メモリの大革命
3次元NANDフラッシュ

厚木エレクトロニクス / 加藤　俊夫

1.はじめに ～メモリの全般状況～

本レポートは、3次元NANDフラッシュ･メモリ（以下、3D-NANDフラッシュ）について詳しく説明するのが目的であるが、メモリに詳しくない方のために、まず最初に半導体メモリ全般について簡単に述べておく。

現在、半導体メモリといえば、DRAM（Dynamic Random Access Memory）、SRAM（Static Random Access Memory）、フラッシュの3種類が主なタイプで、それ以外にFeRAM（Ferro-electric RAM）などもあるが、まだ主流製品ではないので、ここでは省略する。

1. DRAM

DRAMの1ビットは図1のように、MOSトランジスタとキャパシタで構成されている。MOSは、ビット線とワード線がアクセスされた時に導通してON状態になるスイッチの役目を行っている。その時、キャパシタに電荷が貯まっているかどうかでONかOFFと検知される。MOSが導通していない時もキャパシタは完全には絶縁されていないので、電荷が逃げていく。そこで、電荷が失われる前に再度書き込む必要がある。これをリフレッシュと呼んでおり、1秒間の数十～数百回も行う必要があることから、Dynamic RAMと呼んでいる。キャパシタの容量が大きければ、頻繁にリフレッシュする必要がなくなるが、1ビットの面積を小さくして集積度を上げるためには、キャパシタの面積をできるだけ小さくする必要があり、トレンチを掘って側面を利用するトレンチ型や、基板の上部にキャパシタを重ねるスタック型が用いられ、さらに誘電率の高い材料も用いられて、プロセスが複雑になってきている。

2. SRAM

SRAMは、図2のようにフリップフロップ回路である。フリップフロップというのは、ぎったんばったんと動くシーソーのことで、図のMOS･Q1がONになれば、MOS･Q2がOFFになり、Q1がOFFなら、Q2がONになる。このように、どちらが「ぎったん」で、どちらが「ばったん」かによってON-OFFを決めるメモリである。動作速度が速く安定した動作が期待されるので、キャッシュメモリとして良く用いられている。最近のロジック系のLSIでは、チップ面積の半分ぐらいがSRAMで占められている場合があるようである。SRAMの問題点は、MOS6個で1ビットであるから、ビットあたりの面積が大きく、集積度を問題にする用途には向いていないという点である（MOS4個と抵抗2個からなるSRAMもある）。

以上、DRAMとSRAMの説明で分かるように、どちらも電源が繋がって動作している場合はメモリの状態が保たれるが、電源を切るとメモリが消えてしまう。これを揮発性と呼んでいる。

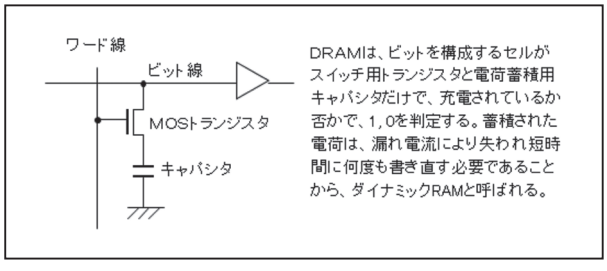


図1　DRAMのビット構成

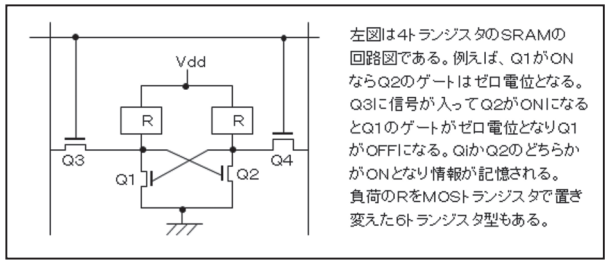


図2　SRAMのビット構成

3. フラッシュ・メモリ

フラッシュ・メモリは、現在、おもに用いられている構造は図3のようなフローティング・ゲート（Floating Gate＝浮遊ゲート、以下、FG）型である。通常のMOSのゲート電極とSi基板（チャンネル）との間に、どこにも繋がっていないFGがあり、このFGに電荷を貯めると、周りがSiO_2の絶縁物なので電荷が逃げる心配がない。電源を切ってもメモリが消えないので、不揮発性メモリと呼ばれている（Non Volatile＝以下、NV）。FGに電荷を貯める方法は、Si基板とFGの間は非常に薄いSiO_2膜（数nm以下）なので、高電圧を掛けてトンネル効果で電子を移動させる。電荷を引き抜く時も同様である。絶縁物なのに電流が流れるのは不思議であるが、ファウラー・ノルドハイム・トンネリング（Fowler-Nordheim Tunneling）と呼ばれる量子トンネル効果を利用しているからである。

