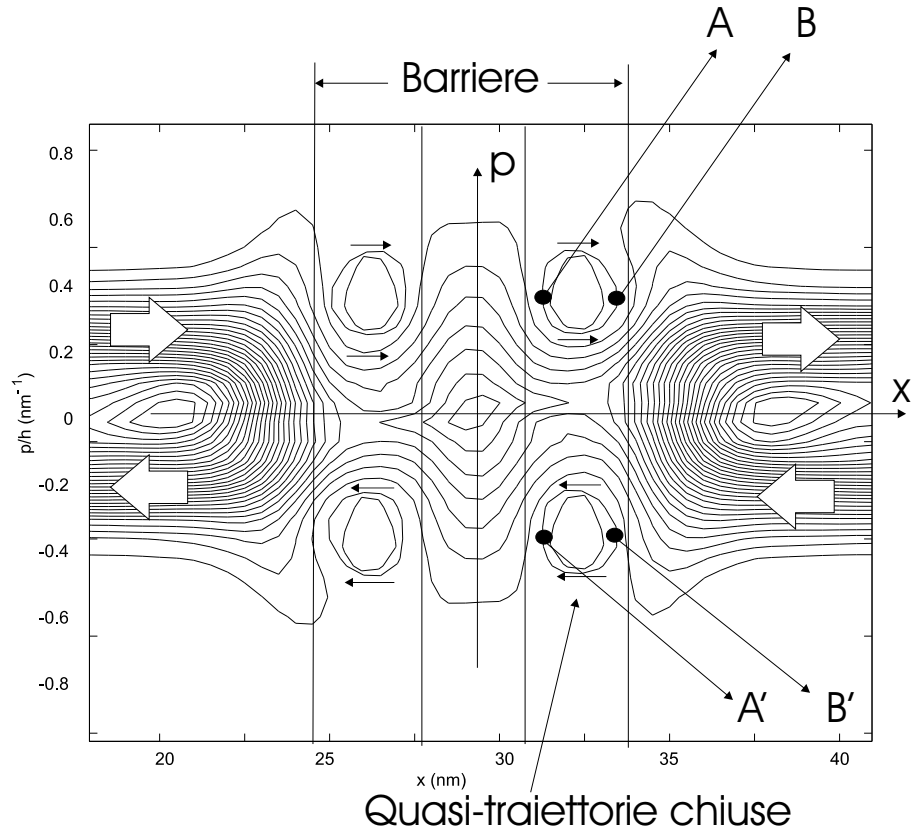


Figura 4.31: Traiettorie f_w per RTD non polarizzato con $T = 300$ K

fatti la presenza di traiettorie di “tipo classico”, ovvero conformi a quelle classiche, relative ad un profilo di potenziale avente le stesse caratteristiche di quello descritto; ma vi saranno anche traiettorie che definiremo “risonanti”, del tutto estranee al mondo classico, e che saranno appunto strettamente connesse col fenomeno del tunnel quantistico. Dall’osservazione delle quasi-traiettorie, sarà possibile sia prevedere il comportamento globale del dispositivo, sia trarre conclusioni sulla dinamica del tunnel. Iniziamo con la discussione del grafico delle traiettorie riportato in fig.4.31. Notiamo che gli stati a bassa energia sono riflessi prima di raggiungere il bordo esterno della barriera, mostrando così il carattere di non località degli effetti quantistici. Traiettorie chiuse esistono per basse quantità di moto, all’interno delle due barriere. Altre “quasi-traiettorie chiuse” si generano *al di sopra e al di sotto* dell’asse $p = 0$, (senza però intercettarlo), e corrispondono alla regione in cui f_w è negativa. Tali traiettorie, prettamente quantistiche, non tagliano l’asse $p = 0$ e sono simmetriche rispetto ad esso. Possono essere interpretate, solo se immaginiamo uno stretto legame di complementarità fra loro. Ovvero, se ammettiamo che la parte con $p > 0$ formi un tutt’uno con quella a $p < 0$, possiamo vederle come regioni in cui la “particella” percorra la traiettoria (nel senso della freccia) fino al suo limite estremo, e poi, istantaneamente, *cambi segno alla propria velocità* (nel disegno di fig. 4.31 ad esempio la particella salta alternativamente da A' a A e da B a B'); tutto avviene come se la particella fosse intrappolata fra due muri impenetrabili. Come anticipato in sez.1.11, i punti di riflessione (A, A') sono proprio quelli che annullano il denominatore della forza quantistica e che creano zone in cui le linee di flusso alternativamente nascono e terminano (rispettivamente sorgenti e pozzi)

