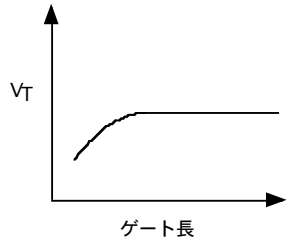
第 5 回　MOSFET　スケーリング

いままでの話ではチャネルを短くしたほうが、また電圧を上げたほうが早く動く事になっていた。しかし実際はそれほど簡単ではない。今日は短チャネル効果と呼ばれるチャネルを短くしたり、電圧を高くかけすぎたときに起こる問題点を述べたいと思う。

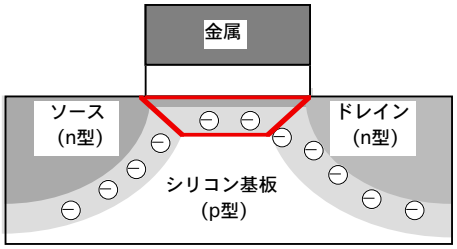
リソグラフィによるサイズ縮小が、MOS-トランジスタの微細化に伴い、短チャネル効果を健在化させてきた。MOS トランジスタの歴史は、短チャネル効果との戦いの歴史でもある。

短チャネル効果

短チャネル効果はゲート長が短くなると、しきい値が小さくなる現象である。しかし本来同じゲート長なら、しきい値電圧変動は回路全体で変わるだけで問題ないはずだが、作成時にゲート長には 3σ でだが 10%程度の誤差が生じる。正規分布だとすれば、確率的には 3σ 内に 99.7%で入るが、トランジスタが 300 万個あれば、1 万個が 3σ 以上でばらつく。ということは、この誤差でもしきい値電圧がバラつくと論理動作しなくなってしまう。したがってゲート長の変化に対してしきい値は変化しないことが好ましい。



前回話した様にしきい値は、仕事関数などの物質によって決まるフラットバンド電圧を除くと、フェルミ準位と真性フェルミ準位との差(真性フェルミ準位からのフェルミポテンシャル)の 2 倍の項と空乏層中のアクセプタの電荷がだす電気力線による項の二つで決まる。ここで、暗黙のうちに、空乏層からの電気力線はまっすぐ垂直に向かうとしていた。しかしながら、もし、その近傍にドレイン/ソースがあったらどうなるか。ドレイン・ソースも空乏層を持っており、その両者は結合すると区別できず、近いほうを選ぶ。ここでこの減少を記述しようとしてチャージシェアモデル(Yu, 1974)がよく用いられる。



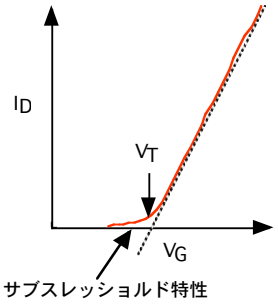
チャネル長を短くすると、絶縁体を通過する電気力線の密度(=電界)がへり、しきい値電圧が下がる。上の絵では本来長方形になるはずが、台形になってしまうことによる。　台形になったことによる面積の変化を小さくするためには、ゲート長に対して空乏層厚が小さければよい。通常 1/2 から 1/3 程度とする。空乏層厚を薄くするためには基板のキャリヤ濃度を上げるべきである。(定量的な計算を行うと酸化膜厚も薄くすることが要求される。)

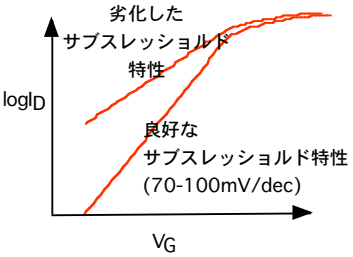
空乏層を薄くすることで次の弊害が出てくる。

チャネルはゲート電位のみでなく、基板電位とも容量結合をすることから、空乏層厚をいままでより薄くすると、基板チャネル間容量が大きくなるに従い、ゲートチャネル間の容量も大きくさせねばならない。するとゲートによる制御の効きが悪くなる。空乏層厚が薄くなると、前回定義したボディ効果係数 m が 1 に較べて大きくなり、基板との容量結合がチャネルのゲートとの容量結合が較べて無視できなくなり、制御性の低下がすると言っても良い。

