(19)日本国特許庁(JP)　(12)公開特許公報(A)　(11)公開番号
特開2023-67686
(P2023-67686A)
(43)公開日　令和5年5月16日(2023.5.16)

(51)国際特許分類		FI			テーマコード（参考）
H10B 99/00 (2023.01)		H01L	27/10	495	5F083
H10B 43/27 (2023.01)		H01L	27/11582		5F101
H10B 43/50 (2023.01)		H01L	27/11575		
H01L 21/336 (2006.01)		H01L	29/78	371	

審査請求　有　請求項の数　18　ＯＬ　外国語出願　（全70頁）

(21)出願番号　特願2022-18575(P2022-18575)
(22)出願日　令和4年2月9日(2022.2.9)
(31)優先権主張番号　63/273,876
(32)優先日　令和3年10月29日(2021.10.29)
(33)優先権主張国・地域又は機関
米国(US)
(31)優先権主張番号　17/569,419
(32)優先日　令和4年1月5日(2022.1.5)
(33)優先権主張国・地域又は機関
米国(US)

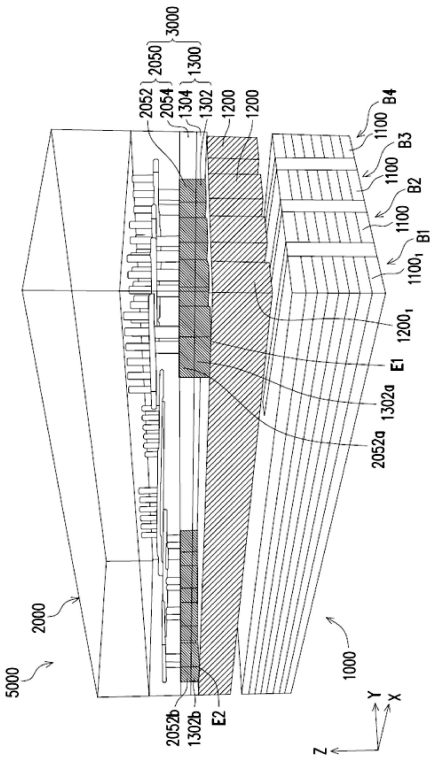
(71)出願人　599129074
旺宏電子股▲ふん▼有限公司
台湾新竹科學工業園區力行路１６號
(74)代理人　100107423
弁理士　城村 邦彦
(74)代理人　100120949
弁理士　熊野 剛
(74)代理人　100093997
弁理士　田中 秀佳
(72)発明者　葉 騰豪
台湾新竹▲縣▼竹北市莊敬五街１６１號
３樓
(72)発明者　呂 函庭
台湾新竹市清華大學東院１９號

Ｆターム（参考）　5F083　EP18 EP22 EP76 ER21

最終頁に続く

(54)【発明の名称】　３Ｄフラッシュメモリモジュールチップおよびその製造方法

(57)【要約】　（修正有）
【課題】フラッシュメモリの性能は、複数回の動作後に著しく低下するため、フラッシュメモリに対して修復プロセスを実行する３Ｄフラッシュメモリモジュール及びその製造方法を提供する。
【解決手段】３Ｄフラッシュメモリモジュールチップ５０００は、メモリチップ１０００及び制御チップ２０００を含む。メモリチップは、複数のタイル及び複数のヒータ１２００を含む。タイルはそれぞれ、複数の３Ｄフラッシュメモリ構造を含む。ヒータは、タイルのそれぞれの３Ｄフラッシュメモリ構造の周りに配置されている。制御チップは、ヒータうちの少なくとも１つを駆動するためにメモリチップと接合されている。メモリチップでは、制御チップがヒータを駆動してメモリチップのローカルセクタに対して修復プロセスを行うことができるように制御チップと接合されている。
【選択図】図１Ａ



【特許請求の範囲】
【請求項１】
メモリチップであって、
それぞれが複数の３Ｄフラッシュメモリ構造を含む複数のタイル、および
前記タイルのそれぞれの前記３Ｄフラッシュメモリ構造の周りに配置された複数のヒータ、
を含む、メモリチップと、
前記ヒータのうちの少なくとも１つを駆動するために前記メモリチップに接合された制御チップと、
を含む、３Ｄフラッシュメモリモジュールチップ。
【請求項２】
前記ヒータが前記３Ｄフラッシュメモリ構造の上方に配置され、前記制御チップに隣接している、請求項１に記載の３Ｄフラッシュメモリモジュールチップ。
【請求項３】
前記ヒータが前記３Ｄフラッシュメモリ構造間の複数のスリットトレンチ内に配置されている、請求項１に記載の３Ｄフラッシュメモリモジュールチップ。
【請求項４】
前記メモリチップが、
第１の基板上に配置された複数の第１のトランジスタと、
前記第１のトランジスタの上方に位置する前記３Ｄフラッシュメモリ構造と、
第１の相互接続構造であって、前記３Ｄフラッシュメモリ構造が前記第１の相互接続構造内に埋め込まれている、第１の相互接続構造と、
をさらに含む、請求項１に記載の３Ｄフラッシュメモリモジュールチップ。
【請求項５】
前記第１の相互接続構造が、
前記３Ｄフラッシュメモリ構造と前記第１のトランジスタとの間に位置し、前記３Ｄフラッシュメモリ構造と前記第１のトランジスタとを電気的に接続する下部相互接続構造と、
前記３Ｄフラッシュメモリ構造上に位置し、前記３Ｄフラッシュメモリ構造に電気的に接続する上部相互接続構造と、
を含む、請求項４に記載の３Ｄフラッシュメモリモジュールチップ。
【請求項６】
前記制御チップが、
それぞれが、
第２の基板上に位置する第２のトランジスタであって、前記第２のトランジスタのソース領域がグローバル電源に電気的に接続されている、第２のトランジスタと、
前記第２のトランジスタのドレイン領域に電気的に接続され、前記ヒータのうちの１つのヒータの第１の端部に電気的に接続された第１のパッドと、
接地され、前記ヒータのうちの前記１つのヒータの第２の端部に電気的に接続された第２のパッドと、
を含む、複数の駆動行
を含む、請求項４に記載の３Ｄフラッシュメモリモジュールチップ。
【請求項７】
前記制御チップが、
前記駆動行の前記第２のトランジスタの複数のゲート層に電気的に結合された行デコーダと、
前記第２のトランジスタの複数のソース領域および前記グローバル電源に電気的に接続された列デコーダと、
をさらに含む、請求項６に記載の３Ｄフラッシュメモリモジュールチップ。
【請求項８】

前記制御チップがアレイ状に配列された複数のタイルを含み、同じ列内の前記タイルの前記第２のトランジスタの前記ソース領域が互いに電気的に接続されている、請求項６に記載の３Ｄフラッシュメモリモジュールチップ。
【請求項９】
前記制御チップと前記メモリチップとが接合構造によって接合されている、請求項６に記載の３Ｄフラッシュメモリモジュールチップ。
【請求項１０】
前記複数の３Ｄフラッシュメモリ構造が複数の３Ｄ　ＡＮＤフラッシュメモリ構造、複数の３Ｄ　ＮＡＮＤフラッシュメモリ構造、または複数の３Ｄ　ＮＯＲフラッシュメモリ構造を含む、請求項６に記載の３Ｄフラッシュメモリモジュールチップ。
【請求項１１】
メモリチップを形成するステップであって、
第１の基板上に複数のタイルを形成することであり、前記タイルのそれぞれが複数の３Ｄフラッシュメモリ構造を含む、第１の基板上に複数のタイルを形成すること、および
前記タイルのそれぞれの前記３Ｄフラッシュメモリ構造の周りに複数のヒータを形成すること、
を含む、メモリチップを形成するステップと、
制御チップを形成するステップと、
前記制御チップと前記メモリチップを接合するステップであって、前記制御チップが前記ヒータを駆動するように構成されている、前記制御チップと前記メモリチップを接合するステップと、
を含む、３Ｄフラッシュメモリモジュールチップの製造方法。
【請求項１２】
前記ヒータが前記３Ｄフラッシュメモリ構造の上方に形成されている、請求項１１に記載の３Ｄフラッシュメモリモジュールチップの製造方法。
【請求項１３】
前記ヒータが前記３Ｄフラッシュメモリ構造の周りの複数のスリットトレンチ内に形成されている、請求項１１に記載の３Ｄフラッシュメモリモジュールチップの製造方法。
【請求項１４】
前記メモリチップを形成する前記ステップが、
前記第１の基板上に複数の第１のトランジスタを形成するステップと、
前記第１のトランジスタの上方に前記３Ｄフラッシュメモリ構造を形成するステップと、
を含む、請求項１１に記載の３Ｄフラッシュメモリモジュールチップの製造方法。
【請求項１５】
前記制御チップを形成する前記ステップが、
複数の駆動行を形成するステップであって、前記駆動行のそれぞれの形成が、
第２の基板上に第２のトランジスタを形成することと、
前記第２のトランジスタ上に第２の相互接続構造を形成することであり、前記第２のトランジスタのソース領域が前記第２の相互接続構造を介してグローバル電源に電気的に結合される、第２の相互接続構造を形成することと、
前記第２の相互接続構造上に第１のパッドを形成することであり、前記第１のパッドが前記第２の相互接続構造を介して前記第２のトランジスタのドレイン領域に電気的に接続される、第１のパッドを形成することと、
前記第２の相互接続構造上に第２のパッドを形成することであり、前記第２のパッドが前記第２の相互接続構造を介して接地に電気的に接続される、第２のパッドを形成することと、
を含む、複数の駆動行を形成するステップ、
を含む、請求項１４に記載の３Ｄフラッシュメモリモジュールチップの製造方法。
【請求項１６】

前記第１のパッドを前記ヒータのうちの１つのヒータの第１の端部に電気的に接続するステップと、
前記第２のパッドを前記ヒータのうちの前記１つのヒータの第２の端部に電気的に接続するステップと、
をさらに含む、請求項１５に記載の３Ｄフラッシュメモリモジュールチップの製造方法。
【請求項１７】
前記制御チップと前記メモリチップとが接合構造によってハイブリッド接合されている、請求項１１に記載の３Ｄフラッシュメモリモジュールチップの製造方法。
【請求項１８】
前記複数の３Ｄフラッシュメモリ構造が複数の３Ｄ　ＡＮＤフラッシュメモリ構造、複数の３Ｄ　ＮＡＮＤフラッシュメモリ構造、または複数の３Ｄ　ＮＯＲフラッシュメモリ構造を含む、請求項１１に記載の３Ｄフラッシュメモリモジュールチップの製造方法。
【発明の詳細な説明】
【技術分野】
【０００１】
本開示の実施形態は、半導体モジュールおよびその製造方法に関し、詳細には３Ｄフラッシュメモリモジュールおよびその製造方法に関する。
【背景技術】
【０００２】
不揮発性メモリは、電源オフ時に記憶データが消失しないという利点を有するため、パーソナルコンピュータまたは他の電子機器に広く使用されるメモリとなっている。現在、業界で一般的に使用されている３次元（３Ｄ）メモリは、ＮＯＲフラッシュメモリおよびＮＡＮＤフラッシュメモリを含む。加えて、別のタイプの３Ｄメモリは、ＡＮＤフラッシュメモリであり、これは、高集積および高面積利用率を有する多次元メモリアレイに適用することができ、動作速度が速いという利点を有する。したがって、３Ｄメモリデバイスの開発が徐々に現在のトレンドになってきている。
【発明の概要】
【発明が解決しようとする課題】
【０００３】
本開示は、フラッシュメモリに対して局所修復プロセスを行うことができる３Ｄフラッシュメモリモジュールチップおよびその製造方法を提供する。
【課題を解決するための手段】
【０００４】
本開示の一実施形態において、３Ｄフラッシュメモリモジュールチップは、メモリチップおよび制御チップを含む。メモリチップは、複数のタイルおよび複数のヒータを含む。タイルはそれぞれ、複数の３Ｄフラッシュメモリ構造を含む。ヒータは、タイルのそれぞれの３Ｄフラッシュメモリ構造の周りに配置されている。制御チップは、ヒータのうちの少なくとも１つを駆動するためにメモリチップと接合されている。
【０００５】
本開示の一実施形態において、３Ｄフラッシュメモリモジュールチップを製造する方法は、以下のステップを含む。メモリチップが形成され、本ステップは、第１の基板上に、それぞれが複数の３Ｄフラッシュメモリ構造を含む複数のタイルを形成すること、およびタイルのそれぞれの３Ｄフラッシュメモリ構造の周りに複数のヒータを形成すること、を含む。制御チップが形成される。制御チップとメモリチップとが接合され、制御チップは、ヒータを駆動するように構成される。
【発明の効果】
【０００６】
上記に基づいて、本開示による３Ｄフラッシュメモリモジュールチップおよびその製造方法では、追加の制御チップを使用してヒータを駆動し、フラッシュメモリの各セクタに

対して局所修復プロセスを実行する。制御チップは、ヒータコントローラがメモリチップの面積を占有することを防止するために別個に製造されてもよく、制御チップは、プロセスのコストを削減するためにそれほど高度でないプロセスによって製造されてもよい。

【図面の簡単な説明】

【０００７】

【図１Ａ】本開示の一実施形態による３Ｄフラッシュメモリモジュールチップの概略斜視図である。

【図１Ｂ】本開示の一実施形態による３Ｄフラッシュメモリモジュールチップの概略斜視図である。

【図２Ａ】本開示の一実施形態によるメモリチップの３Ｄフラッシュメモリ構造の部分上面図である。

【図２Ｂ】図２ＡのＩ－Ｉ’線に沿った断面図である。

【図３Ａ】本開示の別の実施形態によるヒータを有するメモリチップの部分上面図である。

【図３Ｂ】図３ＡのＩ－Ｉ’線に沿った断面図である。

【図４Ａ】本開示の別の実施形態によるヒータを有するメモリチップの部分上面図である。

【図４Ｂ】本開示の別の実施形態によるメモリチップのヒータおよびパッドの部分上面図である。

【図４Ｃ】図４ＢのＩＩ－ＩＩ’線に沿った断面図である。

【図５Ａ】本開示の一実施形態による制御チップの概略斜視図である。

【図５Ｂ】本開示の一実施形態による制御チップの概略斜視図である。

【図５Ｃ】本開示の一実施形態による制御チップの概略斜視図である。

【図５Ｄ】本開示の一実施形態による制御チップの概略斜視図である。

【図５Ｅ】本開示の一実施形態による制御チップの概略斜視図である。

【図６Ａ】本開示の一実施形態によるメモリチップおよび制御チップの概略斜視図である。

【図６Ｂ】図６Ａの概略回路図である。

【図７Ａ】本開示の３Ｄフラッシュメモリモジュールチップを製造するプロセスの概略断面図である。

【図７Ｂ】本開示の３Ｄフラッシュメモリモジュールチップを製造するプロセスの概略断面図である。

【図７Ｃ】本開示の３Ｄフラッシュメモリモジュールチップを製造するプロセスの概略断面図である。

【図８】本開示の別の実施形態による制御チップの概略斜視図である。

【発明を実施するための形態】

【０００８】

フラッシュメモリの性能は、複数回の動作後に著しく低下するため、フラッシュメモリに対して修復プロセスを実行する必要がある。修復プロセスでは、フラッシュメモリの電荷蓄積構造（例えば、窒化物層）を修復するために、ヒータを使用してフラッシュメモリを加熱することがある。現在の技術では、ワード線がヒータとして最も一般的に使用されている。しかしながら、ワード線の数が多く、他の構成要素（例えば、ワード線デコーダ）との構成関係が複雑であるため、フラッシュメモリ構造のレイアウト設計がより困難になる場合がある。

【０００９】

本開示の実施形態は、いくつかの３Ｄフラッシュメモリモジュールチップを提供しており、メモリチップの３Ｄフラッシュメモリ構造の上方または側壁の周りにヒータが配置され、メモリチップは、制御チップがヒータを駆動してメモリチップのローカルセクタに対して修復プロセスを行うことができるように制御チップと接合されている。

【００１０】

図１Ａおよび図１Ｂは、本開示の一実施形態による３Ｄフラッシュメモリモジュールチップの概略斜視図である。図２Ａは、本開示の一実施形態によるメモリチップの３Ｄフラッシュメモリ構造の部分上面図である。図２Ｂは、図２ＡのＩ－Ｉ’線に沿った断面図である。図３Ａは、本開示の別の実施形態によるヒータを有するメモリチップの部分上面図である。図３Ｂは、図３ＡのＩ－Ｉ’線に沿った断面図である。

【００１１】

図１Ａおよび図１Ｂを参照すると、本開示の一実施形態による３Ｄフラッシュメモリモジュールチップ（３Ｄ集積回路（３Ｄ　ＩＣ）とも呼ばれる）５０００は、メモリチップ１０００および制御チップ２０００を含む。メモリチップ１０００は、複数の３Ｄフラッシュメモリ構造１１００および複数のヒータ１２００を含む。ヒータ１２００は、３Ｄフラッシュメモリ構造１１００の周りに配置されている。一部の実施形態では、ヒータ１２００は、図１Ａに示すように、３Ｄフラッシュメモリ構造１１００の上方に配置されている。他の実施形態では、ヒータ１２００は、図１Ｂに示すように、３Ｄフラッシュメモリ構造１１００間のスリットトレンチ１１１０内に配置されている。制御チップ２０００は、メモリチップ１０００の上方に配置され、メモリチップ１０００内のヒータ１２００を駆動する。制御チップ２０００とメモリチップ１０００は、接合構造３０００によって互いに接合されてもよい。

【００１２】

図１Ａおよび図１Ｂを参照すると、メモリチップ１０００の３Ｄフラッシュメモリ構造１１００は、（図２Ａおよび図２Ｂに示すような）３Ｄ　ＡＮＤフラッシュメモリ構造、３Ｄ　ＮＡＮＤフラッシュメモリ構造（図示せず）、または３Ｄ　ＮＯＲフラッシュメモリ構造（図示せず）であってもよい。３Ｄ　ＡＮＤフラッシュメモリ構造を例として、本開示の３Ｄフラッシュメモリ構造１１００を説明するが、本開示の実施形態はこれに限定されない。

【００１３】

図２Ａおよび図２Ｂを参照すると、メモリチップ１０００は、複数のタイルＴを含むことができる。タイルＴは、複数の列および複数の行を含むアレイ状に配置されてもよい。本実施形態では、例示のために４つのタイルＴ（例えば、Ｔ１～Ｔ４）が示されている。４つのタイルＴのうち、タイルＴ１とタイルＴ２とが１つの行に配置され、タイルＴ３とタイルＴ４とが別の行に配置されている。タイルＴ１とタイルＴ３とが１つの列に配置され、タイルＴ２とタイルＴ４とが別の列に配置されている。各タイルＴは、複数のセクタＢ（例えば、Ｂ１～Ｂ４）を含むことができる。各セクタＢは、３Ｄフラッシュメモリ構造１１００を含む。３Ｄフラッシュメモリ構造１１００は、Ｘ方向に延在し、Ｙ方向に配列されている。２つの隣接する３Ｄフラッシュメモリ構造１１００は、スリットトレンチ１１１０によって互いに分離されている。

【００１４】

図２Ｂを参照すると、３Ｄフラッシュメモリ構造１１００のそれぞれは、複数のメモリセルによって形成されたメモリアレイを少なくとも含むことができる。具体的には、３Ｄフラッシュメモリ構造１１００は、第１の基板（例えば、半導体基板）１０１０上の１つまたは複数の能動デバイス（例えば、第１のトランジスタ１０２０）の上方に配置されてもよい。第１のトランジスタ１０２０は、例えば、相補型金属酸化膜半導体（ＣＭＯＳ）電界効果トランジスタである。したがって、本アーキテクチャは、アレイ下相補型金属酸化膜半導体電界効果トランジスタ（ＣＭＯＳ　ｕｎｄｅｒ　Ａｒｒａｙ（ＣＵＡ））アーキテクチャと呼ばれることもある。

【００１５】

図２Ｂを参照すると、３Ｄフラッシュメモリ構造１１００は、半導体ダイのバックエンドオブライン（ＢＥＯＬ）において配置されてもよい。例えば、３Ｄフラッシュメモリ構造１１００は、第１の相互接続構造１０３０に埋め込まれてもよい。第１の相互接続構造１０３０は、例えば、下部相互接続構造１０３２および上部相互接続構造１０３４を含む。下部相互接続構造１０３２は、第１の基板（例えば、半導体基板）１０１０上の１つま

たは複数の能動デバイス（例えば、第１のトランジスタ１０２０）の上方で、３Ｄフラッシュメモリ構造１１００のメモリアレイの下方に配置されている。上部相互接続構造１０３４は、３Ｄフラッシュメモリ構造１１００のメモリアレイの上方に配置されている。下部相互接続構造１０３２は、例えば、下部第１金属層ＢＭ１、下部第２金属層ＢＭ２、および下部第３金属層ＢＭ３、ならびにそれらの間のビアＢＶ１およびＢＶ２を含む。上部相互接続構造１０３４は、例えば、上部第１金属層ＴＭ１および上部第２金属層ＴＭ２、ならびにそれらの間のビアＴＶ１を含む。下部相互接続構造１０３２および上部相互接続構造１０３４の金属層とビアの数は、上記に限定されない。

【００１６】

図２Ｂを参照すると、３Ｄフラッシュメモリ構造１１００は、複数のゲートスタック構造５２を含む。ゲートスタック構造５２のそれぞれは、下部相互接続構造１０３２上に形成されている。ゲートスタック構造５２のそれぞれは、第１の基板１０１０のアレイ領域ＡＲから階段領域ＳＲまでＸ方向に延在する。ゲートスタック構造５２は、第１の基板１０１０の表面上に垂直に積み重ねられた複数のゲート層（ワード線とも呼ばれる）３８および複数の絶縁層５４を含む。Ｚ方向において、ゲート層３８は、間に配置された絶縁層５４によって互いに電気的に絶縁されている。ゲート層３８は、タングステンなどの金属層を含む。一部の実施形態では、ゲート層３８は、チタン（Ｔｉ）、窒化チタン（ＴｉＮ）、タンタル（Ｔａ）、窒化タンタル（ＴａＮ）、またはそれらの組合せなどのバリア層３７をさらに含む。絶縁層５４は、例えば酸化ケイ素である。

【００１７】

ゲート層３８は、（図２Ｂに示す）第１の基板１０１０の表面に平行な方向に延在する。階段領域ＳＲ内のゲート層３８は、（図２Ｂに示す）階段構造ＳＣを有することができ、それにより、下側ゲート層３８は、上側ゲート層３８よりも長く、下側ゲート層３８の端部は、上側ゲート層３８の端部を越えて横方向に延在する。ゲート層３８を接続するためのコンタクトＣ１は、階段領域ＳＲ内のゲート層３８の端部上に着地して、ゲート層３８のそれぞれを、コンタクトＣ１および上部相互接続構造１０３４を介して下部相互接続構造１０３２の導電線（例えば、下部第３金属層ＢＭ３の導電線）に接続することができる。