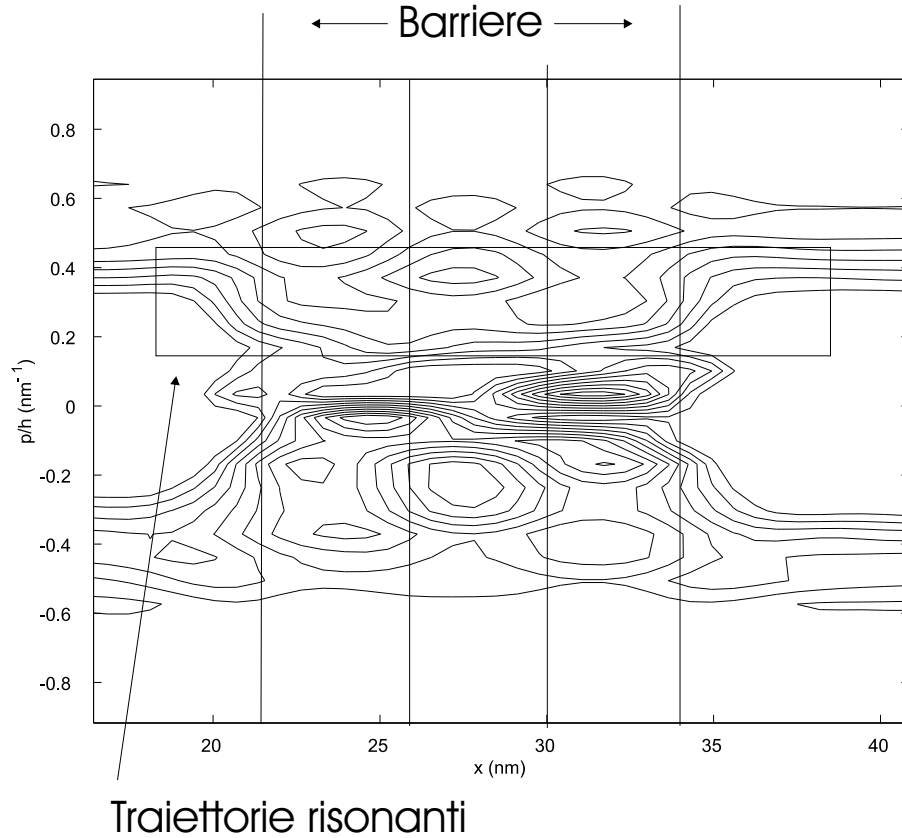


Figura 4.32: Traiettorie f_w per RTD con $V_{Bias} = 0.03$ Volt e $T = 300$ K

4.7.1 Traiettorie risonanti

Analizziamo adesso le traiettorie aperte che connettono anodo con catodo: come ci aspettiamo, sono rallentate all'interno delle barriere ed accelerate nel pozzo quantistico. Tali traiettorie, che definiremo “non risonanti”, “vedono” quindi due distinte barriere poste in sequenza e le attraversano rapidamente. Il loro tempo di permanenza all'interno del dispositivo, che qui riportiamo

$$t(x) = \int_{x_0}^x \frac{m^*}{p} dx$$

è plottato in fig. 4.35, in cui le curve sono parametrizzate rispetto alla quantità di moto iniziale della particella.

All'opposto le "traiettorie risonanti" (assenti nel diodo non polarizzato), spendono molto più tempo all'interno del pozzo quantistico, e vedono la struttura a doppia barriera nella sua "interezza". Le traiettorie risonanti si manifestano non appena si applichi una tensione di polarizzazione al diodo e sono mostrate in fig. 4.32: in contrasto con quanto avviene nella fig. 4.31, gli elettroni che le percorrono, non vengono accelerati quando attraversano il pozzo quantistico. E' interessante analizzare l'effetto dello scattering sulle traiettorie, che ci darà una conferma dell'utilità della precedente classificazione. La presenza di collisioni infatti, ci immaginiamo interferisca fortemente con le traiettorie risonanti poiché esse sono sia lente, sia legate ad una certa coerenza di moto (non attraversano le barriere in maniera sequenziale, ma le passano come fossero un tutt'uno). Le collisioni, che termalizzano la distribuzione, ed il cui raggio d'azione è dell'ordine del libero cammino medio (o equivalentemente del tempo di rilassamento), dovrebbero rompere tale coerenza, costringendo perciò gli elettroni all'attraversamento sequenziale (più rapido) delle due barriere, che abbiamo definito "non risonante"; quindi l'effetto globale dovrà essere quello di aumentare le traiettorie non risonanti a scapito delle altre. Questo è appunto ciò che si verifica (vedi fig.4.33 e fig.4.34) e spiega bene l'apparente paradosso menzionato in sez.4.4.1, ovvero che la collisioni diminuiscono i tempi di transito degli elettroni nel dispositivo nella regione NDR. Inoltre giustifica l'osservazione che le collisioni diminuiscono anche la quantità di carica intrappolata fra le barriere, in quanto aumentando le traiettorie veloci non risonanti, gli elettroni tendono a non soffermarsi dentro al pozzo quantistico. Come già osservato in sez.4.4.1 il tempo di switch ne risulterà diminuito, essendo soprattutto legato necessità di rimuovere la carica dal pozzo. In fig.4.36 riportiamo la forza quantistica sperimentata dalle particelle che seguono una traiettoria non risonante (fig.4.36/A) e una

