



(10) **DE 11 2018 005 614 B4** 2023.01.19

(12)

Patentschrift

(21) Deutsches Aktenzeichen: **11 2018 005 614.7**
(86) PCT-Aktenzeichen: **PCT/IB2018/060321**
(87) PCT-Veröffentlichungs-Nr.: **WO 2019/123303**
(86) PCT-Anmeldetag: **19.12.2018**
(87) PCT-Veröffentlichungstag: **27.06.2019**
(43) Veröffentlichungstag der PCT Anmeldung
in deutscher Übersetzung: **02.07.2020**
(45) Veröffentlichungstag
der Patenterteilung: **19.01.2023**

(51) Int Cl.: **H01L 29/772 (2006.01)**
H01L 21/335 (2006.01)
H01L 29/82 (2006.01)
H01L 29/417 (2006.01)
H01L 29/49 (2006.01)
H01L 29/788 (2006.01)
H10B 41/00 (2023.01)
H10B 51/00 (2023.01)

Innerhalb von neun Monaten nach Veröffentlichung der Patenterteilung kann nach § 59 Patentgesetz gegen das Patent Einspruch erhoben werden. Der Einspruch ist schriftlich zu erklären und zu begründen. Innerhalb der Einspruchsfrist ist eine Einspruchsgebühr in Höhe von 200 Euro zu entrichten (§ 6 Patentkostengesetz in Verbindung mit der Anlage zu § 2 Abs. 1 Patentkostengesetz).

(30) Unionspriorität:
15/850,098 21.12.2017 US

(73) Patentinhaber:
**International Business Machines Corporation,
Armonk, NY, US**

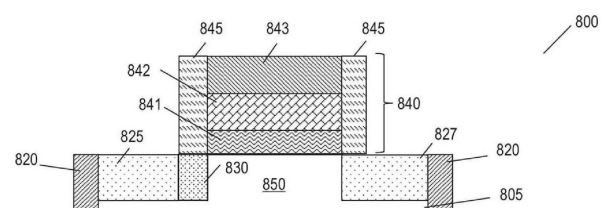
(74) Vertreter:
**Richardt Patentanwälte PartG mbB, 65185
Wiesbaden, DE**

(72) Erfinder:
**Li, Yulong, Yorktown Heights, NY, US; Solomon,
Paul Michael, Yorktown Heights, N.Y., US;
Koswatta, Siyuranga, Yorktown Heights, NY, US**

(56) Ermittelter Stand der Technik:
JP 2016- 171 216 A

(54) Bezeichnung: **Verfahren zum Bilden einer Halbleitereinheit**

(57) Hauptanspruch: Verfahren zum Bilden einer Halbleitereinheit (800), wobei das Verfahren aufweist:
Bilden (704) eines Source-Anschlusses (825) und eines Drain-Anschlusses (827) eines Feldeffekttransistors (FET) auf einem Substrat (805), wobei der Source-Anschluss und der Drain-Anschluss auf jeder Seite eines Kanalbereichs (850) durch Ätzen von Aussparungen in dem Substrat und Abscheiden von dotierten Bereichen in den Aussparungen gebildet werden; und
Bilden (706) einer Energiebarriere (830) zwischen dem Source-Anschluss und dem Kanalbereich durch Ätzen einer Aussparung in dem Substrat und Abscheiden eines dotierten Bereich in dieser Aussparung; und
Bilden (708) einer leitfähigen Gate-Elektrode oberhalb des Kanalbereichs.



Beschreibung**HINTERGRUND DER ERFINDUNG**

[0001] Die vorliegende Erfindung betrifft allgemein Herstellungsverfahren und dabei entstehende Strukturen für Halbleitereinheiten. Genauer gesagt, die vorliegende Erfindung betrifft Feldeffekttransistoren (FETs) mit steuerbarem Widerstand.

[0002] Die vorliegende Erfindung betrifft allgemein auch Herstellungsverfahren und dabei entstehende Strukturen für eine Halbleitereinheit zum Verwenden in künstlichen neuronalen Netzwerken (artificial neural networks, ANNs), die aus Kreuzschienenanordnungen von resistiven Verarbeitungseinheiten (RPU) mit zwei Anschlüssen gebildet sind, die lokale Datenspeicherung und lokale Datenverarbeitung bereitstellen, ohne zusätzliche Verarbeitungselemente über die RPU mit zwei Anschlüssen hinaus zu benötigen und dadurch die Algorithmen zum Umsetzen der ANN-Fähigkeit wie beispielsweise Matrixmultiplikation und Ähnliches zu beschleunigen.

[0003] Der Begriff „maschinelles Lernen“ wird verwendet, um allgemein eine grundlegende Funktion elektronischer Systeme zu beschreiben, die aus Daten lernen. Beim maschinellen Lernen und in der Kognitionswissenschaft stellen ANNs eine Familie statistischer Lernmodelle dar, die auf biologische neuronale Netzwerke von Tieren und insbesondere des Gehirns zurückgehen. ANNs können zum Abschätzen oder Annähern von Systemen und Funktionen verwendet werden, die von einer großen Anzahl Eingabewerten abhängen und im Allgemeinen unbekannt sind. Kreuzschienenanordnungen sind Schaltkreisarchitekturen mit hoher Dichte und geringen Kosten, die zum Bilden einer Vielfalt elektronischer Schaltkreise und Einheiten verwendet werden, darunter ANN-Architekturen, neuromorphe Mikro-Chips und nichtflüchtige Speicher mit ultrahoher Dichte. Eine grundlegende Form der Kreuzschienenanordnung enthält einen Satz leitfähiger Zeilenleitungen und einen Satz leitfähiger Spaltenleitungen, die so gebildet sind, dass sie den Satz leitfähiger Zeilenleitungen kreuzen. Die Schnittpunkte zwischen den beiden Sätzen von Leitungen sind durch so genannte Koppereinheiten voneinander getrennt, die aus Dünnschichtmaterial gebildet werden können.

[0004] Die JP 2016- 171 216 A offenbart einen Tunnel-Feldeffekttransistor, aufweisend einen Source-Bereich, bei dem es sich um einen p-Typ-Bereich hoher Konzentration handelt, einen ersten Kanalbereich, bei dem es sich um einen p-Typ-Bereich niedriger Konzentration handelt, der mit dem Source-Bereich verbunden ist, einen zweiten Kanalbereich, der ein n-Typ-Bereich mit niedriger Konzentration ist, und einen Drain-Bereich, der ein n-Typ-Bereich mit

hoher Konzentration ist, der auf einer anderen Oberfläche als der Verbindungsoberfläche des zweiten Kanalbereichs mit dem ersten Kanalbereich verbunden ist. Der zweite Kanalbereich ist mit einer Gate-Elektrode versehen, die über einen Isolierfilm in Kontakt mit einem stromführenden Bereich steht, der den Source-Bereich, den ersten Kanalbereich, den zweiten Kanalbereich und den Drain-Bereich umfasst, wobei der erste Kanalbereich aus einem Silizium-Einkristall und einem Indium-Einkristall besteht. In der ersten Kanalzone werden Indium und Stickstoff als Additive zu Silizium hinzugefügt.

KURZDARSTELLUNG

[0005] Die der Erfindung zugrundeliegenden Aufgaben werden jeweils mit den Merkmalen des unabhängigen Patentanspruchs gelöst. Ausführungsformen der Erfindung sind Gegenstand der abhängigen Patentansprüche.

[0006] Ausführungsformen der vorliegenden Erfindung betreffen ein Verfahren zum Herstellen einer Halbleitereinheit. Ein nicht als Einschränkung zu verstehendes Beispiel des Verfahrens zum Bilden einer Halbleitereinheit umfasst Bilden eines Source-Anschlusses und eines Drain-Anschlusses eines Feldeffekttransistors (FET) auf einem Substrat. Der Source-Anschluss und der Drain-Anschluss werden auf beiden Seiten eines Kanalbereichs gebildet. Nahe dem Source-Anschluss und dem Kanalbereich wird eine Energiebarriere gebildet. Oberhalb des Kanalbereichs wird eine leitfähige Gate-Elektrode gebildet.

[0007] Ausführungsformen der vorliegenden Erfindung betreffen ein Verfahren zum Herstellen einer Halbleitereinheit. Ein nicht als Einschränkung zu verstehendes Beispiel des Verfahrens umfasst Bilden eines Source-Anschlusses eines Halbleitersteges (semiconductor fin) auf einem Substrat. Auf einer Oberfläche des Source-Anschlusses wird eine Energiebarriere gebildet. Auf einer Oberfläche der Energiebarriere wird ein Kanal gebildet, und auf einer Oberfläche des Kanals wird ein Drain-Anschluss gebildet. Der Drain-Anschluss und der Kanal werden auf beiden Seiten des Kanals vertieft, und die Energiebarriere wird in Aussparungen geätzt, die durch Vertiefen gebildet werden. Der Source-Anschluss wird durch zeitlich abgestimmtes Ätzen vertieft, um einen Teil des Source-Anschlusses in den Aussparungen zu entfernen, die durch Ätzen der Energiebarriere gebildet wurden. Auf einer Oberfläche des Source-Anschlusses und einer Seitenwand des Halbleitersteges wird eine erste untere Abstandsschicht gebildet, und auf der Oberfläche der ersten unteren Abstandsschicht wird ein Gate-Elektrodenstapel gebildet.

[0008] Ausführungsformen der vorliegenden Erfindung betreffen eine Halbleitereinheit. Ein nicht als Einschränkung zu verstehendes Beispiel der Halbleitereinheit enthält einen Source-Anschluss und einen Drain-Anschluss. Der Source-Anschluss und der Drain-Anschluss sind auf beiden Seiten eines Kanalbereichs gebildet, der auf einem Substrat vorgesehen ist. Nahe dem Source-Anschluss und dem Kanalbereich befindet sich eine Energiebarriere, und oberhalb des Kanalbereichs ist ein leitfähiger Gate-Elektrodenstapel gebildet.

[0009] Ausführungsformen der Erfindung betreffen eine Halbleitereinheit. Ein nicht als Einschränkung zu verstehendes Beispiel der Halbleitereinheit enthält einen auf einem Substrat gebildeten Halbleitersteg. Der Halbleitersteg enthält eine Energiebarriere zwischen einem Source-Anschluss und einem Kanalbereich des Halbleitersteges. Oberhalb des Kanalbereichs des Halbleitersteges ist eine leitfähige Gate-Elektrode gebildet, und zwischen der leitfähigen Gate-Elektrode und dem Substrat ist eine untere Abstandsschicht gebildet.

[0010] Ausführungsformen der Erfindung betreffen eine Halbleitereinheit. Ein nicht als Einschränkung zu verstehendes Beispiel der Halbleitereinheit enthält einen auf einem Substrat gebildeten Halbleitersteg. Der Halbleitersteg enthält einen Source-Anschluss, der unter Verwendung eines ersten Dotanden dotiert ist, einen Drain-Anschluss, der unter Verwendung eines zweiten Dotanden mit einer ersten Konzentration dotiert ist, und einen Kanal, der unter Verwendung des zweiten Dotanden mit einer zweiten Konzentration dotiert ist. Die zweite Konzentration ist niedriger als die erste Konzentration.

[0011] Weitere technische Merkmale und Vorteile werden durch die Techniken der vorliegenden Erfindung realisiert. Ausführungsformen und Aspekte der Erfindung werden hierin ausführlich beschrieben und als Bestandteil des beanspruchten Gegenstandes angesehen. Zum besseren Verständnis wird auf die detaillierte Beschreibung und die Zeichnungen verwiesen.

Figurenliste

[0012] Die Einzelheiten der hierin beschriebenen exklusiven Rechte werden in den Ansprüchen am Ende der Beschreibung eingehend dargelegt und ausdrücklich beansprucht. Die obigen sowie weitere Merkmale und Vorteile der Ausführungsformen der Erfindung werden aus der folgenden detaillierten Beschreibung in Verbindung mit den beiliegenden Zeichnungen deutlich, wobei:

Fig. 1 ein vereinfachtes Schaubild der Eingangs- und Ausgangsverbindungen eines biologischen Neurons zeigt;

Fig. 2 ein bekanntes vereinfachtes Modell des in **Fig. 1** gezeigten biologischen Neurons zeigt;

Fig. 3 ein bekanntes vereinfachtes Modell eines ANN zeigt, welches das biologische Neuron des in **Fig. 2** gezeigten Modells enthält;

Fig. 4 ein vereinfachtes Blockschaubild einer bekannten Verfahrensweise zum Aktualisieren von Gewichten zeigt;

Fig. 5 ein Schaubild eines ANN gemäß einer oder mehreren Ausführungsformen ist, das Gewichts-Arrays enthält;

Fig. 6 eine Kreuzschienenanordnung von RPU-Einheiten gemäß Ausführungsformen der vorliegenden Erfindung zusammen mit Spannungsverläufen zeigt, die den Betrieb der RPU veranschaulichen;

Fig. 7 einen Ablaufplan eines beispielhaften Verfahrens gemäß einer oder mehreren Ausführungsformen der vorliegenden Erfindung zum Herstellen eines FET 700 mittels Planardiffusion zeigt;

Fig. 8 eine Querschnittsansicht eines beispielhaften SOI-Wafers 800 gemäß einer oder mehreren Ausführungsformen der vorliegenden Erfindung zeigt;

Fig. 9 eine beispielhafte FET-Struktur gemäß einer oder mehreren Ausführungsformen der vorliegenden Erfindung zeigt;

Fig. 10 eine beispielhafte FET-Struktur gemäß einer oder mehreren Ausführungsformen der vorliegenden Erfindung nach Bilden der Source- und des Drainelektrode und der Energiebarriere zeigt;

Fig. 11 eine Auswirkung der Energiebarriere in der hierin beschriebenen FET-Struktur auf die Stromstärke gemäß einer oder mehreren Ausführungsformen der vorliegenden Erfindung zeigt;

Fig. 12 eine Auswirkung der Energiebarriere in der hierin beschriebenen FET-Struktur auf die Stromstärke gemäß einer oder mehreren Ausführungsformen der vorliegenden Erfindung zeigt;

Fig. 13 eine Auswirkung der Energiebarriere in der hierin beschriebenen FET-Struktur auf die Stromstärke gemäß einer oder mehreren Ausführungsformen der vorliegenden Erfindung zeigt;

Fig. 14 eine beispielhafte FET-Struktur gemäß einer oder mehreren Ausführungsformen der vorliegenden Erfindung nach Bilden des Gate-Elektrodenstapels zeigt;

Fig. 15 eine Ausführungsform zum Speichern des Gewichts in der FET 800 gemäß einer oder mehreren Ausführungsformen der vorliegenden Erfindung zeigt;

Fig. 16 eine Ausführungsform zum Speichern des Gewichts in der FET 800 gemäß einer oder mehreren Ausführungsformen der vorliegenden Erfindung zeigt;

Fig. 17 eine andere Ausführungsform zum Speichern des Gewichts in der FET 800 gemäß einer oder mehreren Ausführungsformen der vorliegenden Erfindung zeigt;

Fig. 18 eine beispielhafte Halbleiterstruktur für den FET gemäß einer oder mehreren Ausführungsformen der vorliegenden Erfindung zeigt;

Fig. 19 eine Auswirkung der Energiebarriere in der hierin beschriebenen III-V-HBFET-Struktur 1800 auf die Stromstärke gemäß einer oder mehreren Ausführungsformen der vorliegenden Erfindung zeigt;

Fig. 20 eine Auswirkung der Energiebarriere in der hierin beschriebenen SiGe-Si-HBFET-Struktur 1800 auf die Stromstärke gemäß einer oder mehreren Ausführungsformen der vorliegenden Erfindung zeigt;

Fig. 21 eine Ausführungsform zum Speichern des Gewichts in dem HBFET 1800 gemäß einer oder mehreren Ausführungsformen der vorliegenden Erfindung zeigt;

Fig. 22 eine Ausführungsform zum Speichern des Gewichts in dem HBFET 1800 gemäß einer oder mehreren Ausführungsformen der vorliegenden Erfindung zeigt;

Fig. 23 eine andere Ausführungsform zum Speichern des Gewichts in dem HBFET 1800 gemäß einer oder mehreren Ausführungsformen der vorliegenden Erfindung zeigt;

Fig. 24 ein beispielhaftes RPU-Array 600 unter Verwendung einer vertikalen Halbleiterstruktur für den HBFET zeigt;

Fig. 25 eine Substratvorbereitung des HBFET 2400-Bereichs zeigt;

Fig. 26 eine darauf folgende Zwischenstruktur des HBFET 2400 während der Herstellung zeigt;

Fig. 27 eine darauf folgende Zwischenstruktur des HBFET 2400 während der Herstellung zeigt;

Fig. 28 eine darauf folgende Zwischenstruktur des HBFET 2400 während der Herstellung zeigt;

Fig. 29 eine darauf folgende Zwischenstruktur des HBFET 2400 während der Herstellung zeigt;

Fig. 30 eine Querschnittsansicht einer VFET-Struktur nach dem Öffnen der oberen Abstandsschicht und dem Bilden der oberen Drain-Elektrode während eines Zwischenschritts des beispielhaften Verfahrens zum Herstellen einer Halbleitereinheit zeigt;

Fig. 31 eine Querschnittsansicht einer vertikalen HBFET-Struktur während der Herstellung zeigt;

Fig. 32 eine Querschnittsansicht einer vertikalen HBFET-Struktur während der Herstellung zeigt;

Fig. 33 eine Querschnittsansicht einer vertikalen HBFET-Struktur während der Herstellung zeigt;

Fig. 34 eine Querschnittsansicht einer vertikalen HBFET-Struktur während der Herstellung zeigt;

Fig. 35 Verbindungen zwischen dem HBFET 2400 und dem RPU-Array 600 zeigt;

Fig. 36 eine Querschnittsansicht einer vertikalen HBFET-Struktur zum Verbinden des HBFET 2400 mit dem RPU-Array 600 während der Herstellung zeigt;

Fig. 37 eine Ausführungsform eines Tunnel-FET 3700 zeigt;

Fig. 38 eine Auswirkung der hierin beschriebenen Tunnel-FET-Struktur auf die Stromstärke zeigt;

Fig. 39 eine Auswirkung der hierin beschriebenen Tunnel-FET-Struktur auf die Stromstärke zeigt;

Fig. 40 eine Auswirkung der Energiebarriere in der hierin beschriebenen Tunnel-FET-Struktur auf die Stromstärke zeigt;

Fig. 41 eine Ausführungsform zum Speichern des Gewichts in dem Tunnel-FET 3700 zeigt;

Fig. 42 eine Ausführungsform zum Speichern des Gewichts in dem Tunnel-FET 3700 zeigt;

Fig. 43 eine andere Ausführungsform zum Speichern des Gewichts in dem Tunnel-FET 3700 zeigt;

Fig. 44 eine Verarbeitung in der Eingabephase (forward pass) unter Verwendung des RPU-Arrays mit den hierin beschriebenen asymmetrischen FET-Strukturen gemäß einer oder mehreren Ausführungsformen der vorliegenden Erfindung zeigt;

Fig. 45 eine Verarbeitung in der Lernphase (backward pass) unter Verwendung des RPU-Arrays mit den hierin beschriebenen asymmetrischen FET-Strukturen gemäß einer oder mehreren Ausführungsformen der vorliegenden Erfindung zeigt;

Fig. 46 eine beispielhafte symmetrische Halbleiterstruktur für den Diffusions-FET gemäß einer oder mehreren Ausführungsformen der vorliegenden Erfindung zeigt; und

Fig. 47 eine beispielhafte symmetrische Halbleiterstruktur für den HBFET gemäß einer oder mehreren Ausführungsformen der vorliegenden Erfindung zeigt.

[0013] Die hierin gezeigten Schaubilder dienen der Veranschaulichung. Es kann viele Varianten des Schaubildes oder der hierin beschriebenen Arbeitsschritte geben, ohne vom Wesensgehalt der Erfindung abzuweichen. Zum Beispiel können die Aktionen in einer anderen Reihenfolge ausgeführt werden, oder es können Aktionen hinzugefügt, weggelassen oder verändert werden.

[0014] In den beiliegenden Figuren und der folgenden detaillierten Beschreibung der beschriebenen Ausführungsformen sind die verschiedenen in den Figuren dargestellten Elemente mit zwei-, drei- oder vierstelligen Bezugsnummern bezeichnet. Bis auf wenige Ausnahmen entspricht (entsprechen) die am weitesten links stehenden Ziffer(n) jeder Bezugsnummer der Figur, in der dieses Element zuerst dargestellt wird.

DETAILLIERTE BESCHREIBUNG

[0015] Verschiedene Ausführungsformen der Erfindung werden hierin unter Bezugnahme auf die zugehörigen Zeichnungen beschrieben. Es können alternative Ausführungsformen der Erfindung dargelegt werden, ohne vom Schutzzumfang dieser Erfindung abzuweichen. Verschiedene Verbindungen und Positionsbeziehungen (z.B. oberhalb, unterhalb, nahe usw.) zwischen Elementen werden in der folgenden Beschreibung und in den Zeichnungen dargelegt. Diese Verbindungen und/oder Positionsbeziehungen können, sofern nicht anderweitig angegeben, direkt oder indirekt sein, und die vorliegende Erfindung soll in dieser Beziehung keine Einschränkung darstellen. Demgemäß kann ein Verbinden von Einheiten entweder ein direktes oder ein indirektes Verbinden betreffen, und eine Positionsbeziehung zwischen Einheiten kann eine direkte oder eine indirekte Positionsbeziehung sein. Außerdem können die verschiedenen hierin beschriebenen Aufgaben und Prozessschritte in eine umfangreichere Prozedur oder einen Prozess mit zusätzlichen Schritten oder Funktionalitäten einbezogen sein, die hierin nicht ausführlich beschrieben sind.

[0016] Die folgenden Definitionen und Abkürzungen dienen zum Auslegen der Ansprüche und der Beschreibung. Die hierin verwendeten Begriffe „weist auf“, „aufweisend“, „enthält“, „enthaltend“, „hat“, „habend“, „enthält“ oder „enthaltend“ oder beliebige andere Varianten derselben sollen eine nichtexklusive Einbeziehung bedeuten. Zum Beispiel ist eine Zusammensetzung, ein Gemisch, ein Prozess, ein Verfahren, ein Artikel oder eine Vorrichtung, die eine Liste von Elementen aufweist, nicht notwendigerweise allein auf diese Elemente beschränkt, sondern kann auch andere Elemente enthalten, die nicht ausdrücklich aufgeführt oder einer solchen Zusammensetzung, einem solchen Gemisch, Prozess, Verfahren, Artikel oder einer solchen Vorrichtung eigen sind.

[0017] Außerdem wird der Begriff „beispielhaft“ hierin in der Bedeutung „als Beispiel, Einzelfall oder Erläuterung“ verwendet. Jede hierin als „beispielhaft“ beschriebene Ausführungsform oder Bauform ist nicht unbedingt als gegenüber anderen Ausführungsformen oder Bauformen bevorzugt oder vorteilhaft zu verstehen. Die Begriffe „mindestens ein“ und „ein oder mehrere“ können in dem Sinne verstanden werden, dass sie eine beliebige ganze Anzahl größer als eins, d.h. eins, zwei, drei, vier usw. enthalten. Der Begriff „eine Mehrzahl“ kann so verstanden werden, dass er eine beliebige ganze Anzahl größer als oder gleich zwei, drei, vier, fünf usw. enthält. Der Begriff „Verbindung“ kann sowohl eine indirekte „Verbindung“ als auch eine direkte „Verbindung“ umfassen.

[0018] Die Begriffe „ungefähr“, „im Wesentlichen“, „annähernd“ und deren Varianten sollen ein Fehlermaß enthalten, das mit der Messung einer bestimmten Größe mittels einer zum Zeitpunkt des Einreichens der Anmeldung verfügbaren Ausrüstung verbunden ist. Zum Beispiel kann „ungefähr“ einen Bereich von $\pm 8\%$ oder 5% oder 2% eines bestimmten Wertes umfassen.

[0019] Der Kürze halber werden übliche Techniken zur Herstellung von Halbleitereinheiten und integrierten Schaltkreisen (IC) hierin gegebenenfalls nicht ausführlich beschrieben. Darüber hinaus können verschiedene hierin beschriebene Aufgaben und Prozessschritte in eine umfangreichere Prozedur oder einen Prozess mit zusätzlichen hierin nicht beschriebenen Schritten oder Funktionalitäten einbezogen werden. Insbesondere sind verschiedene Schritte zur Herstellung von Halbleitereinheiten und Halbleiter-ICs allgemein bekannt, sodass im Interesse einer gestrafften Darstellung viele übliche Schritte hierin nur kurz erwähnt oder ganz weggelassen werden, ohne die normalerweise bekannten Prozessdetails zu benennen.

[0020] Im Folgenden wird ein Überblick über Technologien gegeben, die für Aspekte der vorliegenden

Erfindung bezüglich moderner Prozesse zur Herstellung von Halbleitereinheiten von besonderer Bedeutung sind, mit denen eine große Anzahl Halbleitereinheiten wie beispielsweise Feldeffekttransistoren (VFETs) auf einem einzigen Wafer hergestellt werden. Ferner werden bei einigen nicht in Planartechnik hergestellten Transistorarchitekturen wie beispielsweise vertikalen Feldeffekttransistoren (VFETs) Halbleiterstege und seitliche Gate-Elektroden verwendet, die außerhalb des aktiven Bereichs angeschlossen werden können, was zu einer höheren Einheitendichte und bestimmten verbesserten Leistungsparametern als bei planaren Einheiten führt. In VFETs fließt der Source/Drain-Strom in einer Richtung senkrecht zu einer Hauptfläche des Substrats. Zum Beispiel ist in einer bekannten VFET-Konfiguration eine Hauptfläche des Substrats horizontal, und ein vertikaler Steg oder Nanodraht erstreckt sich von der Substratoberfläche aus nach oben. Der Steg oder die Nanoleiterbahn bildet einen Kanalbereich des Transistors. Ein Source-Bereich und ein Drain-Bereich sind elektrisch mit dem oberen und dem unteren Ende des Kanalbereichs verbunden, während die Gate-Elektrode auf einer oder mehreren Seitenwänden des Steges oder des Nanodrahtes angeordnet ist.

[0021] Es ist jedoch schwierig, VFETs mit Leistungswerten bereitzustellen, die gleich wie oder besser als bei planaren Einheiten sind. In einem VFET ist der Kontakt zur unteren Source/Drain-Elektrode (S/D) von der Oberseite der Struktur her so gebildet, dass der untere S/D-Kontakt die Gate-Elektrode überlappt. Dieses vertikal gestapelte Array führt in Kombination mit der verringerten Grundfläche von VFETs zu einer hohen parasitären Kapazität zwischen der Gate-Elektrode und dem S/D-Bereich des Substrats. Die parasitäre Kapazität zwischen zwei Leitern (auch als Leiter-Leiter-Kapazität bekannt) ist eine Funktion der Länge und Dicke der Leiter sowie des Abstands zwischen den Leitern. Parasitäre Kapazität trägt in der Einheit zu unerwünschten Effekten wie resistiv-kapazitive (RC) Verzögerung, Energieverluste und Übersprechen bei. Bei RC-Verzögerung handelt es sich um die Verzögerung der Signalgeschwindigkeit oder -weiterleitung in einem Schaltkreis als Funktion des Produktes aus Widerstand und Kapazität der Schaltkreiskomponenten. In dem Streben nach immer kleineren elektronischen Einheiten nimmt die parasitäre Kapazität mit kleiner werden Abmessungen der Einheiten und geringeren Abständen zwischen den Komponenten leider immer mehr zu. Herkömmliche Ansätze zum Verringern der parasitären Kapazität zwischen der Gate-Elektrode und dem S/D-Bereich des VFET-Substrats waren nicht durchweg erfolgreich. In einem üblichen VFET wird zum Beispiel eine dünne untere Abstandsschicht zwischen der Gate-Elektrode und dem unteren S/D-Bereich gebildet, um die parasitäre Kapazität ein wenig zu vermindern.