【００１８】

図２Ｂを参照すると、３Ｄフラッシュメモリ構造１１００は、複数のチャネルピラー１６をさらに含む。チャネルピラー１６は、アレイ領域ＡＲのゲートスタック構造５２を貫いて連続的に延びている。一部の実施形態では、チャネルピラー１６は、上面図においてリング形状のプロファイルを有してもよい。チャネルピラー１６の材料は、ドープされていないポリシリコンなどの半導体であってもよい。

【００１９】

図２Ｂを参照すると、３Ｄフラッシュメモリ構造１１００は、絶縁性充填層２４、絶縁性ピラー２８、複数の導電性ピラー（例えば、ソースピラーとして機能する）３２ａ、および複数の導電性ピラー（例えば、ドレインピラーとして機能する）３２ｂをさらに含む。導電性ピラー３２ａ、３２ｂおよび絶縁性ピラー２８は、チャネルピラー１６内に配置されており、それぞれ、ゲート層３８に対して垂直な方向（Ｚ方向）に延在している。導電性ピラー３２ａおよび３２ｂは、絶縁性充填層２４および絶縁性ピラー２８によって互いに分離され、チャネルピラー１６に電気的に接続されている。導電性ピラー３２ａ、３２ｂは、例えばドープされたポリシリコンである。絶縁性充填層２４は、例えば酸化ケイ素であり、絶縁性ピラー２８は、例えば窒化ケイ素である。

【００２０】

図２Ｂを参照すると、電荷蓄積構造４０は、チャネルピラー１６とゲート層３８との間に配置されている。電荷蓄積構造４０は、トンネル層（またはバンドギャップ工学設計されたトンネル酸化物層と呼ばれる）１４、電荷蓄積層１２、およびブロッキング層３６を含むことができる。電荷蓄積層１２は、トンネル層１４とブロッキング層３６との間に配置されている。一部の実施形態では、トンネル層１４、電荷蓄積層１２、およびブロッキ

ング層３６は、例えば、酸化ケイ素、窒化ケイ素、および酸化ケイ素である。一部の実施形態では、図２Ｂに示すように、電荷蓄積構造４０の一部（例えば、トンネル層１４）は、ゲート層３８に垂直な方向（すなわち、Ｚ方向）に連続的に延在し、電荷蓄積構造４０の他の部分（例えば、電荷蓄積層１２およびブロッキング層３６）は、ゲート層３８を取り囲む。他の実施形態では、電荷蓄積構造４０（例えば、トンネル層１４、電荷蓄積層１２、およびブロッキング層３６）は、ゲート層３８を取り囲む（図示せず）。ゲート層３８のそれぞれと、ゲート層３８によって取り囲まれた電荷蓄積構造４０、チャネルピラー１６、ソースピラー３２ａ、およびドレインピラー３２ｂとがメモリセル２０を画定する。したがって、３Ｄフラッシュメモリ構造１１００のそれぞれは、複数のメモリセル２０から構成されるメモリアレイを少なくとも含む。

【００２１】

３Ｄフラッシュメモリ構造１１００は、ローカルビット線LBL_n、ローカルソース線LSL_n、グローバルビット線GBL_n、およびグローバルソース線GSL_nをさらに含む。ローカルビット線LBL_nおよびローカルソース線LSL_nは、上部相互接続構造１０３４の上部第１金属層ＴＭ１内に位置し、コンタクトＣ２を介してソースピラー３２ａおよびドレインピラー３２ｂにそれぞれ電気的に接続されている。グローバルビット線GBL_nおよびグローバルソース線GSL_nは、上部相互接続構造１０３４内の上部ビア（図示せず）を介して、ローカルビット線LBL_nおよびローカルソース線LSL_nにそれぞれ電気的に接続されている。

【００２２】

異なる動作方法によると、メモリセル２０で１ビット動作または２ビット動作を行うことができる。例えば、ソースピラー３２ａおよびドレインピラー３２ｂに電圧が印加されると、ソースピラー３２ａおよびドレインピラー３２ｂがチャネルピラー１６に接続されているため、電子は、チャネルピラー１６に沿って移動し、電荷蓄積構造４０全体に蓄積され得る。したがって、メモリセル２０で１ビット動作を行うことができる。加えて、ファウラー・ノルトハイムトンネリングを伴う動作の場合、電子または正孔は、ソースピラー３２ａとドレインピラー３２ｂとの間の電荷蓄積構造４０内にトラップされ得る。ソース側注入、チャネルホットエレクトロン注入、またはバンド間トンネリングホットキャリア注入を伴う動作の場合、電子または正孔は、ソースピラー３２ａおよびドレインピラー３２ｂの一方に隣接する電荷蓄積構造４０内に局所的にトラップされ得る。したがって、メモリセル２０で、シングルレベルセル（ＳＬＣ、１ビット）またはマルチレベルセル（ＭＬＣ、２ビット以上）動作を行うことができる。

【００２３】

動作中、選択されたワード線（ゲート層）３８に電圧が印加され、例えば、対応するメモリセル２０の対応する閾値電圧（V_{th}）よりも高い電圧が印加されると、選択されたワード線３８と交差するチャネルピラー１６のチャネル領域がオンになり、電流がローカルビット線LBL_nからドレインピラー３２ｂに入り、オンしたチャネル領域を介してソースピラー３２ａに流れ、最終的にローカルソース線LSL_nに流れるようになる。

【００２４】

図３Ａおよび図３Ｂを参照すると、メモリチップ１０００は、複数のヒータ１２００をさらに含む。ヒータ１２００は、３Ｄフラッシュメモリ構造１１００の上方の誘電体層１０４０内に配置されてもよい。誘電体層１０４０の材料は、例えば、酸化ケイ素である。ヒータ１２００は、銅またはタングステンなどの金属層１２０２を含む。一部の実施形態では、ヒータ１２００は、チタン、タンタル、窒化チタン、窒化タンタル、またはそれらの組合せなどのバリア層１２０４をさらに含む。

【００２５】

図３Ａを参照すると、一部の実施形態では、１つのヒータ１２００が各セクタＢ上に配置され、任意の２つの隣接するセクタＢの２つのヒータ１２００は、互いに分離されている。ヒータ１２００は、Ｘ方向に延在することができる。一実施形態において、ヒータ１２００は、（図３Ａおよび図３Ｂに示すように）アレイ領域ＡＲ内に配置され、階段領域

ＳＲまで延在する。一実施形態において、ヒータ１２００は、アレイ領域ＡＲ内に配置されてもよいが、階段領域ＳＲには配置されない（図示せず）。言い換えれば、ヒータ１２００の長さは、３Ｄフラッシュメモリ構造１１００のＸ方向の長さよりも大きくても、等しくても、または小さくてもよい。

【００２６】

加えて、各セクタＢに複数のヒータ１２００を配置してもよく、例えば、アレイ領域ＡＲと階段領域ＳＲにそれぞれ１つのヒータ１２００を配置してもよく、別々に加熱を行ってもよい（図示せず）。しかしながら、本開示の実施形態は、これに限定されない。別の実施形態では、隣接する２つ、３つ、またはそれ以上のセクタＢの複数のヒータ１２００を１つのヒータ（図示せず）にまとめて、複数のセクタＢの３Ｄフラッシュメモリ構造１１００を同時に加熱することもできる。

【００２７】

図３Ａを参照すると、上面図におけるヒータ１２００の形状は、例えば、矩形または別の形状である。複数のセクタＢ上のヒータ１２００は、同じ幅または異なる幅を有することができる。アレイ領域ＡＲにおけるヒータ１２００の幅Ｗ１は、階段領域ＳＲにおけるヒータ１２００の幅Ｗ２と同じである。しかしながら、本開示は、これに限定されない。ヒータ１２００の形状は、実際の要件または設計に従って変更されてもよい。アレイ領域ＡＲにおけるヒータ１２００の幅Ｗ１は、階段領域ＳＲにおけるヒータ１２００の幅Ｗ２よりも大きくても、等しくても、または小さくてもよい。

【００２８】

図１Ａ、図１Ｂ、および図３Ｂを参照すると、メモリチップ１０００は、接合層１３００をさらに含む。接合層１３００は、パッド１３０２および絶縁層１３０４を有する。絶縁層１３０４は、ヒータ１２００上に配置されている。絶縁層１３０４の材料は、例えば、酸化ケイ素である。パッド１３０２は、各ヒータ１２００の表面の絶縁層１３０４に配置されている。パッド１３０２の材料は、例えば、銅である。パッド１３０２は、パッド１３０２ａ、１３０２ｂを有する。パッド１３０２ａ、１３０２ｂは、ヒータ１２００の第１の端部Ｅ１および第２の端部Ｅ２にそれぞれ接続されている。

【００２９】

上記の実施形態では、３Ｄフラッシュメモリ構造１１００は、３Ｄ　ＡＮＤフラッシュメモリ構造であり、ヒータ１２００は、（図３Ａ、図３Ｂ、および図６Ａに示すように）３Ｄ　ＡＮＤフラッシュメモリ構造の上方に配置されている。他の実施形態では、３Ｄフラッシュメモリ構造１１００は、３Ｄ　ＡＮＤフラッシュメモリ構造であり、ヒータ１２００は、（図４Ａ～図４Ｃに示すように）３Ｄ　ＡＮＤフラッシュメモリ構造間のスリットトレンチ１１１０内に配置されている。

【００３０】

図４Ａは、本開示の別の実施形態によるヒータを有するメモリチップの部分上面図である。図４Ｂは、本開示の別の実施形態によるメモリチップのヒータおよびパッドの部分上面図である。図４Ｃは、図４ＢのＩＩ－ＩＩ’線に沿った断面図である。

【００３１】

図４Ａおよび図４Ｃを参照すると、複数のヒータ１２００が、３Ｄフラッシュメモリ構造１１００間のスリットトレンチ１１１０内に配置されている。ヒータ１２００は、ゲートスタック構造５２の複数のゲート層３８および複数の絶縁層５４の周りに配置されている。ヒータ１２００は、（図４Ｃに示すように）絶縁ライナ層１１１２によってゲート層３８および絶縁層５４から分離されている。絶縁ライナ層１１１２は、酸化ケイ素または窒化ケイ素などの絶縁材料を含む。ヒータ１２００は、銅またはタングステンなどの金属層１２０２を（図４Ｃに示すように）含む。一部の実施形態では、ヒータ１２００は、（図４Ｃに示すように）バリア層１２０４をさらに含む。バリア層１２０４は、絶縁ライナ層１１１２と金属層１２０２との間に配置されている。バリア層１２０４は、例えば、チタン、タンタル、窒化チタン、窒化タンタル、またはそれらの組合せである。

【００３２】

一部の実施形態では、１つのヒータ１２００がスリットトレンチ１１１０のそれぞれに配置されている。例えば、ヒータ１２００は、Ｘ方向に延在することができる。一実施形態において、ヒータ１２００は、（図４Ａおよび図４Ｂに示すように）アレイ領域ＡＲ内に配置され、階段領域ＳＲまで延在する。一実施形態において、ヒータ１２００は、アレイ領域ＡＲ内に配置されてもよいが、階段領域ＳＲには配置されない（図示せず）。言い換えれば、ヒータ１２００の長さは、３Ｄフラッシュメモリ構造１１００のＸ方向の長さよりも大きくても、等しくても、または小さくてもよい。

【００３３】

あるいは、複数のヒータ１２００がスリットトレンチ１１１０のそれぞれに配置されてもよい。例えば、アレイ領域ＡＲおよび階段領域ＳＲにそれぞれ１つのヒータ１２００を設けてもよく、別々に加熱が行われてもよい（図示せず）。しかしながら、本開示の実施形態は、これに限定されない。

【００３４】

加えて、図４Ａを参照すると、上面図におけるヒータ１２００の形状は、例えば、矩形または別の形状である。複数のスリットトレンチ１１１０内のヒータ１２００は、同じ幅または異なる幅を有することができる。しかしながら、本開示は、これに限定されない。ヒータ１２００の形状は、実際の要件または設計に従って変更されてもよい。

【００３５】

図４Ｂおよび図４Ｃを参照すると、コンタクトＣ３が各ヒータ１２００の２つの端部（すなわち、Ｅ１およびＥ２）の表面上にそれぞれ配置されている。コンタクトＣ３は、上部相互接続構造１０３４を介して上方のパッド１３０２ａおよび１３０２ｂに接続されてもよく、それにより、メモリチップ１０００のヒータ１２００は、上部相互接続構造１０３４ならびにパッド１３０２ａおよび１３０２ｂを介して制御チップ２０００に電気的に接続され得る。パッド１３０２ａ、１３０２ｂの材料は、例えば、銅である。

【００３６】

図５Ａ～図５Ｅは、本開示の一実施形態による制御チップの概略斜視図である。図６Ａは、本開示の一実施形態によるメモリチップおよび制御チップの概略斜視図である。図６Ｂは、図６Ａの概略回路図である。

【００３７】

図５Ａを参照すると、制御チップ２０００は、複数のタイルＴ’を含むことができる。タイルＴ’は、アレイ状に配置されてもよい。本実施形態では、４つのタイルＴ’（例えば、Ｔ１’～Ｔ４’）が説明のための例として取り上げられる。４つのタイルＴ’のうち、タイルＴ１’とタイルＴ２’は、１つの行に配置され、タイルＴ３’とタイルＴ４’は、別の行に配置されている。タイルＴ１’とタイルＴ３’は、１つの列に配置され、タイルＴ２’とタイルＴ４’は、別の列に配置されている。

【００３８】

図５Ａおよび図５Ｅを参照すると、各タイルＴ’は、複数の駆動行２０００Ｒおよび駆動列２０００Ｃを含む。図５Ｅに示すように、駆動行２０００Ｒのそれぞれは、第２のトランジスタ２０２０と、第２の相互接続構造２０３０と、パッド２０５２と、を含む。第２のトランジスタ２０２０は、第２の基板２０１０の活性領域２０１２上に配置されている。第２の基板２０１０は、シリコン基板などの半導体基板であってもよい。第２のトランジスタ２０２０は、相補型金属酸化膜半導体（ＣＭＯＳ）トランジスタであってもよい。第２のトランジスタ２０２０は、（図５Ａ～図５Ｅに示すような）プレーナトランジスタ、または（図８に示すような）フィントランジスタであってよい。

【００３９】

図５Ｅおよび図８を参照すると、第２のトランジスタ２０２０は、ゲート誘電体層２０２４、ゲート層２０２８、ソース領域２０２２ａ、およびドレイン領域２０２２ｂを含む。ゲート誘電体層２０２４は、例えば、酸化ケイ素または高誘電率材料である。ゲート層２０２８は、例えば、ドープされたポリシリコンまたはタングステンである。ゲート層２０２８は、ゲート誘電体層２０２４上に位置する。ゲート層２０２８は、ストリップ形状

を有し、その延在方向は、例えば、図６Ａに示すように、（例えば、Ｘ方向に延在する）ヒータ１２００の延在方向と同じである。一部の実施形態では、図５Ａに示すように、２つの隣接する行（例えば、タイルＴ１’とＴ２’、またはタイルＴ３’とＴ４’）の第２のトランジスタ２０２０のゲート層２０２８は、電気的に接続されていてもよい。

【００４０】

図５Ｃおよび図５Ｅを参照すると、第２のトランジスタ２０２０のソース領域２０２２ａおよびドレイン領域２０２２ｂは、ゲート層２０２８の両側の活性領域２０１２内に配置されている。ソース領域２０２２ａおよびドレイン領域２０２２ｂは、Ｎ型またはＰ型ドーパントなどのドーパントを含む。一部の実施形態では、２つの隣接する第２のトランジスタ２０２０は、ソース領域２０２２ａを共有する。

【００４１】

図５Ｂおよび図５Ｃを参照すると、第２の相互接続構造２０３０は、第２のトランジスタ２０２０上に配置されている。第２の相互接続構造２０３０は、（図５Ｃに示すような）誘電体層２０３１と、誘電体層２０３１内に位置する複数のコンタクト２０３２および２０３４と、複数の導電線２０３６および２０４０と、複数のビア２０３８および２０４２と、を含む。コンタクト２０３２は、ソース領域２０２２ａおよびドレイン領域２０２２ｂ上にそれぞれ着地し、ソース領域２０２２ａおよびドレイン領域２０２２ｂと電気的に接続されている。コンタクト２０３４は、ゲート層２０２８上に着地し、ゲート層２０２８に電気的に接続されている。コンタクト２０３２は、図５Ｂおよび図５Ｄに示すように、Ｘ方向に沿って延在するストリップ形状を有し、ゲート層２０２８と実質的に平行である。コンタクト２０３４の形状は、コンタクト２０３２の形状とは異なり、例えば、図５Ｂに示すように、柱状形状であってもよい。（図５Ｃに示すような）導電線２０３６および２０４０は、それぞれコンタクト２０３２および２０３４上に配置されている。導電線２０３６および導電線２０４０は、ビア２０３８によって互いに電気的に絶縁されている。ビア２０４２は、導電線２０４０上に配置され、導電線２０４０を上方の接合層２０５０に電気的に接続する。誘電体層２０３１は、例えば、酸化ケイ素である。コンタクト２０３２および２０３４、導電線２０３６および２０４０、ならびにビア２０３８および２０４２は、タングステンまたは銅などの金属層を含む。コンタクト２０３２および２０３４、導電線２０３６および２０４０、ならびにビア２０３８および２０４２は、チタン、タンタル、窒化チタン、窒化タンタル、またはそれらの組合せなどのバリア層（図示せず）をさらに含むことができる。

【００４２】

図５Ｃを参照すると、駆動行２０００Ｒのそれぞれのパッド２０５２は、制御チップ２０００の接合層２０５０の一部である。接合層２０５０は、パッド２０５２および絶縁層２０５４を有する。絶縁層２０５４は、第２の相互接続構造２０３０上に位置する。パッド２０５２は、絶縁層２０５４内に位置し、第２の相互接続構造２０３０のビア２０４２に電気的に接続されている。パッド２０５２の材料は、例えば、銅である。絶縁層２０５４の材料は、例えば、酸化ケイ素である。

【００４３】

図５Ａおよび図５Ｅを参照すると、パッド２０５２は、パッド２０５２ａおよびパッド２０５２ｂを含む。具体的には、駆動行２０００Ｒのそれぞれは、Ｘ方向に沿って配置された一対のパッド２０５２ａ、２０５２ｂを含む。図１Ａ、図１Ｂ、および図６Ａに示すように、パッド２０５２ａは、ヒータ１２００の第１の端部Ｅ１に電気的に接続され、パッド２０５２ｂは、ヒータ１２００の第２の端部Ｅ２に電気的に接続され、接地されている。図５Ｃおよび図５Ｄを参照すると、各パッド２０５２ａは、ビア２０４２ａを介して下方の導電線２０４０ａに電気的に接続されている。同じタイルＴ’内の導電線２０４０ａは、図５Ａおよび図５Ｃに示すように、第２のトランジスタ２０２０のドレイン領域２０２２ｂにそれぞれ電気的に接続されるように、互いに分離され、電気的に絶縁されている。各パッド２０５２ｂは、図５Ｄに示すように、ビア２０４２ｂを介して下方の導電線２０４０ｂに電気的に接続されている。同じ列内のタイルＴ’（例えば、タイルＴ１’とＴ

３’、またはタイルＴ２’とＴ４’）のパッド２０５２ｂは、図５Ａおよび図５Ｄに示すように、同じ導電線２０４０ｂを介して接地に電気的に接続されている。

【００４４】

図５Ｃ、図１Ａ、および図１Ｂを参照すると、制御チップ２０００の接合層２０５０とメモリチップ１０００の接合層１３００とが互いに接合されて、接合構造３０００が形成されている。具体的には、制御チップ２０００の絶縁層２０５４の位置と、メモリチップ１０００の絶縁層１３０４の位置とが互いに対応し、互いに接合されている。制御チップ２０００のパッド２０５２ａ、２０５２ｂの位置と、メモリチップ１０００のパッド１３０２ａ、１３０２ｂの位置とが互いに対応し、互いに接合されている。

【００４５】

図５Ａ、図５Ｃ、および図５Ｄを参照すると、制御チップ２０００の列２０００Ｃは、同じ列内のタイルＴ’（例えば、タイルＴ１’とＴ３’、またはタイルＴ２’とＴ４’）の複数の第２のトランジスタ２０２０の複数の共有ソース領域２０２２ａを、導電線２０４０ｃを介してグローバル電源２１００に電気的に結合している。

【００４６】

図５Ａ、図５Ｃ、および図５Ｄを参照すると、制御チップ２０００の第２のトランジスタ２０２０のドレイン領域２０２２ｂは、図５Ｃに示すように、第２の相互接続構造２０３０および接合層２０５０のパッド２０５２ａに接続されている。パッド２０５２ａは、図６Ａに示すように、メモリチップ１０００のヒータ１２００の第１の端部Ｅ１に接続されたパッド１３０２ａと電気的に接続されている。一実施形態において、制御チップ２０００の駆動行２０００Ｒのそれぞれは、図６Ａおよび図６Ｂに示すように、メモリチップ１０００の１つの対応するセクタＢの１つのヒータ１２００を制御することができる。

【００４７】

図５Ｅを参照すると、一部の実施形態において、制御チップ２０００は、列デコーダ２３００および行デコーダ２２００をさらに含む。列デコーダ２３００は、グローバル電源２１００に電気的に接続されている。列アドレス信号Ａ３およびＡ４を受信した後、列デコーダ２３００は、１つの列（例えば、図５Ａの列$２０００Ｃ_{1}$）の複数のタイル（本例では２つのタイル、例えば、図５Ａのタイル Ｔ１’およびＴ３’）を選択する。したがって、グローバル電源２１００は、第２の相互接続構造２０３０の導電線２０４０ｃ（図５Ａに示す）を介して、選択された列（例えば、図５Ａの列$２０００Ｃ_{1}$）のタイル（例えば、図５Ａのタイル Ｔ１’およびＴ３’）のそれぞれの第２のトランジスタ２０２０の共有ソース領域２０２２ａに供給される。行デコーダ２２００は、駆動行２０００Ｒの第２のトランジスタ２０２０のゲート層２０２８に電気的に接続されている。行デコーダ２２００は、行アドレス信号Ａ０～Ａ２（または制御信号と呼ばれる）を受信した後、入力された行アドレス信号をデコードして、第２のトランジスタ２０２０のうちの１つ（例えば、図５Ａの第２のトランジスタ２０２０_{1}）または複数を選択してオンにする。

【００４８】

一般に、メモリチップ１０００は、メモリアレイを制御するための制御ロジックユニットを含み、制御ロジックユニット内のレジスタは、各セクタＢのメモリアレイの消去回数のステータス信号を記憶する。消去回数が所定の回数に達すると、ステータス信号が制御チップ２０００に送られる。