risonante (fig.4.36/B): è evidente come in questo secondo caso gli elettroni

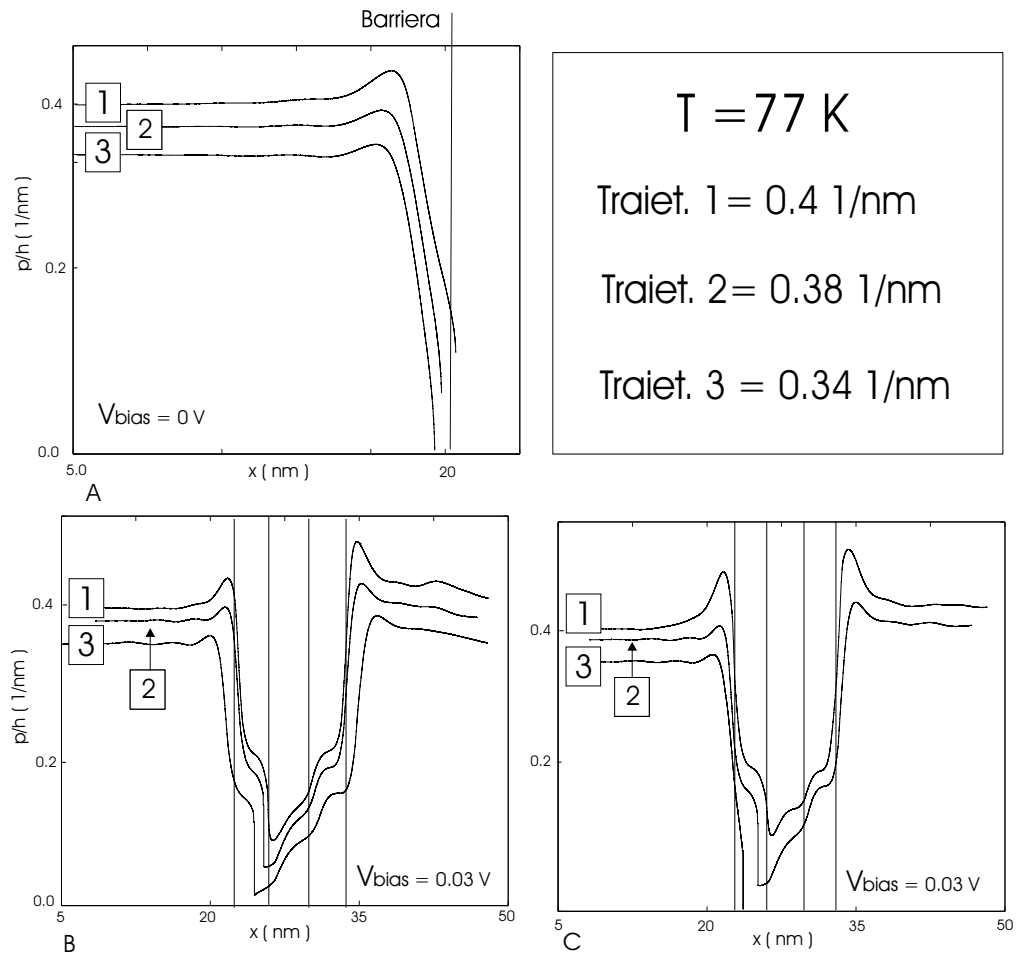


Figura 4.33: Riportiamo un'indagine più approfondita di alcune traiettorie significative per illustrare le conclusioni del testo. Si sono scelte tre traiettorie, numerate progressivamente, le cui quantità di moto iniziale (corrispondente alla velocità con cui la particella entra nel dispositivo) sono indicate nel riquadro in alto. In assenza di polarizzazione (fig.A) le tre traiettorie rimbalzano e non riescono ad attraversare le barriere. In fig.B sono mostrate le stesse traiettorie con una tensione di polarizzazione di 0.03 Volt, ma *senza* l'inclusione delle collisioni: come appare evidente diventano tutte traiettorie risonanti e attraversano il dispositivo. Infine in fig.C sono state incluse le collisioni che hanno degenerato le traiettorie, diminuendone la velocità, e permettendo a due sole di attraversare il dispositivo.

“sentano” esclusivamente una singola barriera e non una successine di due barriere.

Infine, poiché si è verificato che la differenza dell’energia cinetica fra il lato in

