

第13回 MOSFET スケーリング

いままでの話ではチャン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。今日は短チャンネル効果と呼ばれるチャンネルを短くしたり、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、短チャンネル効果を健在化させてきた。MOS トランジスタの歴史は、短チャンネル効果との戦いの歴史でもある。

ただ、その前に言い忘れていたチャンネル長変調効果の話をする。

チャンネル長変調効果

ここで、横方向の電界の話を考えよう。

ドレインソース間に電圧がかかる時に、ピンチオフ点はチャンネル/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、チャンネル長は短くなる。これはチャン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 $I_D = \frac{\mu W \epsilon_{ox}}{L - \Delta L} \frac{(V_G - V_{T0})^2}{2m}$ となる。等価回路向け

の一番簡単な近似として、チャンネル長変調パラメータ λ を用いて ΔL が V_{DS} に比例して $\Delta L = L \lambda V_{DS}$ と表される

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

とすると $\cong \frac{\mu W \epsilon_{ox}}{L} \frac{(V_G - V_{T0})^2}{2m} (1 + \lambda V_{DS})$ となる。

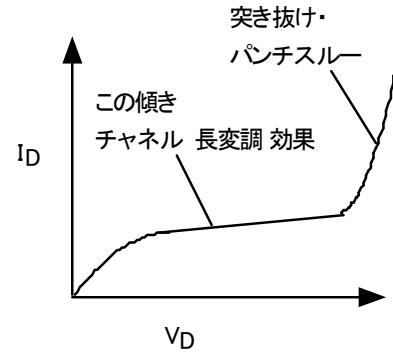
$$\cong \frac{\mu W \epsilon_{ox}}{L} \frac{(V_G - V_{T0})^2}{2m} (1 + \frac{V_{DS}}{V_A})$$

ここで V_A は第三回のバイポーラトランジスタで話したアーリー電圧と同じものである。ここで出力コンダクタ

ンス $g_{ds} = \frac{\partial I_D}{\partial V_{DS}}$ を定義すると、 $g_{ds} = \frac{I_{D0}}{V_A}$ とも表せる。こ

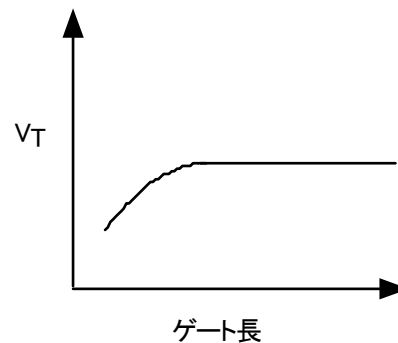
こで I_{D0} はチャンネル長変調効果を無視したときのドレイン電流とした。 ΔL は V_{DS} の関数であるがゲート長 L の関数では（通常）ないので、 L を変化させると λ は反比例（ V_A は比例）する。出力コンダクタンスは電圧利得を制限し、アナログ回路応用時に重要な指標である。

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ようになる。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。

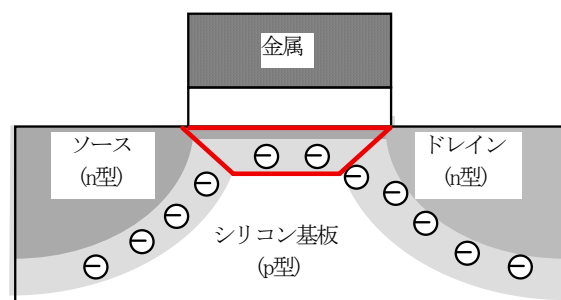


短チャンネル効果

短チャン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。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 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。全部で数十億個あるので、この誤差でもしきい値電圧がバラつく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ア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。)

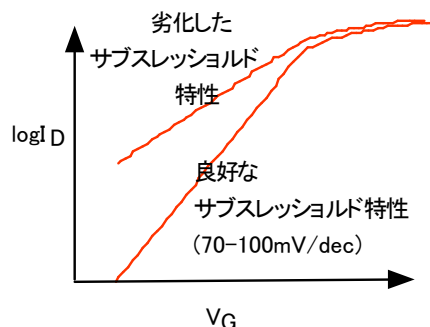
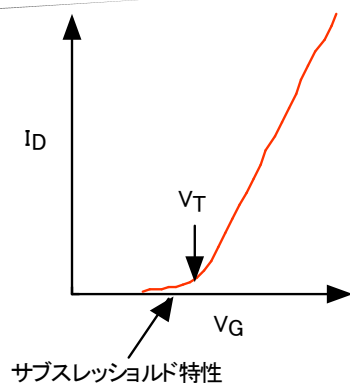
空乏層を薄くすることで次の弊害が出てくる。

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最近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、これはキャリア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アが存在することによる。

これがなぜ問題かといえ、オフ電流を決めるからである。すなわち、CMOS 回路は理想的にオフ時に電流が流れることがないので、消費電力を大きく下げた。ところがサブスレッショルド特性が劣化すると、オフ時(ドレインに対してゲート電圧がゼロ)でも CMOSFET のどこかには電源電圧がかかっている、リーク電流が無視できない。そのリーク電流が消費電力を決める。