フラッシュ・メモリには、NOR型とNAND型があるが、本稿では省略する。NORとNANDの言葉自体についてご存じない方は、ぜひデジタル回路の基礎を勉強していただきたい。

2.NANDフラッシュ・メモリ

1. NANDフラッシュの構造と特徴

まず、NANDフラッシュのビット構成を図4に示す。ワード線が隣接するMOSでソースとドレインを共有され、長く繋がっている。この繋がりはString（ストリング）と呼ばれており、数十個のMOSからなっている。動作については図5に示す。ビット線を選択するには、まず、同図の選択線に繋がっているMOSをONにする。ついで、FGをもっているNV-MOSのうち、2～Nまでのゲートに電圧を加えてONにする。このON状態を太線で表したのが図5である。この状態はMOS1がアクセスされたことを意味しており、書き込みや読み出しを行うことができる。次に、MOS2にアクセスするためには、MOS2以外のすべてのMOSをON状態にすればいいわけである。このように、順次送っていけば、すべてのMOSにアクセスすることができる。

ここで、図5の回路図をよく眺めていただきたい。通常、MOSはソース、ドレイン、ゲートの3つの端子が出ているが、このNANDフラッシュの回路では、ソースとドレインが全MOSに共通で、それぞれのMOSにはFGのみの電極が繋がっている。したがって、1ビットに1配線ですむので、構造がきわめて簡単なものになり、集積度を上げるには理想的な構造である。微細化が進むとDRAMよりも集積度が上がり、今や集積度ではナンバーワンのメモリとなっている。さらに、FGに蓄えた電荷量は「有、無」の2値ではなく、「多い、少ない、ゼロ」のように、多値の情報を記憶することができる。これにより、1個のMOSを2ビットとして働かせれば、集積度が2倍に増えたことになる。これをML（Multi Level）と呼ぶ。このような努力により、最近は1枚のチップで128Gbit（Giga bit、1280億ビット）という高集積度の製品が生産されている。