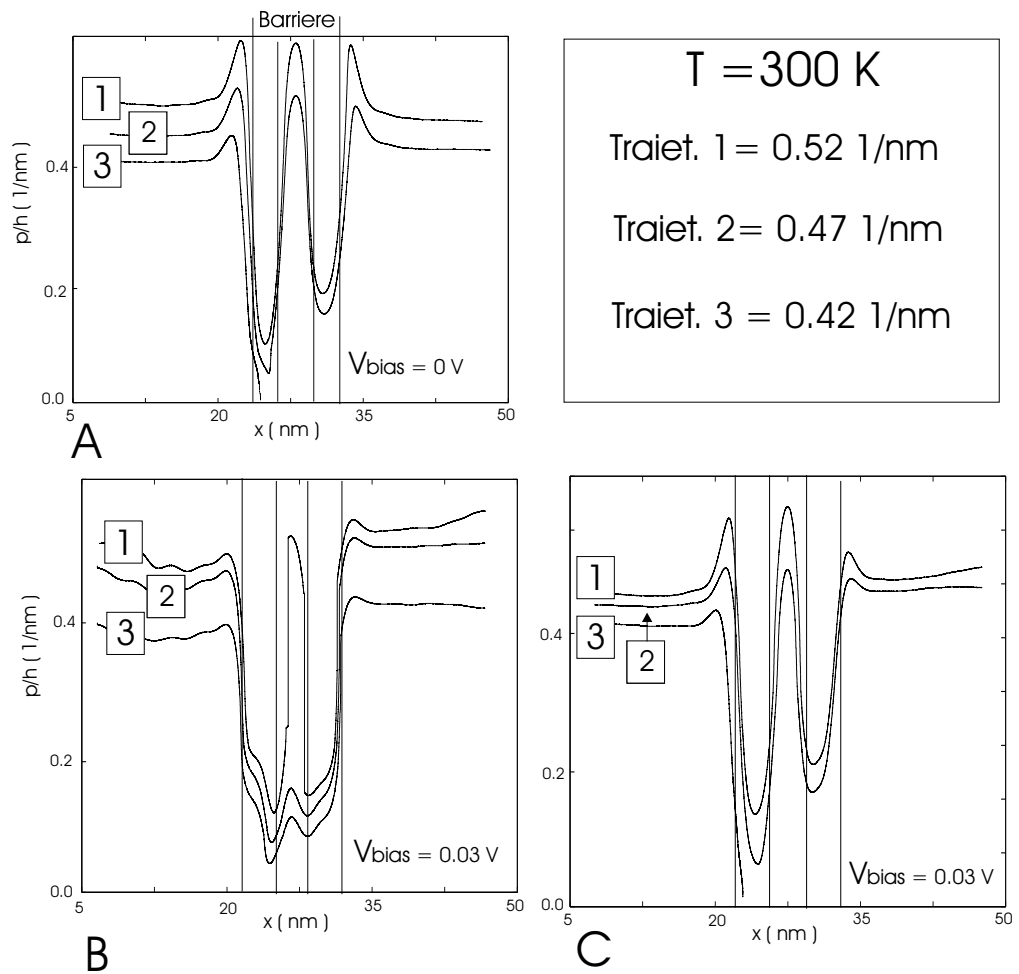


Figura 4.34: Ripetiamo le operazioni descritte in fig.4.33 ma adesso alla temperatura di 300 K: in fig.A appaiono chiaramente le traiettorie non risonanti, presenti anche in assenza di polarizzazione, grazie alla maggior dispersione termica della distribuzione di elettroni entranti nel dispositivo, rispetto al caso di 77 K. Applicando una tensione di 0.03 Volt, ma in assenza di urti, le traiettorie si modificano e diventano risonanti. L’inclusione degli urti, come rimarcato, interagisce con le traiettorie risonanti distruggendole e trasformandole in “non risonanti”.

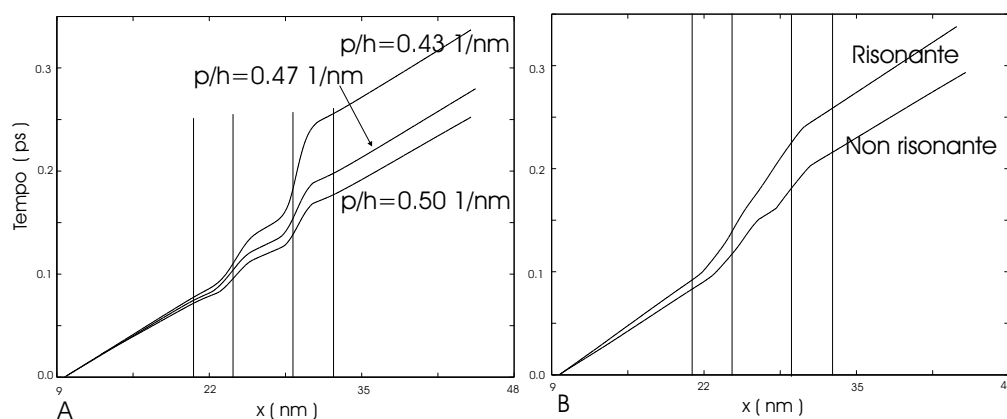


Figura 4.35: Riportiamo i tempi di attraversamento del dispositivo, parametrizzati rispetto alla quantità di moto iniziale dell'elettrone (fig.A) per l'RTD non polarizzato a 300 K. In fig.B confrontiamo i tempi di attraversamento fra le traiettorie risonanti e quelle non risonanti: per esempio sono state prese le traiettorie (1) e (2) di fig.4.34/A

cui le particelle entrano nella traiettoria e quello da cui ne escono è pari alla caduta di potenziale fra anodo e catodo, esse esibiscono un comportamento di “conservazione dell'energia”.