Die Dicke dieser dünnen unteren Abstandsschicht in herkömmlichen VFETs ist jedoch durch Anforderungen seitens der Kanallänge beschränkt. Demzufolge bleibt die parasitäre Kapazität relativ hoch, und es müssen bessere Lösungen gefunden werden.

[0022] Im Folgenden wird eine Übersicht über Aspekte der vorliegenden Erfindung gegeben, indem gemäß einer oder mehreren Ausführungsformen der Erfindung Verfahren und Strukturen zum Bereitstellen eines FET mit steuerbarem Widerstand bereitgestellt werden. Mittels solcher FETs werden Systeme wie beispielsweise elektronische Schaltkreise und Einheiten verbessert, die zum Implementieren von künstlichen neuronalen Netzwerken (artificial neural network, ANN) verwendet werden. Eine technische Herausforderung beim Implementieren von ANNs stellt insbesondere die Anforderung an die Lerngeschwindigkeit für (später beschriebene) Anwendungen von neuronalen Netzwerken für tiefes Lernen dar. Die Aspekte der vorliegenden Erfindung stellen technische Lösungen bereit, die solche technischen Probleme beheben, indem sie eine analoge Komponente zum Aktualisieren von Gewichten bereitstellen, die die Anforderungen an die Lerngeschwindigkeit erfüllt.

[0023] Ferner wird eine kurze Beschreibung von ANN-Implementierungen unter Verwendung von Arrays von resistiven Verarbeitungseinheiten (resistive processing unit, RPU) bereitgestellt. Durch die Aspekte der vorliegenden Erfindung wird ein Bereitstellen von Halbleitereinheiten ermöglicht, die in solchen ANN-Implementierungen verwendet werden können, zum Beispiel zum Speichern von Gewichten an Kreuzungspunkten des RPU-Arrays. Es sollte klar sein, dass die hierin beschriebenen Halbleitereinheiten auch anders als bei dem(den) obigen Beispiel(en) verwendet werden können.

[0024] Es ist zwar von vornherein klar, dass gemäß einer oder mehreren Ausführungsformen der Erfindung in Zusammenhang mit biologischen neuronalen Netzwerken unter besonderer Betonung auf das Modellieren von Hirnstrukturen und -funktionen beschrieben werden, aber die Umsetzung der hierin angegebenen Lehren nicht auf das Modellieren einer bestimmten Umgebung beschränkt ist. Vielmehr sind die Implementierungen der vorliegenden Erfindung in der Lage, eine beliebige Art von Umgebung zu modellieren, darunter zum Beispiel Wettererscheinungen, beliebige aus dem Internet gesammelte Daten und dergleichen, soweit die verschiedenen in die Umgebung eingegebenen Werte in einen Vektor umgewandelt werden können. Demgemäß betreffen Ausführungsformen der vorliegenden Erfindung zwar elektronische Systeme, jedoch werden zum leichteren Vergleichen und Erläutern verschiedene Aspekte der elektronischen Systeme unter Verwendung neurologischer Begriffe

wie zum Beispiel Neuronen, Plastizität und Synapsen beschrieben. Es ist klar, dass bei jeder Erörterung oder Darstellung eines elektronischen Systems hierin durch das Verwenden neurologischer Begriffe oder neurologischer Kurzbegriffe das Vergleichen erleichtert und die neuromorphe(n) bzw. ANN-Entsprechung(en) der beschriebenen neurologischen Funktion oder neurologischen Komponente ausgedrückt werden sollen.

[0025] Künstliche neuronale Netzwerke (ANNs) können zum Abschätzen oder Annähern von Systemen und Funktionen verwendet werden, die von einer großen Anzahl Eingangswerte abhängen und im Allgemeinen unbekannt sind. Neuronale Netzwerke verwenden eine Algorithmenklasse auf der Grundlage eines Konzeptes miteinander gekoppelter „Neuronen“. In einem typischen neuronalen Netzwerk haben Neuronen eine vorgegebene Aktivierungsfunktion, die die Eingangswerte umsetzt. Durch Ermitteln geeigneter Verbindungsgewichte (dieser Prozess wird auch als „Training“ bezeichnet), erlangt ein neuronales Netzwerk die Fähigkeit zum sicheren Erkennen eines gewünschten Musters wie beispielsweise Bilder und Zeichen. Oftmals sind diese Neuronen in „Schichten“ zusammengefasst, um Verbindungen zwischen Gruppen und zu jeder Berechnung von Werten klarer hervortreten zu lassen. Das Training des neuronalen Netzwerkes ist ein rechenaufwändiger Prozess.

[0026] ANNs werden oft als so genannte „neuromorphe“ Systeme von miteinander gekoppelten Prozessorelementen realisiert, die wie simulierte „Neuronen“ fungieren und untereinander „Nachrichten“ in Form von elektronischen Signalen austauschen. Ähnlich der so genannten „Plastizität“ von synaptischen Neurotransmitter-Verbindungen, die Nachrichten zwischen biologischen Neuronen übertragen, sind die Verbindungen in ANNs, die elektronische Nachrichten zwischen simulierten Neuronen übertragen, mit numerischen Gewichten ausgestattet, die der Stärke oder Schwäche einer bestimmten Verbindung entsprechen. Die Gewichte können auf der Grundlage von Erfahrung angepasst und abgestimmt werden, sodass die ANNs an Eingabewerte angepasst und zu lernen in die Lage versetzt werden können. Zum Beispiel ist ein ANN zum Erkennen von Handschriften durch einen Satz von Eingangs-Neuronen definiert, die durch die Pixel eines Eingangsbildes aktiviert werden können. Nach dem Gewichten und Umsetzen durch eine vom Netzwerkentwickler festgelegte Funktion werden die Aktivierungen dieser Eingangs-Neuronen dann an die anderen nachfolgenden Neuronen weitergegeben, die oft auch als „verdeckte“ Neuronen bezeichnet werden. Dieser Prozess wird so lange wiederholt, bis ein Ausgangs-Neuron aktiviert wird. Das aktivierte Ausgangs-Neuron legt dann fest, welches Zeichen gelesen wurde.

[0027] Kreuzschienen-Arrays, auch als Kreuzungspunkt-Arrays oder Querdrahtarrays bekannt, sind preiswerte Schaltkreisarchitekturen mit hoher Dichte, die in einer Vielfalt elektronischer Schaltkreise und Einheiten verwendet werden, darunter ANN-Architekturen, neuromorphe Mikro-Chips und nicht flüchtige Speicher mit extrem hoher Dichte. Eine grundlegende Kreuzschienen-Array-Konfiguration enthält eine Satz leitfähige Zeilenleitungen und einen Satz leitfähige Spaltenleitungen, die den Satz leitfähiger Zeilenleitungen kreuzen. Die Schnittpunkte zwischen den beiden Sätzen von Leitungen sind durch so genannte Koppereinheiten voneinander getrennt, die aus einem Dünnschichtmaterial gebildet sein können.

[0028] Koppereinheiten fungieren im Grunde wie gewichtete Verbindungen des ANN zwischen Neuronen. Einheiten im Nanometermaßstab, zum Beispiel Memristoren mit „idealen“ Schalteigenschaften zwischen Leitungszuständen, werden oft als Koppereinheiten verwendet, um energiesparend synaptische Plastizität zu emulieren. Der Leitungszustand (z.B. der Widerstand) des idealen Memristormaterials kann durch Steuern der zwischen einzelnen Leitungen der Zeile und Spaltenleitungen angelegten Spannungen verändert werden. Digitale Daten können durch Verändern des Leitungszustands des Memristormaterials an dem Schnittpunkt gespeichert werden, um einen Zustand hoher Leitfähigkeit oder einen Zustand niedriger Leitfähigkeit zu erreichen. Das Memristormaterial kann auch so programmiert werden, dass es durch selektives Einstellen des Leitungszustands des Materials zwei oder mehr verschiedene Leitungszustände einhält. Der Leitungszustand des Memristormaterials kann abgelesen werden, indem eine Spannung an das Material angelegt und der durch die betreffende Koppereinheit fließende Strom gemessen wird.

[0029] Zum Senken des Energieverbrauchs sind die Koppereinheiten von ANN-Chip-Architekturen oft so ausgestaltet, dass sie Lerntechniken offline nutzen, wobei sich die Näherung der Zielfunktion nicht ändert, sobald die erste Trainingsphase abgeschlossen worden ist. Durch Offline-Lernen lassen sich die Koppereinheiten von ANN-Architekturen vom Kreuzschienentyp so vereinfachen, dass sie nur sehr wenig Strom verbrauchen.

[0030] Abgesehen von der Möglichkeit, weniger Strom zu verbrauchen, kann ein Durchführen von Offline-Training schwierig sein und umfangreiche Ressourcen in Anspruch nehmen, da während des Trainings üblicherweise eine beträchtliche Anzahl einstellbarer Parameter (z.B. Gewichte) in dem ANN-Modell modifiziert werden muss, um die Eingangs-Ausgangs-Paare an die Trainingsdaten anzupassen. Demgemäß führt das Vereinfachen der Koppereinheiten von ANN-Architekturen mit dem

vorrangigen Ziel der Stromeinsparung dazu, dass die Trainingsgeschwindigkeit und der Trainingserfolg bei Offline-Lerntechniken üblicherweise nicht optimal sind.

[0031] ANNs verwenden nicht das herkömmliche digitale Modell zum Verarbeiten von Nullen und Einsen, sondern erzeugen Verbindungen zwischen Verarbeitungselementen, die im Wesentlichen das Funktionsäquivalent der Hauptfunktionalität des Systems darstellen, die abgeschätzt oder angenähert ermittelt werden soll. Zum Beispiel stellt der IBM™ Computer-Chip SYNAPSE™ die Hauptkomponente einer neuromorphen elektronischen Maschine dar, die eine dem Säugetiergehirn ähnliche Form, Funktion und Architektur anstrebt. Zwar werden durch den IBM SyNapse-Computer-Chip dieselben grundlegenden Transistorkomponenten wie bei herkömmlichen Computer-Chips verwendet, jedoch sind dessen Transistoren so gestaltet, dass sie das Verhalten von Neuronen und deren synaptische Verbindungen nachbilden. Der IBM SyNapse-Computer-Chip verarbeitet Informationen unter Verwendung eines Netzwerkes von mehr als einer Million simulierter „Neuronen“, die unter Verwendung elektrischer Impulse wie bei den synaptischen Übertragungen zwischen biologischen Neuronen Informationen untereinander austauschen. Die IBM SyNapse-Architektur enthält einen Verbund von Prozessoren (d.h. simulierte „Neuronen“), die einen Speicher (d.h. eine simulierte „Synapse“) lesen und einfache Operationen ausführen. Das Übertragen von Informationen zwischen diesen Prozessoren, die üblicherweise in verschiedenen Kernen angeordnet sind, erfolgt mittels auf einem Chip untergebrachter Netzwerkrouter.

[0032] Im Folgenden wird unter Bezugnahme auf die **Fig. 1**, **Fig. 2** und **Fig. 3** allgemein beschrieben, wie ein typisches ANN arbeitet. Oben wurde bereits erwähnt, dass ein typisches ANN das menschliche Gehirn nachbildet, das ungefähr einhundert Milliarden Zellen enthält, die als Neuronen bezeichnet werden. **Fig. 1** zeigt ein vereinfachtes Schaubild eines biologischen Neurons 102 mit Verbindungen 104, 106, 108, 110, die dieses mit vorhergehenden Eingängen 112, 114, nachfolgenden Ausgängen 116 und nachfolgenden „anderen“ Neuronen 118 verbinden, die wie gezeigt miteinander verbunden und angeordnet sind. Jedes biologische Neuron 102 sendet und empfängt über die Verbindungen 104, 106, 108, 110 elektrische Impulse. Die Art dieser elektrischen Impulse und wie sie in dem biologischen Neuron 102 verarbeitet werden, ist in erster Linie für die Gesamtfunktionalität des Gehirns maßgeblich. Die Verbindungen zwischen biologischen Neuronen können stark oder schwach sein. Wenn ein bestimmtes Neuron Eingabeimpulse empfängt, verarbeitet das Neuron den Eingabewert gemäß der Funktion des Neurons und sendet das Ergebnis der Funktion an

nachfolgende Ausgänge und/oder an nachfolgende „andere“ Neuronen.

[0033] Das biologische Neuron 102 ist in **Fig. 2** als Knoten 202 mit einer mathematischen Funktion $f(x)$ modelliert, die durch die in **Fig. 2** gezeigte Gleichung dargestellt ist. Der Knoten 202 empfängt elektrische Signale von den Eingängen 212, 214, multipliziert jeden Eingabewert 212, 214 mit der Stärke seiner jeweiligen Verbindung 204, 206, bildet eine Summe der Eingabewerte, lässt die Summe durch eine Funktion $f(x)$ verarbeiten und erzeugt ein Ergebnis 216, bei dem es sich um einen Endausgabewert oder einen Eingabewert für einen anderen Knoten oder um beides handeln kann. In der vorliegenden Beschreibung wird zum Darstellen einer Multiplikation ein Stern (*) verwendet. Schwache Eingabesignale werden mit einem sehr kleinen Zahlenwert für die Verbindungsstärke multipliziert, sodass die Auswirkung eines schwachen Eingabesignals auf die Funktion sehr gering ist. Desgleichen werden starke Eingabesignale mit einem größeren Zahlenwert für die Verbindungsstärke multipliziert, sodass die Auswirkung eines starken Eingabesignals auf die Funktion stärker ist. Die Funktion $f(x)$ wird vom Entwickler ausgewählt, und es kann eine Vielfalt an Funktionen ausgewählt werden. Beispielsweise kann für $f(x)$ eine Tangenshyperbolicus-Funktion gewählt werden, welche die Funktion der vorhergehenden Summe übernimmt und eine Zahl zwischen minus eins und plus eins ausgibt.

[0034] **Fig. 3** zeigt ein als gewichteten Grafen gestaltetes vereinfachtes ANN-Modell 300, in dem die künstlichen Neuronen Knoten (z.B. 302, 308, 316) sind und in dem die Knoten durch gewichtete gerichtete Kanten (z.B. m1 bis m20) verbunden sind. Das ANN-Modell 300 ist so aufgebaut, dass die Knoten 302, 304, 306 Knoten der Eingangsschicht, die Knoten 308, 310, 312, 314 Knoten der verborgenen Schicht und die Knoten 316, 318 Knoten der Ausgangsschicht sind. Jeder Knoten ist mit jedem Knoten in der benachbarten Schicht durch Verbindungslinien verbunden, die in **Fig. 3** durch Richtungspfeile mit Stärken m1 bis m20 dargestellt sind. Zwar sind nur eine Eingangsschicht, eine verborgene Schicht und eine Ausgangsschicht gezeigt, jedoch können in der Praxis mehrere Eingangsschichten, verborgene Schichten und Ausgangsschichten bereitgestellt werden.

[0035] Ähnlich wie bei der Funktionalität eines menschlichen Gehirns empfängt jeder Knoten 302, 304, 306 der Eingangsschicht des ANN 300 Eingabewerte x_1 , x_2 , x_3 direkt von einer (nicht gezeigten) Quelle, wobei weder Verbindungsstärken angepasst noch Knotensummen gebildet wurden. Demgemäß gilt $y_1 = f(x_1)$, $y_2 = f(x_2)$ und $y_3 = f(x_3)$, was durch die im unteren Teil von **Fig. 3** aufgeführten Gleichungen gezeigt ist. Jeder Knoten 308, 310, 312, 314 der

verborgenen Schicht empfängt Eingabewerte von allen Knoten 302, 304, 306 der Eingangsschicht gemäß den Verbindungsstärken, die den betreffenden Verbindungslinien zugehörig sind. Somit gilt im Knoten 308 der verborgenen Schicht $y_4 = f(m_1 \cdot y_1 + m_5 \cdot y_2 + m_9 \cdot y_3)$, wobei ein Stern * eine Multiplikation darstellt. Eine ähnliche Multiplikation der Verbindungsstärken und Bildung von Knotensummen wird für die Knoten 310, 312, 314 der verborgenen Schicht und die Knoten 316, 318 der Ausgangsschicht durchgeführt, was durch die Gleichungen zum Definieren der Funktionen y_5 bis y_9 im unteren Teil von **Fig. 3** gezeigt ist.

[0036] In dem ANN-Modell 300 werden nacheinander Datensätze verarbeitet, und das ANN „lernt“ durch Vergleichen einer zufälligen Anfangsklassifikation des Datensatzes mit der bekannten tatsächlichen Klassifikation des Datensatzes. Unter Verwendung eines Trainingsverfahrens mit der Bezeichnung „Rückführung“ (backpropagation) (d.h. „Fehlerrückführung“) werden die Fehler aus der Anfangsklassifikation des ersten Datensatzes wieder in das Netzwerk zurückgeführt und dazu verwendet, beim zweiten Durchlauf die gewichteten Verbindungen des Netzwerkes zu verändern, und dieser Rückführungsprozess wird oftmals wiederholt. In der Trainingsphase eines ANN ist die korrekte Klassifizierung jedes Datensatzes bekannt, und den Ausgangsknoten können deshalb „korrekte“ Werte zugewiesen werden. Zum Beispiel ein Knotenwert von „1“ (oder 0,9) für den der korrekten Klasse entsprechenden Knoten und ein Knotenwert von „0“ (oder 0,1) für die anderen Knoten. Somit ist es möglich, die berechneten Werte des Netzwerkes für die Ausgangsknoten mit diesen „korrekten“ Werten zu vergleichen und für jeden Knoten einen Fehlerterm (d.h. die „Delta“-Regel) zu berechnen. Diese Fehlerterme werden dann dazu verwendet, die Gewichte der verborgenen Schichten so zu verändern, dass beim nächsten Durchlauf die Ausgabewerte den „korrekten“ Werten näherkommen.

[0037] Es gibt viele Typen neuronaler Netzwerke, bei den am weitesten verbreiteten Kategorien handelt es sich jedoch um vorwärtsgerichtete Netzwerke und rückgekoppelte/rekurrente Netzwerke. Bei dem ANN-Modell 300 handelt es sich um ein vorwärtsgerichtetes/nichtrekurrentes Netzwerk mit Eingangs-, Ausgangs- und verborgenen Schichten. Die Signale können sich nur in einer Richtung ausbreiten. Eingabedaten werden zu einer Schicht von Verarbeitungselementen weitergeleitet, die Berechnungen ausführen. Jedes Verarbeitungselement führt seine Berechnung auf der Grundlage einer gewichteten Summe seiner Eingabewerte aus. Die neuen berechneten Werte werden dann zu neuen Eingabewerten, die zur nächsten Schicht weitergeleitet werden. Dieser Prozess wird so lange fortgesetzt, bis er alle Schichten durchlaufen hat und der Ausgabewert

ermittelt wurde. Mitunter wird eine Schwellenwert-Übertragungsfunktion verwendet, um den Ausgabewert eines Neurons in der Ausgangsschicht als Zahlenwert auszudrücken.

[0038] Ein rückgekoppeltes/rekurrentes Netzwerk enthält Rückkopplungsleitungen, sodass die Signale unter Verwendung von Schleifen in beiden Richtungen weitergeleitet werden können. Zwischen den Knoten sind alle möglichen Verbindungen zulässig. Da in diesem Netzwerktyp Schleifen vorhanden sind, kann daraus bei bestimmten Operationen ein nichtlineares dynamisches System werden, das sich ständig verändert, bis es einen Gleichgewichtszustand erreicht. Rückgekoppelte Netzwerke werden oft in assoziativen Speichern und für Optimierungsprobleme verwendet, wenn das Netzwerk nach der besten Anordnung miteinander gekoppelter Faktoren sucht.

[0039] Die Geschwindigkeit und die Leistungsfähigkeit des maschinellen Lernens in vorwärtsgerichteten und rekurrenten ANN-Architekturen hängt davon ab, wie wirksam die Koppereinheiten des ANN-Kreuzschienenarrays die zentralen Operationen von typischen Algorithmen für maschinelles Lernen ausführen. Zwar lässt sich eine genaue Definition des maschinellen Lernens nur schwer formulieren, jedoch kann ein Lernprozess in Verbindung mit einem ANN als das Problem des Aktualisierens der Verbindungsgewichte der Koppereinheiten angesehen werden, damit ein Netzwerk eine bestimmte Aufgabe zügig ausführen kann. Die Koppereinheiten lernen normalerweise die erforderlichen Verbindungsgewichte aus den verfügbaren Trainingsmustern. Die Leistungsfähigkeit wird durch häufig wiederholtes Aktualisieren der Gewichte in dem Netzwerk im Laufe der Zeit immer besser. Anstatt einen durch menschliche Fachleute vorgegebenen Satz Regeln zu befolgen, „lernen“ ANNs zugrunde liegende Regeln (wie Eingangs-Ausgangs-Beziehungen) aus dem vorhandenen Vorrat an repräsentativen Beispielen. Demgemäß kann ein Lernalgorithmus im Allgemeinen als Prozedur definiert werden, durch die Lernregeln durch Aktualisieren und/oder Anpassen der betreffenden Gewichte angewendet werden.

[0040] Die drei Hauptformen von Lernalgorithmen sind überwachtes Lernen, nichtüberwachtes Lernen und eine Mischform der beiden. Beim überwachten Lernen oder Lernen mit einem „Lehrer“ erhält das Netzwerk für jedes Eingabemuster eine korrekte Antwort. Es werden Gewichte ermittelt, damit das Netzwerk in die Lage versetzt wird, Antworten zu finden, die den bekannten korrekten Antworten möglichst nahe kommen. Forciertes Lernen stellte eine Variante des überwachten Lernens dar, wobei das Netzwerk lediglich eine Kritik an der Korrektheit der Ausgabewertes des Netzwerkes, nicht jedoch an den

korrekten Antworten selbst erhält. Beim nichtüberwachten Lernen oder Lernen ohne Lehrer hingegen wird keine korrekte Antwort benötigt, die jedem Eingabemuster in dem Satz der Trainingsdaten zugehörig ist. Hierbei werden die zugrundeliegende Struktur in den Daten oder Korrelationen zwischen den Mustern in den Daten untersucht und anhand dieser Korrelationen Muster in Kategorien eingeteilt. Beim hybriden Lernen werden überwachtes und nichtüberwachtes Lernen miteinander kombiniert. Teile der Gewichte werden für gewöhnlich durch überwachtes Lernen und andere Teile durch nichtüberwachtes Lernen gewonnen.

[0041] Um den Stromverbrauch zu begrenzen, werden die Koppereinheiten von ANN-Chip-Architekturen wie oben erwähnt für die Nutzung von Offline-Lerntechniken gestaltet, wobei sich der Näherungswert der Zielfunktion nicht ändert, sobald die anfängliche Trainingsphase absolviert wurde. Zum Offline-Lernen können die Koppereinheiten von ANN-Architekturen vom Kreuzschienentyp so vereinfacht werden, dass sie nur sehr wenig Strom verbrauchen.

[0042] Abgesehen von der Möglichkeit, weniger Strom zu verbrauchen, kann ein Durchführen von Offline-Training schwierig sein und viele Ressourcen erfordern, da während des Trainings eine beträchtliche Anzahl einstellbarer Parameter (d.h. Gewichte) in dem ANN-Modell geändert werden müssen, um die Eingangs-Ausgangs-Paare an die Trainingsdaten anzupassen. **Fig. 4** zeigt eine vereinfachte Darstellung eines typischen Lese-Schreib-Prozesses beim Aktualisieren von Gewichten, wobei CPU/GPU-Kerne (d.h. simulierte „Neuronen“) einen Speicher (d.h. eine simulierte „Synapse“) lesen und Verarbeitungsoperationen zum Aktualisieren von Gewichten ausführen, um dann die aktualisierten Gewichte in den Speicher zurückzuschreiben. Wenn die Koppereinheiten von ANN-Architekturen mit dem vorrangigen Ziel der Stromeinsparung vereinfacht werden, bedeutet dies, dass bei Offline-Lerntechniken die Trainingsgeschwindigkeit und der Trainingserfolg für gewöhnlich nicht optimal sind.

[0043] **Fig. 5** veranschaulicht eine künstliche neuronale Netzwerk- (ANN-) Architektur 500. Im vorwärtsgerichteten Betrieb legt ein Satz von Eingabe-Neuronen 502 parallel jeweils eine Eingangsspannung an eine entsprechende Zeile von Gewichten 504 an. Ein Gewicht 504 ist eine Koppereinheit, beispielsweise eine RPU. Die Gewichte 504 haben jeweils einen einstellbaren Widerstandswert, sodass ein Ausgangsstrom von dem Gewicht 504 zu einem entsprechenden verborgenen Neuron 506 fließt, um den gewichteten Eingabewert darzustellen. Der Ausgangsstrom ist durch ein bestimmtes Gewicht $I = \frac{V}{r}$ gegeben, wobei V die Eingangsspannung aus dem Eingangsneuron 502 und r der eingestellte Wider-

stand des Gewichts 504 ist. Der Strom von jedem Gewicht wird spaltenweise addiert und fließt zu einem verborgenen Neuron 506. Ein Satz Referenzgewichte 507 hat jeweils einen festen Widerstand, und deren Ausgabewerte werden zu einem Referenzstrom kombiniert, der zu jedem der verborgenen Neuronen 506 geleitet wird. Da Leitfähigkeitswerte nur positive Zahlenwerte haben können, wird ein Referenzwert für die Leitfähigkeit benötigt, um sowohl positive als auch negative Werte in der Matrix codieren zu können. Die durch die Gewichte 504 erzeugten Ströme werden ständig neu ermittelt und sind positiv, sodass Referenzgewichte 507 zum Bereitstellen eines Referenzstroms verwendet werden, oberhalb dessen Ströme als positiv und unterhalb dessen Ströme als negativ angesehen werden. Indem die Widerstandswerte (r) von FETs steuerbar gemacht werden und oberhalb eines vorgegebenen Schwellenwertes (z.B. 10 MΩ, 90 MΩ, 99 MΩ, 100 MΩ usw.) steuerbar werden, wird es mittels der hierin beschriebenen technischen Lösungen möglich, die FETs als Gewichtsspeicherkomponenten an einem Kreuzungspunkt in des RPU-Arrays zu verwenden. Alternativ wird das Gewicht in einem Kondensator an jedem Kreuzungspunkt gespeichert, und der FET ermöglicht den Wert aus dem Kondensator zu lesen.

[0044] Die verborgenen Neuronen 506 verwenden die Ströme des Arrays der Gewichte 504 und die Referenzgewichte 507 für Berechnungen. Die verborgenen Neuronen 506 geben dann selbst eine Spannung an ein anderes Array von Gewichten 507 aus. Dieses Array verfährt ebenso mit einer Spalte von Gewichten 504, die eine Spannung von ihrem entsprechenden verborgenen Neuron 506 empfangen, um einen gewichteten Ausgangsstrom zu erzeugen, der zeilenweise addiert und zum Ausgangsneuron 508 geleitet wird.

[0045] Es sollte klar sein, dass eine beliebige Anzahl dieser Stufen verwendet werden kann, indem weitere Schichten von Arrays und verborgenen Neuronen 506 eingefügt werden. Es sollte auch klar sein, dass es sich bei einigen Neuronen um konstante Neuronen 509 handeln kann, die das Array mit einer konstanten Spannung versorgen. Die konstanten Neuronen 509 können zwischen den Eingangsneuronen 502 und/oder den verborgenen Neuronen 506 liegen und werden nur während des vorwärtsgerichteten Betriebs verwendet.