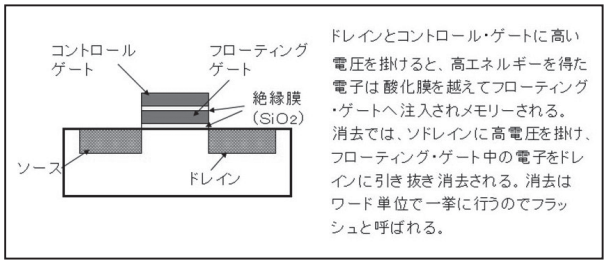


図3　FG（フローティングゲート）型フラッシュメモリ

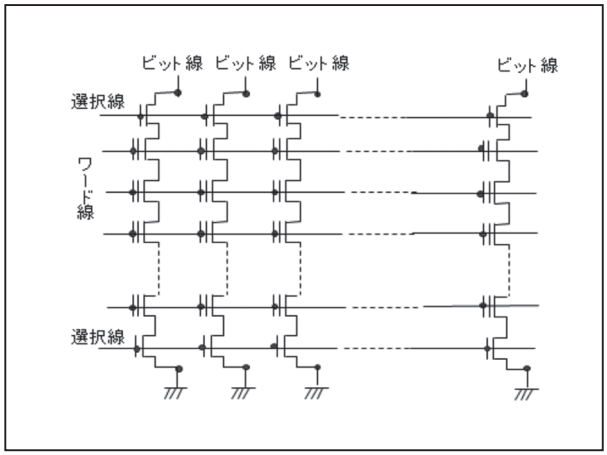


図4　NAND型フラッシュ・メモリの回路図

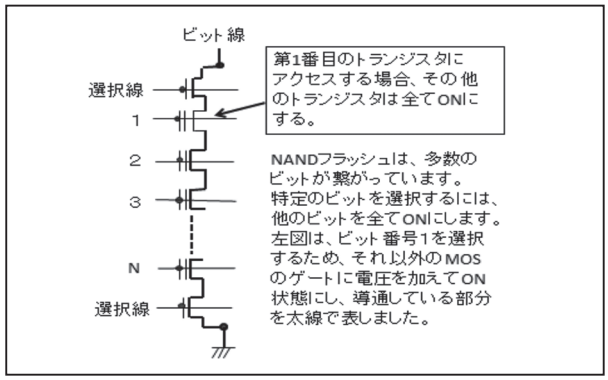


図5　NANDフラッシュの動作

○ちょっと豆情報

フラッシュ･メモリの発明者として、舛岡富士雄氏（当時東芝、現東北大学）が知られているが、半導体エネルギー研究所の山﨑舜平氏が1970年に出願された特許がはるかに早く、「絶縁膜で囲まれたフローティング･ゲートを持つメモリ」となっているので、まさに現在のフラッシュである。舛岡氏がNANDフラッシュを発明されたのは1984年のことである。当時の不揮発性メモリは、フローティング･ゲートに電荷を貯める原理は現在と同じであったが、紫外線消去のEPROM（Erasable Programmable Read Only Memory）で、紫外線光が入るようにパッケージに窓が開いており、消去するのに30秒もかかった。それを電気的に一括消去できるようにした訳で、一括消去が写真のフラッシュのようなので舛岡氏により「フラッシュ」と命名された。当時から、舛岡氏は「将来、フラッシュ･メモリは広く普及し、磁気ディスクを置き換えていく」と主張されていたのが、今や実現しつつある。

2. 微細化の限界とSONOS構造

図3で示した構造のNANDフラッシュは、FG構造のNV-MOSがびっしり並んでいるため、微細化により集積度を上げていくと、隣のビットとの間隔が狭くなってきてしまい、その結果、隣接するセル同士が電気的な干渉を起こし、セル･トランジスタの閾値電圧がシフトしたり、隣のビットの情報との混信が起こるようになる。このため、以前は、NANDフラッシュの微細化は50nmが限度で、それ以上は無理であるといわれていた。しかし、現在は20nm以下の製品も生産されており、15nm程度に微細化しても製品化可能と思われている。このように、限界といわれても、それを突破するのが半導体開発の常である、ということができるだろう。

そのため、FG構造ではない、まったく別の構造のSONOS構造のフラッシュが注目されている。SONOSは、Si-SiO_2-SiN-SiO_2-Poly-Siの略で、その構造は図6に示すとおりである。Poly-Siでできた FGの代わりにSiNが用いられている。SiNは、Stoichiometry（化学量論的）は、Si_3N_4となるが、実際の膜はSi:Nが3対4ではなく、9対10のように、Si richになっている。このため、Siの4本ある結合手のうち、結合にあずかれない手、すなわちダングリング･ボンド（Dangling Bond＝未結合手）が多数生じ、そこに電子がトラップ（Trap＝捕獲）される。SiN中に電荷がトラップされているか否かがON-OFFの情報となる。SiN構造のNVメモリは、FG型より微細化が可能といわれている。

3.NANDフラッシュ･メモリの積層膜による立体化

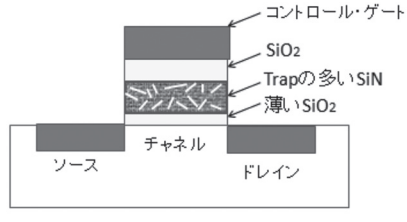
1. V-Channel

これまでのNANDフラッシュは、Siウエハ上にMOSを形成するプレーナ型であったが、積層してMOSを積む巧妙な方法が、2007年のVLSI Tech. IEEEにおいて、（株）東芝がBiCS（Bit Cost Scalable）と呼ばれる新構造のNANDフラッシュ･メモリを発表した。これが本レポートの中心話題である。