4.7.2 Conclusioni e sviluppi futuri

L'algoritmo di calcolo adottato nel presente lavoro si è dimostrato in grado di eseguire un'adeguata modellizzazione di dispositivi ad eterogiunzione “Lattice-matched”. Gli sviluppi futuri nell'ambito della produzione di diodi risonanti, o più in generale di dispositivi quantistici micrometrici, si prevede si orientino verso l'utilizzo di nuovi materiali, ed in particolare delle leghe $Si - Ge$. I motivi sono del tutto pratici e di natura economica in quanto la tecnologia al Silicio permette di ridurre sensibilmente i costi di produzione, rispetto a quella ad esempio che utilizza leghe di Arsenuro di Gallio. In tale ottica risulta particolarmente utile la possibilità di sviluppare attendibili cod-

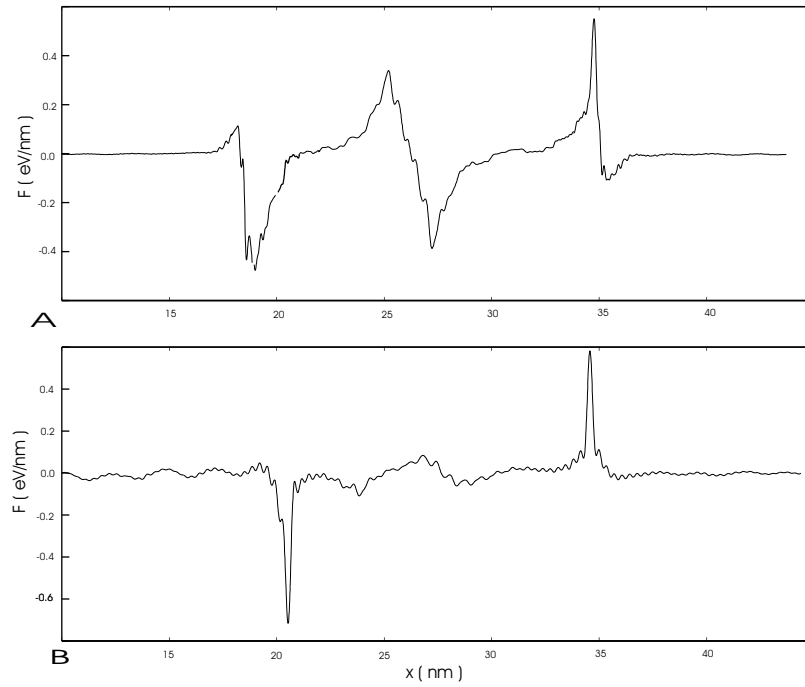


Figura 4.36: Forza quantistica sperimentata dalle traiettorie (1) e (2) di fig.4.34/B

ici di simulazione dei fenomeni di trasporto quantistico. Tuttavia gli attuali codici non sono ancora in grado di simulare dispositivi in tecnologia *Si*, poiché l'adattamento di passo reticolare fra *Si* e *Ge* non è banale e richiede sofisticate tecniche costruttive. Negli ultimi anni sono stati prodotti dispositivi che, allo scopo, utilizzano i cosiddetti “substrati virtuali” e strati “stressati”: tecniche in cui il reticolo cristallino si deforma, creando tensioni interne al materiale con un conseguente aumento del tasso di urti fra elettroni e reticolo. Le modifiche da apportare ai modelli matematici dei semiconduttori per descrivere tali fenomeni non sono ancora chiare ed in letteratura non è reperibile nessun esempio di simulazione di dispositivi su *Si-Ge*. Riguardo al presente lavoro, i risultati commentati nei capitoli 4 sfruttano i risultati un codice implementato in ambiente C++ dal sottoscritto, mentre i disegni,

salvo indicazione contraria, sono stati effettuati in ambiente MATLAB.

Capitolo 5

Appendici

5.1 Sviluppo in serie di potenze dell'operatore pseudo-differenziale

L'operatore pseudo-differenziale, introdotto in sez. (1.8) ed indicato in formula (1.30), che qui riportiamo

$$(\theta_{\hbar}[U])f_w(x, p, t) = \frac{i}{2\pi} \int \frac{U(x + \frac{\hbar}{2}\eta, t) - U(x - \frac{\hbar}{2}\eta, t)}{\hbar} f_w(x, p', t) \exp^{i(p-p')\eta} dp' d\eta \quad (5.1)$$

può essere espresso in forma differenziale. La sua azione sulla f_w è infatti uguale ad un opportuno sviluppo in serie delle derivate successive della f_w rispetto a p . L'utilità di tale sviluppo è sia teorica, perché permette di indagare alcune proprietà dell'equazione di Wigner come l'esistenza e l'unicità della soluzione, sia pratica, perché permette di definire la forza quantistica (1.44), indispensabile nel calcolo delle traiettorie, che è l'analogo della forza classica e vi si riconduce al limite $\hbar \rightarrow 0$.