[0046] Während der Fehlerrückführung schicken die Ausgangsneuronen 508 eine Spannung zurück durch das Array von Gewichten 504. Die Ausgangsschicht vergleicht die erzeugte Netzwerkantwort mit den Trainingsdaten und berechnet einen Fehler. Der Fehler wird als Spannungsimpuls an das Array gesendet, wobei die Höhe und/oder die Dauer des Impulses proportional dem Fehlerwert moduliert wer-

den. Im vorliegenden Beispiel empfängt eine Zeile von Gewichten 504 parallel eine Spannung von einem entsprechenden Ausgangsneuron 508 und setzt diese Spannung in einen Strom um, der spaltenweise addiert wird, um einen Eingabewert für die verborgenen Neuronen 506 bereitzustellen. Die verborgenen Neuronen 506 kombinieren das gewichtete Rückkopplungssignal mit einer Ableitung seiner vorwärtsgerichteten Berechnung und speichern einen Fehlerwert, bevor sie eine Spannung des Rückkopplungssignals an ihre entsprechende Spalte von Gewichte 504 ausgeben. Diese Fehlerrückführung durchquert das gesamte Netzwerk 500, bis in allen verborgenen Neuronen 506 und den Eingangsneuronen 502 ein Fehlerwert gespeichert ist.

[0047] Beim Aktualisieren der Gewichte senden die Eingangsneuronen 502 und die verborgenen Neuronen 506 eine erste Spannung für die Gewichtsaktualisierung in Vorwärtsrichtung, und die Ausgangsneuronen 508 und die verborgenen Neuronen 506 senden eine zweite Spannung für die Gewichtsaktualisierung in Rückwärtsrichtung durch das Netzwerk 500. Die Kombinationen dieser Spannungen bewirken eine Zustandsänderung innerhalb jedes Gewichts 504 und weisen dem Gewicht 504 einen neuen Widerstandswert zu. Auf diese Weise können die Gewichte 504 so trainiert werden, dass sie das neuronale Netzwerk 500 während seiner Verarbeitung an Fehler anpassen. Es sollte klar sein, dass es die drei Betriebsmodi vorwärtsgerichtete Verarbeitung, Fehlerrückführung und Gewichtsaktualisierung gibt, die einander nicht überlappen.

[0048] Die Ausführungsformen der vorliegenden Erfindung stellen wie erwähnt eine analoge Komponente zur Gewichtsaktualisierung bereit, um den Anforderungen an die Lerngeschwindigkeit bei Anwendungen von tiefen neuronalen Netzwerken zu entsprechen. Zum Beispiel kann eine resistive Verarbeitungseinheit (RPU) in einer Kreuzschienenstruktur parallele Matrixmultiplikationen ermöglichen und die Trainingsgeschwindigkeit von neuronalen Netzwerken erhöhen. Bei großen neuronalen Netzwerken, die für gewöhnlich > 1 Million Gewichtselemente enthalten, hat jedes Gewichtselement einen hohen Widerstand in der Größenordnung von $10\text{ M}\Omega$, und der Widerstand muss veränderbar sein. Der Feldeffekttransistor (FET), dessen Widerstand durch das Gate-Potential gut steuerbar ist, kommt als Gewichtselement zum Einrichten eines solchen neuronalen Netzwerks infrage. Allerdings liegt der Widerstand in modernen FETs im Bereich von $10\text{ k}\Omega$, sodass er auf das ungefähr Tausendfache erhöht werden muss, um die Anforderungen zu erfüllen. Die Ausführungsformen der vorliegenden Erfindung beheben solche technischen Probleme und stellen FET-Strukturen bereit, die solche Anforderungen erfüllen und somit das Einrichten von neuronalen

Netzwerken unter Verwendung von RPU-Kreuzschienenstrukturen ermöglichen.

[0049] In einer folgenden Übersicht der vorliegenden Erfindung betreffen eine oder mehrere Ausführungsformen eine programmierbare resistive Koppelkomponente, die hierin als Koppereinheit oder als resistive Verarbeitungseinheit (RPU) bezeichnet wird, die eine Funktionalität zur lokalen Datenspeicherung und eine Funktionalität zur lokalen Datenverarbeitung bereitstellt. Mit anderen Worten, wenn die Datenverarbeitung durchgeführt wird, wird der in jeder RPU gespeicherte Datenwert parallel und lokal aktualisiert, sodass keine relevanten Daten mehr in einen Prozessor und ein separates Speicherelement eingegeben oder aus diesem ausgegeben werden müssen. Außerdem wird die Fähigkeit des ANN zum Ausführen von Algorithmen wie Matrixmultiplikation und dergleichen infolge der durch die beschriebenen RPUs bereitgestellten lokalen Datenspeicherung und lokalen Datenverarbeitung beschleunigt. Demgemäß wird durch das Einrichten einer ANN-Architektur für maschinelles Lernen mit der beschriebenen RPU ein Optimum zwischen Geschwindigkeit, Lernerfolg und Stromverbrauch des ANN erreicht. Mittels der beschriebenen RPU und der resultierenden ANN-Architektur werden die Gesamtleistung des ANN verbessert und ein breiterer Anwendungsbereich von ANN in der Praxis erreicht.

[0050] Durch die RPU in der Kreuzschienenstruktur werden eine parallele Matrixmultiplikation ermöglicht und die Trainingsgeschwindigkeit des neuronalen Netzwerks erhöht. Bei großen neuronalen Netzwerken, die für gewöhnlich mehr als eine Million Gewichtselemente enthalten, muss jedes Gewichtselement einen hohen Widerstand in der Größenordnung von $10\text{ M}\Omega$ haben, und der Widerstand muss veränderbar sein. Zum Beispiel kommen FETs, deren Widerstand durch das Gate-Potenzial gut steuerbar ist, als Gewichtselemente für ein neuronales Netzwerk infrage. Der Widerstand von modernen FETs liegt jedoch in der Größenordnung von $10\text{ k}\Omega$, was wesentlich niedriger als zum Einrichten von großen neuronalen Netzwerken erforderlich ist (fast 1.000 Mal niedriger als erforderlich). Solche technischen Probleme werden durch die hierin beschriebene(n) Erfindung(en) behoben, indem Gewichtselemente bereitgestellt werden, die einen erhöhten steuerbaren Widerstand in wesentlich höheren Größenordnungen haben, beispielsweise $10\text{ M}\Omega$ bis $100\text{ M}\Omega$.

[0051] Gemäß einem oder mehreren Beispielen wird der steuerbare Widerstand eines FET durch zwei Ladungstransportänderungen in der FET-Struktur erhöht, sodass der FET als Gewichtselement in einem RPU-Array verwendbar ist.

[0052] Fig. 6 zeigt ein Kreuzschienen-Array von RPU-Einheiten gemäß einer Ausführungsform der vorliegenden Erfindung zusammen mit Spannungsverläufen, die den Betrieb der RPU veranschaulichen. Fig. 6 ist ein Schaubild eines zweidimensionalen (2D) Kreuzschienen-Arrays 600, das vorwärtsgerichtete Matrixmultiplikation und Gewichtsaktualisierungen gemäß Ausführungsformen der vorliegenden Erfindung ausführt. Das Kreuzschienen-Array 600 ist aus einem Satz leitfähiger Zeilenleitungen 802, 804, 806 und aus einem Satz leitfähiger Spaltenleitungen 808, 810, 812 und 814 gebildet, die den Satz leitfähiger Zeilenleitungen 802, 804 und 806 kreuzen. Die Schnittpunkte zwischen dem Satz Zeilenleitungen und dem Satz Spaltenleitungen werden durch RPUs getrennt, die in Fig. 6 als resistive Elemente mit jeweils einem eigenen einstellbaren/aktualisierbaren resistiven Gewicht gezeigt sind, die mit σ_{11} , σ_{21} , σ_{31} , σ_{41} , σ_{12} , σ_{22} , σ_{32} , σ_{42} , σ_{13} , σ_{23} , σ_{33} beziehungsweise σ_{43} bezeichnet sind. Zur einfacheren Darstellung ist in Fig. 6 nur eine einzige RPU 820 mit einer Bezugsnummer gekennzeichnet. Bei der vorwärtsgerichteten Matrixmultiplikation kann der Leitungszustand (d.h. die gespeicherten Gewichte) durch Anlegen einer Spannung an die RPU und Messen des durch die RPU fließenden Stroms gelesen werden.

[0053] Eingangsspannungen V_1 , V_2 , V_3 werden an Zeilenleitungen 802, 804 beziehungsweise 806 angelegt. In jeder Spaltenleitung 808, 810, 812, 814 werden die durch jede RPU entlang der betreffenden Spaltenleitung erzeugten Ströme I_1 , I_2 , I_3 , I_4 summiert. Zum Beispiel zeigt Fig. 6, dass der Strom I_4 durch die Spaltenleitung 814 gemäß der Gleichung $I_4 = V_1\sigma_{41} + V_2\sigma_{42} + V_3\sigma_{43}$ erzeugt wird. Somit führt das Array 600 die vorwärtsgerichtete Matrixmultiplikation durch Multiplizieren der in den RPUs gespeicherten Werte mit den Eingabewerten der Zeilenleitungen aus, die durch die Spannungen V_1 , V_2 , V_3 definiert sind. Die rückgekoppelte Matrixmultiplikation ist sehr ähnlich. Bei der rückgekoppelten Matrixmultiplikation werden Spannungen an die Spaltenleitungen 808, 810, 812, 814 angelegt und dann von den Zeilenleitungen 802, 804, 806 gelesen. Für Gewichtsaktualisierungen, die im Folgenden näher beschrieben werden, werden gleichzeitig Spannungen an die Spaltenleitungen und Zeilenleitungen angelegt, und alle in den betreffenden RPU-Einheiten gespeicherten Leitfähigkeitswerte werden parallel aktualisiert. Demgemäß werden Multiplikations- und Additionsoperationen, die zum Aktualisieren der Gewichte erforderlich sind, in jeder RPU 820 des Array 600 unter Verwendung der RPU-Einheit selbst sowie der betreffenden Zeilen- oder Spaltenleitung des Array 600 lokal ausgeführt.

[0054] Ferner zeigt Fig. 6 gemäß einer oder mehreren Ausführungsformen den Ablauf eines Verfahrens zum Aktualisieren von positiven Gewichten für die

RPU 820 und deren entsprechendes Gewicht σ_{33} am Schnittpunkt zwischen der leitfähigen Zeilenleitung 806 und der leitfähigen Spaltenleitung 812. Am Rand des Kreuzschienen-Array 600 ist eine (nicht gezeigte) Aktualisierungsgenerator-Schaltlogik bereitgestellt, die als periphere „Übersetzungseinheit“ zum Erzeugen der erforderlichen Spannungsimpulse in Form von stochastischen Bitfolgen verwendet wird, die allen RPUs des 2D-Kreuzschienen-Array 600 zugeleitet werden.

[0055] Bei dem unter Verwendung eines Kreuzschienen-Arrays mit hierin beschriebenen RPUs aufgebauten ANN entscheidet demgemäß in dem Array der Widerstands- (oder Leitfähigkeits-) Wert jedes Knotens über die Kopplung zwischen den Knoten, wobei ein Knoten durch eine RPU-Einheit in dem Array dargestellt ist. Ferner unterscheidet sich nach dem Training des Kopplungs-Arrays gemäß des ANN der Widerstand (oder die Leitfähigkeit) je nach der gewünschten Kopplung von Einheit zu Einheit. Zum Trainieren eines neuronalen Netzwerks müssen die Widerstandswerte aktiv eingestellt werden. Nachdem das Training abgeschlossen ist, bleiben die Widerstandswerte während des Betriebs des Schaltkreises des Kreuzschienen-Arrays konstant, bis das Training für eine neue Aufgabe beginnt.

[0056] Im Folgenden werden Verfahren zum Bilden einer Halbleitereinheit und von Halbleitereinheiten gemäß Ausführungsformen der Erfindung unter Bezugnahme auf die beiliegenden Zeichnungen ausführlich beschrieben.

[0057] Hierin werden verschiedene Ausführungsformen der vorliegenden Erfindung unter Bezugnahme auf die zugehörigen Zeichnungen beschrieben. Es können alternative Ausführungsformen entwickelt werden, ohne vom Schutzbereich dieser Erfindung abzuweichen. Es wird darauf hingewiesen, dass in der folgenden Beschreibung und in den Zeichnungen verschiedene Verbindungen und Positionsbeziehungen (z.B. oberhalb, unterhalb, nahe usw.) zwischen Elementen dargelegt werden. Diese Verbindungen und/oder Positionsbeziehungen können direkt oder indirekt sein, sofern nichts anderes angegeben ist, und die vorliegende Erfindung soll in dieser Beziehung keine Einschränkung darstellen. Demgemäß kann sich eine Kopplung von Einheiten entweder auf eine direkte oder eine indirekte Kopplung beziehen, und eine Positionsbeziehung zwischen Einheiten kann eine direkte oder eine indirekte Positionsbeziehung sein. Gemäß einem Beispiel einer indirekten Positionsbeziehung beziehen sich Hinweise in der vorliegenden Beschreibung auf das Bilden einer Schicht „A“ oberhalb einer Schicht „B“ auf Fälle, in denen zwischen der Schicht „A“ und der Schicht „B“ eine oder mehrere Zwischenschichten (z.B. eine Schicht „C“) liegen, solange die betreffenden Eigenschaften und Funktionalitäten der Schicht

„A“ und der Schicht „B“ durch die Zwischenschicht (en) nicht wesentlich verändert werden.

[0058] Im Rahmen der folgenden Beschreibung beziehen sich die Begriffe „oben“, „unten“, „rechts“, „links“, „vertikal“, „horizontal“ und deren Ableitungen auf die beschriebenen Strukturen und Verfahren gemäß den Figuren. Die Begriffe „darüberliegend“, „oberhalb“, „auf“, „positioniert oberhalb“ oder „positioniert auf“ bedeuten, dass ein erstes Element wie beispielsweise eine erste Struktur auf einem zweiten Element wie beispielsweise einer zweiten Struktur liegt, wobei sich zwischen dem ersten Element und dem zweiten Element dazwischen liegende Elemente wie beispielsweise eine Grenzflächenstruktur befinden können. Unter dem Begriff „direkter Kontakt“ ist zu verstehen, dass ein erstes Element wie beispielsweise eine erste Struktur und ein zweites Element wie beispielsweise eine zweite Struktur ohne dazwischen liegende leitende, isolierende oder Halbleiterschichten an der Grenzflächen der beiden Elemente miteinander verbunden sind. Es sollte klar sein, dass unter dem Begriff „selektiv in Bezug auf“ wie beispielsweise „ein erstes Element selektiv in Bezug auf ein zweites Element“ zu verstehen ist, dass ein erstes Element geätzt werden und das zweite Element als Ätzstopp dienen kann. Der Begriff „ungefähr“ soll eine Fehlergröße in Bezug auf eine Messung des betreffenden Wertes auf der Grundlage der zum Zeitpunkt des Einreichens der Anmeldung verfügbaren Ausrüstung umfassen. Zum Beispiel kann der Begriff „ungefähr“ einen Bereich von $\pm 8\%$ oder 5% oder 2% eines bestimmten Wertes umfassen.

[0059] Im Folgenden wird als Hintergrundinformation eine allgemeinere Beschreibung von Prozessen zur Herstellung von Halbleitereinheiten bereitgestellt, die zum Umsetzen einer oder mehrerer Ausführungsformen der vorliegenden Erfindung genutzt werden können. Zwar sind bestimmte zum Umsetzen einer oder mehrerer Ausführungsformen der vorliegenden Erfindung verwendete Fertigungsschritte möglicherweise im Einzelfall bekannt, jedoch ist die beschriebene Kombination von Operationen und/oder resultierenden Strukturen der vorliegenden Erfindung einzigartig. Somit verwendet die einzigartige Kombination der in Verbindung mit der Herstellung einer Halbleitereinheit mit dicht gepackten vertikalen Transistoren mit verringertem Kontaktwiderstand gemäß der vorliegenden Erfindung beschriebenen Operationen eine Vielfalt im Einzelfall bekannter physikalischer und chemischer Prozesse, die auf einem Halbleitersubstrat (z.B. Silicium) durchgeführt werden, von denen einige in den unmittelbar folgenden Absätzen beschrieben werden.

[0060] Im Allgemeinen werden die verschiedenen zum Bilden eines in einen IC integrierten Mikro-Chips verwendeten Prozesse in vier wesentliche

Kategorien eingeteilt, Dünnschichtabscheidung, Entfernen/Ätzen, Halbleiterdotierung und Strukturieren/-Fotolithografie. Bei Abscheidung handelt es sich um einen beliebigen Prozess, bei dem ein Material auf den Wafer aufgewachsen, abgeschieden oder anderweitig übertragen wird. Als verfügbare Technologien kommen unter anderem physikalische Dampfabscheidung (PVD), chemische Dampfabscheidung (CVD), plasmagestützte chemische Dampfabscheidung (PEVD), elektrochemische Abscheidung (ECD), Molekularstrahl-Epitaxie (MBE) und neuerdings auch Atomlagenabscheidung (ALD) infrage.

[0061] Unter Entfernen/Ätzen ist ein beliebiger Prozess zu verstehen, bei dem ein Material von dem Wafer entfernt wird. Als Beispiele kommen Ätzprozesse (entweder Nass- oder Trockenätzen) und chemisch-mechanisches Planarisieren (CMP) und dergleichen infrage. Bei einem Nassätzprozess, beispielsweise beim Ätzen mit einer gepufferten Flusssäure (BHF) handelt es sich um einen Prozess zum Entfernen von Material, bei dem flüssige Chemikalien oder Ätzmittel zum Entfernen von Materialien von einer Oberfläche verwendet werden. Bei einem Trockenätzprozess wie beispielsweise beim reaktiven Ionenätzen (RIE) wird zum Entfernen eines Materials aus einer maskierten Struktur eines Halbleitermaterials ein chemisch reaktives Plasma verwendet, indem das Material einem Ionenbeschuss ausgesetzt wird, um Teile des Material von der freiliegenden Oberfläche zu entfernen. Das Plasma wird durch ein elektromagnetisches Feld bei niedrigem Druck (Vakuum) erzeugt.

[0062] Unter Halbleiterdotierung ist die Veränderung elektrischer Eigenschaften durch Dotieren von zum Beispiel Source- und Drain-Elektroden von Transistoren im Allgemeinen durch Diffusion und/oder Ionenimplantation zu verstehen. An diese Dotierungsprozesse schließen sich Tempern und/oder Schnelltempern (RTA) an. Tempern dient dem Aktivieren der implantierten Dotanden. Schichten sowohl aus Leitern (z.B. Polysilicium, Aluminium, Kupfer usw.) als auch Isolatoren (z.B. verschiedene Formen von Siliciumdioxid, Siliciumnitrid usw.) werden zum Verbinden und Isolieren von Transistoren und deren Komponenten verwendet. Durch selektives Dotieren verschiedener Bereiche des Halbleitersubstrats kann die Leitfähigkeit des Substrats durch Anlegen einer Spannung verändert werden. Durch Schaffen von Strukturen aus diesen verschiedenen Komponenten können Millionen von Transistoren erzeugt und miteinander elektrisch verbunden werden, um die komplexe Schaltlogik einer modernen Mikroelektronik-Einheit zu bilden.

[0063] Unter Halbleiterlithografie ist das Bilden von Bildern oder Mustern mit einem dreidimensionalen Relief auf dem Halbleitersubstrat zu verstehen, wobei das Muster anschließend auf das Substrat

übertragen wird. In der Halbleiterlithografie werden die Muster durch ein lichtempfindliches Polymer mit der Bezeichnung Fotolack gebildet. Zum Schaffen der komplexen Strukturen, aus denen ein Transistor besteht, und der vielen Leiterbahnen, die die Millionen von Transistoren eines Schaltkreises miteinander verbinden, werden die Schritte der Lithografie und zum Übertragen der Ätzmuster mehrmals wiederholt. Jedes auf den Wafer zu übertragende Muster wird auf die zuvor gebildeten Muster ausgerichtet, und nach und nach werden die Leiterbahnen, Isolatoren und selektiv dotierten Bereiche gebildet, um die fertige Einheit zu bilden.

[0064] Gemäß einer oder mehreren Ausführungsformen werden resistive Elemente unter Verwendung von Halbleiterstreifen beispielsweise aus Polysilizium gebildet. Die Streifen können dotiert werden, um den spezifischen Widerstand des Halbleiters zu steuern. Üblicherweise kann der Widerstandswert unter Verwendung von Streifen unterschiedlicher Abmessungen variiert werden. Bei dem üblichen Verfahren können jedoch mehrere Lithografie-Masken aufeinander abgestimmt werden, damit Kontakte mit den Enden der Streifen hergestellt werden.

[0065] Gemäß einer Übersicht über Aspekte der vorliegenden Erfindung stellen eine oder mehrere Ausführungsformen Verfahren zum Herstellen einer Halbleitereinheit mit einem Koppeleinheiten-Array bereit, das mehrere Koppeleinheiten mit hohen steuerbaren Widerstandswerten ($>10\text{ M}\Omega$) enthält. Gemäß einer oder mehreren Ausführungsformen werden Fotolithografie und/oder Elektronenstrahl-Lithografie mit nachfolgendem selektivem Plasmaätzen zum Herstellen des Koppeleinheiten-Arrays angewendet, das mehrere Koppeleinheiten enthält.

[0066] Fig. 7 zeigt einen Ablaufplan für ein beispielhaftes Verfahren zum Herstellen eines Planar-Diffusions-FET gemäß einer oder mehreren Ausführungsformen der vorliegenden Erfindung. Das Verfahren umfasst Bereitstellen eines Substrats 805, beispielsweise eines Silizium-auf-Isolator- (SOI-) Wafers oder eines beliebigen anderen Halbleiter-substrats (702).

[0067] Fig. 8 zeigt eine Querschnittsansicht eines beispielhaften Wafers 805, der zum Herstellen einer Halbleitereinheit 800 gemäß einer oder mehreren Ausführungsformen der vorliegenden Erfindung verwendet wird. Gemäß einem oder mehreren Beispielen ist der Wafer 800 schwach dotiert. Im Falle des Wafers 800 handelt es sich um einen SOI-Wafer, der ferner eine Isolatorschicht (z.B. eine vergrabene Oxid- (BOX-) Schicht oder eine andere geeignete Isolatorschicht) auf dem Substrat 805 enthalten kann.

[0068] Ebenfalls unter Bezugnahme auf Fig. 7 umfasst das Verfahren zum Herstellen des Planar-FET ferner Strukturieren eines Source-Bereichs 825 und eines Kanalbereichs 827 für eine FET-Struktur in dem Substrat 805 (704).

[0069] Fig. 9 zeigt eine beispielhafte FET-Struktur gemäß einer oder mehreren Ausführungsformen der vorliegenden Erfindung. Die FET-Struktur enthält flache Isolationsgraben- (STI-) Bereiche 820 zum Definieren eines Einheitenbereichs innerhalb der Halbleiterschicht 805. Gemäß einem oder mehreren Beispielen wird der Strukturierungs- und Dotierungsprozess durchgeführt, um den Source-Bereich und den Drain-Bereich 825 beziehungsweise 827 zu bilden. Zwischen dem Source-Bereich und dem Drain-Bereich befindet sich dann der oben angegebene Kanalbereich 850. Der Strukturierungsprozess wird mittels Fotolithografie und der Dotierungsprozess mittels Ionenimplantation und Hochtemperaturtempern durchgeführt.

[0070] Ferner umfasst das Herstellungsverfahren (Fig. 7) Bilden eines Energiebarrierebereichs 830 (706).

[0071] Fig. 10 zeigt eine beispielhafte FET-Struktur gemäß einer oder mehreren Ausführungsformen der vorliegenden Erfindung nach dem Bilden der Energiebarriere 830. Zusätzlich zur Source-Elektrode 825 und zur Drain-Elektrode 827 wird zwischen die Source-Elektrode 825 und den Kanal 850 durch Strukturieren und teilweises Dotieren des Kanals 850 die Energiebarriere 830 eingefügt. In einem Bereich zwischen der Source-Elektrode 825 (oder der Drain-Elektrode 827) und dem Kanal 850 wird die Energiebarriere 830 mit der vorgegebenen Breite L_d erzeugt.

[0072] Die Energiebarriere 830 wird durch stärkeres Dotieren des Bereichs der Energiebarriere 830 im Vergleich mit der Dotierung des Kanals 850 erzeugt. Zum Beispiel wird im Falle der Ionenimplantation die Energiebarriere 830 stärker dotiert (P^+) als der Kanal 850 (P^-), und die Source-Elektrode 825 und die Drain-Elektrode 827 werden mit N dotiert, sodass ein N - P - N -FET gebildet wird. Gemäß anderen Ausführungsformen wird ein P - N - P -FET erzeugt, indem durch Implantieren von Ionen die Energiebarriere 830 stärker dotiert wird (N^+) als der Kanal 850 (N^-) und die Source-Elektrode 825 und die Drain-Elektrode 827 mit P dotiert werden.

[0073] Die dotierten Bereiche (825, 827 und 830) können mittels einer Vielfalt an Verfahren wie beispielsweise Implantation und Plasmadotierung in dem Substrat 815 gebildet werden. Gemäß einem oder mehreren Beispielen können sie abgeätzt und mittels verschiedener Verfahren wieder abgeschieden werden, zum Beispiel mittels Epitaxie mit in

situ-Dotierung, Dotierung nach Epitaxie und dergleichen. Die dotierten Bereiche können durch einen beliebigen geeigneten Prozess gebildet werden, darunter, ohne darauf beschränkt zu sein, chemische Dampfabcheidung im Ultrahochvakuum (UHVCVD), schnelle thermische chemische Dampfabcheidung (RTCVD), chemische Dampfabcheidung von metallorganischen Verbindungen (MOCVD), chemische Dampfabcheidung bei niedrigem Druck (LPCVD), CVD mit eingeschränkter Reaktion (LRPCVD) und Molekularstrahl-Epitaxie (MBE). Gemäß einigen Ausführungsformen enthalten die dotierten Bereiche Halbleitermaterialien, die aus gasförmigen oder flüssigen Ausgangsstoffen epitaxial abgeschieden wurden.

[0074] Fig. 11 zeigt eine Auswirkung der Energiebarriere in der hierin beschriebenen FET-Struktur gemäß einer oder mehreren Ausführungsformen der vorliegenden Erfindung auf die Stromstärke. Die Kurve 1110 zeigt einen Zustand EIN eines Diffusions-FET, und die Kurve 1120 zeigt einen Zustand AUS. In beiden Kurven sind ein Valenzband 1135 und ein Leitungsband 1145 gezeigt. Normalerweise wird in einem Diffusions-FET die Stromstärke des Zustandes „EIN“ durch den Diffusionsstrom von der Source-Elektrode 825 zur Drain-Elektrode 827 durch den Kanal 850 bestimmt. Bei der hierin beschriebenen FET-Struktur wird die Stromstärke durch die Höhe (H) der Energiebarriere zwischen der Source-Elektrode 825 und dem zusätzlichen Dotierungsbereich 830 bestimmt. Diese Barrierehöhe kann mittels der Gate-Spannung durch die Abstandsschicht an der Seitenwand moduliert werden. Bei der Abstandsschicht handelt es sich um eine isolierende Schicht, vorzugsweise aus einem Dielektrikum mit hoher Dielektrizitätskonstante. Durch eine Ladung an der Gate-Elektrode wird in der Abstandsschicht ein Randfeld erzeugt, durch das die Höhe der Energiebarriere moduliert wird.

[0075] Fig. 12 zeigt eine Auswirkung der Energiebarriere in der hierin beschriebenen FET-Struktur gemäß einer oder mehreren Ausführungsformen der vorliegenden Erfindung auf die Stromstärke. Die dargestellte Kurve zeigt einen Vergleich einer ersten Übertragungskennlinie 1210 eines typischen FET ohne die Energiebarriere 830 und eine zweite Übertragungskennlinie 1220 eines Diffusions-FET mit der Energiebarriere 830. Es zeigt sich, dass die Stromstärke bei dem Diffusions-FET mit der Energiebarriere 830 niedriger ist. Dabei ist zu beachten, dass die Kurven in Fig. 12 Ergebniswerte eines bestimmten beispielhaften Szenarios veranschaulichen und dass die Kurven bei anderen Beispielen von den dargestellten Kurven abweichen können.