その後、サムスンはTCAT（Terabit Cell Array Transistor）や、ハイニックスがSMArT（Stacked Memory Array Transistor）と名付けた同様のNANDフラッシュ･メモリを発表した。V-Channel（Vertical Channel＝縦方向チャネル）と呼ばれる構造である。この構造は、これまでのNANDフラッシュはもちろんのこと、それ以外のMOS LSIとも大きく異なり、同じ建物でも平屋の我が家とスカイツリータワーほどの違いがある。以下に、ていねいに説明したいと思う。

2. BiCSの構造

図7に概略の構造を示す。Si基板上に、SiO_2／Poly-Siの層を連続的に積層する。積層数は、最近のサムスンの発表では24層となっているが、さらに多い方がビット数が多くなる。平板状のPoly-Siは、NV-MOSのコントロール電極になり、SiO_2はその間の絶縁物となる。その積層膜の上から下までホール（Hole＝孔）をエッチングする。ホールの開口寸法は50nm程度と思われ、深さとの比（Aspect Ratio）は、50程度になっていると推定している。次に、ホールの中に、図8のようにSONOS構造を形成する。SiN膜は、先に述べたSONOS構造の電荷をトラップする膜となり、SiO_2膜は、トンネル電流が流れるように10nm



チャネルからトラップの多いSiN膜へ、薄いSiO2膜をトンネル電流となって、電子が注入される

図6　SONOS構造のフラッシュ･メモリ

以下の薄膜や、コントロール･ゲートからの電界を強めるため極めて薄い膜になっている。図の縦方向のPoly-Siが、MOSのチャンネルになる部分で、通常の平面MOSとは異なりホールの中で縦になっているので、V－Channel（縦チャンネル）と呼ばれる。また、コントロール電極からの導通を上面に取り出すため、チップ端に階段状にエッチングし電極を取り出す。

このNV-MOSの動作を説明すると、Poly-Siの柱と板状の電極の交点が、Poly-SiをチャンネルとするSONOSのMOSとなる。１本のPoly-Si柱には、このMOSが多数直列に連続して接続されてNANDストリングとなる。最近のサムスンの120GbitのNANDフラッシュの発表では、このPoly-Si柱が25億個(2.5Giga個)あり、ゲート電極が24層で、２ビットのML（Multi　Level）であるから、2.5×24×2＝120Gbitのチップになる。V-Channelでは、チャンネルがN型不純物をわずかにドーピングしたPoly-Siであり、粒界のトラップ密度が高いため電子の移動度は低く、MOSの閾値特性（Vth）がばらつくことになり、この対策としてPoly-Si層を10nm程度と非常に薄くしてトラップ数を減少させている。

これは中空円筒の構造となることから、マカロニ型と呼んでいるようである。マカロニであれば中心部に空洞があれば料理の味付けに最適であるが、LSIでは後の工程に支障をきたすため、SiO_2などの絶縁物で埋めている。平面状に積層されたゲート電極となるPoly-Siは、金属電極の役割なので抵抗を下げるため、高濃度のP型不純物をドーピングしていると思われる。