Indicando con $r' = \eta\hbar$, sviluppiamo in serie la differenza che compare nella

(5.1):

$$U(x + r'/2) - U(x - r'/2) = \sum_{\lambda} \frac{1}{\lambda!} \frac{d^{\lambda} U(x)}{dx^{\lambda}} \left[\left(\frac{r'}{2} \right)^{\lambda} - \left(\frac{-r'}{2} \right)^{\lambda} \right] .$$

I termini pari si cancellano, mentre i dispari si sommano

$$U(x + r'/2) - U(x - r'/2) = 2 \sum_{\lambda \text{ disp.}} \frac{1}{\lambda!} \frac{d^{\lambda} U(x)}{dx^{\lambda}} \left(\frac{r'}{2} \right)^{\lambda} .$$

L'operatore pseudo-differenziale si scrive quindi

$$\begin{aligned} (\theta_{\hbar}[U]) f_w(x, p, t) = \\ \int dp' \frac{2}{2\pi} \frac{i}{\hbar^2} \int dr' e^{i(p-p')r'/\hbar} \sum_{\lambda \text{ disp.}} \frac{1}{\lambda!} \frac{d^{\lambda} U(x)}{dx^{\lambda}} \left(\frac{r'}{2} \right)^{\lambda} f_w(x, p', t) = \\ \frac{2}{2\pi} \frac{i}{\hbar^2} \sum_{\lambda \text{ disp.}} \frac{1}{\lambda!} \frac{d^{\lambda} U(x)}{dx^{\lambda}} \int dr' e^{ipr'/\hbar} \left(\frac{r'}{2} \right)^{\lambda} \int dp' f_w(x, p', t) e^{-ip'r'/\hbar} \end{aligned} \quad (5.2)$$

Focalizziamo l'attenzione sull'ultimo termine della formula precedente,

$$\left(\frac{r'}{2} \right)^{\lambda} \int dp' f_w(x, p', t) e^{-ip'r'/\hbar}$$

che è la trasformata di Fourier della f_w da p' in r' , moltiplicata per $\left(\frac{r'}{2}\right)^{\lambda}$; nella notazione usata nel capitolo 1, quest'ultima si scrive

$$\left(\frac{r'}{2} \right)^{\lambda} \int dp' f_w(x, p', t) e^{-ip'r'/\hbar} = \left(\frac{r'}{2} \right)^{\lambda} (\mathcal{F} f_w)(r')$$

cui possiamo applicare la nota formula

$$ir'(\mathcal{F} f)(r') = \left(\hbar \mathcal{F} \frac{\partial f}{\partial p'} \right) (r') .$$

Esplicitandola nella (5.2) si ottiene

$$-\frac{2i}{2\pi\hbar} \sum_{\lambda \text{ disp.}} \frac{1}{\lambda!} \left(\frac{\hbar}{2i} \right)^{\lambda} \frac{d^{\lambda} V(r)}{dr^{\lambda}} \int dp' \frac{\partial^{\lambda} f_w(x, p', t)}{\partial p^{\lambda}} \int dr' e^{-i(p-p')r'}$$

da cui, ricordando che

$$\int dr' e^{-i(p-p')r'} = \delta_{p,p'} 2\pi$$

segue

$$(\theta_{\hbar}[V])f_w(x, p, t) = -\frac{1}{\hbar} \sum_{\lambda \text{ disp.}} \frac{1}{\lambda!} \left(\frac{\hbar}{2i} \right)^{\lambda-1} \frac{d^{\lambda} V(r)}{dr^{\lambda}} \frac{\partial^{\lambda}}{\partial p^{\lambda}} f_w(x, p, t)$$

che è la forma differenziale dell'operatore pseudo-differenziale.

5.2 Operatore collisionale e valutazione di τ

Nella presente approssimazione a tempo di rilassamento dell'operatore di Boltzmann, l'inclusione delle collisioni nel moto degli elettroni, è descritta per mezzo di due sole grandezze: la funzione di equilibrio termodinamica, e il tempo di rilassamento τ . Dobbiamo precisare fin da subito che τ è un parametro che dipende unicamente dalla temperatura, a differenza con quanto avviene nella descrizione di Schrödinger, in cui i tempi di rilassamento dipendono anche dal numero d'onda k (o equivalentemente dalla quantità di moto p). Il motivo della maggior semplicità della formulazione di Wigner a dispetto della precedente, è da ricercarsi nella seguente considerazione: nella meccanica di Schrödinger, abbiamo già rimarcato in sez. 2.3.1, il punto di partenza di ogni calcolo pratico, rimane la simulazione del moto di uno stato puro, ovvero di singola particella; è chiaro quindi come il tempo medio fra due urti debba dipendere dalla velocità dell'onda considerata: onde veloci (con grandi k) avranno tempi di rilassamento più brevi di quelle lente. L'effetto complessivo delle collisioni infine sarà mediato, una volta nota la funzione di distribuzione delle particelle rispetto al numero d'onda. Nell'ambito della formulazione Wigneriana, ciò non è più applicabile, poiché uno dei