【００４９】

図６Ａおよび図６Ｂを参照すると、修復プロセス中に、制御チップ２０００は、受信したステータス信号に基づいて、修復を必要とするタイルＴおよびセクタＢ（例えば、図６Ａのタイル Ｔ１のセクタＢ１）に対応する行アドレス信号および列アドレス信号を生成し、行アドレス信号および列アドレス信号をそれぞれ行デコーダ２２００および列デコーダ２３００に送信することができる。列デコーダ２３００は、受信した列アドレス信号に従って１つの列（例えば、図６Ａの列$２０００Ｃ_{1}$）を選択して、この列（例えば、図６Ａの列$２０００Ｃ_{1}$）に位置するタイル（例えば、図５Ａのタイル Ｔ１’およびＴ３’）の導電線２０４０ｃにグローバル電源２１００を供給する。行デコーダ２２００は、受信した

行アドレス信号に従って、１つの駆動行２０００Ｒ$_1$の第２のトランジスタ２０２０$_1$を選択してオンにする。したがって、電流は、グローバル電源２１００から導電線２０４０ｃを介して第２のトランジスタ２０２０$_1$のソース領域２０２２ａに流れ、第２のトランジスタ２０２０$_1$のチャネルおよびドレイン領域２０２２ｂを通過し、次いで、第２の相互接続構造２０３０およびパッド２０５２ａを介してメモリチップ１０００のパッド１３０２ａに流れ込み、次いで、ヒータ１２００（例えば、１２００$_1$）の第１の端部Ｅ１に入ることができる。その後、電流は、ヒータ１２００$_1$内を流れ、メモリチップ１０００のパッド１３０２ｂを介してヒータ１２００$_1$の第２の端部Ｅ２から流れ出て、次いで、制御チップ２０００のパッド２０５２ｂに入り、次いで、導電線２０４０ｂを介して接地に電気的に接続される。本開示の実施形態では、制御チップ２０００の第２のトランジスタ（ドライバ）２０２０（例えば、２０２０$_1$）を用いて、特定のヒータ１２００（例えば、１２００$_1$）に高い駆動電流を供給することが可能であり、その結果、ヒータ１２００（例えば、１２００$_1$）として機能する導体が加熱されて、特定のタイルＴ（例えば、Ｔ１）内の特定のセクタＢ（例えば、Ｂ１）の３Ｄフラッシュメモリ構造１１００内の電荷蓄積層を修復する。

【００５０】

図１Ａおよび図１Ｂを参照すると、一部の実施形態では、修復プロセスでは、制御チップ２０００は、１つのヒータ１２００（例えば、１２００$_1$）を駆動して、１つのセクタＢ（例えば、Ｂ１）の３Ｄフラッシュメモリ構造１１００（例えば、１１００$_1$）内の電荷蓄積層を修復することができる。図１Ｂを参照すると、他の実施形態では、修復を行う際に、制御チップ２０００は、２つのヒータ１２００（例えば、１２００$_2$および１２００$_3$）を同時に駆動して、１つのセクタＢ（例えば、Ｂ２）の３Ｄフラッシュメモリ構造１１００（例えば、１１００$_2$）内の電荷蓄積層を修復することもできる。

【００５１】

図７Ａ～図７Ｃは、本開示の３Ｄフラッシュメモリモジュールチップを製造するプロセスの概略断面図を示す。

【００５２】

図７Ａを参照すると、ウエハ１０１０Ｗが用意され、複数のメモリチップ１０００がウエハ１０１０Ｗ上に形成される。メモリチップ１０００間には、スクライブラインＳＬが設けられている。メモリチップ１０００の形成方法は、以下の通りである。図３Ｂを参照すると、１つまたは複数の能動デバイス（例えば、第１のトランジスタ）１０２０が、最初にウエハ１０１０Ｗ上に形成される。次に、下部相互接続構造１０３２が能動デバイス１０２０上に形成される。下部相互接続構造１０３２は、ダマシン、デュアルダマシンなどの任意の既知の方法によって形成することができる。その後、１つの絶縁層（例えば、酸化ケイ素）５４と別の絶縁層（図示せず、例えば、窒化ケイ素）を交互に積み重ねることによって形成された絶縁スタック構造（図示せず）が、下部相互接続構造１０３２上に形成される。次に、任意の既知の方法に従って、電荷蓄積構造４０のトンネル層１４、チャネルピラー１６、ならびに導電性ピラー３２ａおよび３２ｂが絶縁スタック構造内に形成される。トンネル層１４の材料は、酸化ケイ素などの誘電体材料であってもよい。チャネルピラー１６の材料は、ドープされていないポリシリコンなどの半導体であってもよい。導電性ピラー３２ａ、３２ｂは、例えばドープされたポリシリコンである。

【００５３】

次いで、リソグラフィおよびエッチングプロセスを実行して、絶縁スタック構造内にスリットトレンチ１１１０を形成し、絶縁スタック構造を複数のセクタＢに分割する。

【００５４】

その後、ゲート置換プロセスを実行して、ゲートスタック構造５２を形成する。最初に、エッチングプロセスを実行して、エッチング液をスリットトレンチ１１１０に注入して絶縁スタック構造内の別の絶縁層を除去し、複数の水平開口部３４を形成し、次いで水平開口部３４内にゲート層３８を形成する。一部の実施形態では、ゲート層３８が形成される前に、電荷蓄積層１２およびブロッキング層３６も水平開口部３４内に形成される。電

荷蓄積層１２は、例えば、窒化ケイ素である。ブロッキング層３６の材料は、例えば、酸化アルミニウム（Al_2O_3）、酸化ハフニウム（HfO_2）、酸化ランタン（La_2O_5）、遷移金属酸化物、ランタニド酸化物、またはそれらの組合せなどの、７以上の誘電率を有する高誘電率材料である。ゲート層３８は、例えば、タングステンである。一部の実施形態では、ゲート層３８が形成される前に、バリア層３７が形成される。バリア層３７の材料は、例えば、チタン（Ｔｉ）、窒化チタン（ＴｉＮ）、タンタル（Ｔａ）、窒化タンタル（ＴａＮ）、またはそれらの組合せである。

【００５５】

次に、スリットトレンチ１１１０にスリットＳＬＴを形成する。スリットＳＬＴを形成する方法は、ゲートスタック構造５２上およびスリットトレンチ１１１０内に絶縁性充填材料を充填し、次いでエッチバックプロセスまたは平坦化プロセスによってゲートスタック構造５２上の過剰な絶縁性充填材料を除去することを含む。絶縁性充填材料は、例えば、酸化ケイ素である。

【００５６】

その後、上部相互接続構造１０３４（ローカルビット線LBL_n、ローカルソース線LSL_n、グローバルビット線GBL_n、およびグローバルソース線GSL_nを含む）が、ゲートスタック構造５２上に形成される。上部相互接続構造１０３４は、ダマシン、デュアルダマシンなどの任意の既知の方法によって形成することができるが、本明細書では詳細に説明しない。

【００５７】

図３Ａおよび図３Ｂを参照すると、本実施形態では、上部相互接続構造１０３４（ローカルビット線LBL_n、ローカルソース線LSL_n、グローバルビット線GBL_n、およびグローバルソース線GSL_nを含む）が形成された後、上部相互接続構造１０３４の上方にヒータ１２００がさらに形成される。ヒータ１２００を形成する方法は、例えば、最初に上部相互接続構造１０３４の上方に誘電体層１０４０を形成することを含む。誘電体層１０４０の材料は、例えば、酸化ケイ素である。一部の実施形態では、誘電体層１０４０が平坦な表面を有するように、化学機械平坦化プロセスなどの平坦化プロセスがさらに実行される。その後、リソグラフィおよびエッチングプロセスを実行して、誘電体層１０４０内に複数の溝ＯＰ１を形成する。次いで、誘電体層１０４０上および溝内にバリア材料層および金属材料層を順次形成する。次に、化学機械平坦化プロセスなどの平坦化プロセスを実行して、誘電体層１０４０の表面上のバリア材料層および金属材料層を除去し、溝内にバリア層１２０４および金属層１２０２を形成する。金属材料層は、例えば、銅またはタングステンである。バリア材料層は、例えば、チタン、タンタル、窒化チタン、窒化タンタル、またはそれらの組合せである。

【００５８】

図３Ｂを参照すると、ヒータ１２００を形成した後、接合層１３００が形成される。接合層１３００の形成方法は、以下の通りである。最初に、ヒータ１２００および誘電体層１０４０上に絶縁層１３０４を形成した後、リソグラフィおよびエッチングプロセスを実行して、絶縁層１３０４に複数のパッド開口部ＯＰ２を形成する。パッド開口部ＯＰ２の底部は、ヒータ１２００を露出させる。その後、絶縁層１３０４上およびパッド開口部ＯＰ２内に導電層を形成する。次いで、化学機械平坦化プロセスなどの平坦化プロセスを実行して、絶縁層１３０４上の導電層を除去し、パッド開口部ＯＰ２内にパッド１３０２を形成する。

【００５９】

上記の実施形態では、メモリチップ１０００のヒータ１２００は、上部相互接続構造１０３４が形成された後に形成される。他の実施形態では、メモリチップ１０００のヒータ１２００は、上部相互接続構造１０３４が形成される前に形成されてもよい。

【００６０】

図４Ｃを参照すると、メモリチップ１０００のヒータ１２００は、３Ｄフラッシュメモリ構造１１００のゲートスタック構造５２が形成された後、上部相互接続構造１０３４（

ローカルビット線ＬＢＬ$_n$、ローカルソース線ＬＳＬ$_n$、グローバルビット線ＧＢＬ$_n$、およびグローバルソース線ＧＳＬ$_n$を含む）が形成される前に、ゲートスタック構造５２間のスリットトレンチ１１１０内に形成される。

【００６１】

図４Ａおよび図４Ｃを参照すると、ヒータ１２００を形成する方法は、例えば、最初にスリットトレンチ１１１０内にライナ材料層を形成することを含む。ライナ材料層は、例えば、酸化ケイ素または窒化ケイ素である。次に、ゲートスタック構造５２上およびスリットトレンチ１１１０内にバリア材料層および金属材料層を順次形成する。次いで、化学機械平坦化プロセスなどの平坦化プロセスを実行して、ゲートスタック構造５２の表面上のバリア材料層および金属材料層を除去し、スリットトレンチ１１１０内に絶縁ライナ層１１１２、バリア層１２０４、および金属層１２０２を形成する。金属材料層は、例えば、銅またはタングステンである。バリア材料層は、例えば、チタン、タンタル、窒化チタン、窒化タンタル、またはそれらの組合せである。

【００６２】

図４Ｂおよび図４Ｃを参照すると、ヒータ１２００が形成された後、上部相互接続構造１０３４（ローカルビット線ＬＢＬ$_n$、ローカルソース線ＬＳＬ$_n$、グローバルビット線ＧＢＬ$_n$、およびグローバルソース線ＧＳＬ$_n$を含む）が形成される。その後、接合層１３００が、上述の方法に従って上部相互接続構造１０３４上に形成される。

【００６３】

図７Ａを参照すると、複数の制御チップ２０００が設けられている。制御チップ２０００の形成方法は、以下の通りである。図５Ｃを参照すると、第２のトランジスタ２０２０が第２の基板（ウエハ）２０１０上に形成される。次いで、第２の相互接続構造２０３０が第２のトランジスタ２０２０上に形成される。第２の相互接続構造２０３０は、ダマシン、デュアルダマシンなどの任意の既知の方法によって形成することができる。その後、接合層２０５０が、上述の方法に従って第２の相互接続構造２０３０上に形成される。次に、ダイシングを行って複数の制御チップ２０００を形成する。

【００６４】

図７Ｂを参照すると、制御チップ２０００の接合層２０５０とメモリチップ１０００の接合層１３００とが接合されて、接合構造３０００が形成されている。接合方法は、例えば、ハイブリッド接合プロセスである。一部の実施形態では、制御チップ２０００がウエハ１０１０W上のメモリチップ１０００と接合された後、制御チップ２０００の側壁の周りに封止層（図示せず）がさらに形成される。

【００６５】

図７Ｃを参照すると、ダイシングプロセスを行って互いに独立した複数の３Ｄフラッシュメモリモジュールチップ５０００を形成する。

【００６６】

以上をまとめると、本開示では、メモリチップと制御チップを接合して３Ｄフラッシュメモリモジュールチップを形成する。制御チップのドライバが高い駆動電流を供給してメモリチップ内のヒータを加熱することで、フラッシュメモリの電荷蓄積構造を修復して、より高い消去速度を達成し、フラッシュメモリの耐久性を向上させることが可能である。さらに、制御チップは、メモリチップの制御ロジックユニットのステータス信号に従って、対応するセクタを局所的に加熱することができる。加えて、接合によって形成された３Ｄフラッシュメモリモジュールチップでは、制御チップを別個に製造することができ、メモリチップ内に大面積のヒータコントローラを形成する必要がない。したがって、ヒータコントローラがメモリチップの面積を占有することを防止することができ、制御チップをそれほど高度でないプロセスで製造することができ、プロセスのコストを削減することができる。

【産業上の利用可能性】

【００６７】

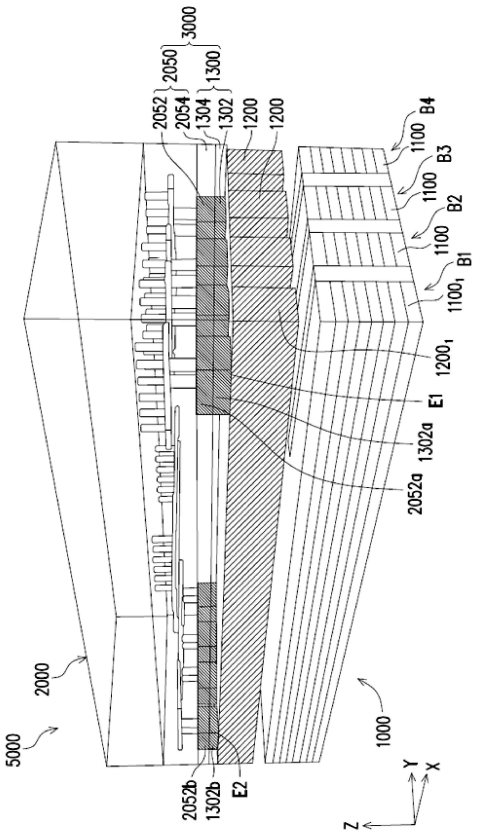
本発明の３Ｄフラッシュメモリモジュールチップおよびその製造方法は、３Ｄメモリデ

バイスおよびその製造方法に適用することができる。
【符号の説明】
【００６８】
１４：トンネル層
１６：チャネルピラー
２０：メモリセル
２４：絶縁性充填層
２８：絶縁性ピラー
３２ａ：導電性ピラー／ソースピラー
３２ｂ：導電性ピラー／ドレインピラー
３４：水平開口部
３６：ブロッキング層
３７，１２０４：バリア層
３８，２０２８：ゲート層
４０：電荷蓄積構造
５２：ゲートスタック構造
５４，１３０４，２０５４：絶縁層
１０００：メモリチップ
１０１０：第１の基板
１０１０Ｗ：ウエハ
１０２０：能動デバイス（第１のトランジスタ）
１０３０：第１の相互接続構造
１０３２：下部相互接続構造
１０３４：上部相互接続構造
１０４０，２０３１：誘電体層
１１００，１１００_{1}，１１００_{2}：３Ｄフラッシュメモリ構造
１１１０：スリットトレンチ
１１１２：絶縁ライナ層
１２００，１２００_{1}，１２００_{2}，１２００_{3}：ヒータ
１２０２：金属層
１３００，２０５０：接合層
３０００：接合構造
１３０２，１３０２ａ，１３０２ｂ，２０５２，２０５２ａ，２０５２ｂ：パッド
２０００：制御チップ
２０００Ｃ，$２０００Ｃ_{1}$：列
２０００Ｒ，$２０００Ｒ_{1}$：駆動行
２０１０：第２の基板
２０１２：活性領域
２０２０，２０２０_{1}：第２のトランジスタ
２０２２ａ：ソース領域
２０２２ｂ：ドレイン領域
２０２４：ゲート誘電体層
２０３０：第２の相互接続構造
２０３２，２０３４，Ｃ１，Ｃ２，Ｃ３：コンタクト
２０３６，２０４０，２０４０ａ，２０４０ｂ，２０４０ｃ：導電線
２０３８，２０４２，２０４２ａ，２０４２ｂ：ビア
２１００：グローバル電源
２２００：行デコーダ
２３００：列デコーダ
５０００：３Ｄフラッシュメモリモジュールチップ

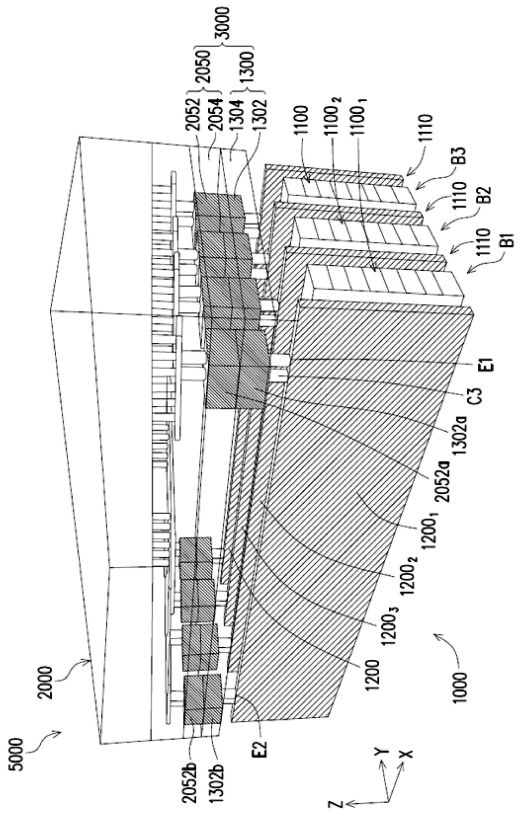
ＡＯ，ＡＩ，ＡＡ２：行アドレス信号
ＡＡ３，ＡＡ４：列アドレス信号
ＡＲ：アレイ領域
ＢＢ，ＢＢ１，ＢＢ２，ＢＢ３，ＢＢ４：セクタ
ＢＭ１：下部第１金属層
ＢＭ２：下部第２金属層
ＢＭ３：下部第３金属層
ＢＶ１，ＢＶ２，ＴＶ１：ビア
ＴＭ１：上部第１金属層
ＴＭ２：上部第２金属層
ＥＥ１：第１の端部
ＥＥ２：第２の端部
ＯＰ１：溝
ＯＰ２：パッド開口部
ＳＣ：階段構造
ＳＬ：スクライブライン
ＳＬＴ：スリット
ＳＲ：階段領域
ＴＴ，ＴＴ’，ＴＴ１，ＴＴ１’，ＴＴ２，ＴＴ２’，ＴＴ３，ＴＴ３’，ＴＴ４，ＴＴ４’：タイル
ＷＷ１，ＷＷ２：幅
Ｉ－Ｉ’，ＩＩ－ＩＩ’：線
ＸＸ，ＹＹ，ＺＺ：方向