説明が少々込み入って分かりにくかったかもしれないが、皆さんの頭脳は、筆者と違って一度読んだところは消えない不揮発性なので大丈夫だろう。

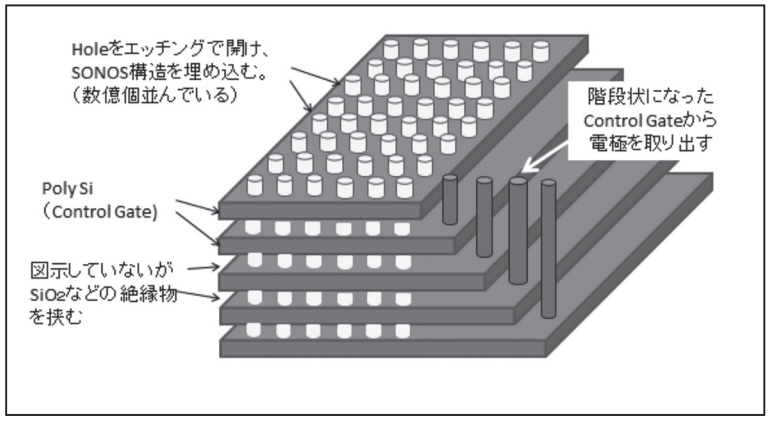


図7　3D-NANDフラッシュの概略図

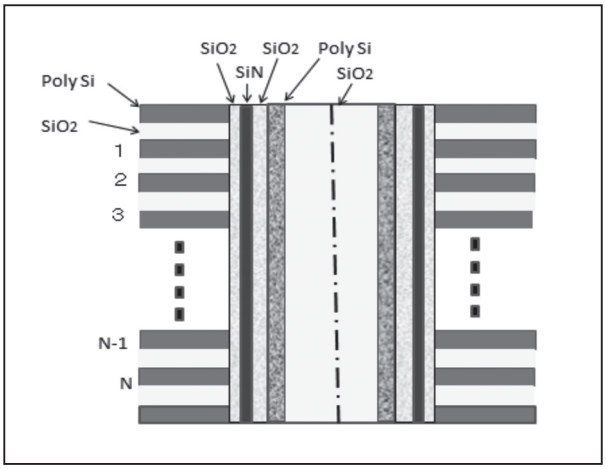


図8　3D-NANDフラッシュ、SONOS構造
（Si-SiO2-SiN-SiO2-Si）

3. 3D-NANDフラッシュの製造プロセスと歩留まり

東芝は、2007年にBiCSの発表を行い、その後も三重工場で量産すると何度か発表したが、６年後の今になっても量産に至っておらず、やはり量産のためのプロセス技術の開発が思うように進んでいないものと思われる。この間、サムスンが一足先に量産開始を発表したが、本当に歩留まりが上がり、適正な価格で大量の３D-NANDが出荷されるのか、まだ疑問視する意見もある。それだけ、歩留まりは大問題であると思われる。

(1)多層膜の生成

SiO_2とPoly-Siを順に積んでいくわけであるが、その厚みはSiO_2が60nm、Poly-Siが40nm程度だろうと筆者は推定している。SiO_2は厚いほど絶縁が完全に行われるが、ホールの深さが深くなりアスペクト比が大きくなるので、エッチングや穴埋めCVDが大変難しくなる。コントロール電極用Poly-Siの厚さはチャンネル長さを決めることになり、微細化することも可能と思われるが、MOSの安定動作には、40nm程度が適当だろうと思う。この結果、60＋40＝100nmとなり、24層なら2400nmがホールの深さとなる。多層膜の平坦性を保つことや、異物混入を防ぐことなど、単層膜のCVDにはない難しさがあると思われる。

コントロール電極は、抵抗を下げるため、NiやCoのシリサイドが用いられる可能性がある。将来的にはグラフェンも考えられ、検討されているようである（グラフェンは、カーボンが平面状に並んだ単結晶で、電気抵抗が金属よりはるかに低いので、電極として用いればメリットがある）。

(2)アスペクト比の大きいホールのエッチング

ホールの形状を推定すると、深さは24層なら2400nm程度と推定され、ホール径は50nmとすると、アスペクト比（Aspect Ratio＝深さ対開口の比）は48となる。トレンチ型DRAMでは、この程度のホールをエッチングした経験があるかもしれないが、3D-NANDフラッシュの場合は、被エッチング材料が均一な結晶ではなく異物質の多層膜であるため、その難しさは比較にならない。筆者の推定するところでは、図9のように正常な形状ばかりでなく、いろいろな不良形状がありうる。SiO_2とSiNの積層物質のエッチングであるから、単一のエッチングガスで均一にエッチングできるとは限らないので、実際のエッチング形状は図9のようになっていると思われる。また、300mmウエハ全面にわたってエッチングガス（プラズマの荷電粒子）が垂直に入射しなければならないから、この制御もかなり難しく、ひとつ間違えると斜めエッチングになってしまう。

(3)ONO膜の製作と問題点

ONO膜の生成は、NANDフラッシュの性能を決める、もっとも重要なプロセスである。重要なノウハウの部分であるため各社のプロセスの詳細は不明だが、筆者は次のように推定している。