この制御性の低下の中でもっとも嫌われているのは、サブスレッショルド特性の劣化である。(しきい値の基板電圧依存性も大きくなるが、これはアナログ回路では特に重要であり考慮すべきである。)

サブスレッショルド特性とは、オーバードライブ電圧が負の時に流れる電流のことであり、これはキャリヤの数がアクセプタと等しいことをしきい値の定義としたが、指数関数的に少なくなるものの、しきい値に達する前からキャリヤが存在することによる。





しきい値電圧以下での電流であるサブスレッショルド電流はゲート電圧に対して指数関数的に変化するが、指数関数の肩に　1/m の係数がかかっている。サブスレッショルド特性が劣化すると、オフ時の消費電力があがり、CMOS 最大の利点が消えていく。(しきい値を大きくすれば消費電力は下がるが、駆動能力が下がり速度は落ちる。)　従って酸化膜を薄くして m を小さく保ちたい。せいぜい 1.5 程度までに抑える。

(なお、サブスレッショルド特性は、空乏層厚が広い場合の短チャネル時にも、ドレイン電流が大きくなり劣化する。これはドレイン誘起障壁低下 Drain Induced Barrier Lowering と呼ばれる。)

結果として酸化膜厚を薄くすると、耐圧に問題が出てくる。歩留まりと信頼性を向上させるためには、最大でも 5MV/cm 以下に抑える必要がある。何故なら 10MV/cm 付近から絶縁破壊を起こすからである。

定電界スケーリング

この短チャンネル効果効果の抑制には、空乏層を薄くし(即ちキャリヤ濃度を上げる)、ゲート酸化膜も薄くし、またドレインやソース領域も薄くすることで避けられる。また、電界を一定に保つために、電圧も下げることが重要である。

このように横方向サイズに併せて、電圧・縦方向サイズを同時に小さくする(空乏層を薄くするためにキャリヤ濃度は上げる)概念をスケーリングと呼ぶ。即ちゲートを短くした分、他も小さくする。

理論的にきちんと提唱され、判りやすいのは定電界スケーリング(Dennard, 1974)である。即ちゲートサイズが 1/k に小さくなっても、内部電界が一定になるようにするため、電圧も 1/k に下げる。縦方向のゲート酸化膜も、空乏層厚も 1/k と薄くする。また電流は、ゲート長とゲート幅がキャンセル、電圧縮小で $1/k^2$ に、酸化膜縮小で k 倍になって全体で 1/k 倍になる。容量はゲート長の縮小と酸化膜厚の縮小がキャンセルして、ゲート幅分小さくなる。回路遅延は容量と電圧の積を電流で割ると出るので、容量 1/k 倍、電圧 1/k 倍、電流 1/k 倍で、遅延も 1/k になる。回路当たりの電力消費は、$1/k^2$ になる。面積当たりの集積度は k^2 なるので、面積当たりの消費電力はあがらない。ここで考えている空乏層厚は、ドレイン電圧に対する空乏層厚で、内蔵電位に較べて大きなドレイン電圧で決まることから、キャリヤ濃度は k 倍にすれば良いことに注意しよう。

定電界スケーリングは他にもメリットがある。例えば、高いドレイン電圧により広がるピンチオフ点・ドレイン間の空乏層の広がりによるチャネル長変調効果の悪化、高い電界により起こる移動度一定条件から離れた場合（速度飽和）の特性変化等も同時に解決できることから、非常に重要な指針となった。