【図面】

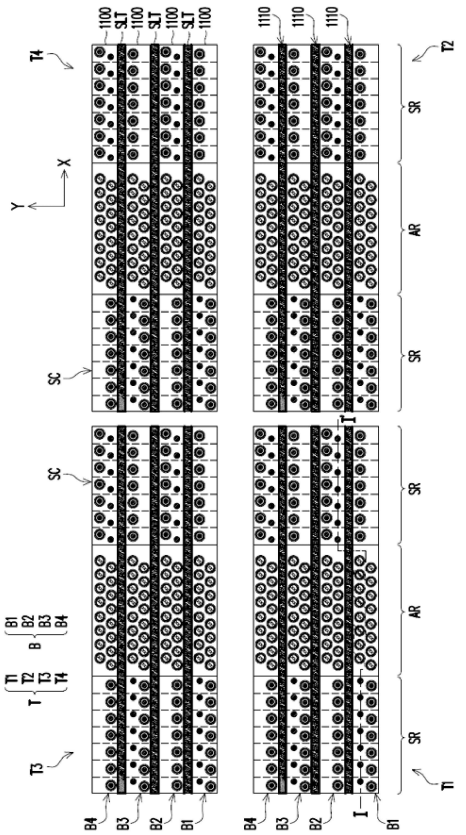
【図１Ａ】



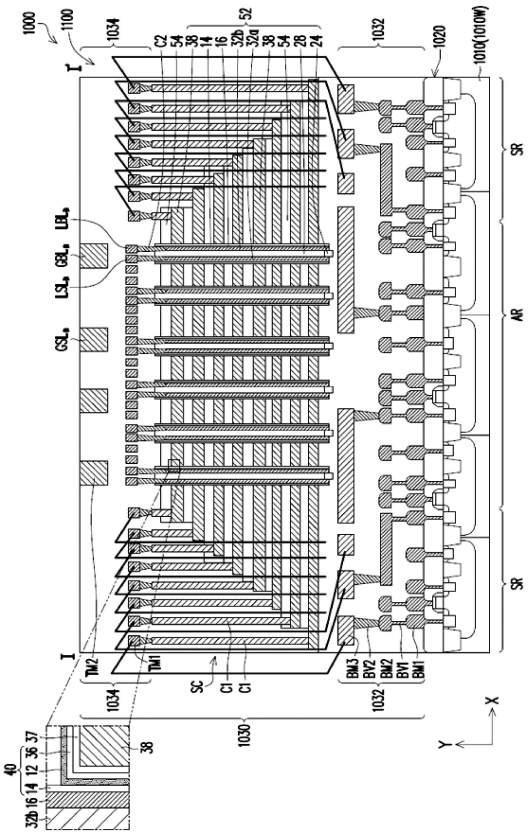
【図１Ｂ】



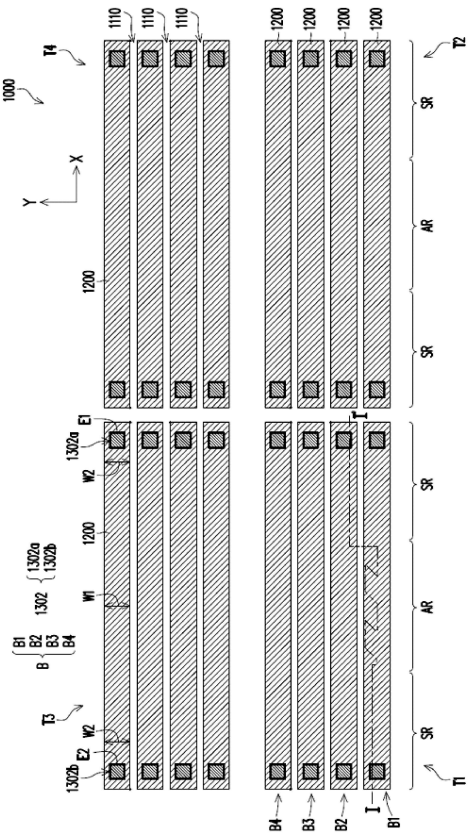
【図２Ａ】



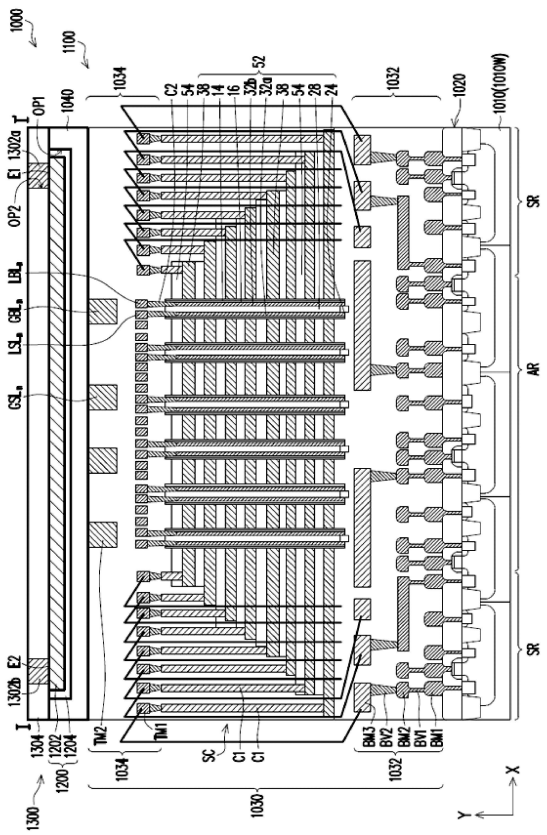
【図２Ｂ】



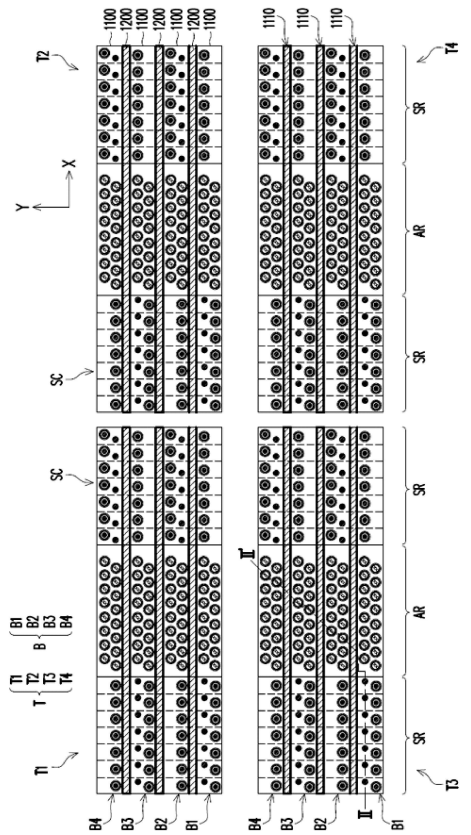
【図３Ａ】



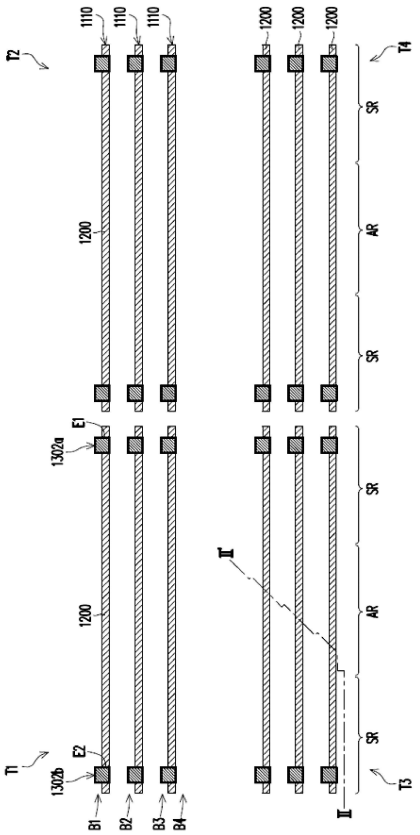
【図３Ｂ】



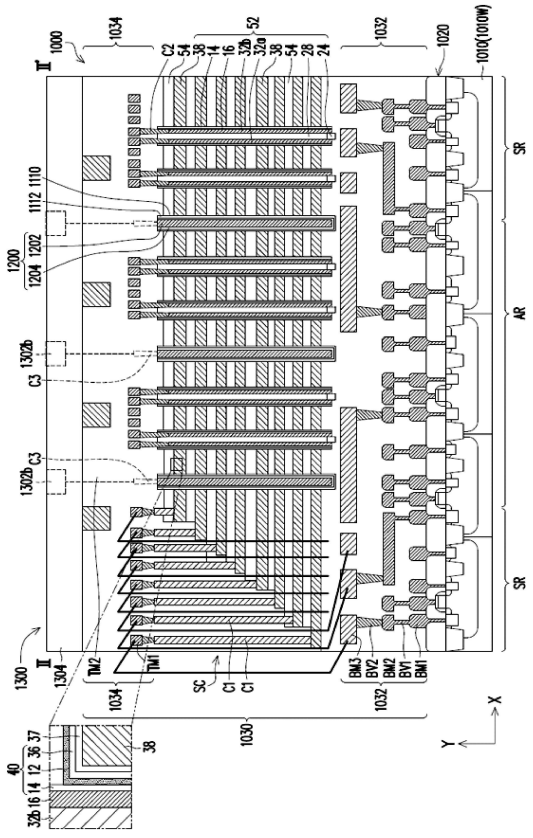
【図４Ａ】



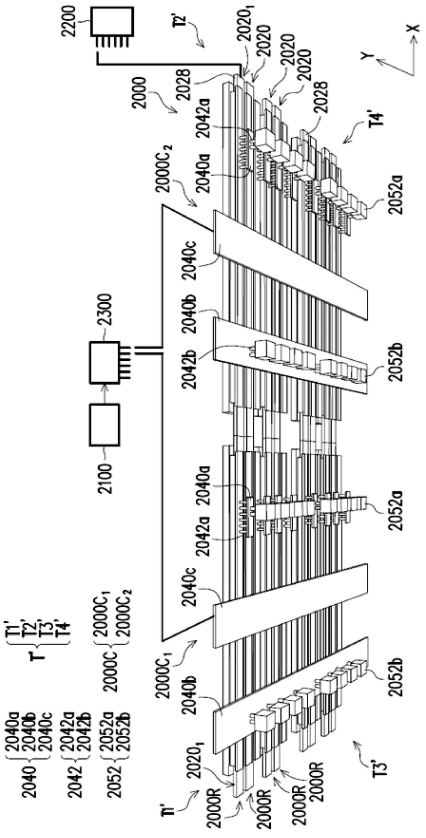
【図４Ｂ】



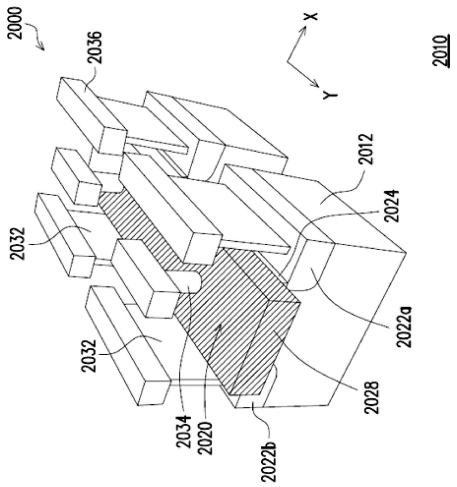
【図４Ｃ】



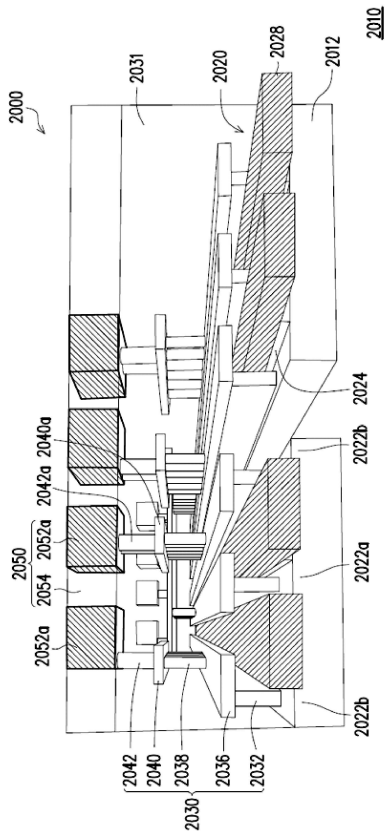
【図５Ａ】



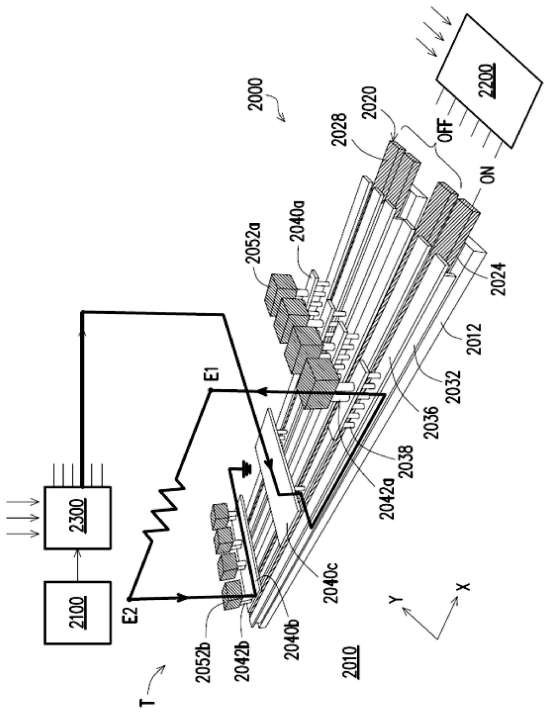
【図５Ｂ】



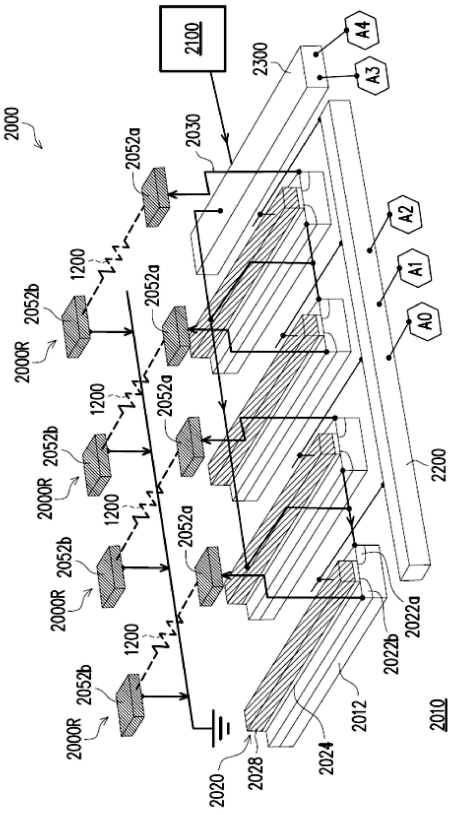
【図５Ｃ】



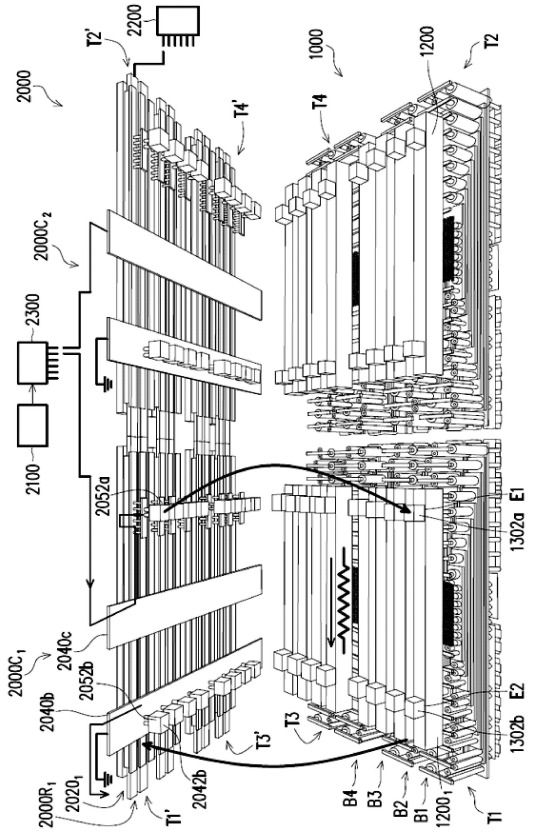
【図５Ｄ】



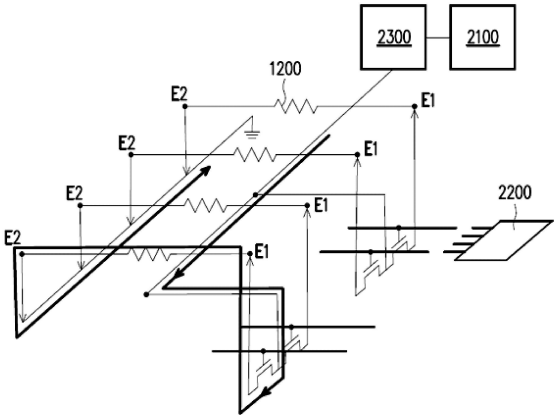
【図５Ｅ】



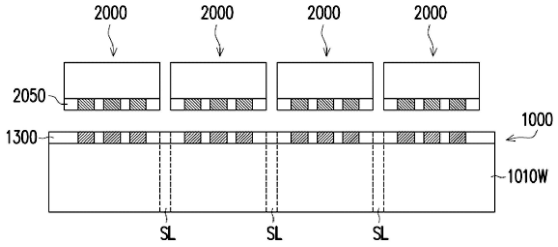
【図６Ａ】



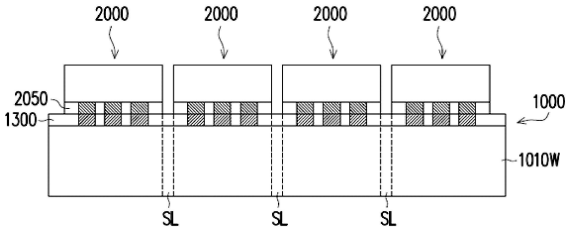
【図６Ｂ】



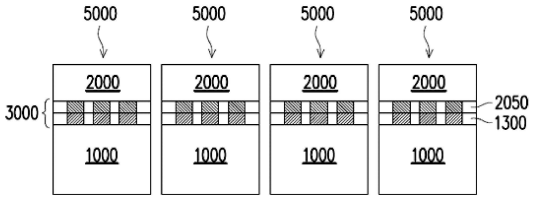
【図７Ａ】



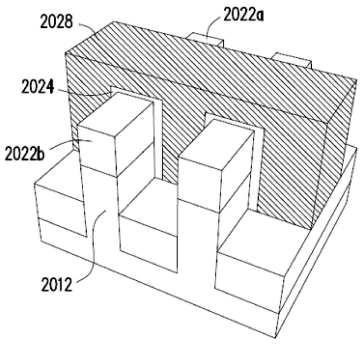
【図７Ｂ】



【図７Ｃ】



【図８】



【外国語明細書】

File: 112718jpf

3D FLASH MEMORY MODULE CHIP AND METHOD OF FABRICATING THE SAME

BACKGROUND

[Technical Field]

[0001] The embodiment of the disclosure relates to a semiconductor module and a method of fabricating the same, and particularly, to a 3D flash memory module and a method of fabricating the same.

[Description of Related Art]

[0002] Since a non-volatile memory has the advantage that stored data does not disappear at power-off, it becomes a widely used memory for a personal computer or other electronics equipment. Currently, the three-dimensional (3D) memory commonly used in the industry includes a NOR flash memory and a NAND flash memory. In addition, another type of 3D memory is an AND flash memory, which can be applied to a multi-dimensional memory array with high integration and high area utilization, and has an advantage of a fast operation speed. Therefore, the development of a 3D memory device has gradually become the current trend.

SUMMARY

TECHNICAL PROBLEM

[0003] The disclosure provides a 3D flash memory module chip and a method of fabricating the same, which can perform a local healing process on a flash memory.

SOLUTION TO PROBLEM

[0004] In an embodiment of the disclosure, a 3D flash memory module chip includes a memory chip and a control chip. The memory chip includes a plurality of tiles and a plurality of heaters. The tiles each include a plurality of 3D flash memory structures. The heaters are disposed around

the 3D flash memory structures of each of the tiles. The control chip is bonded with the memory chip to drive at least one of the heaters.

[0005] In an embodiment of the disclosure, a method of fabricating a 3D flash memory module chip includes the following steps. A memory chip is formed, the step including forming a plurality of tiles on a first substrate, each of the tiles including a plurality of 3D flash memory structures; and forming a plurality of heaters around the 3D flash memory structures of each of the tiles. A control chip is formed. The control chip and the memory chip are bonded, and the control chip is configured to drive the heaters.

EFFECTS OF INVENTION

[0006] Based on the above, in the 3D flash memory module chip and the method of fabricating the same according to the disclosure, an additional control chip is used to drive the heater to perform a local healing process on each sector of the flash memory. The control chip may be manufactured separately to prevent the heater controller from occupying the area of the memory chip, and the control chip may be manufactured by a less advanced process to reduce the cost of the process.

BRIEF DESCRIPTION OF THE DRAWINGS

[0007] FIG. 1A and FIG. 1B are respectively schematic perspective views of a 3D flash memory module chip according to an embodiment of the disclosure.

[0008] FIG. 2A is a partial top view of a 3D flash memory structure of a memory chip according to an embodiment of the disclosure.

[0009] FIG. 2B is a cross-sectional view taken along line I-I' of FIG. 2A.

[0010] FIG. 3A is a partial top view of a memory chip having a heater according to another embodiment of the disclosure.

[0011] FIG. 3B is a cross-sectional view taken along line I-I' of FIG. 3A.

[0012] FIG. 4A is a partial top view of a memory chip having a heater according to another embodiment of the disclosure.

[0013] FIG. 4B is a partial top view of a heater and a pad of a memory chip according to another embodiment of the disclosure.

[0014] FIG. 4C is a cross-sectional view taken along line II-II' of FIG. 4B.

[0015] FIG. 5A to FIG. 5E are schematic perspective views of a control chip according to an embodiment of the disclosure.

[0016] FIG. 6A is a schematic perspective view of a memory chip and a control chip according to an embodiment of the disclosure.

[0017] FIG. 6B is a schematic circuit view of FIG. 6A.

[0018] FIG. 7A to FIG. 7C show schematic cross-sectional views of a process of fabricating a 3D flash memory module chip of the disclosure.

[0019] FIG. 8 is a schematic perspective view of a control chip according to another embodiment of the disclosure.

DESCRIPTION OF THE EMBODIMENTS

[0020] The performance of a flash memory is significantly reduced after multiple operations, so it is necessary to perform a healing process on the flash memory. In the healing process, a heater may be used to heat the flash memory to heal a charge storage structure (e.g., a nitride layer) of the flash memory. In the current art, word lines are most commonly used as the heater. However, due to the large number of word lines and the complicated configuration relationship with other components (e.g., a word line decoder), the layout design of the flash memory structure may be more difficult.

[0021] The embodiments of the disclosure provide several 3D flash memory module chips, in

which a heater is disposed above, or around sidewalls of, a 3D flash memory structure of a memory chip, and the memory chip is bonded with a control chip so that the control chip can drive a heater to perform a healing process on a local sector of the memory chip.

[0022] FIG. 1A and FIG. 1B are respectively schematic perspective views of a 3D flash memory module chip according to an embodiment of the disclosure. FIG. 2A is a partial top view of a 3D flash memory structure of a memory chip according to an embodiment of the disclosure. FIG. 2B is a cross-sectional view taken along line I-I' of FIG. 2A. FIG. 3A is a partial top view of a memory chip having a heater according to another embodiment of the disclosure. FIG. 3B is a cross-sectional view taken along line I-I' of FIG. 3A.

[0023] Referring to FIG. 1A and FIG. 1B, a 3D flash memory module chip (also referred to as a 3D integrated circuit (3D IC)) 5000 according to an embodiment of the disclosure includes a memory chip 1000 and a control chip 2000. The memory chip 1000 includes a plurality of 3D flash memory structures 1100 and a plurality of heaters 1200. The heaters 1200 are disposed around the 3D flash memory structures 1100. In some embodiments, the heaters 1200 are disposed above the 3D flash memory structures 1100, as shown in FIG. 1A. In other embodiments, the heaters 1200 are disposed in slit trenches 1110 between the 3D flash memory structures 1100, as shown in FIG. 1B. The control chip 2000 is disposed above the memory chip 1000 to drive the heaters 1200 in the memory chip 1000. The control chip 2000 and the memory chip 1000 may be bonded to each other by a bonding structure 3000.

[0024] Referring to FIG. 1A and FIG. 1B, the 3D flash memory structure 1100 of the memory chip 1000 may be a 3D AND flash memory structure (as shown in FIG. 2A and FIG. 2B), a 3D NAND flash memory structure (not shown), or a 3D NOR flash memory structure (not shown). The 3D AND flash memory structure will be taken as an example to illustrate the 3D flash memory structure 1100 of the disclosure, but the embodiment of the disclosure is not limited thereto.

[0025] Referring to FIG. 2A and FIG. 2B, the memory chip 1000 may include a plurality of tiles

T. The tiles T may be arranged in an array including a plurality of columns and a plurality of rows. In this embodiment, four tiles T (e.g., T1 to T4) are shown for illustration. Among the four tiles T, the tile T1 and the tile T2 are arranged in a row, and the tile T3 and the tile T4 are arranged in another row. The tile T1 and the tile T3 are arranged in a column, and the tile T2 and the tile T4 are arranged in another column. Each of the tiles T may include a plurality of sectors B (e.g., B1 to B4). Each of the sectors B includes a 3D flash memory structure 1100. The 3D flash memory structures 1100 extend in the X direction and are arranged in the Y direction. Two adjacent 3D flash memory structures 1100 are separated from each other by a slit trench 1110.

[0026] Referring to FIG. 2B, each of the 3D flash memory structures 1100 may include at least a memory array formed by a plurality of memory cells. Specifically, the 3D flash memory structure 1100 may be disposed above one or more active devices (e.g., first transistors 1020) on a first substrate (e.g., a semiconductor substrate) 1010. The first transistor 1020 is, for example, a complementary metal-oxide-semiconductor (CMOS) field-effect transistor. Therefore, this architecture may also be referred to as a complementary metal-oxide-semiconductor field-effect transistor under array (CMOS under Array (CUA)) architecture.