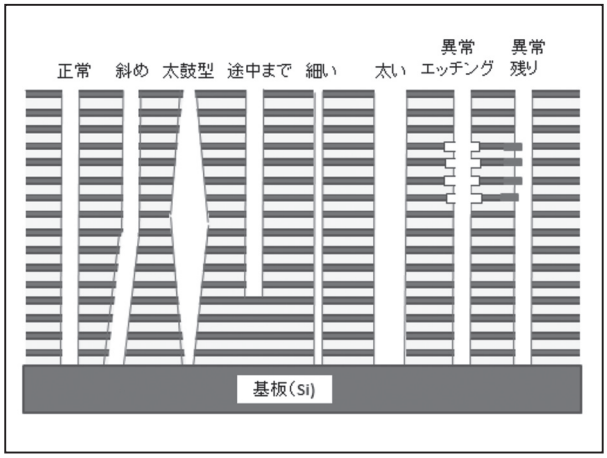


図9　ホール・エッチングの異常の数々

まず、Si側のSiO_2膜は、電子がトンネル効果で注入される膜であるから、膜厚はきわめて薄く、かつ高電圧で破壊が進まないように緻密な膜が要求される。できれば高温熱酸化のような緻密な膜が望ましいと思われるが、微細な膜厚の均一性を考量してALD（Atomic Layer Deposition＝原子層堆積）が用いられるであろう。次いで、SiNは、緻密さではLPCVD（Low pressure CVD＝減圧CVD）が勝るが、ダングリング・ボンドを多く形成するため、通常のPCVD（Plasma CVD）が用いられると思われる。しかし、ホールの側面に均一な膜を形成するのはそれなりの高度な技術が要求される。

(4)チャンネル

MOSのチャンネルはPoly-Siを用いることになるが、半導体は歴史始まって以来、常に良質な単結晶を求めてきたのと、まるで正反対である。ホールの内部にエピタキシなどで単結晶を作成することはとても考えられないので、チャンネルはPoly-Siとなる。MOSのチャンネルが単結晶でないのは不都合な点がいろいろ出てくる。Poly-Siの厚さを10nm以下に薄くするためALDで正確に膜厚を制御し、できるだけ単結晶に近い性質にしなければならないことは、先に述べた通りである。

(5)電極取り出しの工夫

コントロール・ゲートの電極取り出しは、図7にように数十段の段差から取り出す必要があり、このような構造は過去のLSIでは経験がない。段差を実現するには、厚いフォトレ

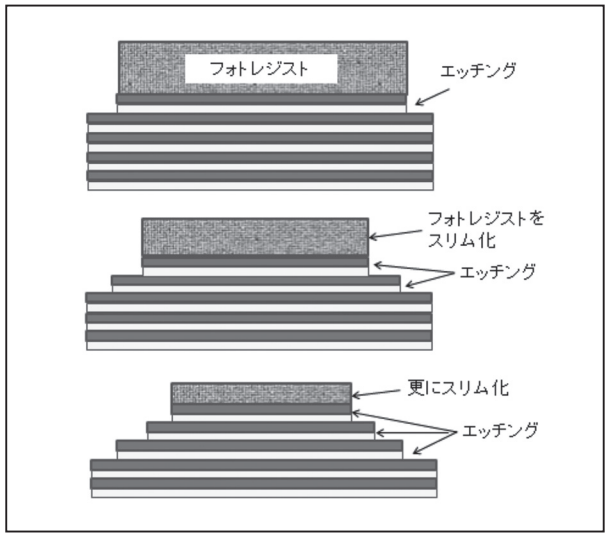


図10　フォトレジストのスリム化により、段差形成

ジストでマスクし、エッチングの度にフォトレジストがサイドエッチングされて細ることにより、繰り返しエッチングで図10のように段差構造を作ることである。この方法だと、フォトレジスト工程が1回ですむ。段差ができると厚い絶縁膜を被せて、上から順にトレンチをエッチングする。厚い絶縁物はウエハ全面に付着するので、CMPで削って平坦化する。次に、ホールのエッチングでは、浅いホールから深いホールまで図11のように数十回のエッチングが必要になり、そのたびにフォトレジストでマスクする必要がある。しかし、図12のような巧妙な案が発表されており、フォトレジストの回数が大幅に減らすことができる。