チャネル長変調効果

ここで、横方向の電界の話を考えよう。

ドレインーソース間に電圧がかかる時に、ピンチオフ点はチャネル/ドレイン境界であるという前提でいままで話を進めてきたが、実際には電圧をかけると空乏層は広がり、チャネル長は短くなる。これはチャネル長変調効果と呼ばれ飽和後もゆっくり電流があがる形となる。

ピンチオフ点からドレインまでの距離を ΔL とすれば、電流は

$$I_D = \frac{\mu W \varepsilon_{ox}}{L - \Delta L} \frac{(V_G - V_{T0})^2}{2m}$$

となる。等価回路向けの一番簡単な近似として、

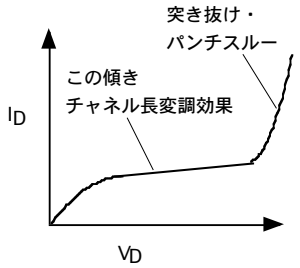
ΔLが、V_{DS}に比例して$\Delta L = L\lambda V_{DS}$と表されるとすると

$$I_D = \frac{\mu W \varepsilon_{ox}}{L(1-\lambda V_{DS})}\frac{(V_G - V_{T0})^2}{2m}$$

となる。

$$\cong \frac{\mu W \varepsilon_{ox}}{L}\frac{(V_G - V_{T0})^2}{2m}(1+\lambda V_{DS})$$

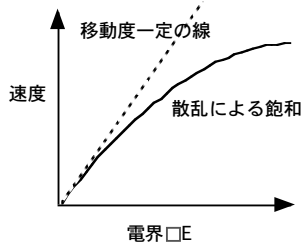
さてもっと電圧をあげると、ゲート下はすべて空乏層になりゲート電圧に関係なく電流が流れる様になる。この状態をバイポーラのベース層が空乏化して電流が制限されなくなったときと同じ、突抜け（パンチスルー）と呼んでいる。



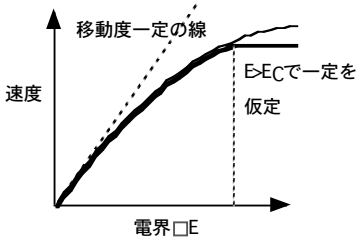
速度飽和

さて、第二回でも述べたが、移動度が一定であるという前提は、ピンチオフ点近傍では成り立たないことがある。

電子がある状態で電界を徐々に大きくしていくと、始め電界に比例して速度は大きくなるが、速度が大きくなりすぎると、一個あたりの電子のエネルギーが大きくなり、光学フォノンをはき出す。結果として速度は頭うちになる。(飽和速度と呼ばれている。)



さて、ゲート電圧一定でゲート長だけを短くしたとする。するとチャネルの電界はどんどん強くなり、移動度がなりたたない領域に入っていき、電荷はほぼ飽和速度に近い状態で走るようになる。また解析手法として、ある一定電界 Ec 以上では飽和速度で一定になるとすると高い電圧をゲート電圧にかけた状態での飽和状態では、E_c×チャネル長程度の電位がチャネルにかかり、残りの電圧は今まで通り最後の薄い空乏層層の部分にかかっているとできる。



このときチャネル下での電荷は$\frac{\varepsilon_{ox}}{d}(V_G-V_T-V(x))$であり、V(x)の範囲は$0<V(x)<E_cL$である。このとき、

$V_G-V_T>>E_cL>V(x)$として考えると、$\frac{\varepsilon_{ox}}{d}(V_G-V_T-V(x))\approx\frac{\varepsilon_{ox}}{d}(V_G-V_T)$となる。この電荷が飽和速度$v_{sat}$で走行するので、電流は$I_D=W\frac{\varepsilon_{ox}v_{sat}}{d}(V_G-V_T)$になり、二乗に比例しない。今までとの違いは、チャネル終端でも、V_G-V_Tとして 0 でない値が存在しうることである。このあたりは、前の説明で$V_G-V_T=0$となって、かつ速度が無限大としていた扱いでの問題点であり、飽和速度の概念を入れるとその辺りが理解できなく成るくらいややこしく成るので逃げている。