[0027] Referring to FIG. 2B, the 3D flash memory structure 1100 may be disposed in a back end of line (BEOL) of a semiconductor die. For example, the 3D flash memory structure 1100 may be embedded in a first interconnect structure 1030. The first interconnect structure 1030 includes, for example, a lower interconnect structure 1032 and an upper interconnect structure 1034. The lower interconnect structure 1032 is disposed above one or more active devices (e.g., the first transistors 1020) on the first substrate (e.g., the semiconductor substrate) 1010 and below the memory array of the 3D flash memory structure 1100. The upper interconnect structure 1034 is disposed above the memory array of the 3D flash memory structure 1100. The lower interconnect structure 1032 includes, for example, a lower first metal layer BM1, a lower second metal layer BM2, and a lower third metal layer BM3, and vias BV1 and BV2 therebetween. The

upper interconnect structure 1034 includes, for example, an upper first metal layer TM1 and an upper second metal layer TM2, and vias TV1 therebetween. The numbers of the metal layers and the vias of the lower interconnect structure 1032 and the upper interconnect structure 1034 are not limited to the above.

[0028] Referring to FIG. 2B, the 3D flash memory structure 1100 includes a plurality of gate stack structures 52. Each of the gate stack structures 52 is formed on the lower interconnect structure 1032. Each of the gate stack structures 52 extends in the X direction, from an array region AR to staircase regions SR of the first substrate 1010. The gate stack structure 52 includes a plurality of gate layers (also referred to as word lines) 38 and a plurality of insulating layers 54 vertically stacked on the surface of the first substrate 1010. In the Z direction, the gate layers 38 are electrically isolated from each other by the insulating layers 54 disposed therebetween. The gate layer 38 includes a metal layer such as tungsten. In some embodiments, the gate layer 38 further includes a barrier layer 37, such as titanium (Ti), titanium nitride (TiN), tantalum (Ta), tantalum nitride (TaN), or a combination thereof. The insulating layer 54 is, for example, silicon oxide.

[0029] The gate layer 38 extends in a direction parallel to the surface of the first substrate 1010 (shown in FIG. 2D). The gate layers 38 in the staircase region SR may have a staircase structure SC (shown in FIG. 2B), so that a lower gate layer 38 is longer than an upper gate layer 38, and the end of a lower gate layer 38 extends laterally beyond the end of an upper gate layer 38. A contact C1 for connecting the gate layer 38 may land on the end of the gate layer 38 in the staircase region SR to connect each of the gate layers 38 to conductive lines of the lower interconnect structure 1032 (e.g., the conductive line of the lower third metal layer BM3) via the contact C1 and the upper interconnect structure 1034.

[0030] Referring to FIG. 2B, the 3D flash memory structure 1100 further includes a plurality of channel pillars 16. The channel pillar 16 continuously extends through the gate stack structure

52 in the array region AR. In some embodiments, the channel pillar 16 may have a ring-shaped profile in a top view. The material of the channel pillar 16 may be a semiconductor, such as undoped polysilicon.

[0031] Referring to FIG. 2B, the 3D flash memory structure 1100 further includes an insulating filling layer 24, an insulating pillar 28, a plurality of conductive pillars (e.g., serving as source pillars) 32a, and a plurality of conductive pillars (e.g., serving as drain pillars) 32b. The conductive pillars 32a and 32b and the insulating pillar 28 are disposed in the channel pillar 16 and each extend in a direction (i.e., the Z direction) perpendicular to the gate layer 38. The conductive pillars 32a and 32b are separated from each other by the insulating filling layer 24 and the insulating pillar 28 and are electrically coupled to the channel pillar 16. The conductive pillars 32a and 32b are, for example, doped polysilicon. The insulating filling layer 24 is, for example, silicon oxide, the insulating pillar 28 is, for example, silicon nitride.

[0032] Referring to FIG. 2B, a charge storage structure 40 is disposed between the channel pillar 16 and the gate layers 38. The charge storage structure 40 may include a tunneling layer (or referred to as a bandgap engineered tunneling oxide layer) 14, a charge storage layer 12, and a blocking layer 36. The charge storage layer 12 is located between the tunneling layer 14 and the blocking layer 36. In some embodiments, the tunneling layer 14, the charge storage layer 12, and the blocking layer 36 are, for example, silicon oxide, silicon nitride, and silicon oxide. In some embodiments, a part (e.g., the tunneling layer 14) of the charge storage structure 40 continuously extends in a direction (i.e., the Z direction) perpendicular to the gate layer 38, and the other part (e.g., the charge storage layer 12 and the blocking layer 36) of the charge storage structure 40 surrounds the gate layer 38, as shown in FIG. 2B. In other embodiments, the charge storage structure 40 (e.g., the tunneling layer 14, the charge storage layer 12, and the blocking layer 36) surrounds the gate layer 38 (not shown). Each of the gate layers 38, and the charge storage structure 40, the channel pillar 16, the source pillar 32a, and the drain pillar 32b that are

surrounded by the gate layer 38 define a memory cell 20. Therefore, each of the 3D flash memory structures 1100 includes at least a memory array composed of a plurality of memory cells 20.

[0033] The 3D flash memory structure 1100 further includes a local bit line LBL_n, a local source line LSL_n, a global bit line GBL_n, and a global source line GSL_n. The local bit line LBL_n and the local source line LSL_n are located in the upper first metal layer TM1 of the upper interconnect structure 1034, and are respectively electrically connected to the source pillar 32a and the drain pillar 32b via contacts C2. The global bit line GBL_n and the global source line GSL_n are respectively electrically connected to the local bit line LBL_n and the local source line LSL_n via upper vias (not shown) in the upper interconnect structure 1034.

[0034] According to different operation methods, a 1-bit operation or a 2-bit operation may be performed on the memory cell 20. For example, when a voltage is applied to the source pillar 32a and the drain pillar 32b, since the source pillar 32a and the drain pillar 32b are connected to the channel pillar 16, electrons may be transferred along the channel pillar 16 and stored in the entire charge storage structure 40. Accordingly, a 1-bit operation may be performed on the memory cell 20. In addition, for an operation involving Fowler-Nordheim tunneling, electrons or holes may be trapped in the charge storage structure 40 between the source pillar 32a and the drain pillar 32b. For an operation involving source side injection, channel-hot-electron injection, or band-to-band tunneling hot carrier injection, electrons or holes may be locally trapped in the charge storage structure 40 adjacent to one of the source pillar 32a and the drain pillar 32b. Accordingly, a single level cell (SLC, 1 bit) or multi-level cell (MLC, greater than or equal to 2 bits) operation may be performed on the memory cell 20.

[0035] During operation, a voltage is applied to a selected word line (gate layer) 38; for example, when a voltage higher than a corresponding threshold voltage (V_{th}) of the corresponding memory cell 20 is applied, a channel region of the channel pillar 16 intersecting the selected word line 38

is turned on to allow a current to enter the drain pillar 32b from a bit line BL_n, flow to the source pillar 32a via the turned-on channel region, and finally flow to a source line SL_n.

[0036] Referring to FIG. 3A and FIG. 3B, the memory chip 1000 further includes a plurality of heaters 1200. The heater 1200 may be disposed in a dielectric layer 1040 above the 3D flash memory structure 1100. The material of the dielectric layer 1040 is, for example, silicon oxide. The heater 1200 includes a metal layer 1202, such as copper or tungsten. In some embodiments, the heater 1200 further includes a barrier layer 1204, such as titanium, tantalum, titanium nitride, tantalum nitride, or a combination thereof.

[0037] Referring to FIG. 3A, in some embodiments, one heater 1200 is disposed on each sector B, and two heaters 1200 of any two adjacent sectors B are separated from each other. The heater 1200 may extend in the X direction. In an embodiment, the heater 1200 is disposed in the array region AR and extends to the staircase regions SR (as shown in FIG. 3A and FIG. 3B). In an embodiment, the heater 1200 may be disposed in the array region AR but is not disposed in the staircase regions SR (not shown). In other words, the length of the heater 1200 may be greater than, equal to, or less than the length of the 3D flash memory structure 1100 in the X direction.

[0038] In addition, multiple heaters 1200 may be disposed on each sector B; for example, one heater 1200 may be disposed in the array region AR and the staircase regions SR respectively and may perform heating separately (not shown). However, the embodiment of the disclosure is not limited thereto. In another embodiment, multiple heaters 1200 of adjacent two, three, or more sectors B may also be combined into one heater (not shown) to simultaneously heat the 3D flash memory structures 1100 of multiple sectors B.

[0039] Referring to FIG. 3A, the shape of the heater 1200 in a top view is, for example, a rectangle or another shape. The heaters 1200 on multiple sectors B may have the same width or different widths. A width W1 of the heater 1200 in the array region AR is the same as a width W2 of the heater 1200 in the staircase region SR. However, the disclosure is not limited thereto.

The shape of the heater 1200 may be changed according to the actual requirements or design. The width W1 of the heater 1200 in the array region AR may be greater than, equal to, or less than the width W2 of the heater 1200 in the staircase region SR.

[0040] Referring to FIG. 1A, FIG. 1B, and FIG. 3B, the memory chip 1000 further includes a bonding layer 1300. The bonding layer 1300 includes a pad 1302 and an insulating layer 1304. The insulating layer 1304 is disposed on the heater 1200. The material of the insulating layer 1304 is, for example, silicon oxide. The pad 1302 is disposed in the insulating layer 1304 on the surface of each of the heaters 1200. The material of the pad 1302 is, for example, copper. The pad 1302 includes pads 1302a and 1302b. The pads 1302a and 1302b are respectively connected to a first end E1 and a second end E2 of the heater 1200.

[0041] In the above embodiment, the 3D flash memory structures 1100 are 3D AND flash memory structures, and the heaters 1200 are disposed above the 3D AND flash memory structures (as shown in FIG. 3A, FIG. 3B, and FIG. 6A). In other embodiments, the 3D flash memory structures 1100 are 3D AND flash memory structures, and the heaters 1200 are disposed in slit trenches 1110 between the 3D AND flash memory structures (as shown in FIG. 4A to FIG. 4C).

[0042] FIG. 4A is a partial top view of a memory chip having a heater according to another embodiment of the disclosure. FIG. 4B is a partial top view of a heater and a pad of a memory chip according to another embodiment of the disclosure. FIG. 4C is a cross-sectional view taken along line II-II' of FIG. 4B.

[0043] Referring to FIG. 4A and FIG. 4C, a plurality of heaters 1200 are disposed in slit trenches 1110 between 3D flash memory structures 1100. The heater 1200 is disposed around a plurality of gate layers 38 and a plurality of insulating layers 54 of a gate stack structure 52. The heater 1200 is separated from the gate layers 38 and the insulating layers 54 by an insulating liner layer 1112 (as shown in FIG. 4C). The insulating liner layer 1112 includes an insulating material such as silicon oxide or silicon nitride. The heater 1200 includes a metal layer 1202 (as shown in FIG.

4C), such as copper or tungsten. In some embodiments, the heater 1200 further includes a barrier layer 1204 (as shown in FIG. 4C). The barrier layer 1204 is located between the insulating liner layer 1112 and the metal layer 1202. The barrier layer 1204 is, for example, titanium, tantalum, titanium nitride, tantalum nitride, or a combination thereof.

[0044] In some embodiments, one heater 1200 is disposed in each of the slit trenches 1110. For example, the heater 1200 may extend in the X direction. In an embodiment, the heater 1200 is disposed in the array region AR and extends to the staircase regions SR (as shown in FIG. 4A and FIG. 4B). In an embodiment, the heater 1200 may be disposed in the array region AR but is not disposed in the staircase regions SR (not shown). In other words, the length of the heater 1200 may be greater than, equal to, or less than the length of the 3D flash memory structure 1100 in the X direction.

[0045] Alternatively, multiple heaters 1200 may be disposed in each of the slit trenches 1110. For example, one heater 1200 may be provided respectively in the array region AR and the staircase region SR, and heating may be performed separately (not shown). However, the embodiment of the disclosure is not limited thereto.

[0046] Referring to FIG. 4A, in addition, the shape of the heater 1200 in a top view is, for example, a rectangle or another shape. The heaters 1200 in multiple slit trenches 1110 may have the same width or different widths. However, the disclosure is not limited thereto. The shape of the heater 1200 may be changed according to the actual requirements or design.

[0047] Referring to FIG. 4B and FIG. 4C, a contact C3 is disposed respectively on surfaces of two ends (i.e., E1 and E2) of each of the heaters 1200. The contact C3 may be connected to pads 1302a and 1302b above via an upper interconnect structure 1034, so that the heater 1200 of the memory chip 1000 can be electrically connected to the control chip 2000 via the upper interconnect structure 1034 and the pads 1302a and 1302b. The material of the pads 1302a and 1302b is, for example, copper.

[0048] FIG. 5A to FIG. 5E are schematic perspective views of a control chip according to an embodiment of the disclosure. FIG. 6A is a schematic perspective view of a memory chip and a control chip according to an embodiment of the disclosure. FIG. 6B is a schematic circuit view of FIG. 6A.

[0049] Referring to FIG. 5A, the control chip 2000 may include a plurality of tiles T'. The tiles T' may be arranged in an array. In this embodiment, four tiles T' (e.g., T1' to T4') will be taken as an example for illustration. Among the four tiles T', the tile T1' and the tile T2' are arranged in a row, and the tile T3' and the tile T4' are arranged in another row. The tile T1' and the tile T3' are arranged in a column, and the tile T2' and the tile T4' are arranged in another column.

[0050] Referring to FIG. 5A and FIG. 5E, each of the tiles T' includes a plurality of driving rows 2000R and columns 2000C. Each of the driving rows 2000R includes a second transistor 2020, a second interconnect structure 2030, and a pad 2052, as shown in FIG. 5E. The second transistor 2020 is disposed on an active region 2012 of a second substrate 2010. The second substrate 2010 may be a semiconductor substrate, such as a silicon substrate. The second transistor 2020 may be a complementary metal-oxide-semiconductor (CMOS) transistor. The second transistor 2020 may be a planar transistor (as shown in FIG. 5A to FIG. 5E) or a fin transistor (as shown in FIG. 8).

[0051] Referring to FIG. 5E, and FIG. 8, the second transistor 2020 includes a gate dielectric layer 2024, a gate layer 2028, a source region 2022a, and a drain region 2022b. The gate dielectric layer 2024 is, for example, silicon oxide or a high dielectric constant material. The gate layer 2028 is, for example, doped polysilicon or tungsten. The gate layer 2028 is located on the gate dielectric layer 2024. The gate layer 2028 has a strip shape, and its extending direction is, for example, the same as the extending direction of the heater 1200 (e.g., extending in the X direction), as shown in FIG. 6A. In some embodiments, the gate layers 2028 of the second

transistors 2020 in two adjacent rows (e.g., the tiles T1' and T2', or the tiles T3' and T4') may be electrically connected, as shown in FIG. 5A.

[0052] Referring to FIG. 5C and FIG. 5E, the source region 2022a and the drain region 2022b of the second transistor 2020 are disposed in the active region 2012 on two sides of the gate layer 1138. The source region 2022a and the drain region 2022b contain a dopant, such as an N-type or P-type dopant. In some embodiments, two adjacent second transistors 2020 share a source region 2022a.

[0053] Referring to FIG. 5B and FIG. 5C, the second interconnect structure 2030 is located on the second transistors 2020. The second interconnect structure 2030 includes a dielectric layer 2031 (as shown in FIG. 5C), and a plurality of contacts 2032 and 2034, a plurality of conductive lines 2036 and 2040, and a plurality of vias 2038 and 2042 which are located in the dielectric layer 2031. The contacts 2032 respectively land on the source region 2022a and the drain region 2022b, and are electrically connected to the source region 2022a and the drain region 2022b. The contact 2034 lands on the gate layer 2028 and is electrically connected to the gate layer 2028. The contact 2032 has a strip shape which extends along the X direction and is substantially parallel to the gate layer 2028, as shown in FIG. 5B and FIG. 5D. The shape of the contact 2034 is different from the shape of the contact 2032 and may be, for example, a columnar shape, as shown in FIG. 5B. The conductive lines 2036 and 2040 (as shown in FIG. 5C) are respectively disposed on the contacts 2032 and 2034. The conductive line 2036 and the conductive line 2040 are electrically insulated from each other by the via 2038. The via 2042 is disposed on the conductive line 2040 and is electrically connected the conductive line 2040 to a bonding layer 2050 above. The dielectric layer 2031 is, for example, silicon oxide. The contacts 2032 and 2034, the conductive lines 2036 and 2040, and the vias 2038 and 2042 include a metal layer such as tungsten or copper. The contacts 2032 and 2034, the conductive lines 2036 and 2040, and the vias 2038 and 2042 may further include a barrier layer (not shown), such as titanium, tantalum, titanium nitride, tantalum

nitride, or a combination thereof.

[0054] Referring to FIG. 5C, the pad 2052 of each of the driving rows 2000R is a part of the bonding layer 2050 of the control chip 2000. The bonding layer 2050 includes the pad 2052 and an insulating layer 2054. The insulating layer 2054 is located on the second interconnect structure 2030. The pad 2052 is located in the insulating layer 2054 and is electrically connected to the via 2042 of the second interconnect structure 2030. The material of the pad 2052 is, for example, copper. The material of the insulating layer 2054 is, for example, silicon oxide.

[0055] Referring to FIG. 5A and FIG. 5E, the pad 2052 includes a pad 2052a and a pad 2052b. Specifically, each of the driving rows 2000R includes a pair of pads 2052a and 2052b disposed along the X direction. The pad 2052a is electrically connected to the first end E1 of the heater 1200; the pad 2052b is electrically connected to the second end E2 of the heater 1200 and is grounded, as shown in FIG. 1A, FIG. 1B and FIG. 6A. Referring to FIG. 5C and FIG. 5D, each pad 2052a is electrically connected to a conductive line 2040a below via a via 2042a. The conductive lines 2040a in the same tile T' are separated and electrically isolated from each other, so as to be respectively electrically connected to the drain region 2022b of the second transistor 2020, as shown in FIG. 5A and FIG. 5C. Each pad 2052b is electrically connected to a conductive line 2040b below via a via 2042b, as shown in FIG. 5D. The pads 2052b of the tiles T' in the same column (e.g., the tiles T1' and T3', or the tiles T2' and T4') are electrically connected to the ground via the same conductive line 2040b, as shown in FIG. 5A and FIG. 5D.

[0056] Referring to FIG. 5C, FIG. 1A, and FIG. 1B, the bonding layer 2050 of the control chip 2000 and the bonding layer 1300 of the memory chip 1000 are bonded to each other to form a bonding structure 3000. Specifically, the position of the insulating layer 2054 of the control chip 2000 and the position of the insulating layer 1304 of the memory chip 1000 correspond to each other and are bonded to each other. The positions of the pads 2052a and 2052b of the control chip 2000 and the positions of the pads 1302a and 1302b of the memory chip 1000 correspond to

each other and are bonded to each other.

[0057] Referring to FIG. 5A, FIG. 5C and FIG. 5D, the column 2000C of the control chip 2000 electrically couples a plurality of shared source regions 2022a of multiple second transistors 2020 of the tiles T' in the same column (e.g., the tiles T1' and T3', or the tiles T2' and T4') to a global power supply 2100 via a conductive line 2040c.

[0058] Referring to FIG. 5A, FIG. 5C and FIG. 5D, the drain region 2022b of the second transistor 2020 of the control chip 2000 is connected to the second interconnect structure 2030 and the pad 2052a of the bonding layer 2050, as shown in FIG. 5C. The pad 2052a is electrically connected to the pad 1302a connected to the first end E1 of the heater 1200 of the memory chip 1000 as shown in FIG 6A. In an embodiment, each of the driving rows 2000R of the control chip 2000 may control one heater 1200 of one corresponding sector B of the memory chip 1000, as shown in FIG. 6A and FIG. 6B.

[0059] Referring to FIG. 5E, in some embodiments, the control chip 2000 further includes a column decoder 2300 and a row decoder 2200. The column decoder 2300 is electrically connected to the global power supply 2100. After receiving column address signals A3 and A4, the column decoder 2300 selects multiple tiles (two tiles in this example, e.g., the tiles T1' and T3' in FIG. 5A) of one column (e.g., a column $2000C_1$ in FIG. 5A). Accordingly, the global power supply 2100 is provided to the shared source regions 2022a of the second transistors 2020 of each of the tiles (e.g., the tiles T1' and T3' in FIG. 5A) of the selected column (e.g., the column $2000C_1$ in FIG. 5A) via the conductive line 2040c (shown in FIG. 5A) of the second interconnect structure 2030. The row decoder 2200 is electrically connected to the gate layers 2028 of the second transistors 2020 of the driving rows 2000R. After receiving row address signals A0 to A2 (or referred to as control signals), the row decoder 2200 decodes the inputted row address signals to select and turn on one (e.g., a second transistor 2020_1 in FIG. 5A) or more of the second transistors 2020.

File: 112718jpf

[0060] Generally, the memory chip 1000 includes a control logic unit for controlling the memory array, and the register in the control logic unit stores a status signal of an erase count of the memory array of each sector B. When the erase count reaches a predetermined count, the status signal is sent to the control chip 2000.

[0061] Referring to FIG. 6A and FIG. 6B, during the healing process, the control chip 2000 may generate a row address signal and a column address signal corresponding to the tile T and the sector B (e.g., the sector B1 of the tile T1 in FIG. 6A) that need healing based on the received status signal, and transmit the row address signal and the column address signal respectively to the row decoder 2200 and the column decoder 2300. The column decoder 2300 selects one column (e.g., the column $2000C_1$ in FIG. 6A) according to the received column address signal to provide the global power supply 2100 to the conductive line 2040c of the tiles (e.g., the tiles T1' and T3' in FIG. 5A) located in this column (e.g., the column $2000C_1$ in FIG. 6A). The row decoder 2200 selects and turns on the second transistor 2020_1 of one driving row $2000R_1$ according to the received row address signal. Therefore, current can flow from the global power supply 2100 into the source region 2022a of the second transistor 2020_1 via the conductive line 2040c, pass through the channel and the drain region 2022b of the second transistor 2020_1, then flow into the pad 1302a of the memory chip 1000 via the second interconnect structure 2030 and the pad 2052a, and then enter a first end E1 of a heater 1200 (e.g., 1200_1). Afterwards, the current flow in the heater 1200_1 and flows out a second end E2 of the heater 1200_1 via the pad 1302b of the memory chip 1000, then enters the pad 2052b of the control chip 2000, and then is electrically connected to the ground via the conductive line 2040b. In the embodiment of the disclosure, with the second transistor (driver) 2020 (e.g., 2020_1) of the control chip 2000, it is possible to provide a high driving current to a specific heater 1200 (e.g., 1200_1), so that the conductor serving as the heater 1200 (e.g., 1200_1) is heated to heal the charge storage layer in the 3D flash memory structure 1100 of a specific sector B (e.g., B1) in a specific tile T (e.g., T1).

File: 112718jpf

[0062] Referring to FIG. 1A and FIG. 1B, in some embodiments, in the healing process, the control chip 2000 may drive one heater 1200 (e.g., 1200_1) to heal the charge storage layer in the 3D flash memory structure 1100 (e.g., 1100_1) of one sector B (e.g., B1). Referring to FIG. 1B, in other embodiments, when healing is performed, the control chip 2000 may also simultaneously drive two heaters 1200 (e.g., 1200_2 and 1200_3) to heal the charge storage layer in the 3D flash memory structure 1100 (e.g., 1100_2) of one sector B (e.g., B2).