[0076] Fig. 13 zeigt eine Auswirkung der Energiebarriere in der hierin beschriebenen FET-Struktur gemäß einer oder mehreren Ausführungsformen

der vorliegenden Erfindung auf die Stromstärke. Die veranschaulichte Kurve zeigt den Widerstand des Diffusions-FET mit der Energiebarriere 830 als Funktion der Gate-Spannung bei verschiedenen Längen (Ld) der Energiebarriere 830 mit der extra starken Dotierung. Zum Beispiel zeigt die Kurve den Widerstand des FET für Längen Ld von 16 nm (1330), 18 nm (1320) beziehungsweise 20 nm (1310). Dabei ist zu beachten, dass der Wert Ld von den oben angegebenen beispielhaften Werten abweichen kann, sodass sich der Widerstand des FET anders als in Fig. 13 gezeigt ändert.

[0077] Bezugnehmend wiederum auf Fig. 7 umfasst das Verfahren zur Herstellung des Planar-FET herkömmliches Verarbeiten, um einen Gate-Elektrodenstapel 840 (708) zu bilden.

[0078] Fig. 14 zeigt eine beispielhafte FET-Struktur gemäß einer oder mehreren Ausführungsformen der vorliegenden Erfindung nach dem Bilden des Gate-Elektrodenstapels. Die FET-Struktur enthält ferner einen Gate-Elektrodenstapel 840 (z.B. eine dielektrische Gate-Elektrodenschicht 841, eine leitfähige Gate-Elektrodenschicht 842 auf der dielektrischen Gate-Elektrodenschicht 841 und eine Deckschicht 843, beispielsweise eine Nitrid-Deckschicht auf der leitfähigen Gate-Elektrodenschicht 842) oberhalb des vorgesehenen Kanalbereichs 850 innerhalb des Einheitenbereichs. Üblicherweise besteht die dielektrische Gate-Elektrodenschicht aus einem Siliciumoxid, jedoch kann ein beliebiges geeignetes Material als Dielektrikum für die Gate-Elektrode verwendet werden. Als Beispiele für andere Dielektrika der Gate-Elektrode kommen HfO_2 und Al_2O_3 infrage. Die leitfähige Gate-Elektrodenschicht 842 kann aus Metall und/oder Polysilicium oder einem oder mehreren beliebigen anderen Materialien bestehen, die als Gate-Elektrode für die Halbleitereinheit 800 geeignet sind.

[0079] Ferner sind auf beiden Seiten des Gate-Elektrodenstapels 840 (710) Abstandsschichten 845 als Seitenwand der Gate-Elektrode gebildet. Gemäß einer oder mehreren Beispielen können (z.B. früher oder später) je nach Integrationskonzept Erweiterungsbereiche und/oder Randbereiche der Source-Elektrode/Drain-Elektrode gebildet werden. Die Einzelheiten der oben erwähnten herkömmlichen Verarbeitung sind allgemein bekannt und werden hier weggelassen, um die Aufmerksamkeit auf die wesentlichen Aspekte der hierin beschriebenen Ausführungsformen zu lenken.

[0080] Ferner wird zum Speichern des Gewichts in dem FET 800 das Gate-Elektroden-Potenzial auf einem bestimmten Wert gehalten. Dies lässt sich durch verschiedene Strukturen erreichen.

[0081] Fig. 15 zeigt eine Ausführungsform zum Speichern des Gewichts in dem FET 800 gemäß einer oder mehreren Ausführungsformen der vorliegenden Erfindung. Der FET 800 ist mit einem Kondensator 1510 verbunden. Die Ladung wird auf der Platte gespeichert, um die Gate-Spannung bereitzustellen, wobei die gespeicherte Ladung dem in der Koppeleinheit des RPU-Array 600 gespeicherten Gewicht entspricht. Zu beachten ist hierbei, dass zum Vereinfachen der Darstellung nur einige Teile der FET-Einheit 800 gezeigt sind.

[0082] Fig. 16 zeigt eine Ausführungsform zum Speichern des Gewichts in dem FET 800 gemäß einer oder mehreren Ausführungsformen der vorliegenden Erfindung. Zu dem Gate-Elektrodenstapel 840 ist eine Ladungsspeicherschicht 1610 hinzugefügt, beispielsweise eine Floating-Gate-Elektrode. Die Ladung kann in der Ladungsspeicherschicht 1610 gespeichert werden, um die Gate-Spannung bereitzustellen. Die Ladungsspeicherschicht 1610 besteht aus Polysilicium. Gemäß einer oder mehreren Ausführungsformen können nach der Ladungsspeicherschicht 1610 eine weitere dielektrische Gate-Elektroden-schicht 841 und anschließend die weiteren Schichten in dem Gate-Elektrodenstapel 840 hinzugefügt werden, beispielsweise die Schicht 842 aus leitfähigem Material für die Elektrode.

[0083] Fig. 17 zeigt eine weitere Ausführungsform zum Speichern des Gewichts in dem FET 800 gemäß einer oder mehreren Ausführungsformen der vorliegenden Erfindung. Hierbei enthält der Gate-Elektrodenstapel die dielektrische Gate-Elektroden-schicht 841 und anschließend die erste Schicht 842 aus leitfähigem Material. Anschließend wird eine Schicht 1710 aus ferroelektrischem Material hinzugefügt, die beispielsweise aus Hafniumdioxid (HfO_2) besteht. Durch die Polarisierung der ferroelektrischen Schicht 1710 wird in der Abstandsschicht ein elektrisches Feld erzeugt, das die Höhe der Energiebarriere in der Schicht 830 moduliert. Ferner enthält der Gate-Elektrodenstapel 840 eine zweite Schicht 842 aus leitfähigem Material, die die Steuerelektrode der Gate-Elektrode bildet. Die erste Schicht 842 aus leitfähigem Material behebt ein technisches Problem der Depolarisation der Schicht 1710 aus ferroelektrischem Material.

[0084] Ferner wird gemäß einem oder mehreren Beispielen die Energiebarriere 830 zwischen der Source-Elektrode 825 und dem Kanal 850 indirekt über die Dielektrika der Seitenwand 845 moduliert. Zum Beispiel können die Seitenwände 845 aus Hafniumdioxid (HfO_2) bestehen, und zum Bilden einer dielektrischen Schicht mit hoher Dielektrizitätskonstante werden abgeschiedene Schichten aus Titan-nitrid (TiN) verwendet. Die ferroelektrische Schicht 1710 und die Seitenwände 845 bilden somit einen

metallischen Gate-Elektrodenstapel mit hoher Dielektrizitätskonstante (MHK).

[0085] Somit bieten die eine oder die mehreren hierin beschriebenen Planar-Diffusions-FET-Halbleiter-einheiten technische Lösungen für die technischen Probleme beim Bereitstellen eines steuerbaren Widerstands, insbesondere oberhalb 10 M Ω . Solche Planar-Diffusions-FETs können in Koppeleinheiten als Bestandteil eines RPU-Array 600 verwendet werden, das Matrixmultiplikationen durchführt, beispielsweise zum Realisieren von ANNs.

[0086] Gemäß einer oder mehreren Ausführungsformen der vorliegenden Erfindung handelt es sich ferner bei dem Halbleiter-FET um einen Heterobarriere-FET (HBFET), zum Beispiel mit einer (III-V-Struktur). **Fig. 18** zeigt eine beispielhafte Halbleiterstruktur für den FET gemäß einer oder mehreren Ausführungsformen der vorliegenden Erfindung. Die (gezeigte) Heterobarriere-FET-(III-V)-Struktur ist dem oben beschriebenen Diffusions-FET ähnlich, wobei der Heterobarriere-FET keinen dotierten Bereich als Energiebarrierebereich 830 nahe der Source-Elektrode verwendet, sondern ein Hetero-übergangsmaterial enthält.

[0087] Zum Beispiel bestehen in dem HBFET 1800 die Source-Elektrode 825, die Drain-Elektrode 817 und der Kanal 850 aus einem III-V-Material mit kleiner Bandlücke wie beispielsweise aus Indium-Gallium-Arsenid (InGaAs) oder Gallium-Arsenid-Antimonid (GaAsSb), um die Schaltgeschwindigkeit des komplementären Metalloxid-Halbleiter-(CMOS-) Transistors zu erhöhen. Ferner besteht die Energiebarriere 830 aus einem Material mit großer Bandlücke wie beispielsweise Indiumphosphid (InP), um eine Hetero-Barriere zu bilden. Der HBFET 1800 enthält ferner den Gate-Elektrodenstapel 840 mit Seitenwänden, die aus einem Abstandsmaterial wie Si_3N_4 bestehen, das eine Modulation der Hetero-Barriere durch Randfelder ermöglicht. Gemäß einem oder mehreren Beispielen sind die Bereiche zum Bilden eines NPN-HBFET so dotiert, dass die Source-Elektrode 825 und die Drain-Elektrode 827 mit N $^+$ und die Energiebarriere 830 und der Kanal 850 mit P $^-$ dotiert sind. Zu beachten ist, dass die Dotierung zum Bilden eines PNP-HBFET gemäß anderen Beispielen verschieden sein kann.

[0088] Fig. 19 zeigt eine Auswirkung der Energiebarriere in der hierin beschriebenen III-V-HBFET-Struktur 1800 gemäß einer oder mehreren Ausführungsformen der vorliegenden Erfindung auf die Stromstärke. Die Kurve 1910 zeigt einen Zustand EIN des HBFET, und Kurve 1920 zeigt einen Zustand AUS des HBFET. In beiden Kurven sind ein Valenzband 1935 und ein Leitungsband 1945 gezeigt. Normalerweise wird in einem Diffusions-FET die Stromstärke des Zustands „EIN“ durch den Diffusionsstrom

von der Source-Elektrode 825 zur Drain-Elektrode 827 durch den Kanal 850 bestimmt. In der hierin beschriebenen HBFET-Struktur wird die Stromstärke durch die Höhe (H) der Hetero-Energiebarriere zwischen der Source-Elektrode 825 und dem Energiebarrierebereich 830 bestimmt. Die Höhe dieser Energiebarriere kann mittels der Gate-Spannung durch die Abstandsschicht der Seitenwand moduliert werden.

[0089] Gemäß einer oder mehreren Ausführungsformen der vorliegenden Erfindung handelt es sich bei dem HBFET 1800 um eine Si-SiGe-Struktur, wobei die Source-Elektrode 825, die Drain-Elektrode 827 und der Kanal 850 aus verspanntem Silicium-Germanium (SiGe) oder Germanium (Ge). Ferner besteht die Energiebarriere 830 aus Si. Der HBFET 1800 enthält ferner einen Gate-Elektrodenstapel 840, dessen Seitenwände aus einem Abstandsmaterial wie Si_3N_4 bestehen. Gemäß einem oder mehreren Beispielen sind die Bereiche zum Bilden eines NPN-HBFET so dotiert, dass die Source-Elektrode 825 und die Drain-Elektrode 827 mit N⁺ und die Energiebarriere 830 und der Kanal 850 mit P⁻ dotiert sind. Zu beachten ist, dass die Dotierung zum Bilden eines PNP-HBFET bei anderen Beispielen verschieden sein kann.

[0090] Fig. 20 zeigt eine Auswirkung der Energiebarriere in der hierin beschriebenen SiGe-Si-HBFET-Struktur 1800 gemäß einer oder mehreren Ausführungsformen der vorliegenden Erfindung auf die Stromstärke. Die Kurve 2010 zeigt einen Zustand EIN des HBFET, und die Kurve 2020 zeigt einen Zustand AUS des HBFET. In beiden Kurven sind ein Valenzband 2035 und ein Leitungsband 2045 gezeigt. Normalerweise wird in einem Diffusions-FET die Stromstärke des Zustands „EIN“ durch den Diffusionsstrom von der Source-Elektrode 825 durch den Kanal 850 zur Drain-Elektrode 827 bestimmt. In der hierin beschriebenen HBFET-Struktur wird die Stromstärke durch die Höhe (H) der Hetero-Energiebarriere zwischen der Source-Elektrode 825 und dem Bereich der Energiebarriere 830 bestimmt. Die Höhe dieser Barriere kann durch die Gate-Spannung moduliert werden.

[0091] Ferner kann das Gate-Potenzial der HBFET-Einheit 1800 zum Speichern des Gewichts in dem HBFET 1800 auf einem bestimmten Wert gehalten werden. Dies kann mittels verschiedener Strukturen erreicht werden.

[0092] Fig. 21 zeigt eine Ausführungsform zum Speichern des Gewichts in dem HBFET 1800 gemäß einer oder mehreren Ausführungsformen der vorliegenden Erfindung. Der HBFET 1800 ist mit einem Kondensator 1510 verbunden. Die Ladung wird auf der Platte gespeichert, um die Gate-Spannung bereitzustellen, wobei die gespeicherte Ladung

dem in der Koppereinheit des RPU-Array 600 gespeicherten Gewicht entspricht. Dabei ist zu beachten, das zur vereinfachten Darstellung in Fig. 21 nur einige Teile der HBFET-Einheit 1800 gezeigt sind.

[0093] Fig. 22 eine Ausführungsform zum Speichern des Gewichts in dem HBFET 1800 gemäß einer oder mehreren Ausführungsformen der vorliegenden Erfindung. Eine Ladungsspeicherschicht 1610 wie beispielsweise eine Floating-Gate-Elektrode ist zu dem Gate-Elektrodenstapel 840 hinzugefügt. Die Ladung kann in der Ladungsspeicherschicht 1610 gespeichert werden, um die Gate-Spannung bereitzustellen. Die Ladungsspeicherschicht 1610 besteht aus Polysilicium. Gemäß einem oder mehreren Beispielen werden nach der Ladungsspeicherschicht 1610 eine weitere dielektrische Gate-Elektrodenschicht 841 und anschließend die weiteren Schichten in dem Gate-Elektrodenstapel 840 hinzugefügt, beispielsweise die Schicht 842 aus leitfähigem Material für die Elektrode.

[0094] Fig. 23 zeigt eine weitere Ausführungsform zum Speichern des Gewichts in dem HBFET 1800 gemäß einer oder mehreren Ausführungsformen der vorliegenden Erfindung. In diesem Fall enthält der Gate-Elektrodenstapel die dielektrische Gate-Elektrodenschicht 841, auf die eine erste Schicht 842 aus leitfähigem Material folgt. Anschließend wird eine Schicht 1710 aus ferroelektrischem Material hinzugefügt, die aus Hafniumdioxid (HfO_2) besteht. Die Polarisierung der ferroelektrischen Schicht 1710 erzeugt in der Abstandsschicht der Seitenwand ein elektrisches Feld, um die Höhe der Energiebarriere der Schicht 830 zu modulieren. Ferner enthält der Gate-Elektrodenstapel 840 eine zweite Schicht 842 aus leitfähigem Material zum Bilden der Steuerelektrode der Gate-Elektrode. Durch die erste Schicht 842 aus leitfähigem Material wird ein technisches Problem der Depolarisierung der Schicht 1710 aus ferroelektrischem Material behoben.

[0095] Ferner wird gemäß einem oder mehreren Beispielen die Energiebarriere 830 zwischen der Source-Elektrode 825 und dem Kanal 850 indirekt über die Dielektrika der Seitenwand 845 moduliert. Zum Beispiel können die Seitenwände 845 aus Hafniumdioxid (HfO_2) bestehen, und zum Bilden einer Schicht mit hoher Dielektrizitätskonstante wird abgeschiedenes Titannitrid (TiN) verwendet. Somit bilden die ferroelektrische Schicht 1710 und die Seitenwände 845 einen Metall-Gate-Elektrodenstapel mit hoher Dielektrizitätskonstante.

[0096] Somit stellen der eine oder mehrere hierin beschriebene Planar-HBFET-Halbleitereinheiten technische Lösungen für die technischen Probleme beim Bereitstellen von steuerbaren Widerständen bereit, insbesondere oberhalb von 10 MΩ. Solche

Planar-HBFETs können in Koppereinheiten als Bestandteil eines RPU-Array 600 verwendet werden, die Matrixmultiplikationen ausführen, um beispielsweise ANNs zu realisieren.

[0097] Ferner kann es sich gemäß einem erläuterten Beispiel bei dem Halbleiter-HBFET um einen vertikalen HBFET handeln. **Fig. 24** zeigt ein beispielhaftes RPU-Array 600, in dem eine vertikale Halbleiterstruktur für den HBFET verwendet wird. Das RPU-Array 600 enthält den vertikalen HBFET 2400 sowie weitere Schaltkreise (z.B. Kondensator, FETs zur Gewichtsaktualisierung usw.) an jedem Kreuzungspunkt. Eine Querschnittsansicht des vertikalen HBFET 2400 entlang der Achse A-A' ist gezeigt. Der vertikale HBFET 2400 kann die Energiebarriere 830 enthalten (III-V oder SiG-Si oder beliebige andere).

[0098] Im vorliegenden Falle sind 802, 804, 806 die oberen Metallleitungen (Zeilenleitungen), die an den oberen Anschlüssen der Einheit anliegen (in diesem Falle die Drain-Elektrode 827). Der untere Kontakt (Source-Elektrode 825) ist durch die Halbleiterschichten gebildet, die beim Erzeugen der aktiven Halbleiterbereiche 808, 810 (spaltenweise) auf dem Substrat abgeschieden wurden. Außerdem enthält der vertikale HBFET 2400 den Stapel aus verschiedenen Halbleiterschichten, die durch die Prozessschritte in den **Fig. 25** bis **Fig. 36** dargestellt sind. Der restliche Bereich 808, 810 außerhalb des HBFET 2400 hingegen wird bis zur (p+)-SiGe-Schicht 2530 abgeätzt und bildet somit den gemeinsamen unteren Leiter (spaltenweise) zu den (nicht gezeigten) HBFETs 808, 810.

[0099] Im Folgenden wird die Herstellung des vertikalen HBFET 2400 in dem RPU-Array 600 beschrieben. In den verschiedenen zur Beschreibung des Herstellungsverfahrens verwendeten Abbildungen werden für jeden Schritt Querschnittsansichten entlang der Achse A-A' verwendet. Der Herstellungsprozess wird für die Struktur auf der Grundlage von Si-SiGe beschrieben, jedoch kann der Fachmann die Beschreibung zum Herstellen anderer Typen als des hierin beschriebenen vertikalen HBFET anwenden.

[0100] **Fig. 25** zeigt eine Substratvorbereitung des Bereichs für den HBFET 2400 (insbesondere in Bezug auf **Fig. 20**). Zuerst wird auf dem Substrat 2510 ein HBFET-Bereich definiert (zum Beispiel Öffnen eines Einheitenbereichs nach dem STI-Verfahren). Bei dem Substrat 2510 kann es sich um ein beliebiges geeignetes Substratmaterial handeln, zum Beispiel um SiC oder Halbleiter-auf-Isolator (SOI) oder dergleichen. Gemäß einem Beispiel enthält das Substrat 2510 eine (nicht gezeigte) vergrabene Oxidschicht. Der aktive Halbleiterbereich kann durch flache Grabenisolation (STI) 2512 von ande-

ren Bereichen des Substrats 2510 elektrisch isoliert werden. Bei der STI 2512 kann es sich um ein beliebiges geeignetes dielektrisches Material wie beispielsweise Siliciumdioxid handeln. Zum Bilden der Isolation 2512 zwischen Einheiten kann unterschiedlich verfahren werden. Gemäß einem Beispiel wird die STI 2512 gebildet, indem zuerst das dielektrische Material (2512) auf dem gesamten Substrat abgeschieden und dann der aktive Halbleiterbereich 808, 810 unter Verwendung eines Ätzprozesses geöffnet wird, worauf innerhalb der oben erwähnten Öffnung weitere Halbleiterschichten 2520, 2530, 2540, 2550, 2560 abgeschieden werden. Gemäß einem Beispiel wird die STI 2512 gebildet, indem zuerst verschiedene Halbleiterschichten 2520, 2530, 2540, 2550, 2560 auf dem gesamten Substrat 2510 abgeschieden und dann zum Bilden eines Grabens bis zum Substrat 2510 abgeätzt werden, der Graben mit dem STI-Material 2512 aufgefüllt und zum Beispiel unter Verwendung eines CMP-Prozesses bis auf eine Oberfläche der Halbleiterschicht 2560 planarisiert wird.

[0101] Demgemäß wird auf das Substrat 2510 ein spannungsgeminderter Pufferbereich (SRB) 2520 abgeschieden. Anschließend wird auf dem SRB-Bereich 2520 (P+)-SiGe (untere Source-Elektrode 825) abgeschieden. Darauf folgt Abscheiden der (n-)-Si-Schicht 2540 (Hetero-Barriere 830). Anschließend wird eine (N-)-SiGe-Schicht 2550 (Kanal 850) abgeschieden. Dann wird eine (P+)-SiGe-Schicht 2560 (untere Drain-Elektrode 827) abgeschieden. Zu beachten ist, dass die Schichten zwar als Source-Elektrode und Drain-Elektrode bezeichnet werden, jedoch können gemäß anderen Beispielen die Drain- und Source-Schichten vertauscht sein.

[0102] Bei dem stark dotierten Bereich 2530 des Substrats kann es sich zum Beispiel um einen Source- oder Drain-Bereich handeln, der durch eine Vielfalt an Verfahren in dem Substrat gebildet sind, zum Beispiel epitaxiale situ-Dotierung, Dotierung nach Epitaxie oder durch Implantation und Plasma-dotierung. Der stark dotierte Bereich 2530 kann mittels eines beliebigen geeigneten Prozesses gebildet werden, darunter, ohne darauf beschränkt zu sein, chemische Dampfabcheidung im Ultrahochvakuum (UHVCVD), schnelle thermische Dampfabcheidung (RTCVD), chemische Dampfabcheidung von metallorganischen Verbindungen (MOCVD), chemische Dampfabcheidung bei niedrigem Druck (LPCVD), CVD mit eingeschränkter Reaktion (LRPCVD) und MBE. Gemäß einem Beispiel enthält der stark dotierte Bereich 2530 aus gasförmigen oder flüssigen Ausgangsstoffen epitaxial abgeschiedene Halbleitermaterialien. Gemäß einem Beispiel werden epitaxiale Bereiche durch Epitaxie auf dem Substrat 2510 abgeschieden. Epitaxiale Halbleitermaterialien können unter Verwendung von Dampfphasenepitaxie (VPE), MBE, Flüssigphasenepitaxie (LPE) oder

anderen geeigneten Prozessen abgeschieden werden. Epitaxial abzuscheidendes Silicium, SiGe und/oder mit Kohlenstoff dotiertes Silicium (Si:C) kann während der Abscheidung (in situ-Dotierung) durch Hinzufügen von Dotanden, n-Typ-Dotanden (z.B. Phosphor oder Arsen) oder p-Typ-Dotanden (z.B. Ga, B, BF₂ oder Al) dotiert werden. Die Dotandenkonzentration in den dotierten Bereichen kann von $1 \times 10^{19} \text{ cm}^{-3}$ bis $2 \times 10^{21} \text{ cm}^{-3}$ reichen oder zwischen $1 \times 10^{20} \text{ cm}^{-3}$ und $1 \times 10^{21} \text{ cm}^{-3}$ liegen.

[0103] Gemäß einem Beispiel enthält die Gasquelle zum Abscheiden von epitaxialem Halbleitermaterial eine Silicium enthaltende Gasquelle, eine Germanium enthaltende Gasquelle oder eine Kombination derselben. Zum Beispiel kann eine epitaxiale Si-Schicht aus einer Silicium enthaltenden Gasquelle abgeschieden werden, die aus der Gruppe ausgewählt wird, die aus Silan, Disilan, Trisilan, Tetrasilan, Hexachlordisilan, Tetrachlorsilan, Dichlorsilan, Trichlorsilan, Methylsilan, Ethylsilan, Methylidisilan, Methylidisilan, Dimethylidisilan, Hexamethylidisilan und deren Kombinationen besteht. Eine epitaxiale Germaniumschicht kann aus einer Germanium enthaltenden Gasquelle abgeschieden werden, die aus der Gruppe ausgewählt wird, die aus German, Digerman, Halogerman, Dichlorgerman, Trichlorgerman, Tetrachlorgerman und deren Kombinationen besteht. Eine epitaxiale Silicium-Germanium-Legierungsschicht kann unter Verwendung einer Kombination solcher Gasquellen gebildet werden. Es können Trägergase wie Wasserstoff, Stickstoff, Helium und Argon verwendet werden. Gemäß einem Beispiel enthalten die dotierten Bereiche mit Kohlenstoff dotiertes Silicium (Si:C). Diese Si:C-Schicht kann in derselben für andere Epitaxieschritte genutzten Kammer oder in einer speziellen Si:C-Epitaxiekammer abgeschieden werden. Das Si:C kann Kohlenstoff im Bereich von ungefähr 0,2 Prozent bis ungefähr 3,0 Prozent enthalten.

[0104] Fig. 26 zeigt eine nachfolgende Zwischenstruktur des HBFET 2400 während des Herstellungsprozesses gemäß einem erläuternden Beispiel. In diesem Schritt wird unter Verwendung lithografischer Techniken ein vertikaler Ätzschritt ausgeführt, um jede der SiGe-Schichten 2550 und 2560 bis zur ersten Si-Schicht 2540 abzuätzen. Zum vertikalen Ätzen wird eine Maske unter Verwendung vorgegebener Abmessungen verwendet, um eine Drain-Elektrode und einen Kanal mit vorgegebenen Abmessungen stehenzulassen.

[0105] Zum Beispiel wird auf einer Oberfläche jedes der Halbleiterstege eine Hartmaske gebildet. Die Hartmaske kann ein Oxid, ein Nitrid, ein Oxynitrid oder eine Kombination derselben enthalten, darunter auch Mehrschichten. Gemäß einem Beispiel kann die Hartmaske Siliciumoxid oder Siliciumnitrid enthalten. Die Hartmaske kann unter Verwendung

eines Abscheidungsprozesses gebildet werden, zum Beispiel mittels chemischer Dampfabscheidung (CVD), plasmagestützter chemischer Dampfabscheidung (PECVD), chemischer Abscheidung aus Lösungen, Verdampfung. Gemäß einem Beispiel kann die Hartmaske durch einen thermischen Prozess gebildet werden, zum Beispiel durch Oxidieren oder Nitridieren auf der Halbleiterschicht. Zum Bilden der Hartmaske kann auch eine beliebige Kombination der oben erwähnten Prozesse verwendet werden. Die Hartmaske kann eine Dicke von 20 nm bis 80 nm, zum Beispiel von 30 nm bis 60 nm haben.

[0106] Gemäß einem Beispiel wird die Hartmaske vor den Halbleiterstegen gebildet. Dann wird die Hartmaske strukturiert, und das Muster wird auf das Substrat 2510 übertragen, um die Halbleiterstege unter Verwendung bekannter lithografischer Prozesse zu bilden. Der lithografische Schritt kann Aufbringen einer (nicht gezeigten) Fotolackschicht auf die Hartmaske, Belichten der Fotolackschicht mit einem gewünschten Strahlungsmuster und Entwickeln der belichteten Fotolackschicht unter Verwendung eines Fotolackentwicklers umfassen. Der Ätzprozess kann Trockenätzen und/oder chemisches Nassätzen umfassen. Als Beispiele für verwendbare Trockenätzprozesse kommen reaktives Ionenätzen (RIE), Ionenstrahlätzen oder Laserablation infrage. Durch den Ätzprozess kann das Muster der strukturierten Fotolackschicht auf die Hartmaske und auf das Substrat 2510 übertragen werden. Gemäß einem Beispiel dient eine (nicht gezeigte) vergrabene Isolatorschicht als Ätzstopp. Nach dem Bilden der Halbleiterstege kann die strukturierte Fotolackschicht unter Verwendung eines Fotolack-Ablöseprozesses (Stripping) entfernt werden, zum Beispiel durch Veraschen. Gemäß einem Beispiel werden die Halbleiterstege unter Verwendung eines (nicht gezeigten) Prozesses zur Seitenwandbildübertragung (SIT) gebildet. In einem SIT-Prozess können Abstandsschichten auf einem verlorenen Zapfen gebildet werden. Dann kann der verlorene Zapfen entfernt werden, und die zurückbleibenden Abstandsschichten können zum Ätzen der oberen Halbleiterschicht als Hartmaske verwendet werden. Nach dem Bilden der Halbleiterstege können dann die Abstandsschichten entfernt werden.