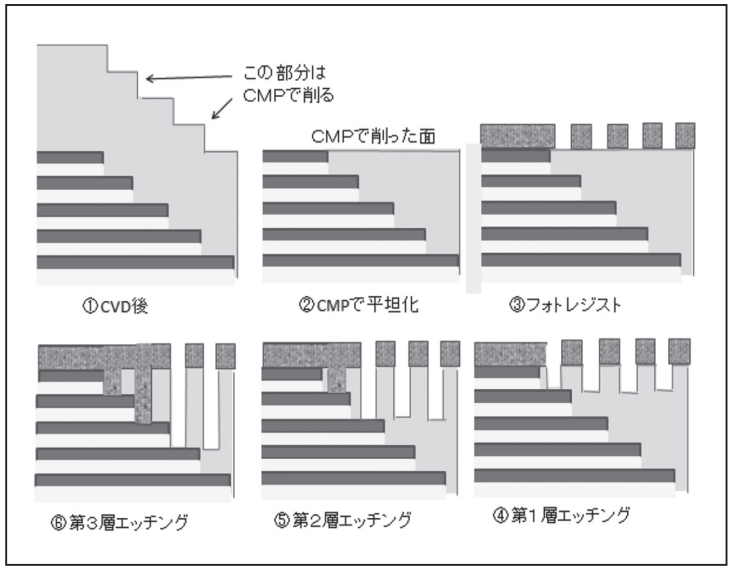


図11　階段状ホールの形成方法（1段ずつエッチングする方法）

4.NANDの将来

1. 3D-NANDフラッシュの性能について

サムスンの発表によると、10nm世代の浮遊ゲートNANDフラッシュに比べて、動作信頼性を2～10倍、書き込み速度を2倍に改善できるとし、セルの寿命を示す書き込み回数（耐久年限）は製品ごとに最低2倍から最大10倍以上に向上する、一方、消費電力は半分に減らせるという。しかし、3D-NANDは、これまで用いられてこなかったSONOS構造であり、用いられてこなかった原因があるわけで、その方がすぐれているという理屈はあり得ない、という反論もある。また、肝心のMOSのチャンネルが単結晶ではないのも不利な点だろう。したがって、性能の優劣を云々するのはまだ早計というべきで、ユーザーの検討を待ちたいと思う。

なお、今回はSONOS構造のみを紹介したが、FG（フローティング･ゲート）型の3D－NANDフラッシュも開発されている。かなりトリッキーなプロセスと思われるので、主流技術となるかどうか疑問に思っているが、図13にその構造図のみを載せておく。

2. コストについて

LSIのコストは、一般に設備投資額と歩留りが大きな影響を与える。3D－NANDフラッシュの場合の投資額について考えると、リソグラフィは数十nm程度のパターンであるか

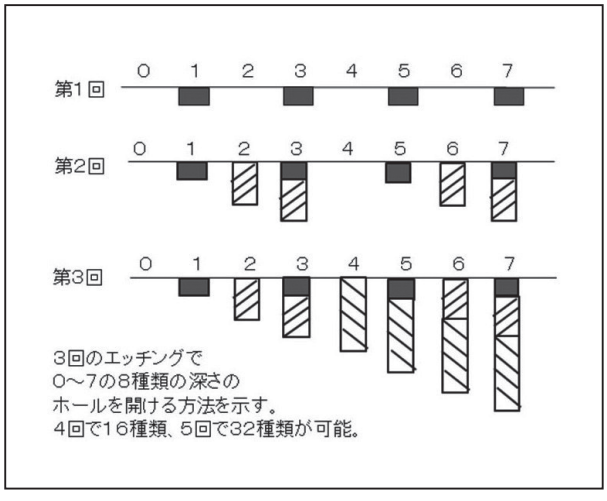
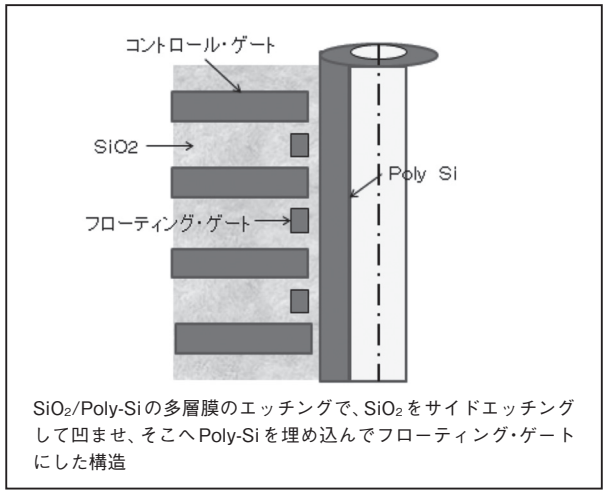