[0063] FIG. 7A to FIG. 7C show schematic cross-sectional views of a process of fabricating a 3D flash memory module chip of the disclosure.

[0064] Referring to FIG. 7A, a wafer 1010W is provided, and a plurality of memory chips 1000 are formed on the wafer 1010W. Scribe lines SL are provided between the memory chips 1000. The method of forming the memory chip 1000 is as follows. Referring to FIG. 3B, one or more active devices (e.g., first transistors) 1020 are first formed on the wafer 1010W. Next, a lower interconnect structure 1032 is formed on the active devices 1020. The lower interconnect structure 1032 may be formed by any known method, such as damascene, dual-damascene, and the like. Afterwards, an insulating stack structure (not shown) formed by alternately stacking one insulating layer (e.g., silicon oxide) 54 and another insulating layer (not shown, e.g., silicon nitride) is formed on the lower interconnect structure 1032 . Next, according to any known method, a tunneling layer 14 of a charge storage structure 40, a channel pillar 16, and conductive pillars 32a and 32b are formed in the insulating stack structure. The material of the tunneling layer 14 may be a dielectric material, such as silicon oxide. The material of the channel pillar 16 may be a semiconductor, such as undoped polysilicon. The conductive pillars 32a and 32b are, for example, doped polysilicon.

[0065] Then, lithography and etching processes are performed to form slit trenches 1110 in the insulating stack structure to divide the insulating stack structure into a plurality of sectors B.

[0066] Afterwards, a gate replacement process is performed to form a gate stack structure 52.

First, an etching process is performed to inject an etchant into the slit trenches 1110 to remove the another insulating layer in the insulating stack structure to form a plurality of horizontal openings 34 and then form gate layers 38 in the horizontal openings 34. In some embodiments, before the gate layer 38 is formed, a charge storage layer 12 and a blocking layer 36 are also formed in the horizontal opening 34. The charge storage layer 12 is, for example, silicon nitride. The material of the blocking layer 36 is, for example, a high dielectric constant material having a dielectric constant greater than or equal to 7, such as aluminum oxide (Al_1O_3), hafnium oxide (HfO_2), lanthanum oxide (La_2O_5), transition metal oxide, lanthanide oxide, or combinations thereof. The gate layer 38 is, for example, tungsten. In some embodiments, before the gate layers 38 are formed, a barrier layer 37 is formed. The material of the barrier layer 37 is, for example, titanium (Ti), titanium nitride (TiN), tantalum (Ta), tantalum nitride (TaN), or a combination thereof.

[0067] Next, slits SLT are formed in the slit trenches 1110. The method of forming the slits SLT includes filling an insulating filling material on the gate stack structure 52 and in the slit trenches 1110, and then removing the excessive insulating filling material on the gate stack structure 52 through an etch-back process or a planarization process. The insulating filling material is, for example, silicon oxide.

[0068] Afterwards, an upper interconnect structure 1034 (including a local bit line LBL_n, a local source line LSL_n, a global bit line GBL_n and a global source line GSL_n) is formed on the gate stack structure 52. The upper interconnect structure 1034 may be formed by any known method, such as damascene, dual-damascene, and the like, which shall not be described in detail herein.

[0069] Referring to FIG. 3A and FIG. 3B, in this embodiment, after the upper interconnect structure 1034 (including the local bit line LBL_n, the local source line LSL_n, the global bit line GBL_n, and the global source line GSL_n) is formed, a heater 1200 is further formed above the upper interconnect structure 1034. The method of forming the heater 1200 includes, for example,

forming a dielectric layer 1040 above the upper interconnect structure 1034 first. The material of the dielectric layer 1040 is, for example, silicon oxide. In some embodiments, a planarization process such as a chemical mechanical planarization process is further performed, so that the dielectric layer 1040 has a flat surface. Afterwards, lithography and etching processes are performed to form a plurality of grooves OP1 in the dielectric layer 1040. Then, a barrier material layer and a metal material layer are sequentially formed on the dielectric layer 1040 and in the grooves. Next, a planarization process such as a chemical mechanical planarization process is performed to remove the barrier material layer and the metal material layer on the surface of the dielectric layer 1040 and form a barrier layer 1204 and a metal layer 1202 in the groove. The metal material layer is, for example, copper or tungsten. The barrier material layer is, for example, titanium, tantalum, titanium nitride, tantalum nitride, or a combination thereof.

[0070] Referring to FIG. 3B, after the heater 1200 is formed, a bonding layer 1300 is formed. The method of forming the bonding layer 1300 is as follows. First, an insulating layer 1304 is first formed on the heater 1200 and the dielectric layer 1040, and then lithography and etching processes are performed to form a plurality of pad openings OP2 in the insulating layer 1304. The bottom of the pad opening OP2 exposes the heater 1200. Afterwards, a conductive layer is formed on the insulating layer 1304 and in the pad openings OP2. Then, a planarization process such as a chemical mechanical planarization process is performed to remove the conductive layer on the insulating layer 1304 and form pads 1302 in the pad openings OP2.

[0071] In the above embodiment, the heater 1200 of the memory chip 1000 is formed after the upper interconnect structure 1034 is formed. In other embodiments, the heater 1200 of the memory chip 1000 may be formed before the upper interconnect structure 1034 is formed.

[0072] Referring to FIG. 4C, the heater 1200 of the memory chip 1000 is formed in the slit trench 1110 between the gate stack structures 52 after the gate stack structure 52 of the 3D flash memory structure 1100 is formed and before the upper interconnect structure 1034 (including the

local bit line LBL_n, the local source line LSL_n, the global bit line GBL_n, and the global source line GSL_n) is formed.

[0073] Referring to FIG. 4A and FIG. 4C, the method of forming the heater 1200 includes, for example, forming a liner material layer in the slit trench 1110 first. The liner material layer is, for example, silicon oxide or silicon nitride. Next, a barrier material layer and a metal material layer are sequentially formed on the gate stack structure 52 and in the slit trench 1110. Then, a planarization process such as a chemical mechanical planarization process is performed to remove the barrier material layer and the metal material layer on the surface of the gate stack structure 52 and form an insulating liner layer 1112, a barrier layer 1204, and a metal layer 1202 in the slit trench 1110. The metal material layer is, for example, copper or tungsten. The barrier material layer is, for example, titanium, tantalum, titanium nitride, tantalum nitride, or a combination thereof.

[0074] Referring to FIG. 4B and FIG. 4C, after the heater 1200 is formed, an upper interconnect structure 1034 (including a local bit line LBL_n, a local source line LSL_n, a global bit line GBL_n, and a global source line GSL_n) is formed. Afterwards, a bonding layer 1300 is formed on the upper interconnect structure 1034 according to the above-mentioned method.

[0075] Referring to FIG. 7A, a plurality of control chips 2000 are provided. The method of forming the control chips 2000 is as follows. Referring to FIG. 5C, second transistors 2020 are formed on a second substrate (wafer) 2010. Then, a second interconnect structure 2030 is formed on the second transistors 2020. The second interconnect structure 2030 may be formed by any known method, such as damascene, dual-damascene, and the like. Afterwards, a bonding layer 2050 is formed on the second interconnect structure 2030 according to the above-mentioned method. Next, dicing is performed to form a plurality of control chips 2000.

[0076] Referring to FIG. 7B, the bonding layer 2050 of the control chips 2000 and the bonding layer 1300 of the memory chips 1000 are bonded to form a bonding structure 3000. The bonding

method is, for example, a hybrid bonding process. In some embodiments, after the control chips 2000 are bonded with the memory chips 1000 on the wafer 1010W, an encapsulation layer (not shown) is further formed around the sidewalls of the control chips 2000.

[0077] Referring to FIG. 7C, a dicing process is performed to form a plurality of mutually independent 3D flash memory module chips 5000.

[0078] In summary of the above, in the disclosure, the memory chip and the control chip are bonded to form the 3D flash memory module chip. With the driver of the control chip providing a high driving current to heat the heater in the memory chip, it is possible to heal the charge storage structure of the flash memory to achieve a higher erase speed and improve the endurance of the flash memory. Furthermore, the control chip can locally heat the corresponding sector according to the status signal of the control logic unit of the memory chip. In addition, in the 3D flash memory module chip formed by bonding, the control chip may be manufactured separately, and it is not required to form a large-area heater controller in the memory chip. Therefore, it is possible to prevent the heater controller from occupying the area of the memory chip, and the control chip may be manufactured by a less advanced process so as to reduce the cost of the process.

INDUSTRIAL APPLICABILITY

[0079] The 3D flash memory module chip and the method of fabricating the same of the present invention may be applied to 3D memory devices and methods of fabricating the same.

REFERENCE SIGNS LIST

[0080] 14: tunneling layer

16: channel pillar

20: memory cell

24: insulating filling layer

28: insulating pillar

32a: conductive pillar/source pillar

32b: conductive pillar/drain pillar

34: horizontal opening

36: blocking layer

37, 1204: barrier layer

38, 2028: gate layer

40: charge storage structure

52: gate stack structure

54, 1304, 2054: insulating layer

1000: memory chip

1010: first substrate

1010W: wafer

1020: active device (first transistor)

1030: first interconnect structure

1032: lower interconnect structure

1034: upper interconnect structure

1040, 2031: dielectric layer

1100, 1100_1, 1100_2: 3D flash memory structure

1110: slit trench

1112: insulating liner layer

1200, 1200_1, 1200_2, 1200_3: heater

1202: metal layer

1300, 2050: bonding layer

3000: bonding structure

1302, 1302a, 1302b, 2052, 2052a, 2052b: pad

2000: control chip

File: 112718jpf

2000C, 2000C_1: column

2000R, 2000R_1: driving row

2010: second substrate

2012: active region

2020, 2020$_1$: second transistor

2022a: source region

2022b: drain region

2024: gate dielectric layer

2030: second interconnect structure

2032, 2034, C1, C2, C3: contact

2036, 2040, 2040a, 2040b, 2040c: conductive line

2038, 2042, 2042a, 2042b: via

2100: global power supply

2200: row decoder

2300: column decoder

5000: 3D flash memory module chip

A0, A1, A2: row address signal

A3, A4: column address signal

AR: array region

B, B1, B2, B3, B4: sector

BM1: lower first metal layer

BM2: lower second metal layer

BM3: lower third metal layer

BV1, BV2, TV1: via

TM1: upper first metal layer

File: 112718jpf

TM2: upper second metal layer

E1: first end

E2: second end

OP1: groove

OP2: pad opening

SC: staircase structure

SL: scribe line

SLT: slit

SR: staircase region

T, T', T1, T1', T2, T2', T3, T3', T4, T4': tile

W1, W2: width

I-I', II-II': line

X, Y, Z: direction

File: 112718jpf

WHAT IS CLAIMED IS:

1. A 3D flash memory module chip comprising:

a memory chip comprising:

a plurality of tiles each comprising a plurality of 3D flash memory structures; and

a plurality of heaters disposed around the 3D flash memory structures of each of the tiles; and

a control chip bonded with the memory chip to drive at least one of the heaters.

2. The 3D flash memory module chip according to claim 1, wherein the heaters are disposed above the 3D flash memory structures and are adjacent to the control chip.

3. The 3D flash memory module chip according to claim 1, wherein the heaters are disposed in a plurality of slit trenches between the 3D flash memory structures.

4. The 3D flash memory module chip according to claim 1, wherein the memory chip further comprises:

a plurality of first transistors located on a first substrate;

the 3D flash memory structures located above the first transistors; and

a first interconnect structure, wherein the 3D flash memory structures are embedded in the first interconnect structure.

5. The 3D flash memory module chip according to claim 4, wherein the first interconnect structure comprises:

a lower interconnect structure located between the 3D flash memory structures and the first transistors and electrically connecting the 3D flash memory structures and the first transistors; and

an upper interconnect structure located on the 3D flash memory structures and electrically connecting the 3D flash memory structures.

6. The 3D flash memory module chip according to claim 4, wherein the control chip

comprises:

a plurality of driving rows each comprising:

a second transistor located on a second substrate, wherein a source region of the second transistor is electrically connected to a global power supply;

a first pad electrically connected to a drain region of the second transistor and electrically connected to a first end of one of the heaters; and

a second pad that is grounded and is electrically connected to a second end of the one of the heaters.

7. The 3D flash memory module chip according to claim 6, wherein the control chip further comprises:

a row decoder electrically coupled to a plurality of gate layers of the second transistors of the driving rows; and

a column decoder electrically coupled to a plurality of source regions of the second transistors and the global power supply.

8. The 3D flash memory module chip according to claim 6, wherein the control chip comprises a plurality of tiles arranged in an array, wherein the source regions of the second transistors of the tiles in a same column are electrically connected to each other.

9. The 3D flash memory module chip according to claim 6, wherein the control chip and the memory chip are bonded by a bonding structure.

10. The 3D flash memory module chip according to claim 6, wherein the plurality of 3D flash memory structures comprising a plurality of 3D AND flash memory structure, a plurality of 3D NAND flash memory structure, or a plurality of 3D NOR flash memory structure.

11. A method of fabricating a 3D flash memory module chip, comprising:

forming a memory chip, comprising:

forming a plurality of tiles on a first substrate, wherein each of the tiles comprises

a plurality of 3D flash memory structures; and

forming a plurality of heaters around the 3D flash memory structures of each of the tiles;

forming a control chip; and

bonding the control chip and the memory chip, wherein the control chip is configured to drive the heaters.

12. The method of fabricating a 3D flash memory module chip according to claim 11, wherein the heaters are formed above the 3D flash memory structures.

13. The method of fabricating a 3D flash memory module chip according to claim 11, wherein the heaters are formed in a plurality of slit trenches around the 3D flash memory structures.

14. The method of fabricating a 3D flash memory module chip according to claim 11, wherein the step of forming the memory chip further comprises:

forming a plurality of first transistors on the first substrate; and

forming the 3D flash memory structures above the first transistors.

15. The method of fabricating a 3D flash memory module chip according to claim 14, wherein the step of forming the control chip comprises:

forming a plurality of driving rows, wherein formation of each of the driving rows comprises:

forming a second transistor on a second substrate;

forming a second interconnect structure on the second transistor, wherein a source region of the second transistor is electrically coupled to a global power supply via the second interconnect structure;

forming a first pad on the second interconnect structure, wherein the first pad is electrically connected to a drain region of the second transistor via the second interconnect structure; and

forming a second pad on the second interconnect structure, wherein the second pad is electrically connected to ground via the second interconnect structure.

16. The method of fabricating a 3D flash memory module chip according to claim 15, further comprising:

electrically connecting the first pad to a first end of one of the heaters; and

electrically connecting the second pad that to a second end of the one of the heaters.

17. The method of fabricating a 3D flash memory module chip according to claim 11, wherein the control chip and the memory chip are hybrid bonded by a bonding structure.

18. The method of fabricating a 3D flash memory module chip according to claim 11, wherein the plurality of 3D flash memory structures comprising a plurality of 3D AND flash memory structure, a plurality of 3D NAND flash memory structure, or a plurality of 3D NOR flash memory structure.

File: 112718jpf

ABSTRACT OF THE DISCLOSURE

A 3D flash memory module chip includes a memory chip and a control chip. The memory chip includes a plurality of tiles and a plurality of heaters. The tiles each include a plurality of 3D flash memory structures. The heaters are disposed around the 3D flash memory structures of each of the tiles. The control chip is bonded with the memory chip to drive at least one of the heaters.

REPRESENTATIVE DRAWING: FIG. 1A

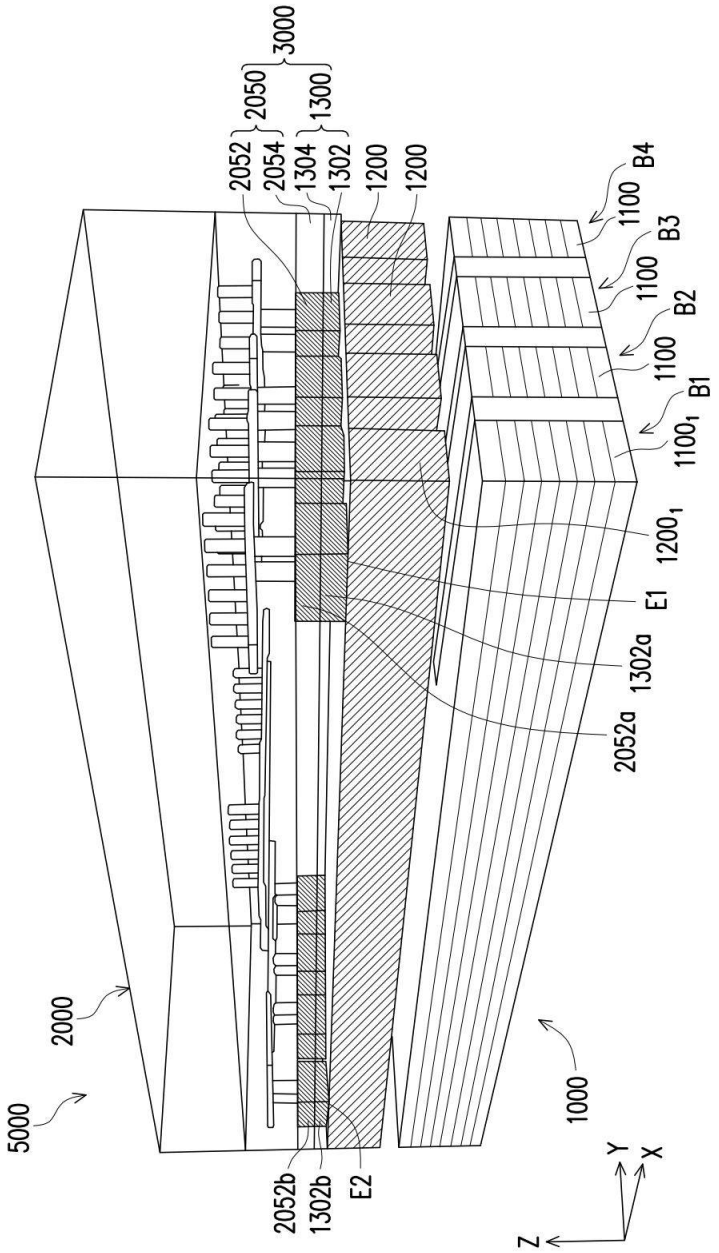

3000
2050
1300
2052
2054
1304
1302
1200
1200
B4
1100
B3
1100
B2
1100
B1
1100₁
1200₁
E1
1302a
2052a
2000
5000
1000
2052b
1302b
E2
Y
X
Z

FIG. 1A

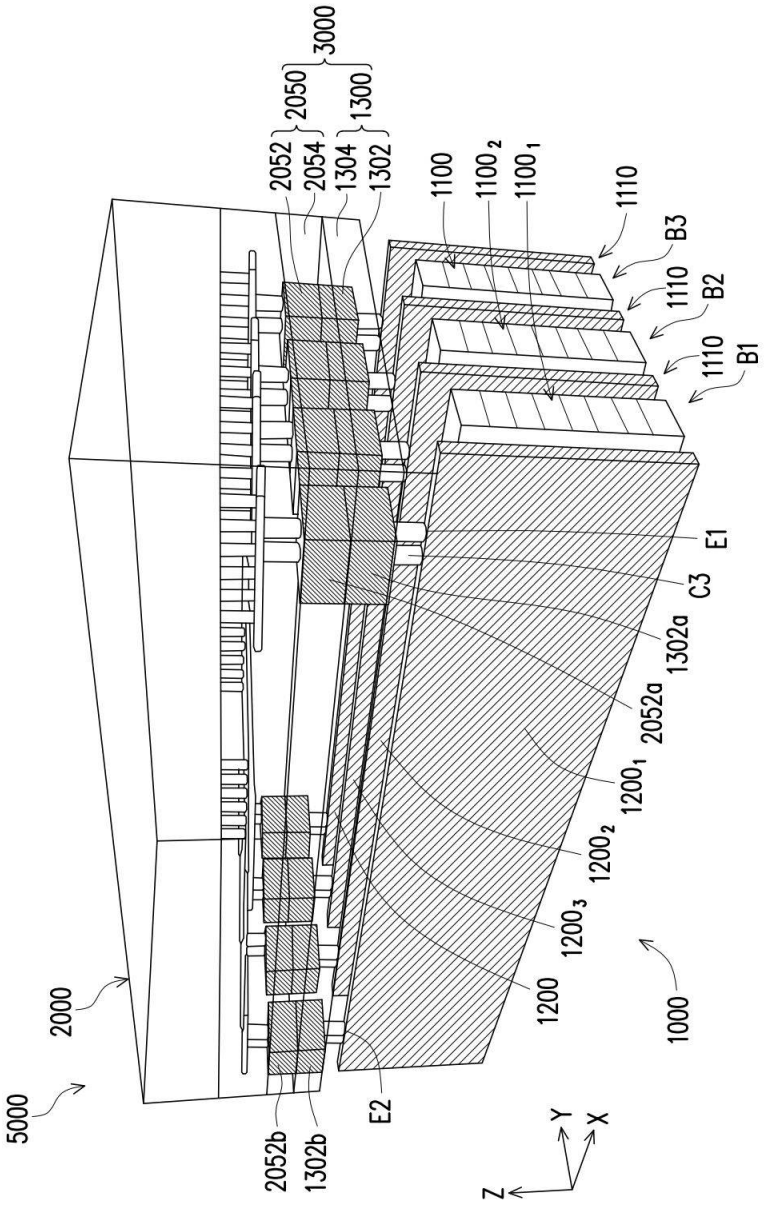

3000
2050
1300
2052
2054
1304
1302
1100
1100₂
1100₁
1110
B3
1110
B2
1110
B1
E1
C3
1302a
2052a
1200₁
1200₂
1200₃
1200
1000
2000
5000
E2
2052b
1302b
Y
X
Z