[0107] Fig. 27 zeigt eine nachfolgende Zwischenstruktur des HBFET 2400 während der Herstellung gemäß einem erläuternden Beispiel. In diesem Schritt wird unter Verwendung lithografischer Techniken ein vertikales Ätzen durchgeführt, um die Si-Schicht 2540 bis zur SiGe-Schicht 2530 abzuätzen. Dann wird die SiGe-Schicht 2530 der Source-Elektrode 825 zeitgesteuert weitergeätzt. Das Ätzen erfolgt während einer vorgegebenen Zeitdauer, oder bis zu einer vorgegebenen Tiefe S der SiGe-Schicht 2530 abgeätzt worden ist. Das Ätzen erfolgt wie hierin beschrieben.

[0108] Fig. 28 zeigt eine nachfolgende Zwischenstruktur des HBFET 2400 einem erläuternden Beispiel während der Herstellung. In diesem Schritt wird das zum Erstellen der Seitenwand-Abstandsschichten 845 des Gate-Elektrodenstapels 840 verwendete Material abgeschieden, um die Schicht 2570 zu bilden. Die Schicht 2570 wird in den vertikal geätzten Gräben auf der Schicht 2530, die die Source-Elektrode 825 bildet, sowie auf der Schicht 2560 abgeschieden, die die Drain-Elektrode 827 bildet. Gemäß einem oder mehreren Beispielen wird ein anisotropes Ätzen durchgeführt, damit die Abstandsschicht 2570 eine vorgegebene Höhe oberhalb der Schicht 2530 erreicht.

[0109] Die erste untere Abstandsschicht 2570 kann ein dielektrisches Material wie Siliciumoxid, Siliciumnitrid, Siliciumoxynitrid oder eine Kombination derselben enthalten und unter Verwendung bekannter Abscheidungsprozesse gebildet werden. Gemäß einem Beispiel wird die erste untere Abstandsschicht 2570 durch Ausführen eines gerichteten Abscheidungsprozesses wie beispielsweise eines Gascluster-Ionenstrahl- (GCIB-) Prozesses gebildet. Der GCIB-Prozess ist ein Abscheidungsprozess, der stark gerichtet wirken kann. Zum Beispiel kann der gerichtete Abscheidungsprozess zum Abscheiden von dielektrischem Material auf den horizontal ausgerichteten Oberflächen der Einheit wie auf der oberen Fläche der Hartmaske und des Substrats 2510 führen, während das Abscheiden einer wesentlichen Menge des dielektrischen Materials auf den vertikal ausgerichteten Flächen der Einheit wie den Seitenwänden der Halbleiterstege vermieden wird.

[0110] Fig. 29 zeigt eine nachfolgende Zwischenstruktur des HBFET 2400 einem erläuternden Beispiel während der Herstellung. In diesem Schritt wird mit einem Dielektrikum aufgefüllt, um ein zwischen Schichten liegendes dielektrisches Material 2580 (ILD) auf freiliegenden Flächen der HBFET-Struktur abzuscheiden, zum Beispiel, um Bereiche zwischen den Halbleiterstegen aufzufüllen. Bei dem ILD 2580 kann es sich um ein beliebiges geeignetes Material wie beispielsweise ein Siliciumoxid handeln.

[0111] Das ILD 2580 kann zum Beispiel unter Verwendung von CMP poliert werden, das selektiv auf die obere Abstandsschicht ausgerichtet ist (zum Beispiel am Nitrid stoppt). Der CMP-Prozess kann zum Entfernen überschüssiger Teile des ILD 2580 verwendet werden, sodass die obere Fläche des ILD 2580 mit der oberen Fläche der oberen Abstandsschicht 2570 koplanar ist. Gemäß einem Beispiel wird das Material des ILD 2580 so gewählt, dass ein Teil der oberen Abstandsschicht 2570 und die Hartmaske während eines nachfolgenden Ätzschrittes (siehe **Fig. 30**) selektiv in Bezug auf das ILD 2580 entfernt werden können.

[0112] Fig. 30 zeigt eine Querschnittsansicht einer VFET-Struktur nach dem Öffnen des ILD während eines Zwischenschrittes des beispielhaften Verfahrens zum Herstellen einer Halbleitereinheit gemäß einem erläuternden Beispiel. Gemäß einem Beispiel wird ein Teil des ILD 2580 entfernt, um Teile der Halbleiterstege freizulegen. Gemäß einem anderen in **Fig. 30** gezeigten Beispiel wird das ILD 2580 bis zum Abstandsmaterial 2540 entfernt, um Teile der Halbleiterstege freizulegen.

[0113] Fig. 31 zeigt eine Querschnittsansicht einer vertikalen HBFET-Struktur einem erläuternden Beispiel während der Herstellung. Eine Schicht 3110 aus einem Material mit hoher Dielektrizitätskonstante wird auf der unteren Abstandsschicht 2570 (845), der aus Dielektrikum bestehenden Seitenwand 2580 und der Seitenwand des Halbleitersteges abgeschieden. Bei dem Material mit hoher Dielektrizitätskonstante kann es sich um ein Dielektrikum wie HfO_2 oder ein beliebiges geeignetes Gate-Material zum Bilden des Gate-Elektrodenstapels 840 handeln. Das Gate-Elektrodenoxid mit hoher Dielektrizitätskonstante kann zum Beispiel durch Abscheiden oberhalb der Kanalbereiche (d.h. Seitenwände) der Halbleiterstege und der unteren Abstandsschicht 845 gebildet werden.

[0114] Die Schicht 3110 mit hoher Dielektrizitätskonstante kann aus einem beliebigen geeigneten Gate-Material gebildet werden, beispielsweise aus einem Material mit einer höheren Dielektrizitätskonstante als Siliciumdioxid. Als beispielhafte Materialien mit hoher Dielektrizitätskonstante kommen zum Beispiel HfO_2 , ZrO_2 , La_2O_3 , Al_2O_3 , TiO_2 , SrTiO_3 , LaAlO_3 , Y_2O_3 , HfO_xN_y , ZrO_xN_y , $\text{La}_2\text{O}_x\text{N}_y$, $\text{Al}_2\text{O}_x\text{N}_y$, TiO_xN_y , SrTiO_xN_y , LaAlO_xN_y , $\text{Y}_2\text{O}_x\text{N}_y$, SiON , SiN_x , ein Silikat derselben und eine Legierung derselben infrage, wobei jeder Wert x unabhängig von 0,5 bis 3 reicht und jeder Wert y unabhängig von 0 bis 2 reicht.

[0115] Gemäß einem oder mehreren Beispielen kann ferner eine dünne Schicht 3120 aus Metall mit einer bestimmten Austrittsarbeit (WF-Metall) abgeschieden werden, worauf eine Schicht 3130 aus leitfähigem Material folgt (Gate-Elektroden-Füllmetall). Die dünne Metallschicht 3120 dient zum Einstellen der Schwellenspannung des FET.

[0116] Die leitfähige Gate-Elektroden-schicht 3130 kann aus Metall und/oder Polysilicium oder einem oder mehreren beliebigen anderen Materialien bestehen, die als Gate-Elektrode für die Halbleitereinheit 2400 dienen. Der leitfähige Kontakt kann aus einem beliebigen geeigneten leitfähigen Material bestehen, beispielsweise aus Metall (z.B. Wolfram, Titan, Tantal, Ruthenium, Zirkonium, Cobalt, Kupfer, Aluminium, Blei, Platin, Zinn, Silber, Gold), leitfähigen Metallverbindungen (z.B. Tantalnitrid, Titannitrid, Tantalcarbid, Titancarbid, Titan-Aluminium-Carbid,

Wolframsilicid, Wolframnitrid, Rutheniumoxid, Cobaltoxid, Cobaltsilicid, Nickelsilicid), Kohlenstoff-Nanoröhren, leitfähigem Kohlenstoff, Graphen oder einer beliebigen geeigneten Kombination dieser Materialien. Das leitfähige Material kann ferner Dotanden enthalten, die während oder nach dem Abscheiden eingebaut werden. Gemäß einem Beispiel kann der leitfähige Kontakt 3130 aus Kupfer bestehen und eine (nicht gezeigte) Metallauskleidung als Diffusionsbarriere enthalten. Durch Diffusionsbarriere der Metallauskleidung wird verhindert, dass das Kupfer in die umgebenden Materialien diffundiert oder diese dotiert, wodurch deren Eigenschaften verschlechtert werden können. Wenn zum Beispiel Silicium mit Kupfer dotiert wird, bildet es tiefe Haftstellen. Eine ideale Metallauskleidung als Diffusionssperre muss das Diffusionsvermögen so weit einschränken, dass sie den Kupferleiter chemisch von den umgebenden Materialien isoliert und zugleich eine hohe elektrische Leitfähigkeit hat, zum Beispiel Tantalnitrid und Tantal (TaN/Ta), Titan, Titanitrid, Cobalt, Ruthenium und Mangan.

[0117] Gemäß einem oder mehreren Beispielen kann das Abscheiden erfolgen, indem der Graben mit den obigen Schichten ausgefüllt und unter Verwendung zum Beispiel eines CMP-Prozesses bis auf die Oberfläche der Nitridschicht 2570 planarisiert wird.

[0118] Fig. 32 zeigt eine Querschnittsansicht einer vertikalen HBFET-Struktur einem erläuternden Beispiel während der Herstellung. Nach dem oben beschriebenen Abscheiden der Schichten des Gate-Elektrodenstapels wird ein anisotropes Ätzen des Gate-Elektrodenstapels (hohe Dielektrizitätskonstante, WE-Metall, Gate-Elektroden-Metallfüllung) durchgeführt. Das Entfernen erfolgt unter Verwendung eines beliebigen lithografischen oder Ätzverfahrens, zum Beispiel wird ein gegenüber lithografischem Hartmasken-Material selektives RIE-Verfahren verwendet.

[0119] Fig. 33 zeigt eine Querschnittsansicht einer HBFET-Struktur gemäß einem erläuternden Beispiel während der Herstellung. Die HBFET-Struktur wird mit dem Material einer oberen Abstandsschicht 3310 ausgefüllt. Bei dem Material der oberen Abstandsschicht kann es sich um dasselbe wie für die Schicht 2570 verwendeten Material handeln, und es kann ein dielektrisches Material wie Siliciumoxid, Siliciumnitrid, Siliciumoxynitrid oder eine Kombination derselben enthalten und unter Verwendung bekannter Abscheidungsprozesse gebildet werden. Gemäß einem Beispiel wird die Abstandsschicht 3310 unter Verwendung eines gerichteten Abscheidungsprozesses, zum Beispiel des GCIB-Prozesses, abgeschieden. Daran schließt sich ein CMP-Prozess an.

[0120] Fig. 34 zeigt eine Querschnittsansicht einer vertikalen HBFET-Struktur gemäß einem erläuternden Beispiel während der Herstellung. Im vorliegenden Falle wird eine Middle-of-line-Verarbeitung verwendet, um die HBFET-Struktur 240 mit den Zeilenleitungen des RPU-Array 600 (gezeigt ist 804) zu verbinden. Die Middle-of-line- (MOL-) Verarbeitung umfasst Ätzen eines Kontakts durch eine Auskleidungsschicht, bei der es sich um das in den Abstandsschichten 2570 und 3310 verwendete Siliciumnitrid handelt. Die Kontaktätzung verläuft durch die Siliciumoxidschicht und die Siliciumnitridschicht, um einen Kontakt mit dem Halbleiterbereich und insbesondere mit dem aktiven kristallinen Halbleiterbereich 2560 herzustellen, der den Drainbereich bildet. In Bereichen, in denen Kontakte zu der unteren leitfähigen Schicht 2530 hergestellt werden (die die spaltenweisen Leiter 808, 810 bilden), verläuft die Kontaktätzung durch die Siliciumoxidschicht und die Siliciumnitridschicht, um den Kontakt mit der (nicht gezeigten) Halbleiterschicht 2530 herzustellen.

[0121] Gemäß einem oder mehreren Beispielen wird die MOL-Auskleidungsschicht unter Verwendung der plasmagestützten chemischen Dampfabcheidung (PECVD) abgeschieden und überdeckt die Halbleiterstruktur. Unter Verwendung einer plasmagestützten chemischen Dampfabcheidung kann die Dicke der in dem zentralen Bereich und in dem Randbereich abgeschiedenen MOL-Abstandsschicht in Abhängigkeit von den Abständen zwischen den Transistoren in dem zentralen Bereich und in dem Randbereich gesteuert werden.

[0122] Gemäß einem oder mehreren Beispielen wird nach dem Abscheiden der MOL-Auskleidungsschicht ferner eine (nicht gezeigte) isolierende Schicht zwischen dem HBVFET und der Zeilenleitung 802 abgeschieden. Die isolierende Schicht enthält zum Beispiel ein isolierendes Material wie Siliciumnitrid, Siliciumoxid, Siliciumoxynitrid oder Borophosphosilikatglas (BPSG). Die isolierende Schicht kann aus mehr als einer Schicht des isolierenden Materials bestehen.

[0123] Nach dem Abscheiden der isolierenden Schicht wird eine Kontaktätzung durch die MOL-Auskleidungsschicht durchgeführt. Eine Kontaktätzung 3410 wird durch die Nitridschichten 2570 und 3310 durchgeführt. Durch die Kontaktätzung 3410 wird es möglich, dass ein die Oberfläche der Halbleiterstruktur überdeckender aktiver Kontakt (CA) elektrisch mit dem Drainbereich des HBFET verbunden wird. Ferner wird eine vertikale Verbindungsschicht (V0) 3420 gebildet und abgeschieden, um den Kontakt 3410 mit der Zeilenleitung 804 zu verbinden.

[0124] Fig. 35 zeigt Verbindungen zwischen dem HBFET 2400 und dem RPU-Array 600 gemäß einem erläuternden Beispiel. Fig. 36 zeigt eine Quer-

schnittsansicht einer vertikalen HBFET-Struktur gemäß einem erläuternden Beispiel während der Herstellung, um den HBFET 2400 mit dem RPU-Array 600 zu verbinden. Im Folgenden wird beschrieben, wie zwischen der Zeilenleitung 804 und dem Anschluss des HBFET 2400 (**Fig. 34**) ein aktiver Kontakt (CA) hergestellt wird. Durch Strukturieren des leitfähigen Materials und unter Verwendung der MOL-Verarbeitung wird ferner ein Kontakt zum Gatebereich (CB) 3510 hergestellt, um den HBFET mit den übrigen Schaltungselementen 2410 am Kreuzungspunkt des RPU-Array 600, beispielsweise mit dem Kondensator 1510 usw., zu verbinden. Bei dem Kondensator 1510 kann es sich um einen Kondensator zum Speichern eines Gewichts handeln.

[0125] Die VFET-Struktur kann zum Beispiel unter Verwendung des RIE strukturiert werden. Gemäß einem Beispiel wirkt das RIE selektiv in Bezug auf das Substrat 2510. Die dabei entstehende Struktur enthält wie hierin beschrieben eine Energiebarriere 830.

[0126] Die bisher erörterten FET-Strukturen enthalten eine aus Polysilicium und/oder einem Metall bestehende Gate-Struktur, die auf einem Isolator oberhalb der zwischen der Source-Elektrode und der Drain-Elektrode angeordneten Halbleiterschicht gebildet ist und in Kontakt mit dieser steht. Die Halbleiterschicht kann verschiedene Dotanden enthalten, wobei ein Dotierungstyp die Source-Elektrode und die Drain-Elektrode und ein anderer Dotierungstyp den Kanal und die Energiebarriere-Schicht betrifft. Durch Anlegen einer Spannung an die Gate-Struktur kann zwischen dem Source-Anschluss und dem Drain-Anschluss ein elektrisch leitender Kanal innerhalb der Halbleiterschicht erzeugt werden. Der Bereich der Energiebarriere 830 wurde in diesen Strukturen hergestellt, um einen steuerbaren hohen Widerstand zu erzeugen.

[0127] Gemäß einem weiteren erläuternden Beispiel kann für die FET-Struktur eine alternative FET-Struktur genutzt werden, die als Tunnel-FET bekannt ist und einen Gate-Kontakt enthält, der durch eine Oxidschicht von der Halbleiterschicht getrennt ist. Die Halbleiterschicht kann mehrere Typen von Halbleitermaterialien und Dotanden enthalten, sodass durch Steuern einer Spannung an der Gate-Elektrode der Stromfluss zwischen einem Source-Kontakt und einem Drain-Kontakt an zwei Enden der Halbleiterschicht beeinflusst wird. Im Gegensatz zu den Diffusions-FET- und HBFET-Strukturen, bei denen der Source-Bereich und der Drainbereich denselben Dotierungstyp aufweisen, sind der Source-Bereich und der Drain-Bereich eines Tunnel-FET unterschiedlich dotiert.

[0128] **Fig. 37** zeigt einen Tunnel-FET 3700 gemäß einem erläuternden Beispiel. Im Gegensatz zu dem

oben beschriebenen Diffusions-FET (siehe zum Beispiel **Fig. 14**) enthält der Tunnel-FET 3700 keinen Energiebarriere-Bereich 830. Stattdessen ist in dem Tunnel-FET 3700 die gesamte Source-Anschluss 825 mit P+ dotiert, während in dem veranschaulichten Beispiel der Drain-Anschluss 827 mit N+ dotiert ist. Dabei ist zu beachten, dass gemäß anderen Beispielen der Source-Anschluss 825 mit N+ dotiert und der Drain-Anschluss 827 mit P+ dotiert sein kann. Ferner ist der Kanal 850 unter Verwendung desselben Dotanden wie der Drain-Anschluss dotiert, jedoch mit einer geringeren Konzentration als der Drain-Anschluss 827.

[0129] Gemäß einem oder mehreren Beispielen kann das Dotieren unter Verwendung Ionenimplantation unter Verwendung von strukturierten Masken erfolgen, die einen Bereich abdecken, während der andere Bereich dotiert wird. Zum Beispiel kann der Source-Anschluss 825 abgedeckt werden, während der Drain-Anschluss 827 mit N+ dotiert wird, und anschließend wird der Drain-Anschluss 827 abgedeckt und der Source-Anschluss 825 mit P+ dotiert.

[0130] **Fig. 38** zeigt eine Auswirkung der hierin beschriebenen Tunnel-FET-Struktur gemäß einem erläuternden Beispiel auf die Stromstärke. Die Kurve 3810 zeigt einen Zustand EIN eines Diffusions-FET, und die Kurve 3820 zeigt einen Zustand AUS. In beiden Kurven sind ein Valenzband 3835 und ein Leitungsband 3845 gezeigt. Normalerweise wird in einem Tunnel-FET der Stromfluss im Zustand „EIN“ durch das Tunneln von Ladungsträgern zwischen den Bändern bestimmt. Die Stromstärke wird durch die Form der Energiebarriere zwischen dem Source-Anschluss 825 und dem Kanal 850 bestimmt. Der Widerstand eines solchen FET kann durch Dotieren des Source-Bereichs 826 und des Kanals 850 moduliert werden und ist eine Funktion des Spannungsabfalls zwischen Gate-Bereich und Drain-Bereich.

[0131] **Fig. 39** zeigt eine Auswirkung der hierin beschriebenen Tunnel-FET-Struktur gemäß einem erläuternden Beispiel auf die Stromstärke. Die gezeigte Kurve zeigt einen Vergleich zwischen einer ersten Übertragungskennlinie 3910 eines typischen FET und einer zweiten Übertragungskennlinie 3920 des Tunnel-FET 3700, bei dem die Source-Elektrode anders als die Drain-Elektrode und der Kanal dotiert sind. Es zeigt sich, dass der Tunnel-FET 3700 einen niedrigeren Strom hat. Dabei ist darauf zu achten, dass die Kurven in **Fig. 39** Ergebniswerte eines beispielhaften Szenarios veranschaulichen und dass gemäß anderen Beispielen die Kurven von den veranschaulichten abweichen können.

[0132] **Fig. 40** zeigt eine Auswirkung der Energiebarriere in der hierin beschriebenen Tunnel-FET-Struktur gemäß einem erläuternden Beispiel auf die

Stromstärke. Die veranschaulichte Kurve zeigt den Widerstand des Tunnel-FET als Funktion der Gate-Spannung bei unterschiedlichen Dotandkonzentrationen am Source-Anschluss 825. Zum Beispiel zeigt die Kurve die Dotandkonzentration von $4 \times 10^{20} \text{ cm}^{-3}$ (4010), $6 \times 10^{20} \text{ cm}^{-3}$ (4020), $8 \times 10^{20} \text{ cm}^{-3}$ (4030) beziehungsweise $1 \times 10^{21} \text{ cm}^{-3}$ (4040). Dabei ist zu beachten, dass sich der Wert der Konzentration von den obigen beispielhaften Werten unterscheiden kann, sodass sich der FET-Widerstand anders als in Fig. 40 gezeigt verändert.

[0133] Wiederum Bezug nehmend auf Fig. 37 enthält die Tunnel-FET-Struktur ferner einen Gate-Elektrodenstapel 840 (z.B. eine dielektrische Schicht 841 der Gate-Elektrode, eine Leiterschicht 842 der Gate-Elektrode oberhalb der dielektrischen Schicht 841 der Gate-Elektrode und eine Deckschicht 843, beispielsweise eine Nitrid-Deckschicht auf der Leiterschicht 842 der Gate-Elektrode) oberhalb des festgelegten Kanalbereichs 850 innerhalb des Einheitenbereichs. Ferner sind auf entgegengesetzten Seiten des Gate-Elektrodenstapels 840 Abstandsschichten 845 als Seitenwand der Gate-Elektrode gebildet.

[0134] Um das Gewicht in dem Tunnel-FET 3700 zu speichern, wird ferner das Gate-Potenzial der FET-Einheit 3700 auf einem bestimmten Wert gehalten. Das kann mittels verschiedener Strukturen erreicht werden.

[0135] Fig. 41 zeigt eine Veranschaulichung zum Speichern des Gewichts in dem Tunnel-FET 3700 gemäß einem erläuternden Beispiel. Der Tunnel-FET 3700 ist mit einem Kondensator 1510 verbunden, zum Beispiel von den anderen Schaltelementen 2410 an jedem der Kreuzungspunkte in dem RPU-Array 600. Die Ladung wird auf der Platte gespeichert, um die Gate-Spannung bereitzustellen, wobei die gespeicherte Ladung dem an dem Kreuzungspunkt in dem RPU-Array 600 gespeicherten Gewicht entspricht. Dabei ist zu beachten, dass zum Vereinfachen der Darstellung nur wenige Teile der FET-Einheit 3700 in Fig. 41 gezeigt sind.

[0136] Fig. 42 zeigt eine Veranschaulichung zum Speichern des Gewichts in dem Tunnel-FET 3700 gemäß einem erläuternden Beispiel. Eine Ladungsspeicherschicht 1610 wie beispielsweise eine Floating-Gate-Elektrode wird dem Gate-Elektrodenstapel 840 hinzugefügt. Die Ladung kann in der Ladungsspeicherschicht 1610 gespeichert werden, um die Gate-Spannung bereitzustellen. Die Ladungsspeicherschicht 1610 besteht aus Polysilicium. Gemäß einem oder mehreren Beispielen werden nach der Ladungsspeicherschicht 1610 noch eine dielektrische Schicht 841 der Gate-Elektrode und anschließend die weiteren Schichten in dem Gate-Elektrodenstapel 840 wie beispielsweise die Schicht 842

aus leitfähigem Material für die Elektrode hinzugefügt.

[0137] Fig. 43 zeigt eine weitere Veranschaulichung zum Speichern des Gewichts in dem Tunnel-FET 3700 gemäß einem erläuternden Beispiel. Hier enthält der Gate-Elektrodenstapel die dielektrische Schicht 841 der Gate-Elektrode, worauf eine erste Schicht 842 aus leitfähigem Material folgt. Anschließend wird eine Schicht 1710 aus ferroelektrischem Material hinzugefügt, die beispielsweise aus Hafniumdioxid (HfO_2) besteht. Durch die Polarisierung der ferroelektrischen Schicht 1710 wird die Gate-Spannung bereitgestellt. Ferner enthält der Gate-Elektrodenstapel 840 eine zweite Schicht 842 aus leitfähigem Material zum Bilden der Gate-Steuer-elektrode. Durch die erste Schicht 842 aus leitfähigem Material wird ein technisches Problem der Depolarisation der Schicht 1710 aus ferroelektrischem Material behoben.

[0138] Die hierin beschriebenen FET-Strukturen stellen eine asymmetrische Struktur bereit, zum Beispiel infolge der Energiebarriere 830 oder der unterschiedlichen Dotierung von Source-Elektrode und Drain-Elektrode/Kanal. Aufgrund der asymmetrischen Struktur der Einheit erfolgen die Leseoperationen in Vorwärts- und Rückwärtsrichtung unterschiedlich.

[0139] Fig. 44 zeigt eine Eingabephase (forward pass) unter Verwendung des RPU-Arrays mit den hierin beschriebenen asymmetrischen FET-Strukturen gemäß einer oder mehreren Ausführungsformen der vorliegenden Erfindung. Das RPU-Array 600 enthält wie gezeigt einen asymmetrischen FET 4400 (Diffusions-FET 800, HBFET 1400 oder Tunnel-FET 3700) an jedem Kreuzungspunkt. Während der Eingabephase werden an jede Zeile des RPU-Array 600 Spannungsimpulse 4420 angelegt, und die in den Spalten sich daraus ergebenden Ströme werden unter Verwendung von Stromintegratoren 4410 addiert. In der Eingaberichtung ist der Source-Anschluss 825 des FET 440 mit dem Stromintegrator 4410 der entsprechenden Spalte verbunden, und ein positiver Impuls wird an einen Drain-Anschluss 827 des FET 4400 angelegt. Demgemäß wird die Stromstärke während der Eingabephase mittels des FET 440 durch den steuerbaren Widerstand des FET 4400 beeinflusst.

[0140] Fig. 45 zeigt eine Lernphase (backward pass) unter Verwendung des RPU-Arrays mit den hierin beschriebenen asymmetrischen FET-Strukturen gemäß einer oder mehreren Ausführungsformen der vorliegenden Erfindung. Das RPU-Array 600 enthält wie gezeigt einen asymmetrischen FET 4400 (Diffusions-FET 4400, HBFET 1400 oder Tunnel-FET 3700) an jedem Kreuzungspunkt. Während der Lernphase werden an jede Spalte des RPU-Array

600 Spannungsimpulse 4420 angelegt, und die sich daraus ergebenden Ströme in den Zeilen werden unter Verwendung von Stromintegratoren 4410 addiert. Während der Lernphase wird die Basisspannung an der Source-Elektrode 825 und der Drain-Elektrode 827 auf einen positiven Wert +V angehoben. Der Drain-Anschluss 827 des FET 4400 ist mit dem Stromintegrator 4410 der entsprechenden Zeile verbunden, und an den Source-Anschluss 825 des FET 440 wird ein negativer Impuls angelegt. Demgemäß wird die Stromstärke während der Lernphase mittels des FET 4400 durch den steuerbaren Widerstand des FET 440 beeinflusst.

[0141] Ferner kann das RPU-Array 600 gemäß einer oder mehreren Ausführungsformen der vorliegenden Erfindung unter Verwendung symmetrischer FET-Strukturen realisiert werden. Der auf Diffusion beruhende FET 800 und der HBFET 1800, die eine Energiebarriere 830 enthalten, können mit einer im Wesentlichen identischen Energiebarriere 830' auf der Drain-Seite hergestellt werden, um eine symmetrische Struktur der FET-Einheit zu erreichen.