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

この短チャネル効果効果の抑制には、空乏層を薄くし(即ちキャリア濃度を上げる)、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ア濃度は上げる)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 $1/k$ に小さくなくても、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、キャリア濃度は k 倍にすれば良いことに注意しよう。

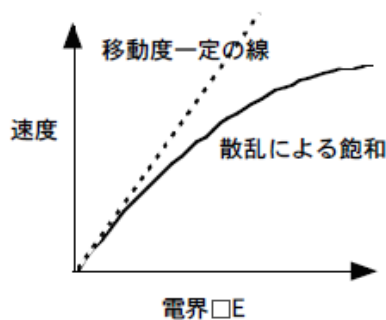
定電界スケーリングは他にもメリットがある。例えば、高いドレイン電圧により広がるピンチオフ点・ドレイン間の空乏層の広がりによるチャネル長変調効果の悪化、高い電界により起こる移動度一定条件から離れた場合（速度飽和）の特性変化等も同時に解決できることから、非常に重要な指針となった。

このスケーリングの概念がサイズ縮小の基本概念であり、そのためにプロセスルール（専門家はテクノロジーノードと呼ぶ）は小さくなり続けている。

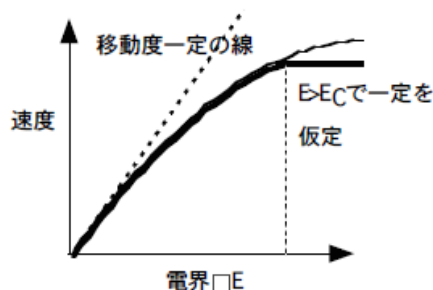
速度飽和

さて、第9回で述べた移動度が一定で電界が速度に比例するという前提は、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く、電荷はほぼ飽和速度に近い状態で走ようになる。また解析手法として、ある一定電界 E_c 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、 $E_c \times$ 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 $\frac{\epsilon_{ox}}{d}(V_G - V_T - V(x))$ であり、 $V(x)$ の範囲は $0 < V(x) < E_c L$ である。このとき、 $V_G - V_T \gg E_c L > V(x)$ として考えると、

$\frac{\epsilon_{ox}}{d}(V_G - V_T - V(x)) \approx \frac{\epsilon_{ox}}{d}(V_G - V_T)$ となる。この電荷が飽和

速度 v_{sat} で走行するので、電流は $I_D = W \frac{\epsilon_{ox} V_{sat}}{d} (V_G - V_T)$ 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 $V_G - V_T = 0$ となつて、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ます。

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なつた。また、電流がチャネル長に依存しなくなつた。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（それを緩和するために歪シリコンが導入された）。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、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（ここまでは定電圧スケーリングと呼ぶ）→耐えなくなつ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（0.5 μ m）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、0.7V 程度となつて止まっている。

（これはサブスレッショルド特性から決まるオフ電流増大の抑制のためである。ある程度のしきい値がないと消費電力が上がってしまう。電源電圧に対して 3 割等のルールよりこちらが優先される。消費電力に厳しい回路の方がしきい値を大きくとるために、電源電圧を大きくしている。）

下に年代とノード、Halfpitch、ゲート長と合わせて電源電圧の推移を示す。なお、電極間周期の半分 (Halfpitch) がもともとは DRAM の微細化が半導体プロセスを推進した時代に使われていたテクノロジーノードの指標となっていた値だが、ゲート長がそれより短い時代にノードと乖離した。さらに電源電圧が下げられないこともあり、ゲート長は最近縮小できなくなり、いまはテクノロジーノードが何を指しているか、誰も説明できなくなっている。

年	1979	1992	1995	1997	1999	2001	2003	2005	2007	2009	2018
ノード(nm)	2000	500	350	250	180	130	100	65	45	32	14
Halfpitch(nm)	2000	500	350	250	230	150	100	90	68	52	30
ゲート長(nm)	2000	500	350	200	140	65	45	32	38	29	24
電源電圧(V)	5	5	3.3	2.5	1.8	1.2	1.1	1.1	1	0.75	0.7

<https://en.wikichip.org/> および J. Image, Graphics and Signal Processing, 2017, 12, 23-29 などによる

スケーリングで解決できないこと・問題解決法

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(High-k)を使われるようになった。

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。

また、ゲート下によるしきい値変動の問題をなくすために、マルチゲート構造は finFET などですでに導入され、また高移動度材料なども試みられている。

さらに、サブスレッショルド特性向上には急峻性を改善するためにトンネル FET や負性容量 FET などが試みられている。

なお、時間がたりず、DRAM, フラッシュメモリ, 光受光器などは説明ができなかったが、昔の講義資料から補遺を作り、OCW に上げたので、参考にされたい。また、メモリについてはマイクロンからの講師を招いての講義があるので、それにも興味を持ってほしい。

それ以外に昔の講義でカバーしていた化合物半導体電子デバイスについては、僕の大学院講義でカバーできることから、そちらを参照されたい。