FIG. 1B

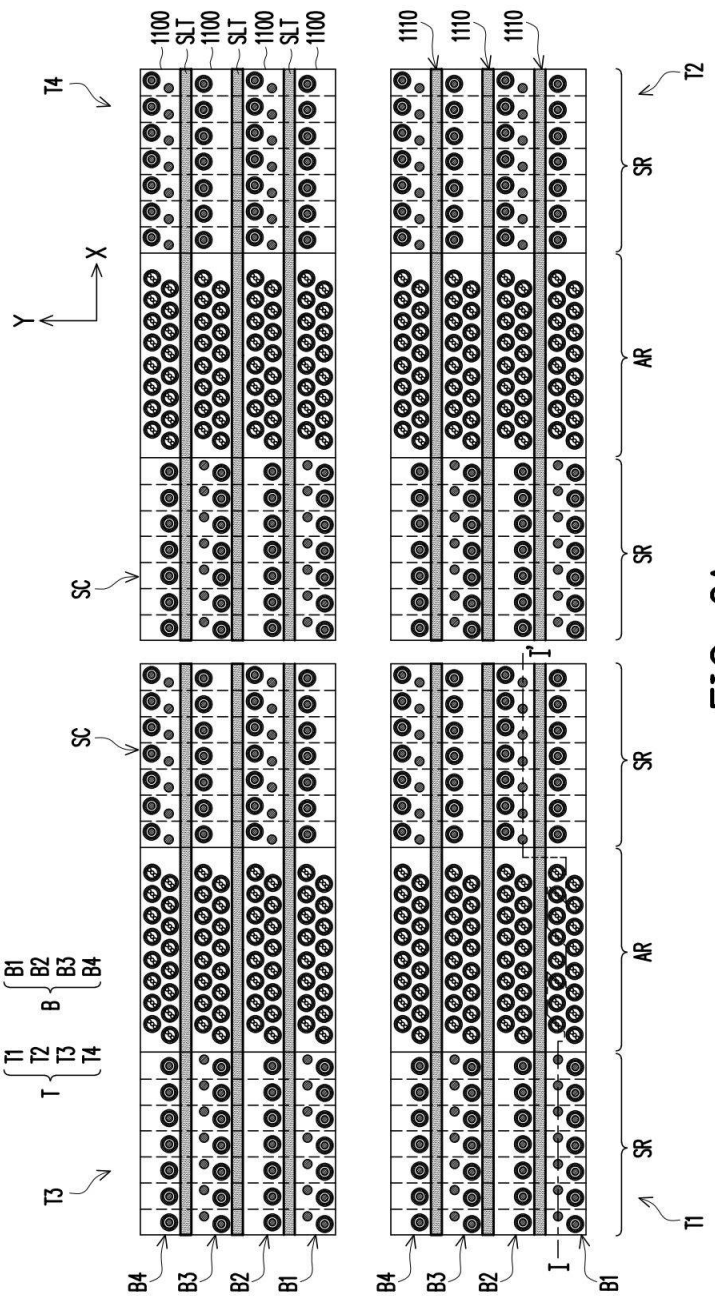


FIG. 2A

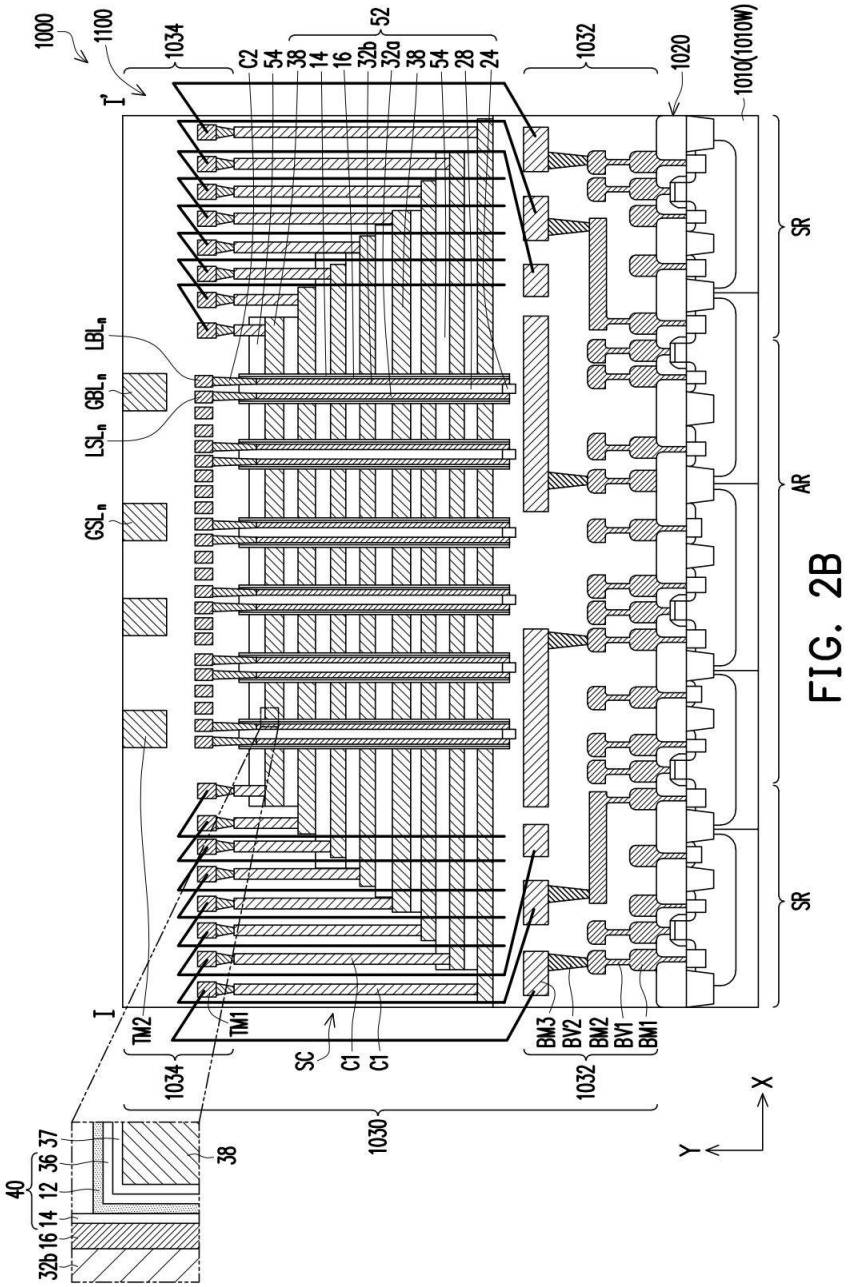

1000
1100
I'
1034
C2
54
38
14
16
32b
32a
52
38
54
28
24
1032
1020
1010(1010W)
SR
AR
SR
LBLn
GBLn
LSLn
GSLn
I
TM2
TM1
SC
C1
C1
BM3
BV2
BM2
BV1
BM1
1034
1032
1030
X
Y
40
37
36
12
14
16
32b
38

FIG. 2B

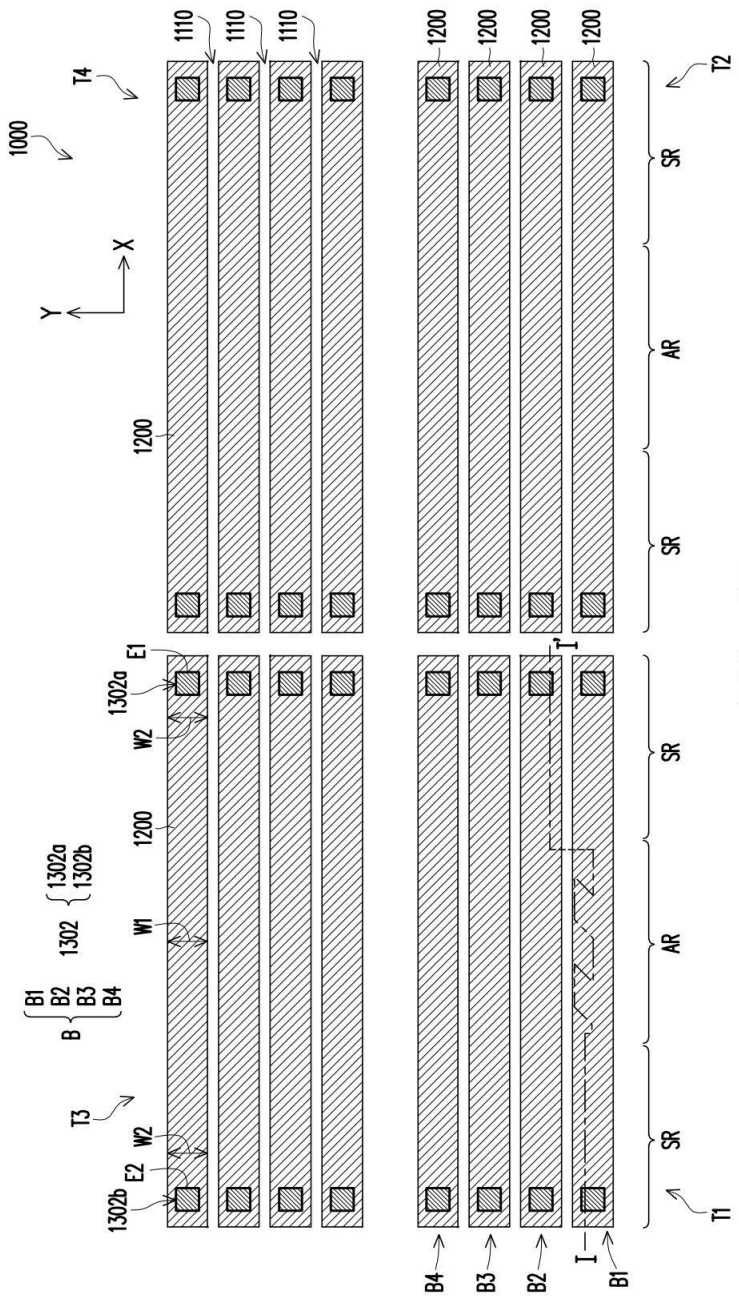

1000
T4
T2
T1
T3
X
Y
1110
1200
SR
AR
E1
E2
W1
W2
1302
1302a
1302b
B
B1
B2
B3
B4
I
I'

FIG. 3A

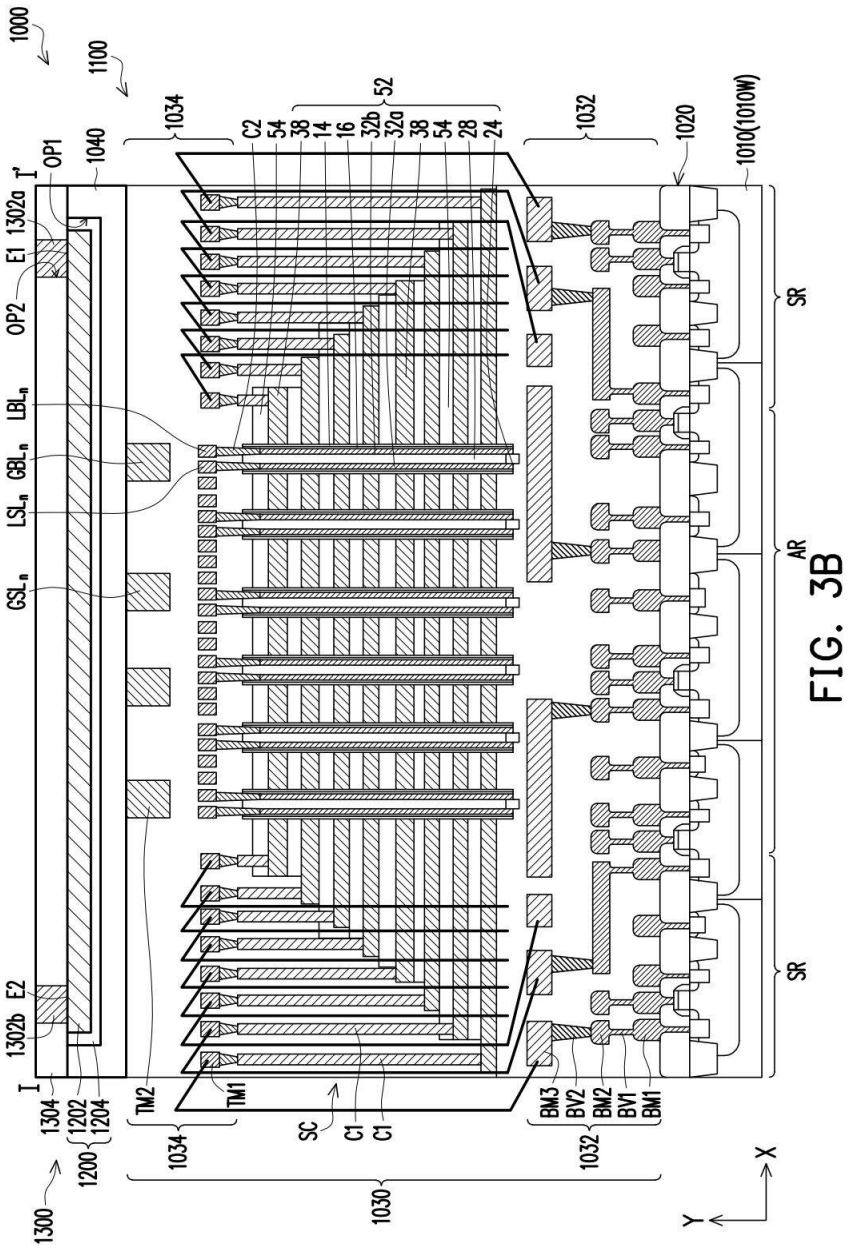

1000
1100
1034
52
C2
54
38
14
16
32b
32a
38
54
28
24
1032
1020
1010(1010W)
1040
OP1
I'
1302a
E1
OP2
LBLn
GBLn
LSLn
GSLn
SR
AR
SR
1302b
E2
I
1304
1202
1204
1200
1300
TM2
TM1
SC
C1
C1
1034
1030
BM3
BY2
BM2
BY1
BM1
1032
X
Y

FIG. 3B

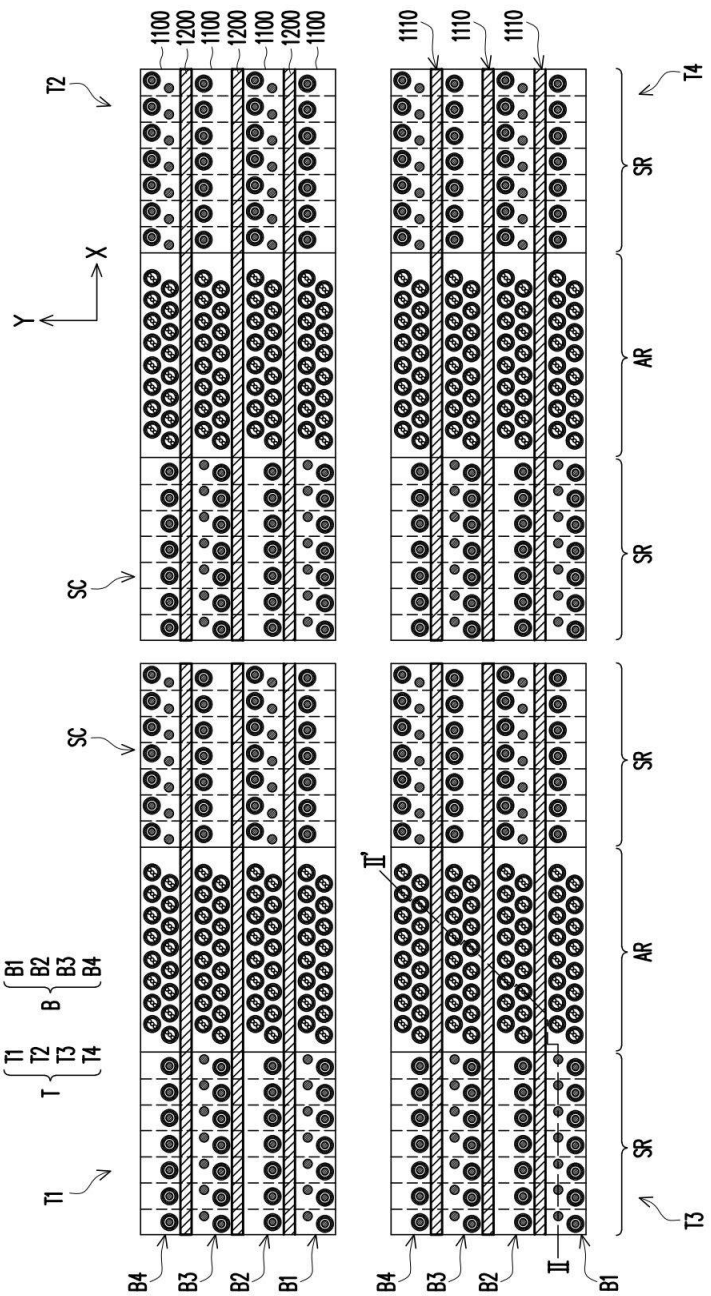

1100
1200
1110
T2
T4
X
Y
SC
SR
AR
B1
B2
B3
B4
B
T1
T2
T3
T4
T
II'
II
T1
T3
B4
B3
B2
B1

FIG. 4A

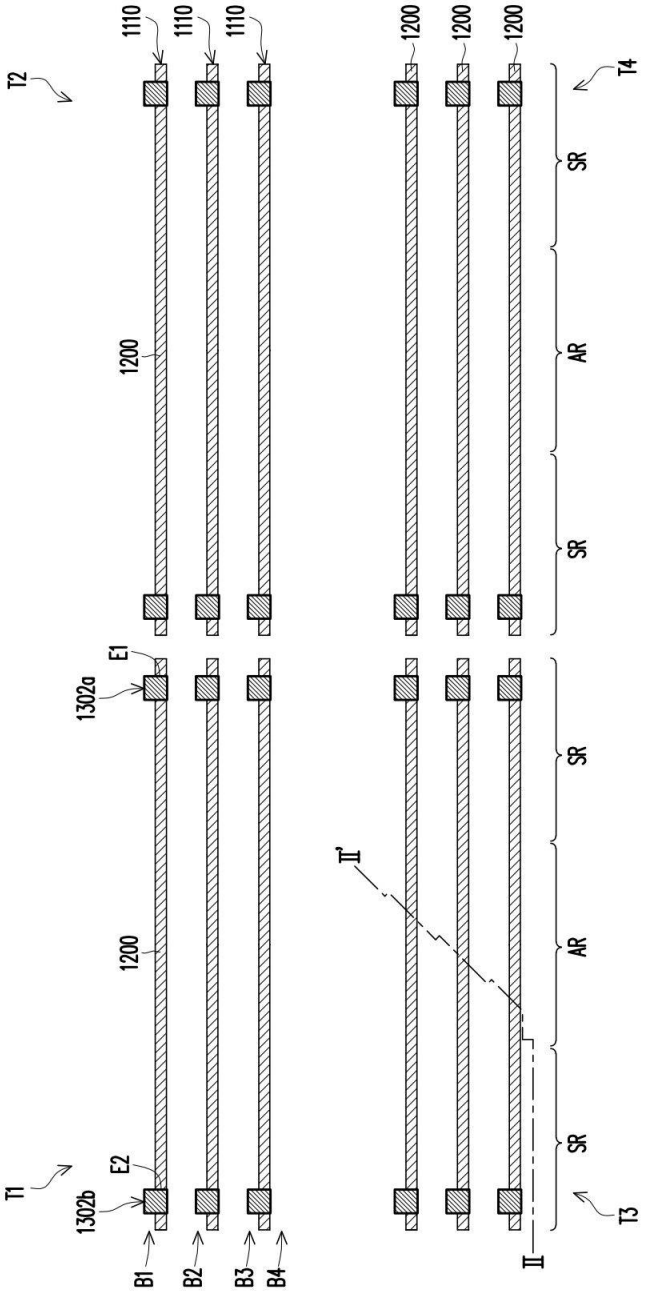

T2
1110
1110
1110
1200
1200
1200
T4
SR
AR
SR
E1
1302a
SR
II'
1200
AR
T1
E2
1302b
SR
T3
B1
B2
B3
B4
II

FIG. 4B

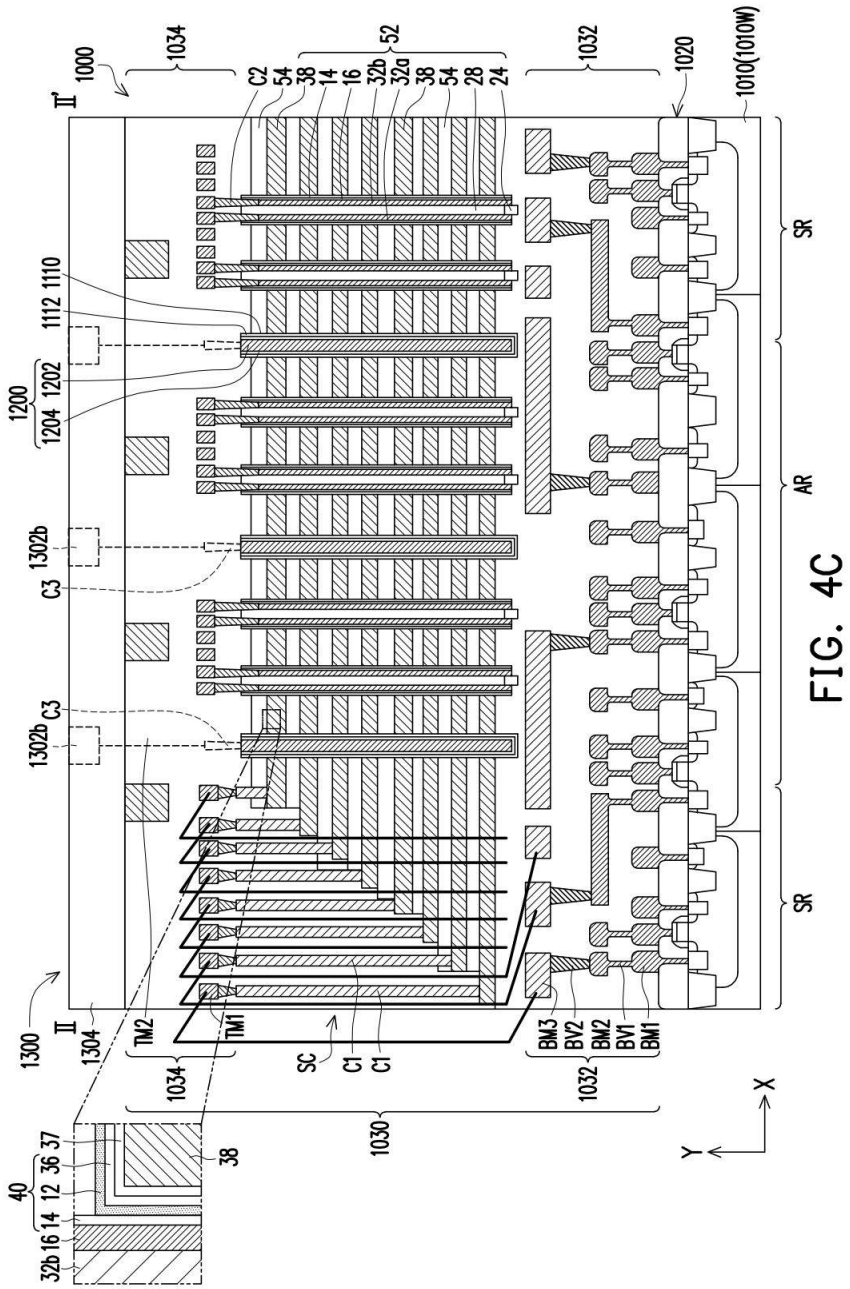

II'
1000
1034
C2
54
38
14
16
32b
52
32a
38
54
28
24
1032
1020
1010(1010W)
SR
AR
SR
1110
1112
1200
1202
1204
1302b
C3
1302b
C3
1300
II
1304
TM2
TM1
SC
C1
C1
BM3
BV2
BM2
BV1
BM1
1034
1032
1030
X
Y
40
37
36
12
14
16
32b
38