[0142] Fig. 46 zeigt eine beispielhafte symmetrische Halbleiterstruktur für den Diffusions-FET gemäß einer oder mehreren Ausführungsformen der vorliegenden Erfindung. Gemäß dieser Ausführungsform enthält der Diffusions-FET ein Paar Energiebarrieren, eine erste Energiebarriere 830 zwischen der Source-Elektrode 825 und dem Kanal 850 und eine zweite Energiebarriere 830' zwischen dem Kanal 850 und der Drain-Elektrode 827. Die beiden Energiebarrierebereiche 830 und 830' bestehen aus demselben Material, das mit einer stärkeren (höheren) Konzentration als der Kanal 850 dotiert ist.

[0143] Fig. 47 zeigt eine beispielhafte symmetrische Halbleiterstruktur für den HBFET gemäß einer oder mehreren Ausführungsformen der vorliegenden Erfindung. Gemäß dieser Ausführungsform enthält der HBFET ein Paar Energiebarrieren, eine erste Energiebarriere 830 zwischen der Source-Elektrode 825 und dem Kanal 850 und eine zweite Energiebarriere 830' zwischen dem Kanal 850 und der Drain-Elektrode 827. Die beiden Energiebarrierebereiche 830 und 830' bestehen aus demselben Material wie zum Bilden einer Heterobarriere, beispielsweise aus Indiumphosphid (InP) oder aus Si, in Abhängigkeit davon, ob eine III-V-Struktur oder eine Si-SiGe-Struktur verwendet wird.

[0144] Hierin werden unter Bezugnahme auf die entsprechenden Zeichnungen Ausführungsformen der vorliegenden Erfindung beschrieben. Es können alternative Ausführungsformen entwickelt werden, ohne vom Schutzzumfang dieser Erfindung abzuweichen. Zwar werden zwischen Elementen in der folgenden Beschreibung und in den Zeichnungen verschiedene Verbindungen und Positionsbeziehungen

(z.B. oberhalb, unterhalb, nahe usw.) dargelegt, jedoch ist dem Fachmann einsichtig, dass viele der hierin beschriebenen Positionsbeziehungen nicht von einer Ausrichtung abhängen, wenn die beschriebene Funktionalität trotz Änderung der Ausrichtung erhalten bleibt. Sofern nichts anderes angegeben ist, können diese Verbindungen und/oder Positionsbeziehungen direkt oder indirekt sein, und die vorliegende Erfindung soll in dieser Beziehung keiner Einschränkung unterliegen. Desgleichen wird mit dem Begriff „gekoppelt“ und dessen Varianten beschrieben, dass zwischen den Elementen ein Datenübertragungspfad vorliegt und nicht von einer direkten Verbindung zwischen den Elementen ohne dazwischenliegende Elemente/Verbindungen auszugehen ist. Alle diese Varianten werden als Bestandteil der Beschreibung angesehen. Demgemäß kann ein Koppeln von Einheiten entweder ein direktes oder ein indirektes Koppeln betreffen, und bei einer Positionsbeziehung zwischen Einheiten kann es sich um eine direkte oder eine indirekte Positionsbeziehung handeln. Gemäß einem Beispiel einer indirekten Positionsbeziehung umfassen Bezugnahmen in der vorliegenden Beschreibung auf Bilden einer Schicht „A“ oberhalb einer Schicht „B“ Fälle, in denen sich zwischen der Schicht „A“ und der Schicht „B“ eine oder mehrere Zwischenschichten (z.B. eine Schicht „C“) befinden, solange die einschlägigen Eigenschaften und Funktionalitäten der Schicht „A“ und der Schicht „B“ durch die Zwischenschicht(en) nicht wesentlich verändert werden.

[0145] Die folgenden Definitionen und Abkürzungen werden zum Auslegen der Ansprüche und der Beschreibung verwendet. Die hierin verwendeten Begriffe „weist auf“, „aufweisend“, „enthält“, „enthaltend“, „hat“, „habend“, „enthält“ oder „enthaltend“ oder deren beliebige Varianten sollen eine nichtexklusive Einbeziehung bedeuten. Zum Beispiel sind eine Zusammensetzung, eine Mischung, ein Prozess, ein Verfahren, ein Artikel oder eine Vorrichtung, die eine Liste von Elementen aufweist, nicht notwendigerweise allein auf diese Elemente beschränkt, sondern können auch andere Elemente enthalten, die nicht ausdrücklich aufgeführt oder einer solchen Zusammensetzung, einer solchen Mischung, einem solchen Prozess, einem solchen Verfahren, einem solchen Artikel oder einer solchen Vorrichtung eigen sind.

[0146] Weiterhin wird der hierin verwendete Begriff „beispielhaft“ in der Bedeutung „als Beispiel, Einzelfall oder Veranschaulichung dienend“ gebraucht. Eine beliebige hierin als „beispielhaft“ beschriebene Ausführungsform oder Bauart ist nicht notwendigerweise als gegenüber anderen Ausführungsformen oder Bauarten bevorzugt oder beispielhaft zu verstehen. Die Begriffe „mindestens ein“ und „ein oder mehrere“ sind so zu verstehen, dass sie eine beliebige ganze Zahl größer als oder gleich eins umfas-

sen, d.h. eins, zwei, drei, vier usw. Der Begriff „eine Mehrzahl“ ist so zu verstehen, dass er eine beliebige ganze Zahl größer als oder gleich zwei umfasst, d.h. zwei, drei, vier, fünf usw. Der Begriff „Verbindung“ kann eine indirekte „Verbindung“ und eine direkte „Verbindung“ umfassen.

[0147] Bezugnahmen in der Beschreibung auf „eine bestimmte Ausführungsform“, „eine Ausführungsform“, „eine beispielhafte Ausführungsform“ usw. zeigen an, dass die beschriebene Ausführungsform ein bestimmtes Merkmal, eine bestimmte Struktur oder Eigenschaft enthalten kann, dass unter Umständen aber jede Ausführungsform das bestimmte Merkmal, die bestimmte Struktur oder Eigenschaft enthalten kann. Außerdem betreffen solche Ausdrücke nicht notwendigerweise ein und dieselbe Ausführungsform. Wenn ein bestimmtes Merkmal, eine bestimmte Struktur oder Eigenschaft in Verbindung mit einer Ausführungsform beschrieben wird, wird ferner davon ausgegangen, dass es vom Wissen des Fachmanns abhängt, ein solches Merkmal, eine solche Struktur oder Eigenschaft in Verbindung mit anderen Ausführungsformen zu beeinflussen, auch wenn dies nicht ausdrücklich beschrieben ist.

[0148] Die zum Zweck der folgenden Beschreibung verwendeten Begriffe „oben“, „unten“, „rechts“, „links“, „vertikal“, „horizontal“, „obere“, „untere“ und deren Ableitungen sollen sich auf die Ausrichtung der beschriebenen Strukturen und Verfahren in den Figuren beziehen. Die Begriffe „darüberliegend“, „oberhalb“, „zuoberst“, „positioniert oberhalb“ oder „positioniert zuoberst“ bedeuten, dass ein erstes Element wie beispielsweise eine erste Struktur oberhalb eines zweiten Elements wie beispielsweise einer zweiten Struktur liegen kann, wobei sich zwischen dem ersten Element und dem zweiten Element dazwischen liegende Elemente wie beispielsweise eine Grenzflächenstruktur befinden können. Unter dem Begriff „direkter Kontakt“ ist zu verstehen, dass ein erstes Element wie eine erste Struktur und ein zweites Element wie eine zweite Struktur miteinander verbunden sind, ohne dass sich an der Grenzfläche der beiden Elemente dazwischen liegende leitende, isolierende oder Halbleiterschichten befinden.

[0149] Die Begriffe „ungefähr“, „im Wesentlichen“, „annähernd“ und deren Varianten sollen eine mit der Messung der betreffenden Größe verbundene Fehlergröße auf der Grundlage der zum Zeitpunkt des Einreichens der Anmeldung verfügbaren Ausrüstung enthalten. Der Begriff „ungefähr“ kann zum Beispiel einen Bereich von $\pm 8\%$ oder 5% oder 2% eines bestimmten Wertes umfassen.

[0150] Der Ausdruck „selektiv in Bezug auf“ wie beispielsweise „ein erstes Element selektiv in Bezug auf ein zweites Element“ bedeutet, dass das erste Ele-

ment geätzt werden und das zweite Element als Ätzstopp dienen kann.

[0151] Der Begriff „oberflächenkonform“ (z.B. eine oberflächenkonforme Schicht) bedeutet, dass die Dicke der Schicht auf allen Oberflächen im Wesentlichen gleich groß ist oder dass die Schichtdickenabweichungen vom Nennwert der Schichtdicke geringer als 15% sind.

[0152] Die Begriffe „epitaxiales Wachstum und/oder epitaxiale Abscheidung“ und „epitaxial gebildet und/oder abgeschieden“ bedeuten, dass das Wachstum eines Halbleitermaterials (kristallines Material) auf einem anderen Halbleitermaterial (kristallines Material), auf dem das Halbleitermaterial abgeschieden wird (kristalline Deckschicht), zu im Wesentlichen dieselben kristallinen Eigenschaften wie im Halbleitermaterial der Abscheidungs Oberfläche (Keimbildungsmaterial) führt. Bei einem epitaxialen Abscheidungsprozess können die durch die chemischen Reaktionspartner bereitgestellten Ausgangsgase gesteuert und die Systemparameter so eingestellt werden, dass die sich abscheidenden Atome mit ausreichend hoher Energie auf der Abscheidungs Oberfläche des Halbleitersubstrats ankommen, sodass die sich abscheidenden Atome sich in das Kristallgitter der Atome auf der Abscheidungs Oberfläche einordnen. Ein epitaxial abgeschiedenes Halbleitermaterial kann im Wesentlichen dieselben kristallinen Eigenschaften wie die Abscheidungs Oberfläche haben, auf denen das epitaxial abgeschiedene Material gebildet wird. Zum Beispiel kann ein auf einer Kristallfläche mit $\{100\}$ -Ausrichtung epitaxial abgeschiedenes Halbleitermaterial eine $\{100\}$ -Ausrichtung annehmen. Gemäß einem Beispiel der Erfindung können epitaxiale Wachstums- und/oder -Abscheidungsprozesse selektiv auf einer Halbleiter Oberfläche gebildet werden, und gegebenenfalls kann Material auf freiliegenden Oberflächen wie Siliciumdioxid- oder Siliciumnitrid-Oberflächen abgeschieden werden.

[0153] Der Kürze halber werden wie oben bereits erwähnt herkömmliche Verfahren zur Herstellung von Halbleitereinheiten und integrierten Schaltkreisen (IC) hierin gegebenenfalls nicht ausführlich beschrieben. Als Hintergrundinformation wird im Folgenden jedoch eine allgemeinere Beschreibung der Prozesse zur Herstellung von Halbleitereinheiten bereitgestellt, die zum Umsetzen einer oder mehrerer Ausführungsformen der vorliegenden Erfindung genutzt werden können. Zwar sind spezielle zum Umsetzen einer oder mehrerer Ausführungsformen der vorliegenden Erfindung möglicherweise im Einzelfall bekannt, jedoch ist die beschriebene Kombination von Arbeitsschritten und/oder sich daraus ergebender Strukturen der vorliegenden Erfindung einzigartig. Somit wird im Rahmen der in Verbindung mit der Herstellung einer Halbleitereinheit gemäß der

vorliegenden Erfindung beschriebenen einzigartigen Kombination von Arbeitsschritten eine Vielfalt im Einzelnen bekannter physikalischer und chemischer Prozesse mit einem Halbleitersubstrat (z.B. Silicium) durchgeführt, von denen einige in den unmittelbar folgenden Absätzen beschrieben werden.

[0154] Im Allgemeinen werden die verschiedenen zum Bilden eines Mikro-Chips verwendeten Prozesse in vier Hauptkategorien eingeteilt, das heißt Dünnschichtabscheidung, Entfernen/Ätzen, Halbleiterdotierung und Strukturieren/Lithografie. Unter Abscheidung ist ein beliebiger Prozess zum Abscheiden, Beschichten oder anderweitigen Übertragen eines Materials auf den Wafer zu verstehen. Verfügbare Technologien umfassen physikalische Dampfabscheidung (PFD), chemische Dampfabscheidung (CVD), elektrochemische Abscheidung (ECD), Molekularstrahl-Epitaxie (MBE), neuerdings Atomlagenabscheidung (ALD) und anderes mehr. Unter Entfernen/Ätzen ist ein beliebiger Prozess zu verstehen, bei dem Material von dem Wafer entfernt wird. Beispiele umfassen Ätzprozesse (entweder Nassätzen oder Trockenätzen), chemisch-mechanisches Planarisieren (CMP) und dergleichen. Reaktives Ionenätzen (RIE) ist zum Beispiel eine Art Trockenätzen, bei dem ein chemisch reaktives Plasma zum Entfernen eines Materials verwendet wird, beispielsweise ein durch Masken vorgegebenes Muster aus Halbleitermaterial, indem das Material einem Beschuss mit Ionen ausgesetzt wird, durch den Teile des Materials von der freiliegenden Oberfläche entfernt werden. Das Plasma wird normalerweise bei niedrigem Druck (Vakuum) durch ein elektromagnetisches Feld erzeugt. Beim Dotieren von Halbleitern handelt es sich um das Verändern der elektrischen Eigenschaften durch Dotieren zum Beispiel der Source-Elektroden und Drain-Elektroden eines Transistors im Allgemeinen durch Diffusion und/oder Ionenimplantation. An diese Dotierungsprozesse schließen sich Tempern oder Schnelltempern (RTA) an. Tempern dient zum Aktivieren der implantierten Dotanden. Dünne Schichten sowohl von Leitern (z.B. Polysilicium, Aluminium, Kupfer usw.) als auch Isolatoren (z.B. verschiedene Formen von Siliciumdioxid, Siliciumnitrid usw.) werden zum Verbinden von Transistoren und deren Komponenten miteinander und zum Isolieren voneinander verwendet. Durch selektives Dotieren verschiedener Bereiche des Halbleitersubstrats ist es möglich, die Leitfähigkeit des Substrats durch Anlegen einer Spannung zu ändern. Durch Erzeugen von Strukturen aus diesen verschiedenen Komponenten können Millionen von Transistoren erzeugt und miteinander elektrisch verbunden werden, um die komplexen Schaltkreise einer modernen Mikroelektronikeinheit zu bilden. Mittels Halbleiterlithografie werden dreidimensionale Reliefbilder oder Muster auf dem Halbleitersubstrat oder zum Übertragen des Musters auf das Substrat gebildet. Bei der Halbleiterlithografie werden die

Muster durch ein lichtempfindliches Polymer mit der Bezeichnung Fotolack gebildet. Zum Aufbauen der komplexen Strukturen, aus denen ein Transistor besteht, und der vielen Leiterbahnen zum Verbinden der Millionen von Transistoren eines Schaltkreises miteinander werden lithografische Schritte und Schritte zum Übertragen von Ätzmustern mehrmals wiederholt. Jedes auf den Wafer zu druckendes Muster ist auf die zuvor gebildeten Muster ausgerichtet, und nach und nach werden die Leiter, Isolatoren und selektiv dotierten Bereiche zusammengefügt, um die fertige Einheit zu bilden.

[0155] Der Ablaufplan und die Blockschaubilder in den Figuren veranschaulichen mögliche Umsetzungen von Herstellungs- und/oder Arbeitsverfahren gemäß verschiedenen Ausführungsformen der vorliegenden Erfindung. In dem Ablaufplan werden verschiedene Funktionen/Arbeitsschritte des Verfahrens durch Blöcke dargestellt. Gemäß alternativen Umsetzungen können die in den Blöcken angegebenen Funktionen in einer von den Figuren abweichenden Reihenfolge auftreten. Zum Beispiel können zwei nacheinander dargestellte Blöcke je nach vorgesehener Funktionalität in Wirklichkeit im Wesentlichen gleichzeitig ausgeführt werden, oder die Blöcke können mitunter in der umgekehrten Reihenfolge ausgeführt werden.

[0156] Die Beschreibungen der verschiedenen Ausführungsformen der vorliegenden Erfindung sind zum Veranschaulichen vorgelegt worden, erheben jedoch nicht den Anspruch auf Vollständigkeit oder auf Beschränkung auf die beschriebenen Ausführungsformen. Dem Fachmann sind viele Modifikationen und Varianten offensichtlich, ohne vom Schutzzumfang und Wesensgehalt der beschriebenen Ausführungsformen abzuweichen. Die hierin verwendeten Begriffe wurden gewählt, um die Grundgedanken der Ausführungsformen, die praktische Anwendung oder technische Verbesserung gegenüber handelsüblichen Technologien bestmöglich zu erläutern oder anderen Fachleuten das Verständnis der hierin beschriebenen Ausführungsformen zu ermöglichen.

Patentansprüche

1. Verfahren zum Bilden einer Halbleitereinheit (800), wobei das Verfahren aufweist:
Bilden (704) eines Source-Anschlusses (825) und eines Drain-Anschlusses (827) eines Feldeffekttransistors (FET) auf einem Substrat (805), wobei der Source-Anschluss und der Drain-Anschluss auf jeder Seite eines Kanalbereichs (850) durch Ätzen von Aussparungen in dem Substrat und Abscheiden von dotierten Bereichen in den Aussparungen gebildet werden; und
Bilden (706) einer Energiebarriere (830) zwischen dem Source-Anschluss und dem Kanalbereich

durch Ätzen einer Aussparung in dem Substrat und Abscheiden eines dotierten Bereich in dieser Aussparung; und
Bilden (708) einer leitfähigen Gate-Elektrode oberhalb des Kanalbereichs.

2. Verfahren nach Anspruch 1, wobei die Energiebarriere und der Kanalbereich unter Verwendung eines ersten Dotanden dotiert sind und die Energiebarriere mit einer höheren Dotierungskonzentration als der Kanalbereich dotiert ist.

3. Verfahren nach Anspruch 2, wobei der Source-Anschluss und der Drain-Anschluss unter Verwendung eines zweiten Dotanden dotiert sind.

4. Verfahren nach Anspruch 1, das ferner Bilden einer Seitenwand-Abstandsschicht nahe der leitfähigen Gate-Elektrode aufweist, wobei durch eine an der Gate-Elektrode gespeicherte Ladung ein Randfeld in der Seitenwand-Abstandsschicht erzeugt wird, durch das eine Höhe der Energiebarriere moduliert wird.

5. Verfahren nach Anspruch 1, wobei die Energiebarriere aus einem Heterobarriere-Material gebildet wird.

6. Verfahren nach Anspruch 1, das ferner Bilden einer Floating-Gate-Elektrode (1610) in der leitfähigen Gate-Elektrode aufweist.

7. Verfahren nach Anspruch 1, das ferner Bilden einer Schicht aus einem ferroelektrischen Material (1710) auf einer Metallschicht in der leitfähigen Gate-Elektrode aufweist.

Es folgen 39 Seiten Zeichnungen

Anhängende Zeichnungen

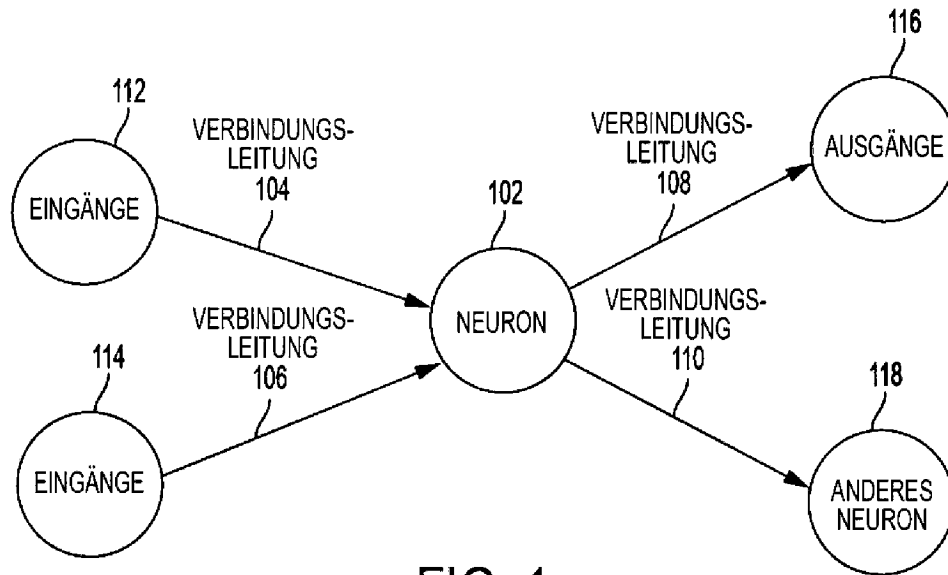
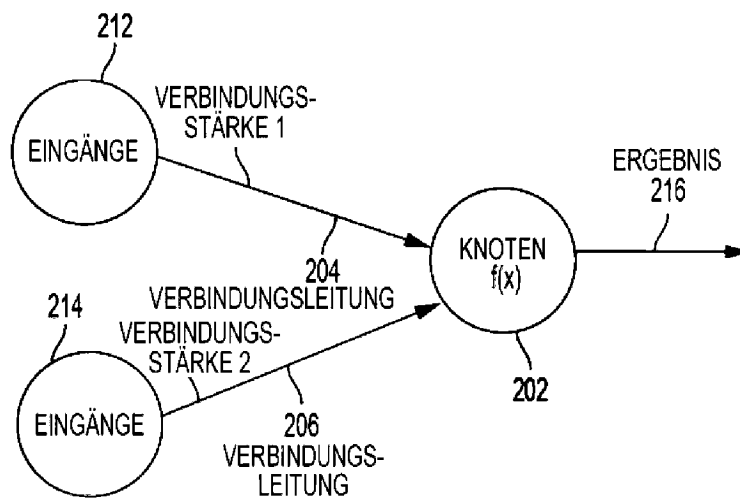


FIG. 1



$$F(x) = f(\text{EINGABEWERT 1} * \text{VERBINDUNGSSTÄRKE 1} + \text{EINGABEWERT 2} * \text{VERBINDUNGSSTÄRKE 2})$$