さて、チャネルを短くして、速度飽和がおこるようにすると、ゲート電圧に対して電流の増幅が比例するように変わった。また、電流がチャネル長に依存しなくなった。またはg_mが一定という表現でも良い。

デジタル回路での充電時間はこの場合、L の二乗で速度があがる有利さを失う。

但し後に述べる様にゲート長が縮小された MOS では、強い縦電界を持っていて、電子は界面に押さえつけられ移動度が通常の場合に比べて相当下がる。この低下の効果で、実は速度飽和は起きにくい。しかしながら、180nm 世代(実際のゲート長は 120nm)より後では、オーバードライブ電圧の 2 乗に電流が比例しなくなっているので、やはり定電界でないことから起こった飽和速度の関係と見て良いと思う。

実際のスケーリング

この様に定電界スケーリングは非常に判りやすく、問題を解決しうる方法である。しかし、実際に行われてきたことは、素子特性向上の為に、ゲート長短縮→縦方向を小さく(ここまでを定電圧スケーリングと呼ぶ)→耐えなくなって電圧縮小という流れであり、理想的な定電界スケーリングは実際には行われていない。例えば素子縮小に伴い電圧を下げることを回路上の要請から避けてきて、電圧 5v 時代が定電圧スケーリングで 15 年以上続き、90 年頃からついに堪らず 3.3V へ移行した。そこから縮小に併せて電圧が下がり始めた。最近は 1V 程度となっている。但し今でも実際今でも厳密な電界一定ではなく、電界はサイズ縮小と共に少しずつ上がり続けている。

きちんと定電界スケーリングするときは、電圧も長さに比例して下がる。したがって、チャネル長さに比例して速度は早くなる。集積度はスケーリングの 2 乗に比例してあがり、(同クロック速度での一回路あたりの)消費電力は、スケーリングの二乗に比例して下がる。

サイズ(um)	2	1.2	0.8	0.5	0.35	0.25	0.1
電圧(V)	5	5	5	3.3	3.3	2.5	1.5
ゲート膜厚(nm)	35	25	18	12	10	7	3
ゲート酸化膜中電界(MV/cm)	1.4	2.0	2.8	2.8	3.3	3.6	5

Y.Taur & T. H. Ning "Fundamental of Modern VLSI Devices", Cambridge, 2009 から

西暦	99	01	03	05	07	09
配線半周期 DRAM Half pitch (nm)	210	150	100	80	68	52
ゲート長(nm)	90	65	45	32	32	29

ITRS 及び平本による

これから先は ITRS と言う団体で討議して 2024 年まではどのような値を取るべきかが予測されている。(興味があれば、最新英語版を http://www.itrs.net/で、日本語訳版(少し古い)は http://strj-jeita.elisasp.net/strj/でみれる。)

また、最近 65nm プロセスと言っても、ゲート長は ITRS に沿うが、DRAM half-pitch は達成できていないものが殆どである。

スケーリングで解決できないこと

残念ながら（たとえ定電界スケーリングでも）スケーリングは万能ではない。

熱によって決まるサブスレッショルド特性は、いくら m が小さくても、ある電圧変化が無いと電流レベルが下げられず、最終的には温度をスケーリングしない限り解決できない。

バンドギャップが変化しないことは、内蔵電位の非スケーリング性をもたらす。物質固有の固溶限等も変わらない。

ゲートのリーク電流はトンネルなので厚さによる変化は薄くした最後は指数関数的であり、ここから高誘電率の新しいゲート材料を使う試みが起こった。

また回路が大きくなり、充電時間が配線容量によって決まり、配線容量が一定(そうなることが多い)ならば、定電界スケーリングでは速度が変わらなくなる。それでも、充電する相手が同じく比例して小さくなれば、早くなる。

従って単純なスケーリングだけでは解決できない問題も多い。