FIG. 4C

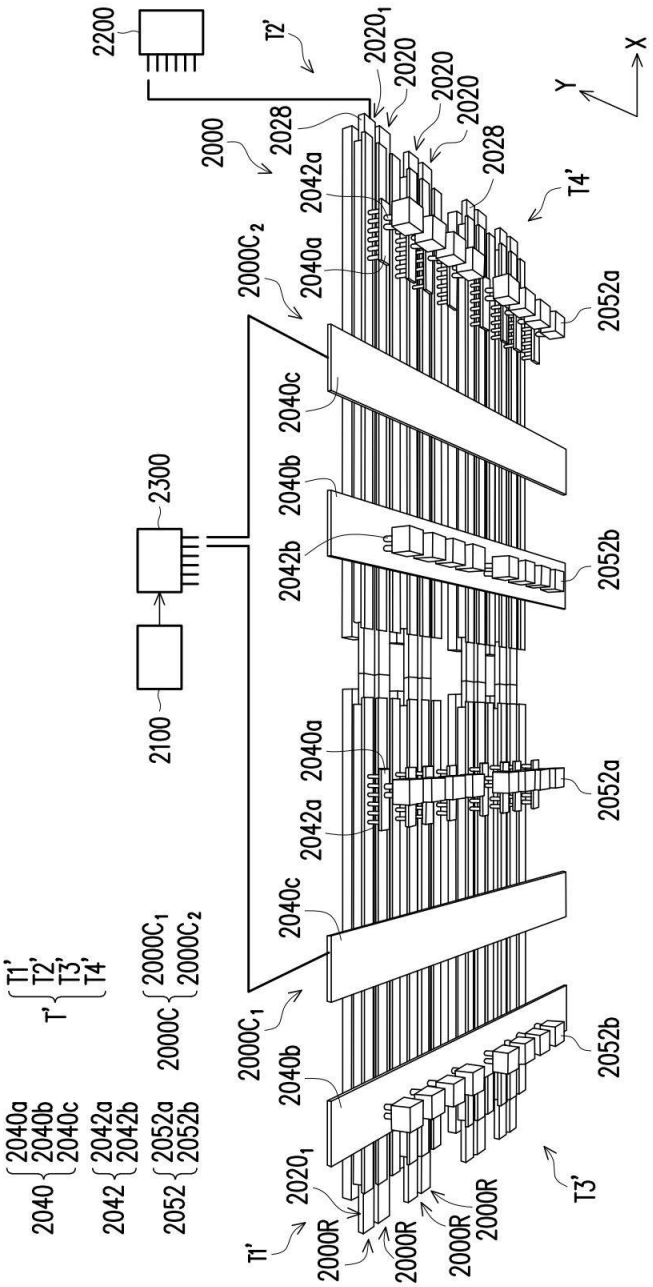

2200
2100
2300
2000
2000C2
2000C1
2020
2028
2042a
2040a
2040c
2040b
2042b
2052a
2052b
T1'
T2'
T3'
T4'
T'
2000C
2040
2042
2052
2000R
X
Y

FIG. 5A

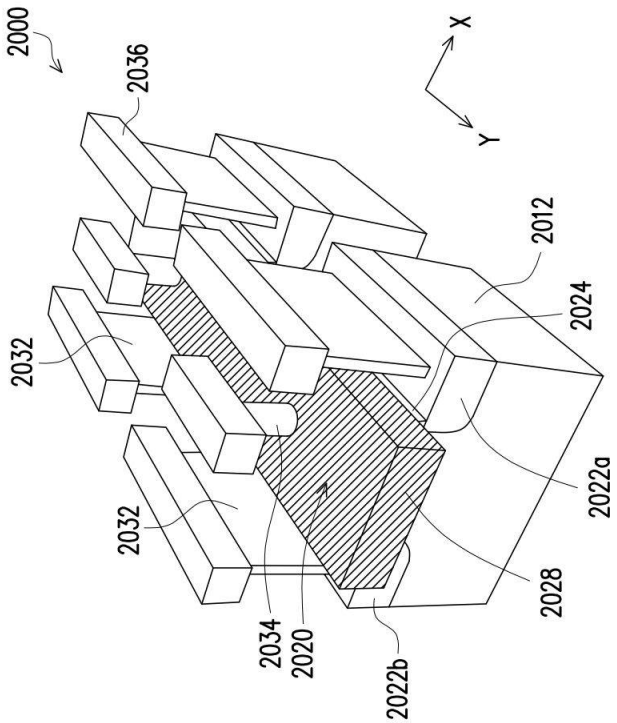

2000
2036
X
Y
2010
2012
2024
2032
2032
2022a
2028
2034
2020
2022b

FIG. 5B

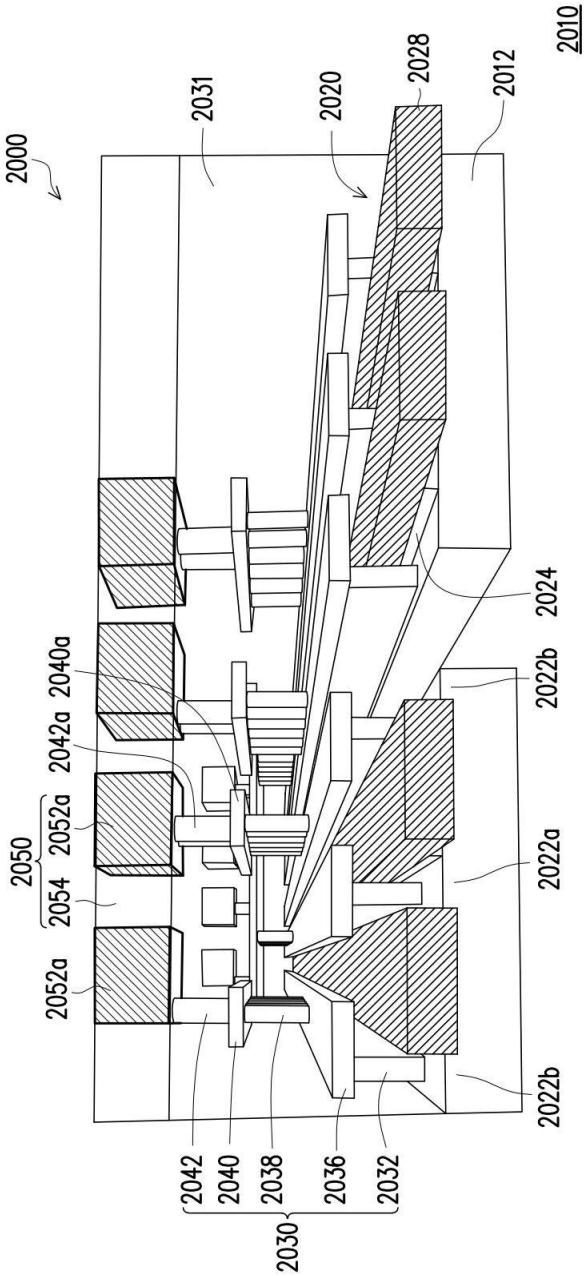

2000
2031
2020
2028
2012
2010
2024
2022b
2040a
2042a
2052a
2050
2054
2022a
2052a
2022b
2042
2040
2038
2036
2032
2030

FIG. 5C

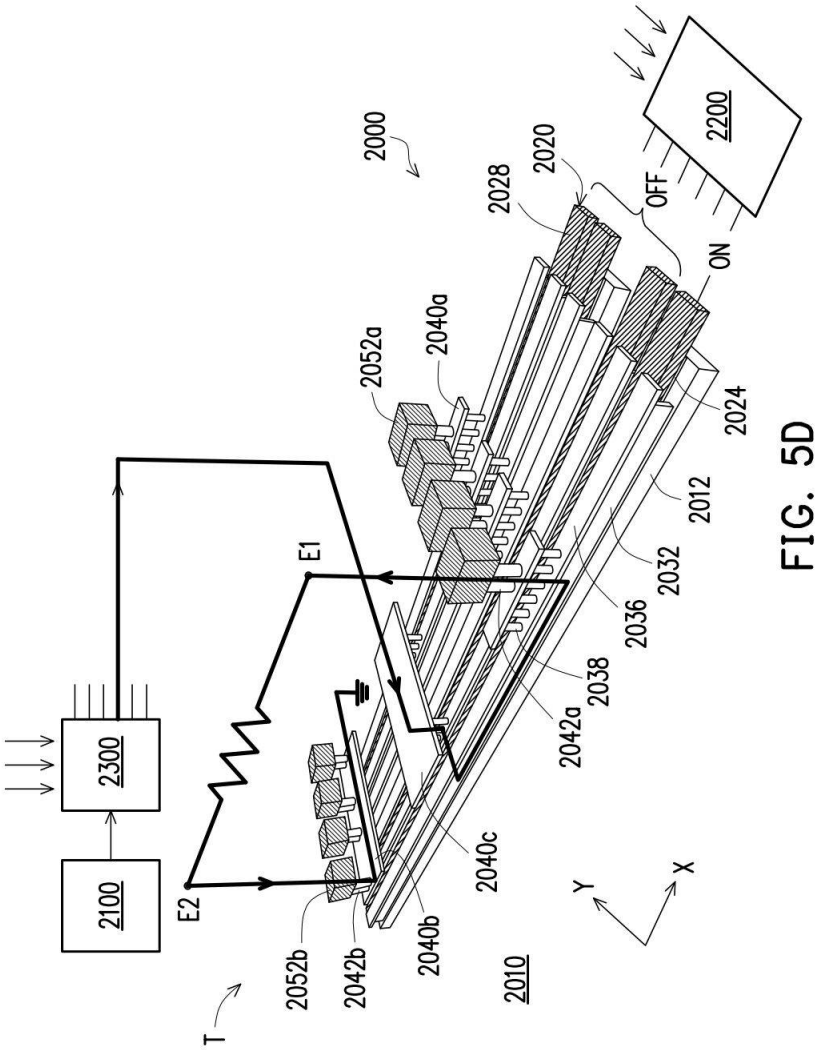

2000
2020
2028
2200
OFF
ON
2052a
2040a
2024
2012
2032
2036
2038
2042a
2040c
E1
E2
2300
2100
2052b
2042b
2040b
2010
T
Y
X

FIG. 5D

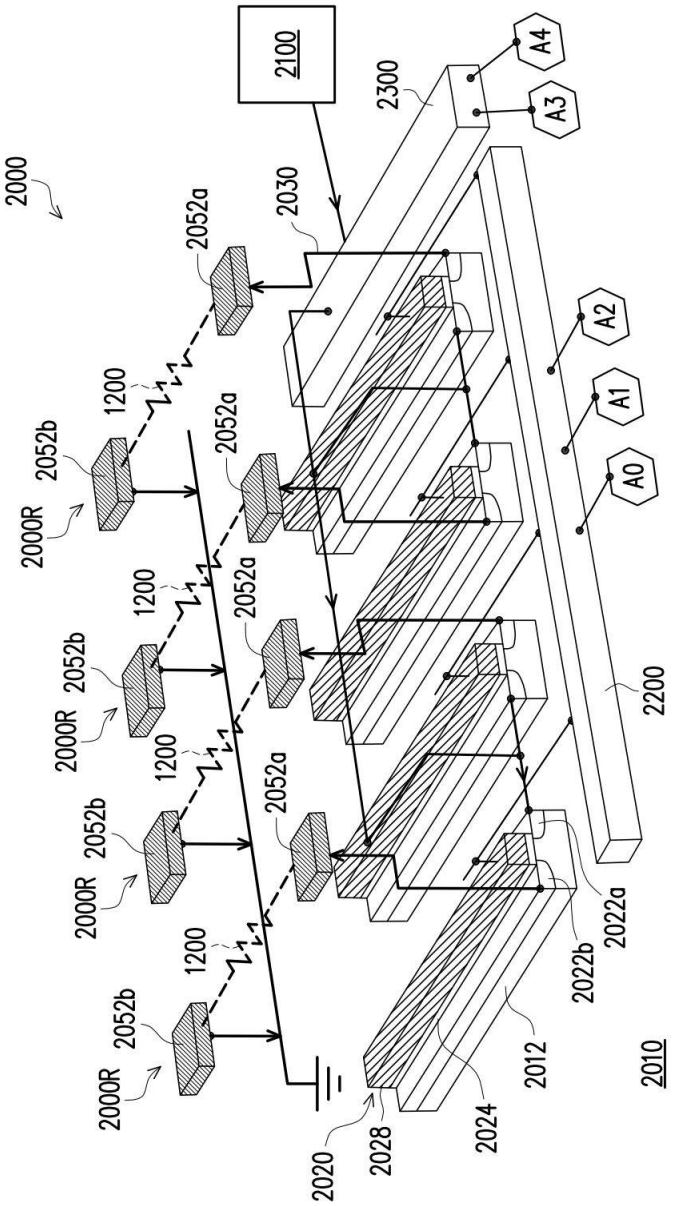

2000
2100
2300
A4
A3
2030
2052a
1200
2052b
2000R
A2
A1
A0
2200
2022a
2022b
2012
2010
2024
2028
2020

FIG. 5E

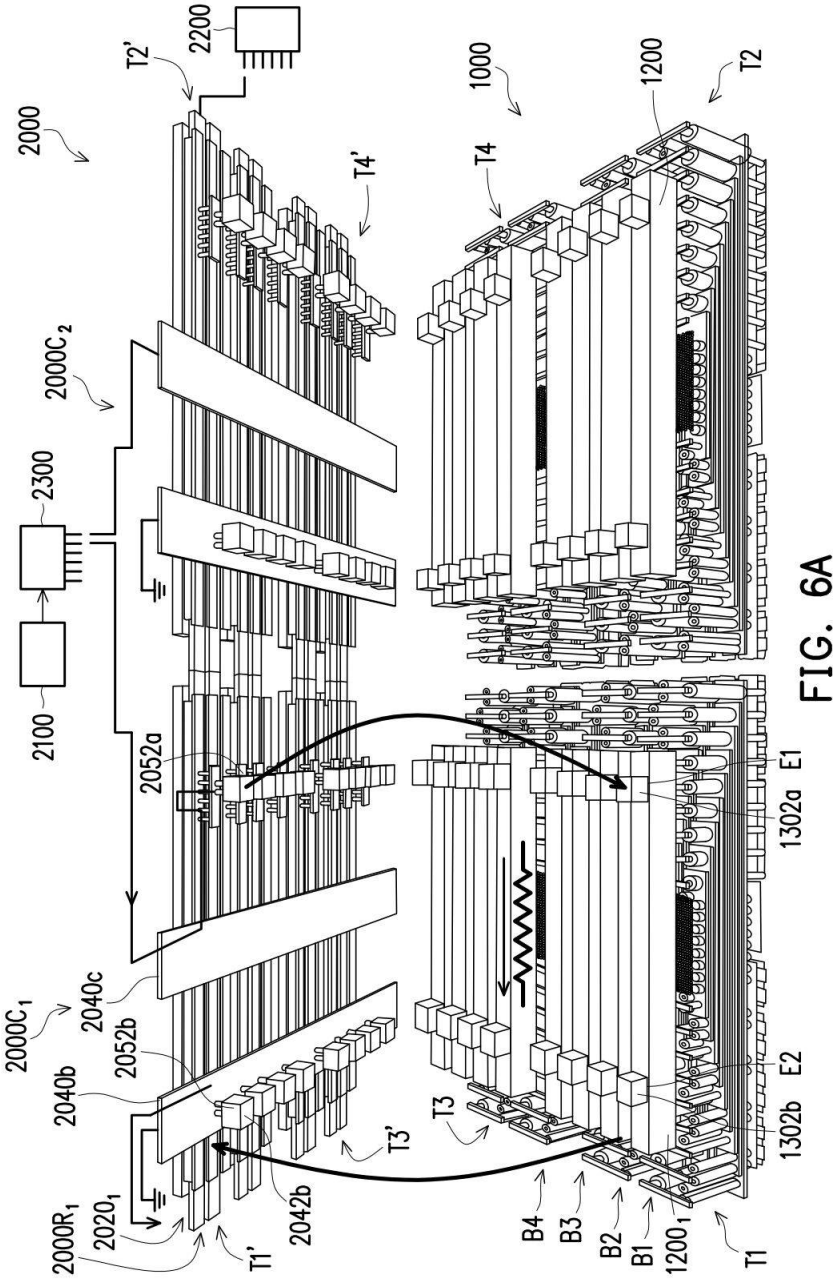

2000
T2'
2200
T4'
1000
1200
T2
T4
2000C2
2300
2100
2052a
1302a E1
2040c
2000C1
2052b
2040b
T3'
T3
1302b E2
2000R1
2020₁
T1'
2042b
B4
B3
B2
B1
1200₁
T1

FIG. 6A

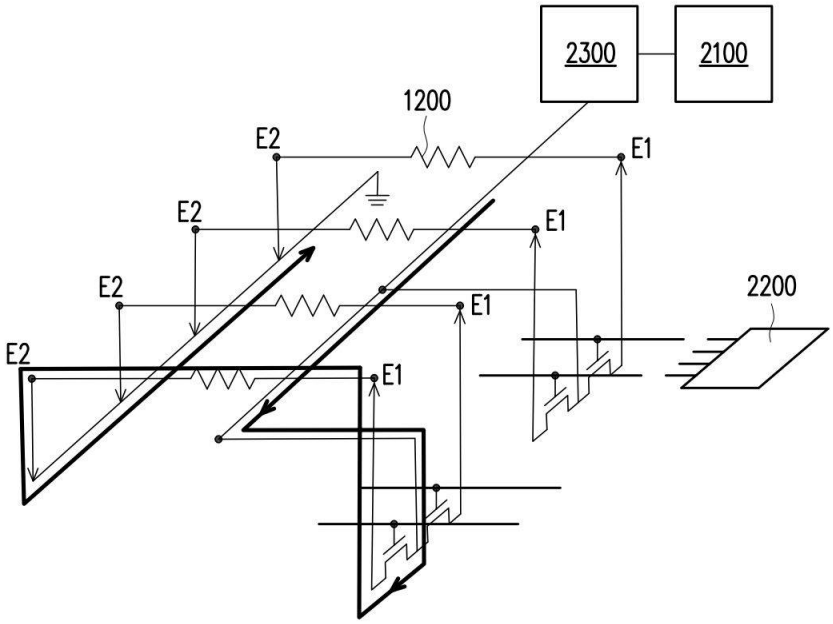

2300
2100
1200
E2
E1
E2
E1
E2
E1
E2
E1
2200

FIG. 6B

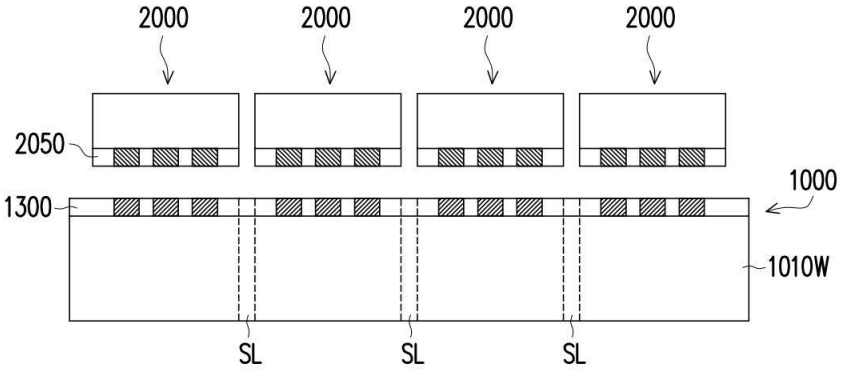

2000
2000
2000
2000
2050
1000
1300
1010W
SL
SL
SL

FIG. 7A

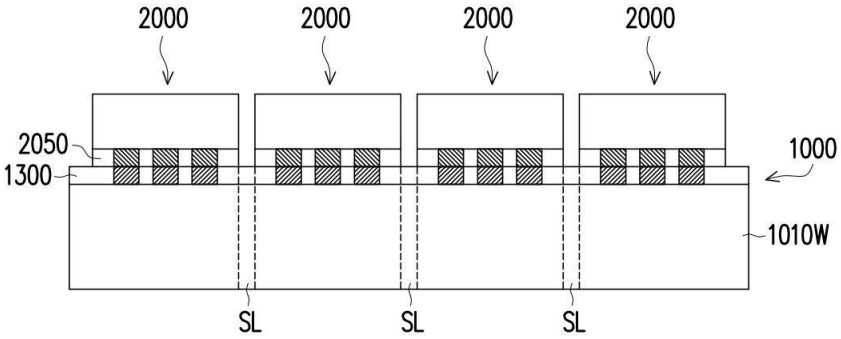

2000
2000
2000
2000
2050
1300
1000
1010W
SL
SL
SL

FIG. 7B

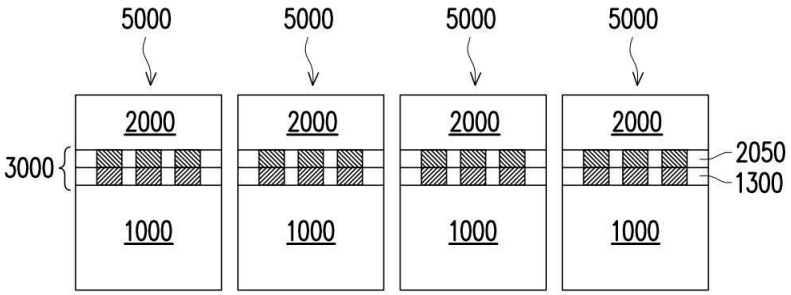

5000
5000
5000
5000
2000
2000
2000
2000
3000
2050
1300
1000
1000
1000
1000

FIG. 7C

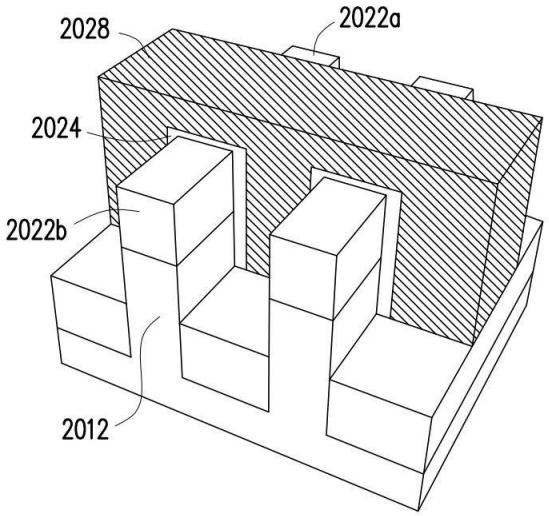


FIG. 8

フロントページの続き

Ｆターム（参考） GA10 JA04 JA19 JA37 JA39 JA40 MA06 MA16 MA19 ZA01
5F101 BA46 BA47 BB02 BD16 BD30 BD34 BE07