sui punti di forza è proprio quello di descrivere l'evoluzione di stati *misti*, ovvero dell'ensemble di elettroni nel suo complesso, e non della singola onda viaggiante. A riprova di ciò ricordiamo che la $f_w(x, p, t)$, diventando anche negativa, non è interpretabile come la probabilità che un elettrone al tempo t , si trovi nel punto x , con velocità p/m ; condizione quest'ultima indispensabile per poter sfruttare la funzione $\tau(p)$, il cui significato fisico è il tempo medio fra due collisioni di un elettrone che abbia quantità di moto p ; l'impiego di $\tau(p)$ richiederebbe quindi la conoscenza del numero di elettroni nel generico punto (x, p) *del piano delle fasi*, mentre dalla f_w possiamo ricavare solo il numero complessivo di elettroni nella coordinata x . L'approccio a tempo di rilassamento, offre quindi il vantaggio sia di semplicità implementativa, sia di poter includere differenti meccanismi di scattering in un solo parametro. La scelta dell'espressione appropriata da dare all'operatore collisionale nell'approssimazione a tempo di rilassamento, è basata sulla necessità di rispettare l'equazione di bilancio dettagliato. Da essa discende che, all'equilibrio, il termine collisionale si annulli; in altre parole, che le particelle che nell'unità di tempo giungono in un generico punto (x, p) , siano tante quante quelle che lo lasciano. Inoltre richiederemo che non venga violata la legge di conservazione della carica. La forma usuale

$$\left. \frac{\partial f}{\partial t} \right|_{coll} = \frac{f - f_0}{\tau} \quad (5.3)$$

deve essere modificata. Determiniamo la corretta approssimazione dell'operatore collisionale di Boltzmann (formula (1.9)) che qui di seguito riportiamo

$$\left. \frac{\partial f}{\partial t} \right|_{coll} = \int_{-\infty}^{\infty} dp' [W_{p \leftarrow p'} f(x, p', t) - W_{p' \leftarrow p} f(x, p, t)]$$

Si definisce

$$\tau(p) = \int_{-\infty}^{+\infty} dp' W_{p' \leftarrow p}$$

La condizione di bilancio dettagliato, e l'utilizzo di una costante di rilassamento τ , suggeriscono di identificare

$$W_{p \leftarrow p'} = \frac{f_0(p, x)}{\rho_0(x)\tau} \quad (5.4)$$

dove $\rho(x) = \int_{-\infty}^{+\infty} dp f(p, x)$. Vediamone il motivo alla luce del significato di W ; indichiamo

$$\left. \frac{\partial f}{\partial t} \right|_{coll}^{in} = \int_{-\infty}^{\infty} dp' [W_{p \leftarrow p'} f(x, p', t)]$$

Il secondo membro rappresenta infatti il numero di particelle che, nell'unità di tempo, cambiano la propria quantità di moto da p' in p . Lo si giustifica in quanto $W_{p \leftarrow p'}$, che per ipotesi non dipende da x , è la probabilità di transizione, nell'unità di tempo, fra lo stato quantistico p' e p ; così $W_{p \leftarrow p'}$ per la densità di particelle nel generico x , contiene il numero di particelle che, nell'unità di tempo, escono da un volumetto infinitesimo centrato in (x, p') e sono scatterate in un volumetto infinitesimo centrato in (x, p) . L'integrale $W_{p \leftarrow p'} f_w(x, p')$, su tutti i possibili stati p' , rappresenta quindi il flusso totale di particelle entranti in (x, p) nell'unità di tempo che, come detto, per soddisfare l'equazione di bilancio dettagliato dovrà essere uguale al numero di particelle che lasciano tale il punto stesso. Nell'ipotesi che la distribuzione non sia troppo dissimile da quella di equilibrio e che gli stati siano tutti equiprobabili, l'usuale approssimazione a tempo di rilassamento, identifica questa grandezza con $\frac{f_0(p, x)}{\tau}$. L'uguaglianza che ne deriverebbe $\left. \frac{\partial f}{\partial t} \right|_{coll}^{in} = \frac{f_0(p, x)}{\tau}$ porta alla classica approssimazione a tempo di rilassamento (5.3) che nel nostro caso non è del tutto soddisfacente infatti, come anticipato, vogliamo che la carica totale in un generico punto x non vari a causa delle collisioni: questo sia perché ipotizziamo che gli urti siano locali (ovvero portino solo una degenerazione della velocità degli elettroni senza che ne cambino istantaneamente la posizione) sia perché vogliamo tutelarci affinché la carica totale all'inter-

no del semiconduttore sia conservata; in altre parole cerchiamo una forma dell'operatore collisionale tale per cui

$$\int_0^L dx \int_{-\infty}^{\infty} dp \frac{\partial f}{\partial t} \Big|_{coll} = \int_0^L dx \frac{\partial \rho(x)}{\partial t} \Big|_{coll} = 0 \quad (5.5)$$

Con la (5.3) la precedente equazione non è in generale soddisfatta (lo sarebbe solo se il sistema fosse chiuso o illimitato). La soluzione che si rivela più indicata è rinormalizzare la funzione di equilibrio rispetto alla densità di portatori in ogni punto, ovvero identificare

$$\frac{\partial f}{\partial t} \Big|_{coll}^{in} = \frac{f_0(p, x) \rho(x)}{\tau \rho_0(x)}$$

che, unito all'ipotesi che la probabilità di transizione non dipenda dallo stato iniziale (stati equiprobabili), porta immediatamente alla (5.4). Quindi infine l'operatore collisionale nell'approssimazione a tempo di rilassamento si scrive

$$\frac{\partial f(x, p, t)}{\partial t} \Big|_{coll} = \mathcal{C}\{f(x, p, t)\} = \frac{1}{\rho_0(x)\tau} [\rho(x, t) f_0(x, p) - \rho_0(x, t) f(x, p, t)]$$

Si verifica immediatamente che così la (5.5) è sempre soddisfatta.