FIG. 2

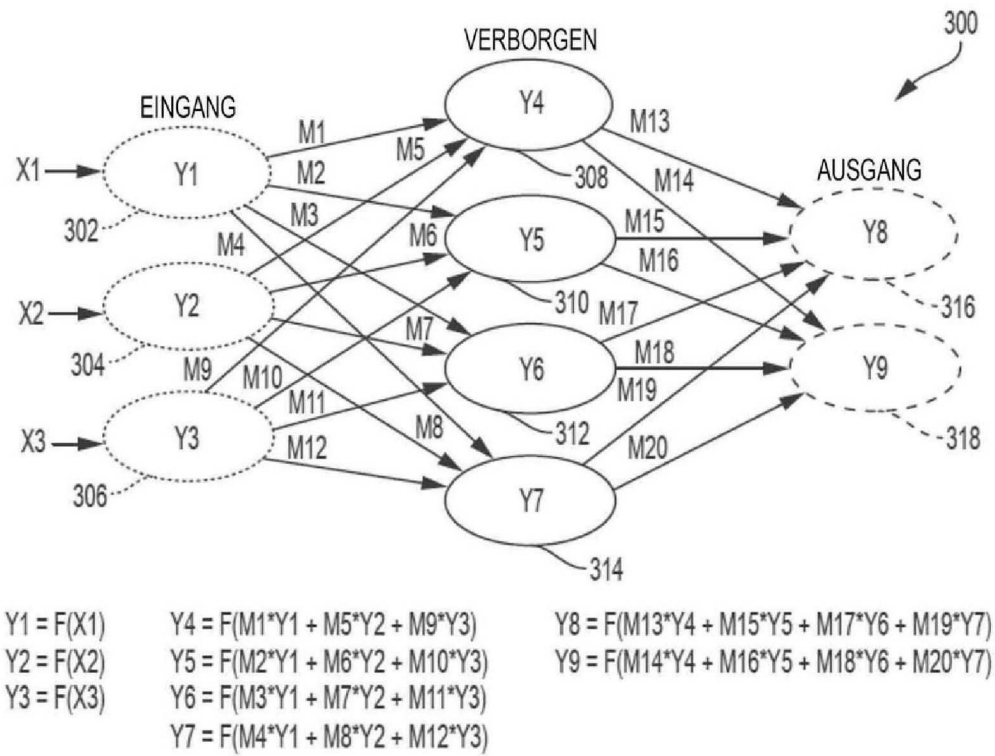


FIG. 3

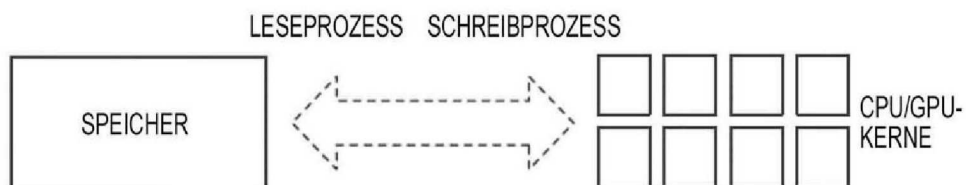


FIG. 4

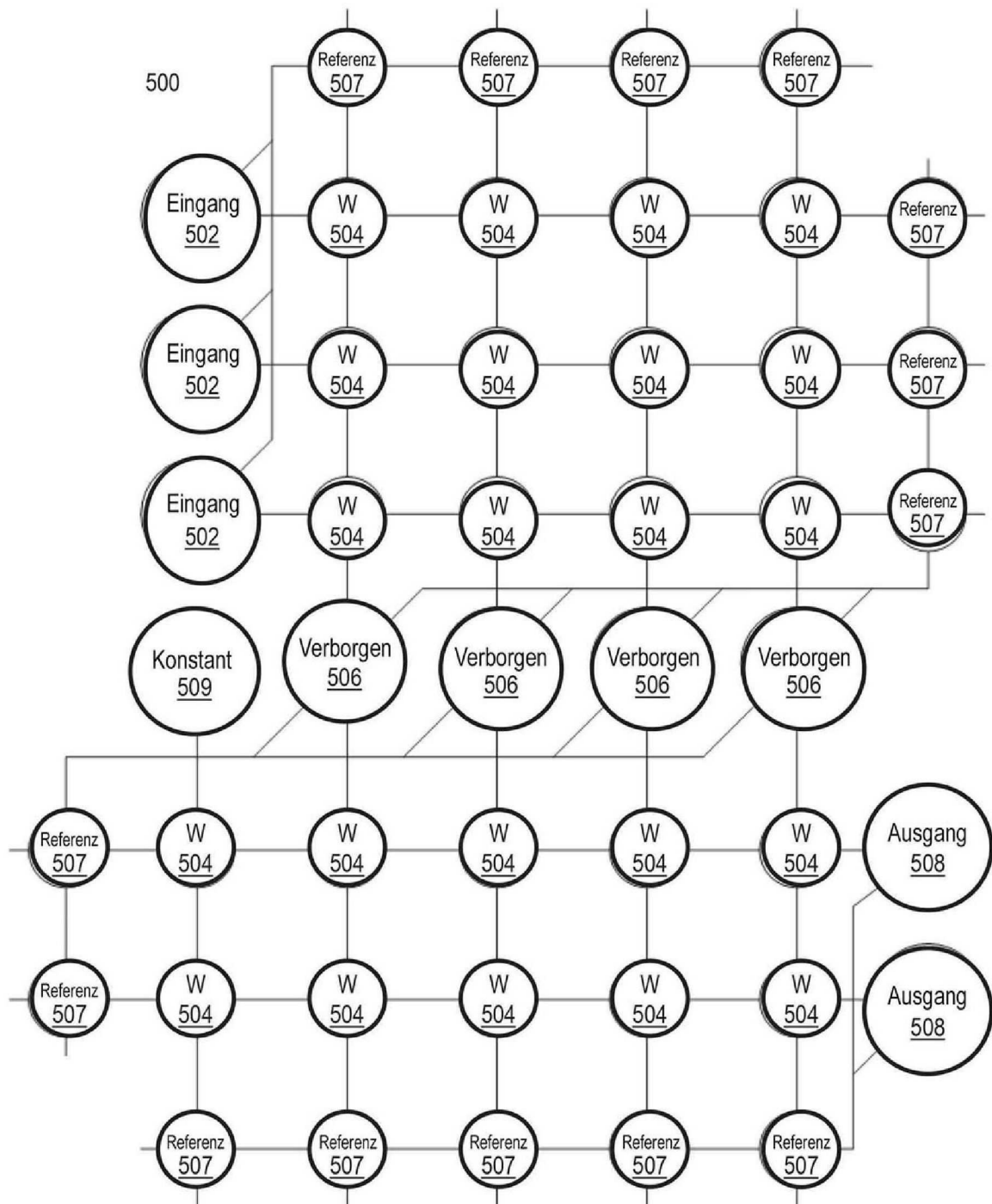


FIG. 5

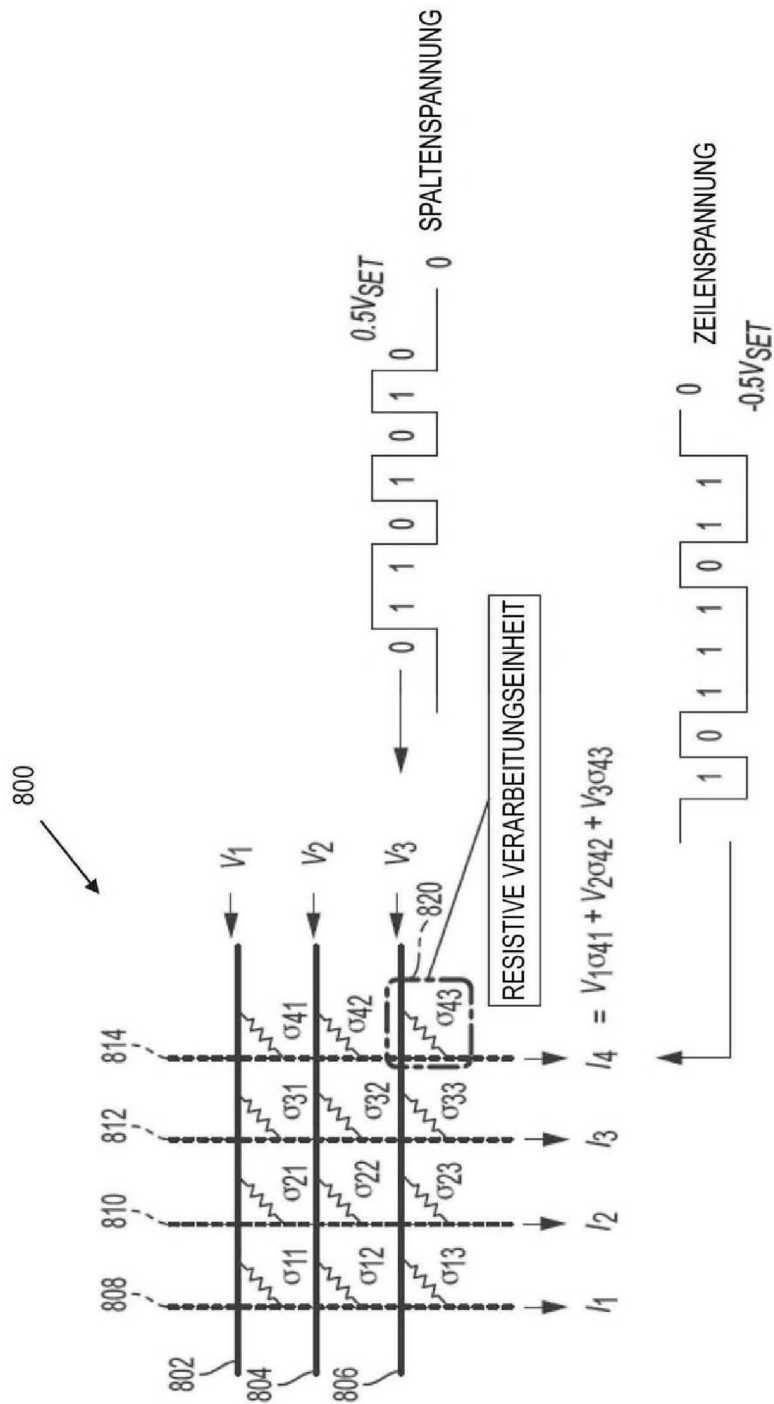


FIG. 6

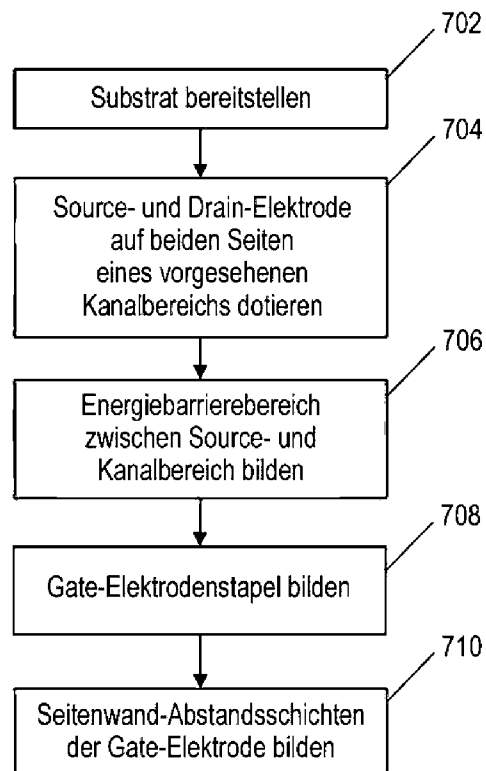


FIG. 7

FIG. 8

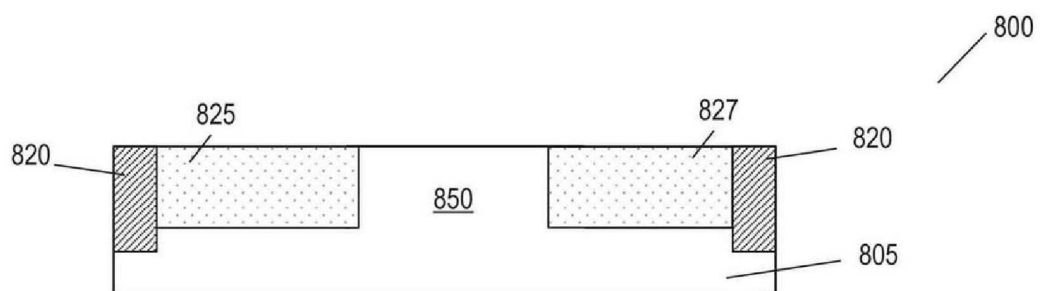
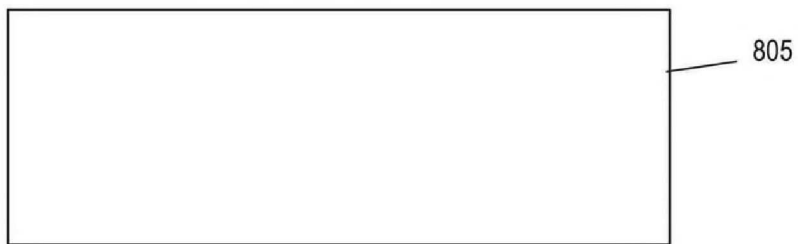


FIG. 9

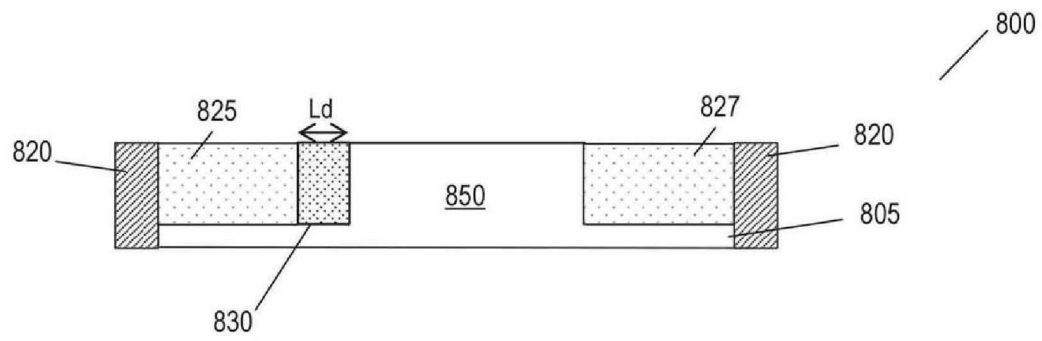


FIG. 10

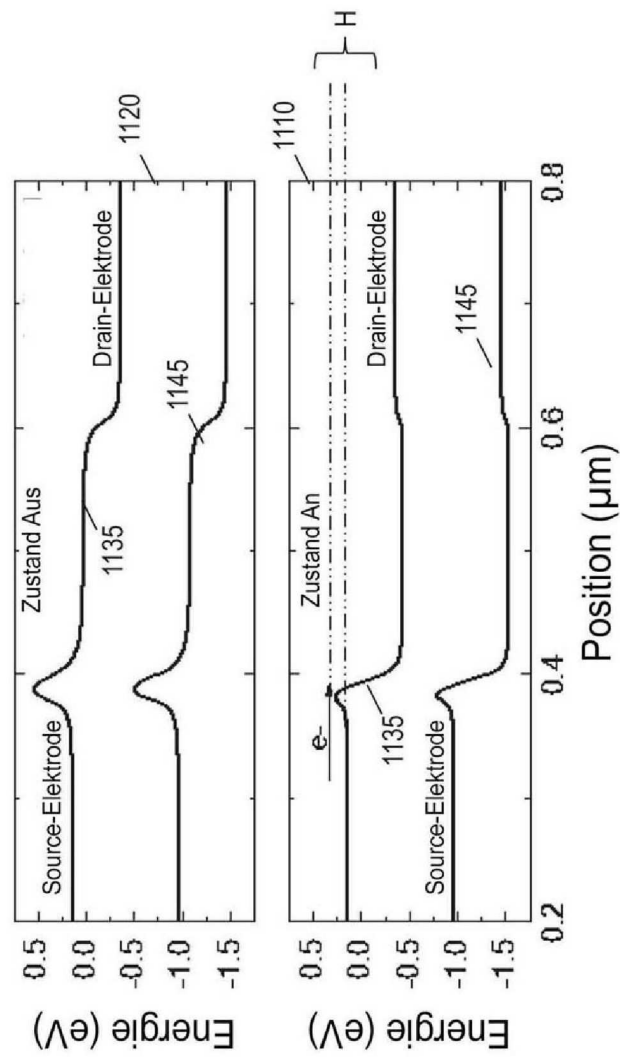


FIG. 11

FIG. 12

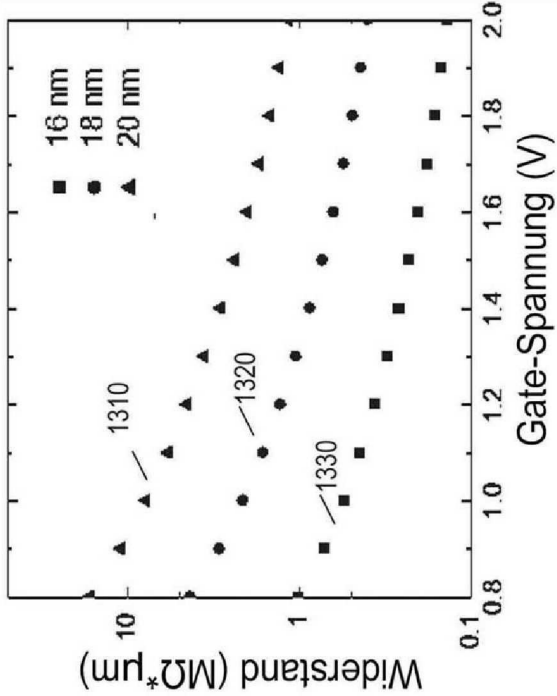
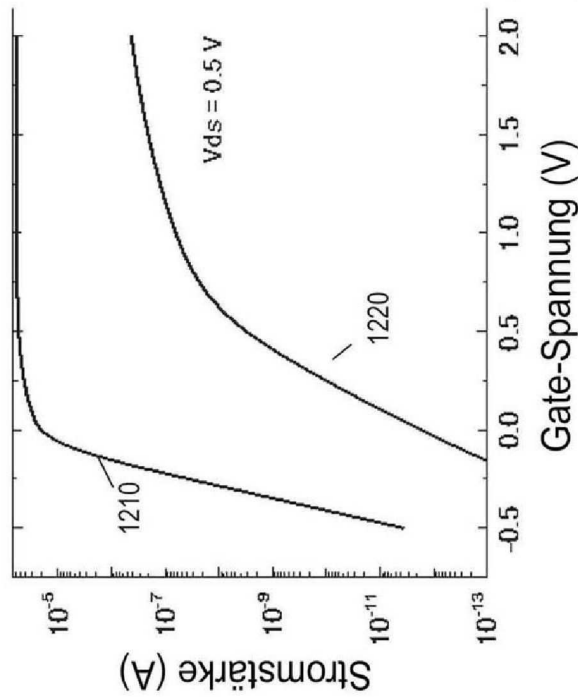