図12　エッチング回数を減らした階段状ホールの形成方法



SiO_2/Poly-Siの多層膜のエッチングで、SiO_2をサイドエッチングして凹ませ、そこへPoly-Siを埋め込んでフローティング･ゲートにした構造

図13フローティング･ゲートの3D-NANDフラッシュ

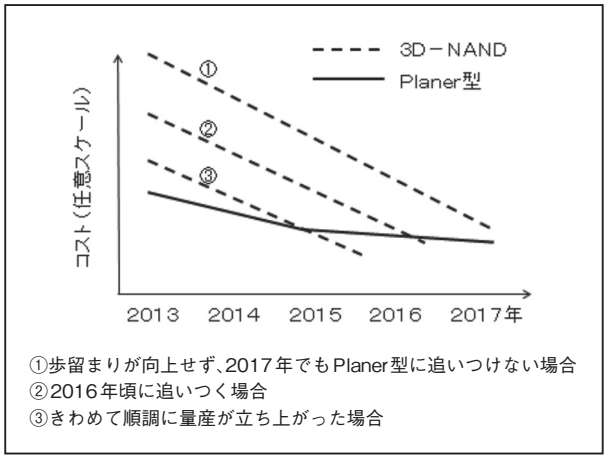
ら巨額な投資が必要な微細加工の必要がない。従来から使われているArFのステッパで十分に間に合い、Double Patterningも必要ない。一方、CVDとエッチングは、従来の装置でも工夫すれば間に合うかもしれないが、多層連続膜付けや、多層膜エッチングなどはこれまでにないプロセスであるから、専用の特殊な装置が必要だろうと思われる。CMPなども特別な仕様の装置が必要かもしれないが、総合すると投資額はリソグラフィが簡単であるから、あまり大きくないと考えられる。

歩留まりの点では、仮にホールが25億個とするとすべて合格するとは考えられないため、冗長回路を設けておき、不良ビットを合格ビットに置き換える方法が行われる。ただし、冗長回路をどの程度の割合で導入するかは、歩留まりに依って決まるが、やたらに多くすることもできない。

歩留まりがこれまでのプレーナ型に追いつくのは容易ではないと思われる。そこで、筆者のまったくのあてずっぽうではあるが、コスト推移について図14のような経過を辿るのではないかと考えている。

3. NANDフラッシュの今後

NANDフラッシュ・メモリのビット需要は、スマートフォンなどの携帯電話機やタブレット端末の二つだけで全体の50%を占め、ますます増えつつある。さらに中長期的にはサーバなどのインフラ側でもNANDフラッシュ・メモリの採用が進むと考えられる。したがって、10年ぐらいのレンジで考えれば総ビット数は桁違いの量が要求されると思われ、生産工場をいくら増やしても追いつかず、3D化などの技術革新で対応する必要がある。以前の東芝の発表では、2015年に512Gビットの大容量品を実現するとしており、サムスンはTera bitを視野に入れた戦略を考えているようである。



①歩留まりが向上せず、2017年でもPlaner型に追いつけない場合
②2016年頃に追いつく場合
③きわめて順調に量産が立ち上がった場合

図14　Planer型と3D-NANDのコスト比較

また、ReRAMへ期待するという意見もある。3D-NANDフラッシュは、SiNに電荷を蓄えるメモリであるが、抵抗の変化を利用したReRAMを用いたフラッシュ・メモリなどが活発に検討されており、SiN型と置き換わる可能性がある。ReRAMについても、いずれ機会があれば本誌で取り上げたいと思う。

いずれにせよ、フラッシュ・メモリは、Giga bit時代からTera bit時代へと大変革が起こると予想される。