5.2.1 Valutazione numerica del tempo di rilassamento

I fononi (quantità di oscillazione del reticolo cristallino) si dividono in fononi acustici ed ottici, a seconda della legge di dispersione che seguono: ad essi sono connessi tempi medi di urto differenti. Nei semiconduttori inoltre sono presenti anche meccanismi di urto estranei ai precedenti, come ad esempio lo scattering con le impurezze, che hanno anch'essi tempi caratteristici propri. Allo scopo di inglobarli in un unico parametro τ , si sfrutta la regola empirica di Matthiessen: detti τ_i i tempi di rilassamento di un qualsiasi fenomeno collisionale, la regola di Matthiessen sancisce che una buona approssimazione

Meccanismi di scattering	Tempo di rilassamento (fs)
fononi acustici	$\tau_{fa} = 32/\sqrt{xt^3}$
fononi ottici	$\tau_{fo} = \frac{998}{\sqrt{xt[n_w\alpha + (n_w+1)\alpha]}}$
impurezze	$\tau_i = \frac{22282\sqrt{(x/t)^3(1+\alpha)}}{\alpha^2 \int_0^1 \exp(\alpha s)(1-s)(1+\alpha s)^{-1} ds}$
$x = \frac{E}{K_B T}$ $n_w = [\exp(E/K_B T) - 1]^{-1}$	$t = \frac{K_B T}{(1\text{eV})}$ $\alpha = (1 + Et/\hbar)^{1/2}$

Tabella 5.1: Tempi di rilassamento [22]

per il tempo di rilassamento di tutti i precedenti fenomeni congiunti, qualora siano meccanismi di scattering indipendenti fra loro è dato da

$$\frac{1}{\tau} = \sum_i \frac{1}{\tau_i}$$

In tab. 5.1 riportiamo i tempi tipici dei fenomeni collisionali determinanti nella dinamica all'interno del semiconduttore: come già detto tali tempi sono funzione dell'energia dei portatori nonché della temperatura. Così il

Temperatura (K)	Tempo di rilassamento (fs)
50	1322
77	525
100	321
200	129
300	97
400	85

Tabella 5.2: Tempi di rilassamento totali [22]

parametro τ_i da inserire nella formula precedente è calcolato come il valor medio su una popolazione distribuita secondo la distribuzione statistica canonica: il sistema fononico, possedendo una maggiore capacità termica, funziona da termostato per gli elettroni (infatti si parla di “bagno fononico”), a cui quindi possono essere applicate i risultati applicabili ai sistemi canonici

$$\tau_i = \frac{1}{\Gamma(5/2)} \int_0^\infty x^{3/2} \tau_i(x) \exp(-x) dx$$

in cui Γ è la funzione gamma di Eulero e $x = \frac{E}{k_B T}$. La tabella 5.2 riporta i risultati finali del valore di τ alle varie temperature.

Bibliografia

- [1] M. Toda, R. Kubo, N. Saito, *“Statistical physics I”*
- [2] A. Fasano, S. Marmi, *“Meccanica analitica”*
- [3] P. A. Markowitch, *“Semiconductor equation”*
- [4] J.J. Sakurai, *“Meccanica quantistica moderna”*
- [5] V. I. Tatarskii, Sov. Phys. Usp. 26,311
- [6] L. Demeio, J. Comp. of Phys. 99,203
- [7] C. Jacoboni, Phys. Rew. B 58,3503
- [8] H.W. Lee, Found. of Phys. 13,61
- [9] D. K. Ferry, Phys. Rew. B 67,7720
- [10] W. R. Frensley, Rew. of Mod. Phys. 62,745
- [11] N. W. Ashcroft, *“Solid state physics”*
- [12] W. R. Frensley, Phys. Rew. B 36,1570
- [13] F. A. Buot, K. L. Jensen, J. Appl. Phys. 65,5248
- [14] F. A. Buot, K. L. Jensen, IEEE Tran. on Elect. Dev. 38,2337

-
- [15] P. Zhao, H. L. Liang, IEEE Tran. on Elect. Dev. 48,614
 - [16] R. W. Dixon, Appl. Phys. Lett. 31,354
 - [17] D. L. Woolard, Appl. Phys. Lett. 46,158
 - [18] W. R. Frensley, Appl. Phys. Let. 51,448
 - [19] B. Biegel, J. Plummer, Phys. Rev. B 54,8070
 - [20] F. A. Buot, K. L. Jensen, Phys. Rev. B 42,9429
 - [21] D. Coleman, Appl. Phys. Let. 67,7602
 - [22] W. Walukiewitc, J. App. Phys. 50,899
 - [23] F. A. Buot, K. L. Jensen, J. Appl. Phys. 67,7602
 - [24] B. Ricco, Phys. Rev. B 29,1970
 - [25] D. C. Tsui, Phys. Rev. Lett. 58,1256