FIG. 13

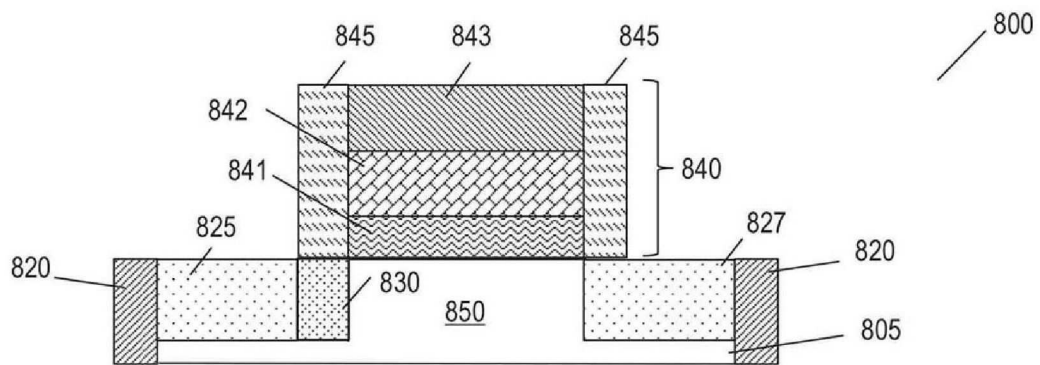


FIG. 14

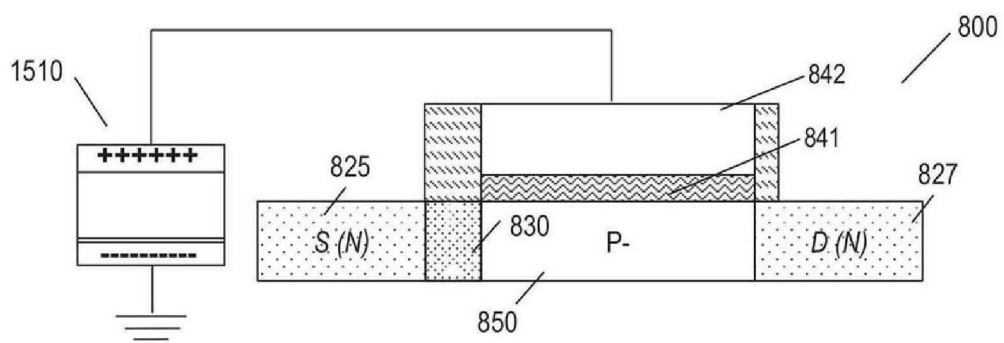


FIG. 15

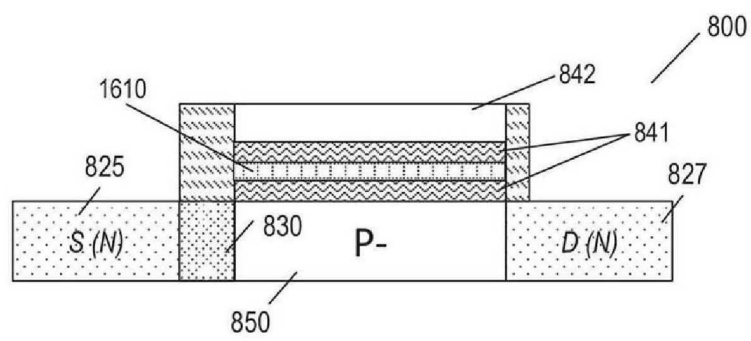


FIG. 16

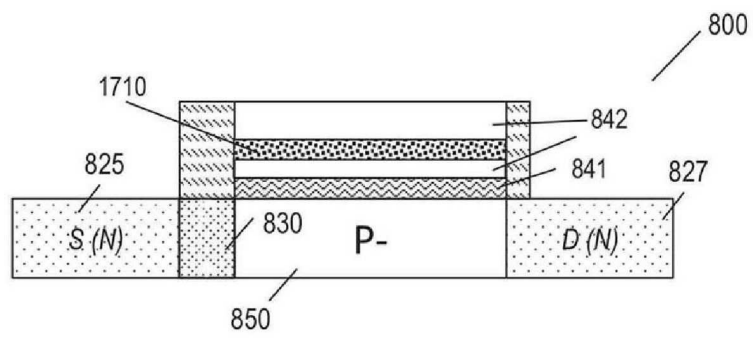


FIG. 17

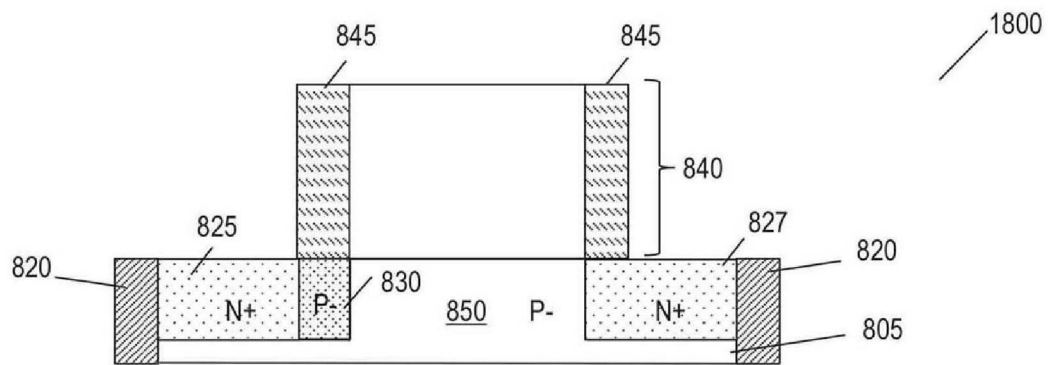


FIG. 18

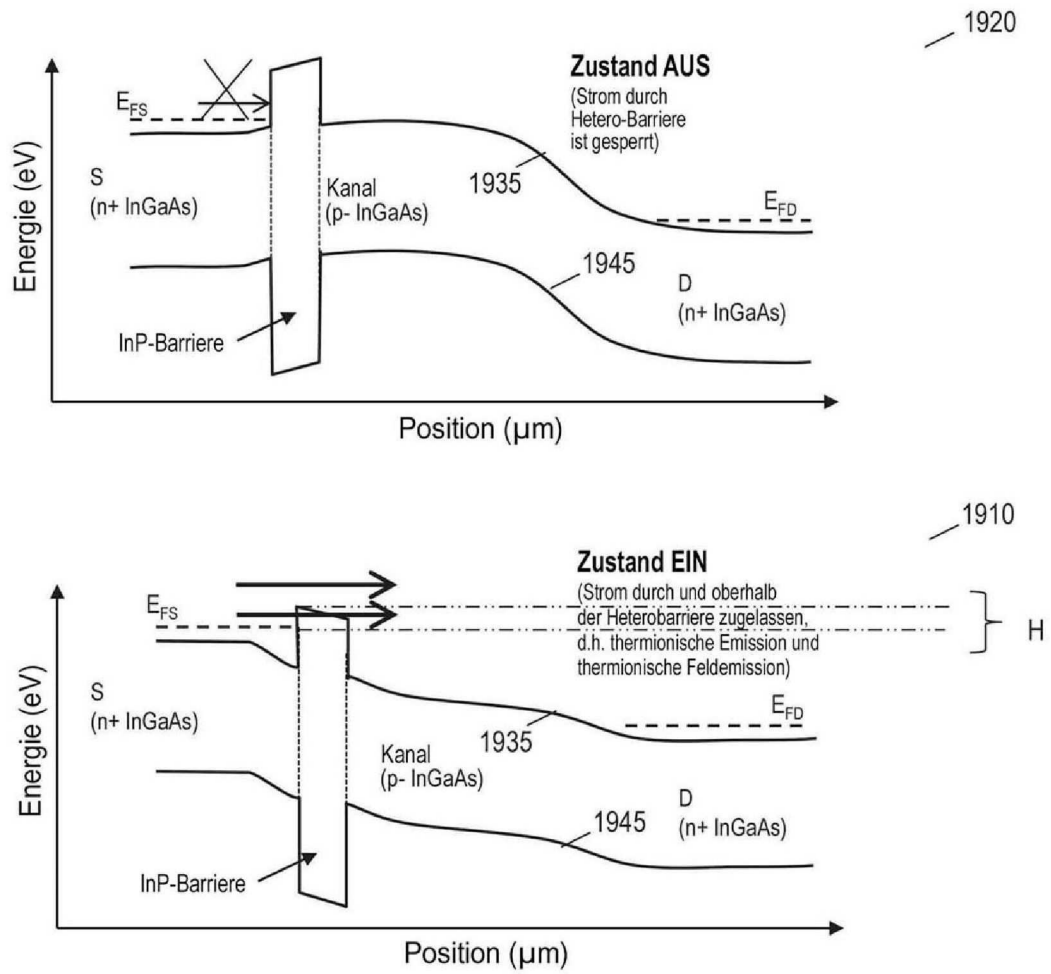


FIG. 19

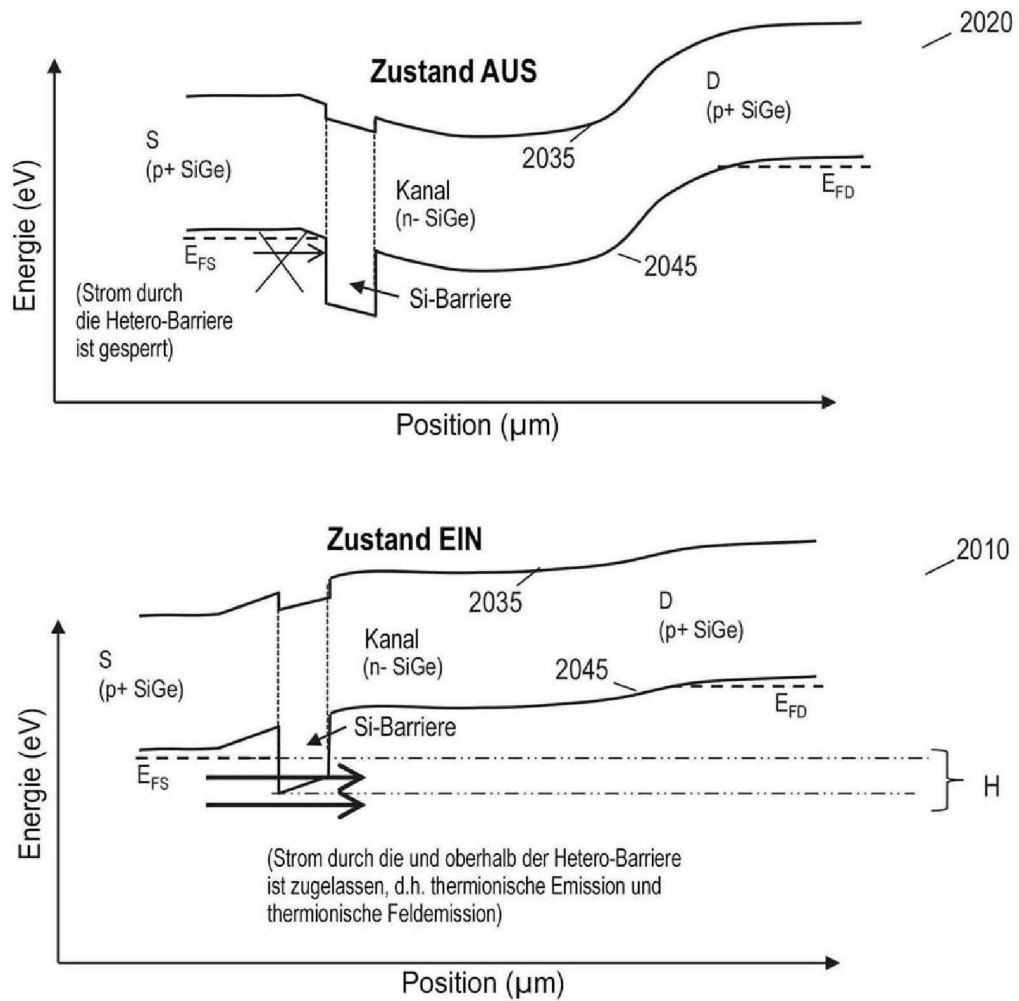


FIG. 20

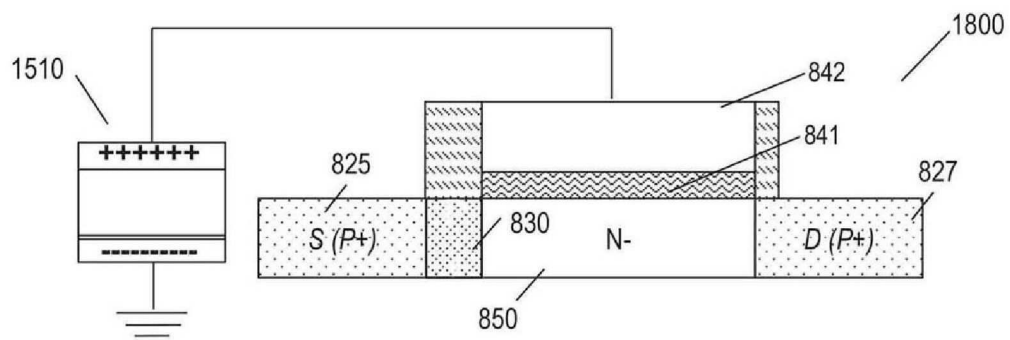


FIG. 21

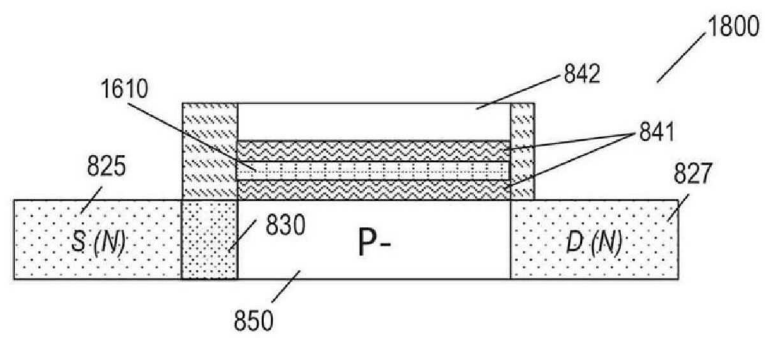


FIG. 22

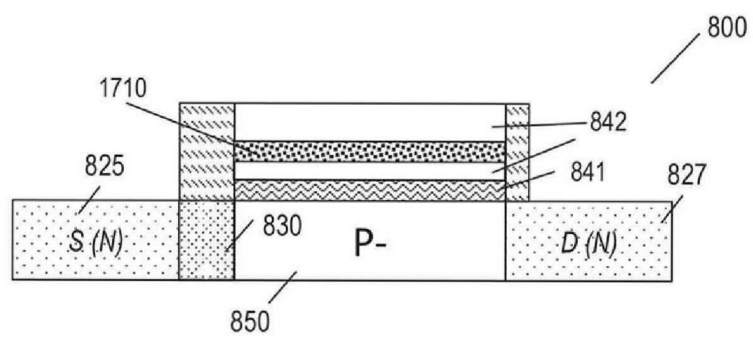


FIG. 23

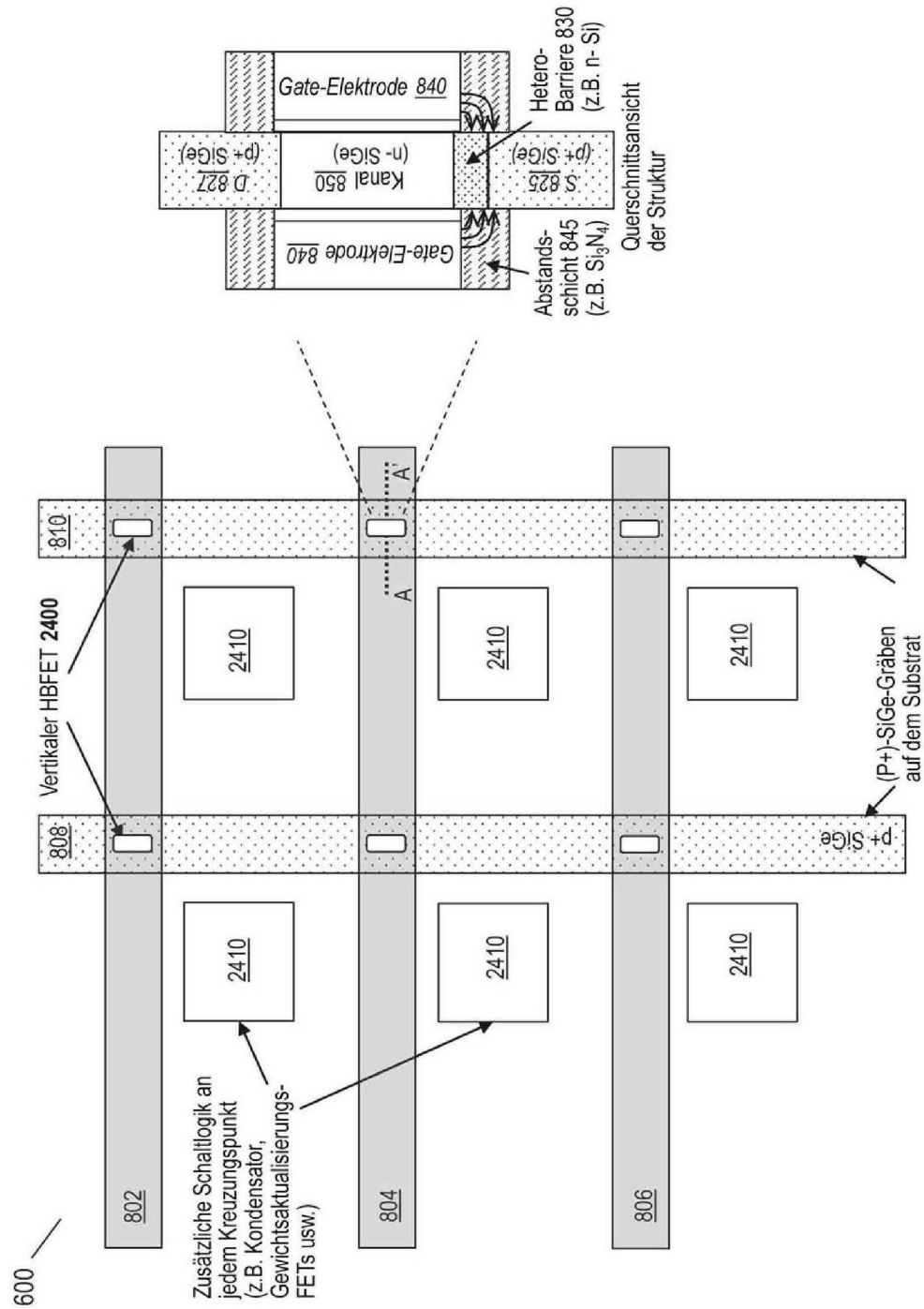


FIG. 24

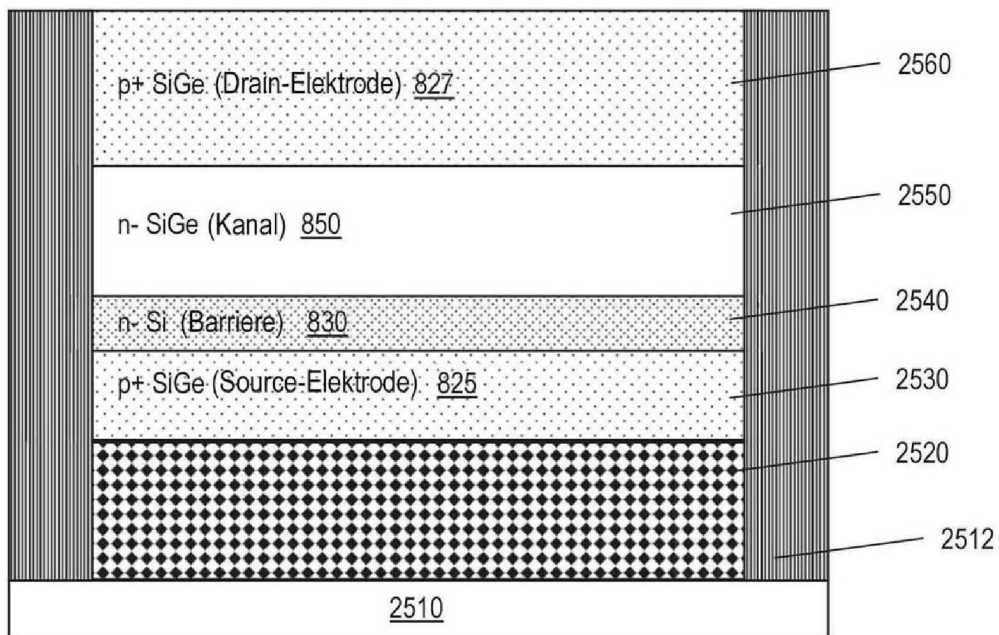


FIG. 25

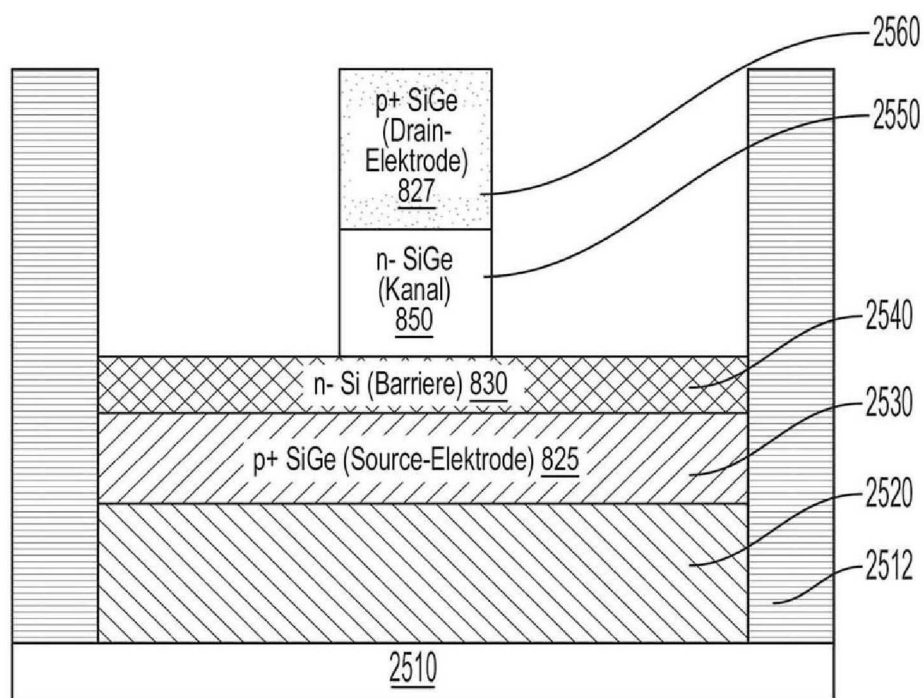


FIG. 26

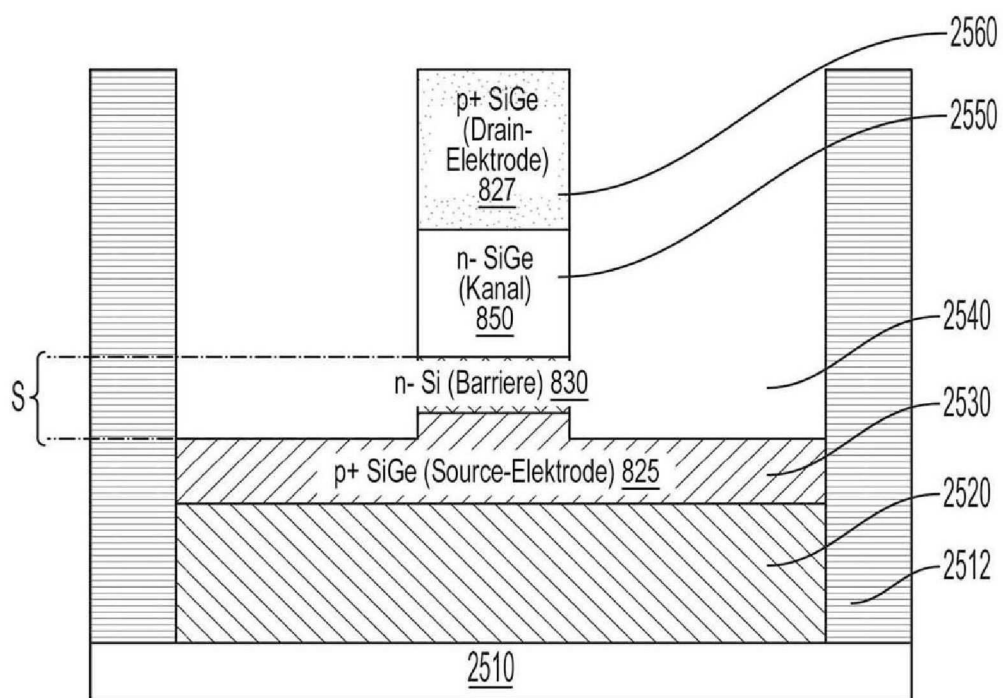


FIG. 27

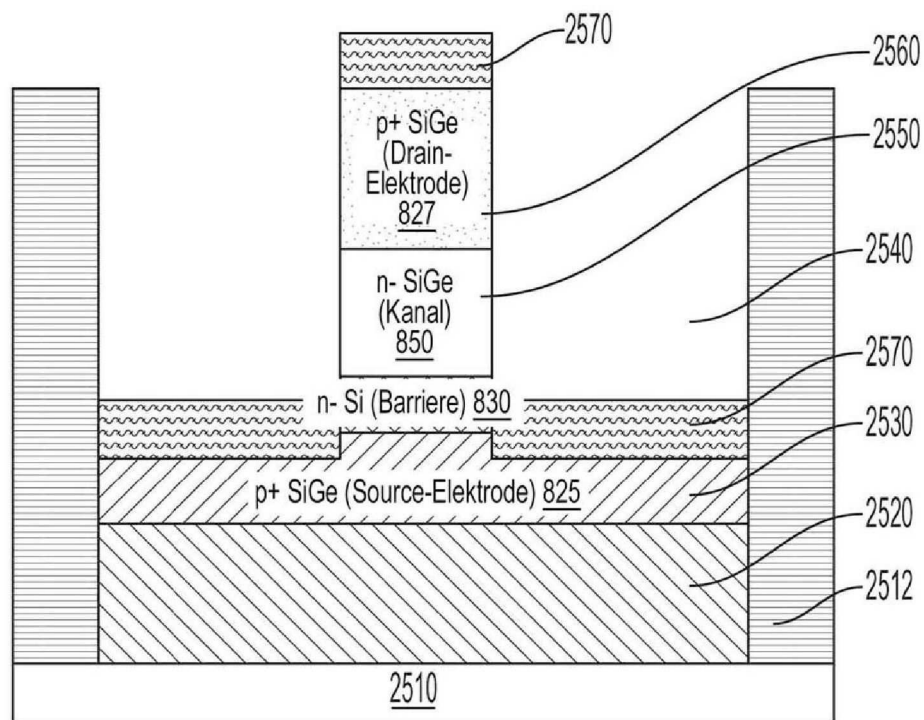


FIG. 28

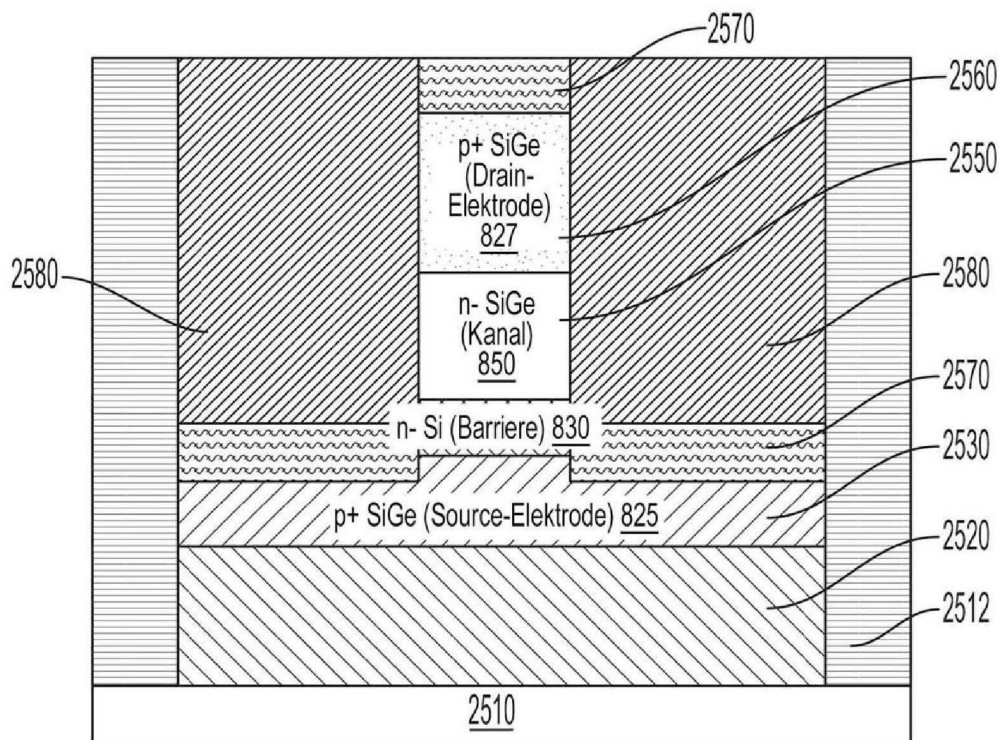


FIG. 29

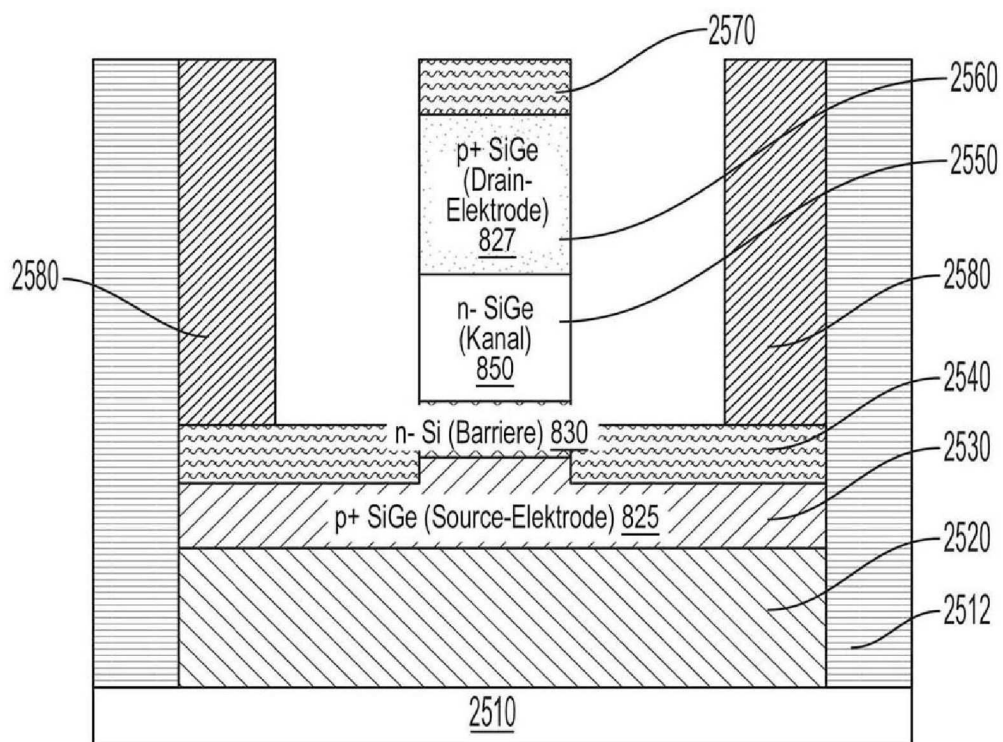


FIG. 30

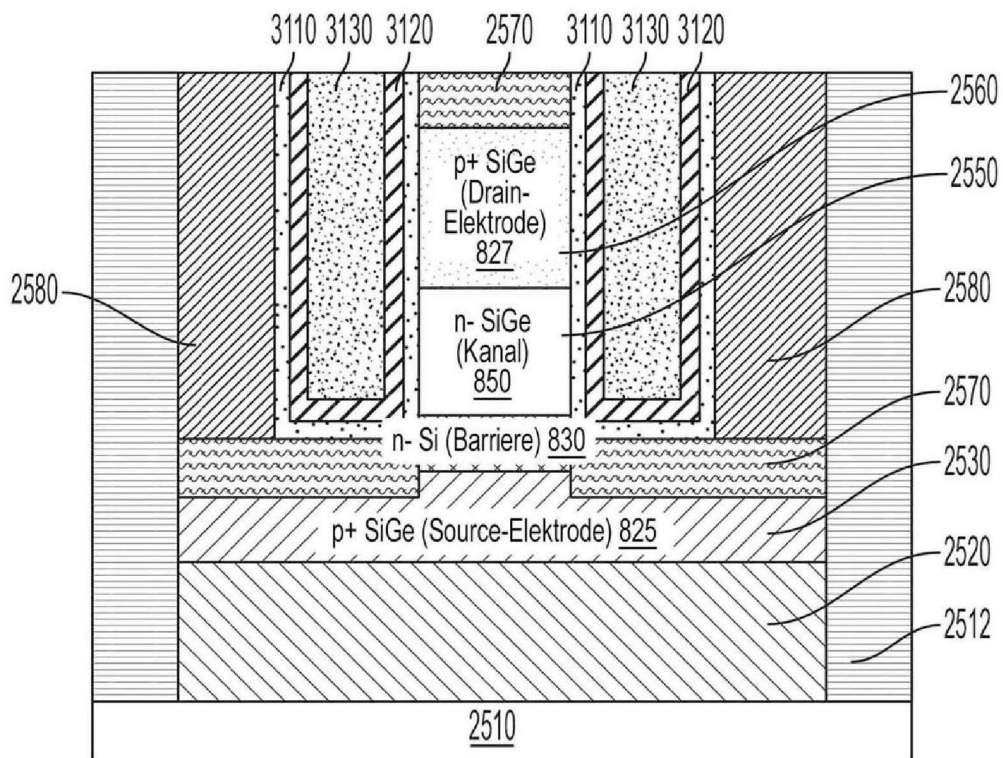


FIG. 31

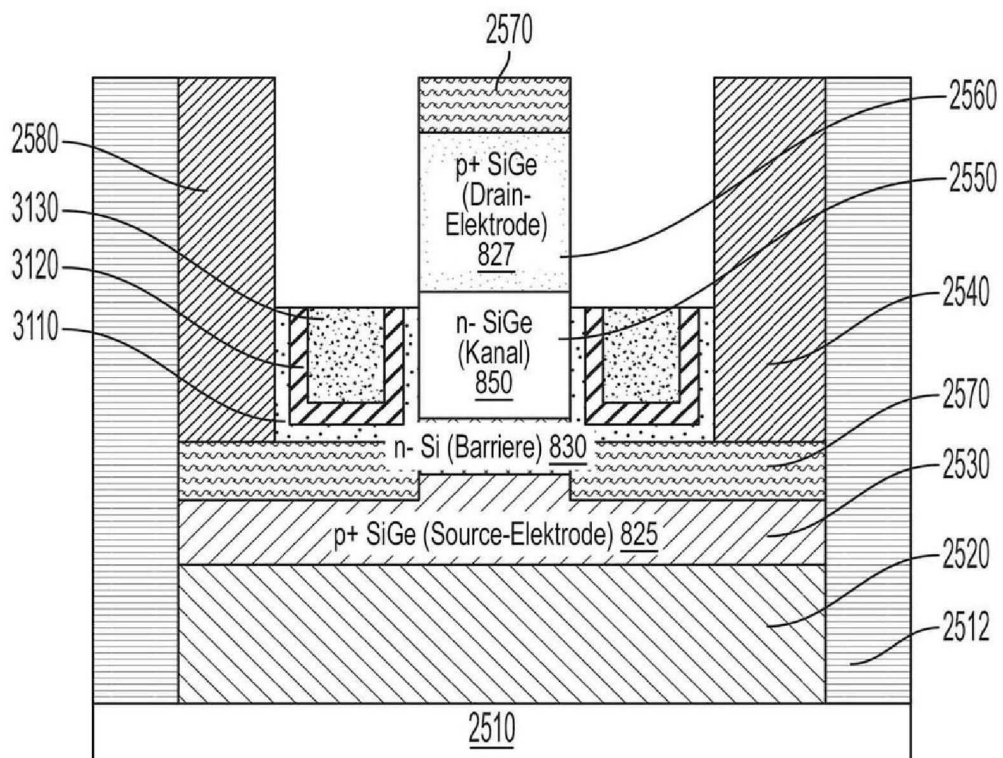


FIG. 32

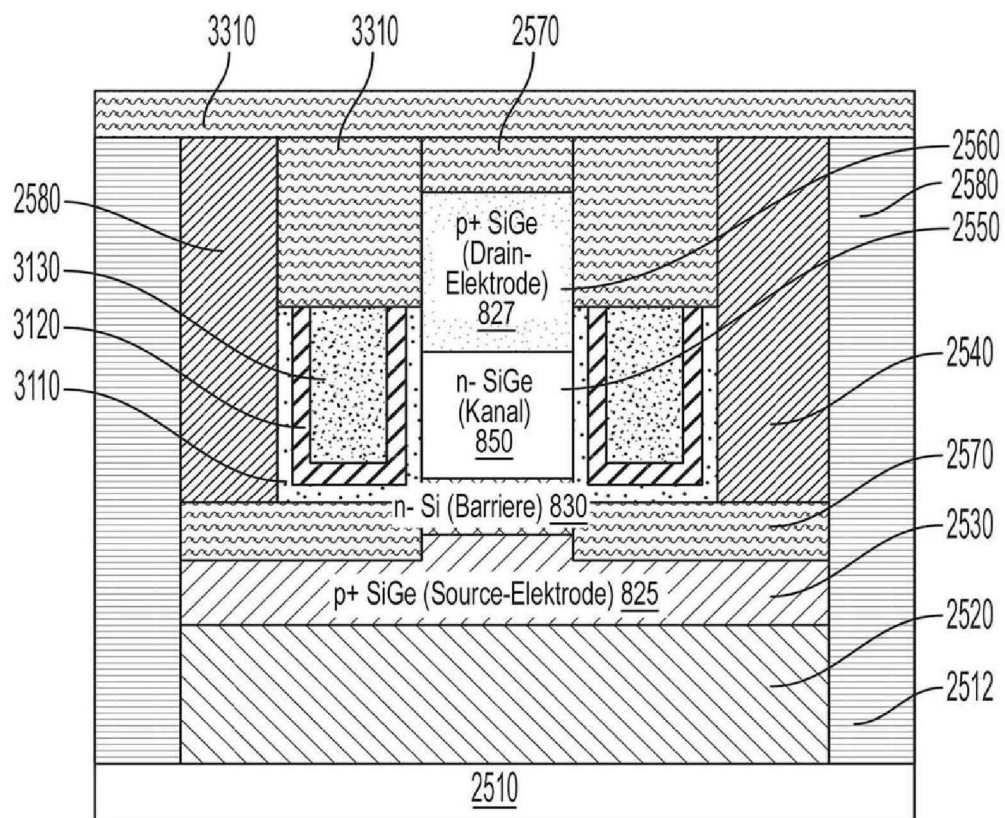


FIG. 33

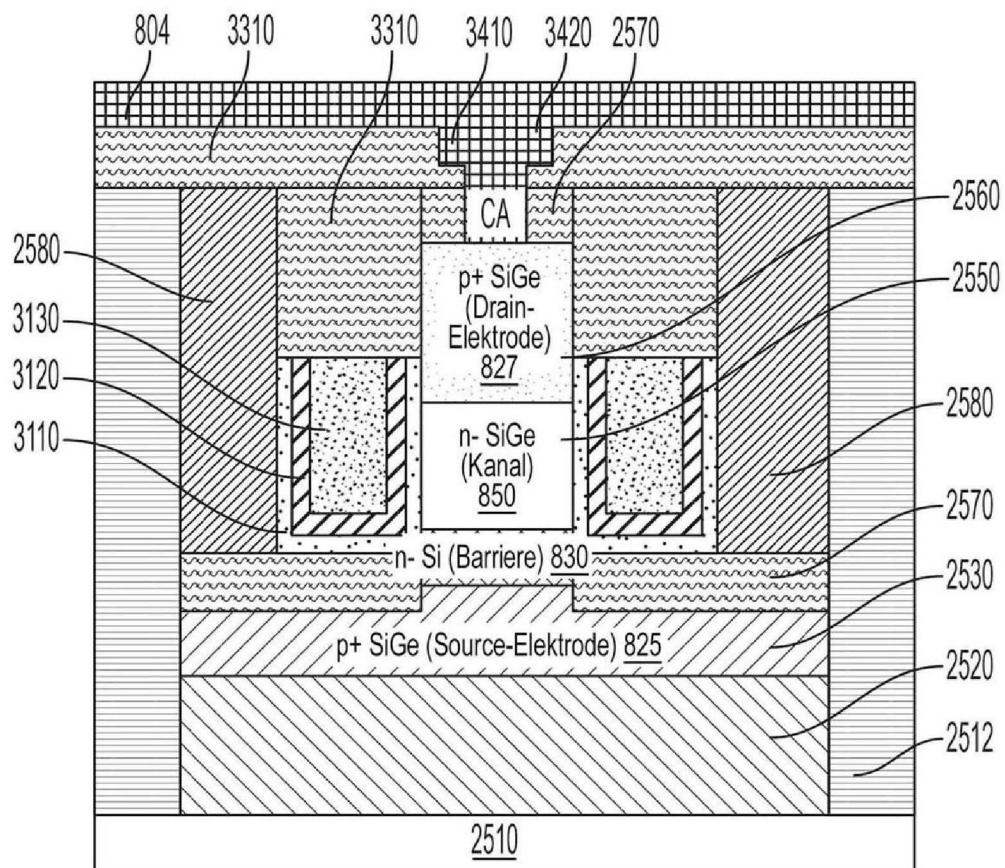


FIG. 34

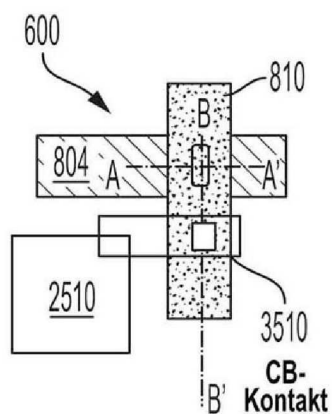


FIG. 35

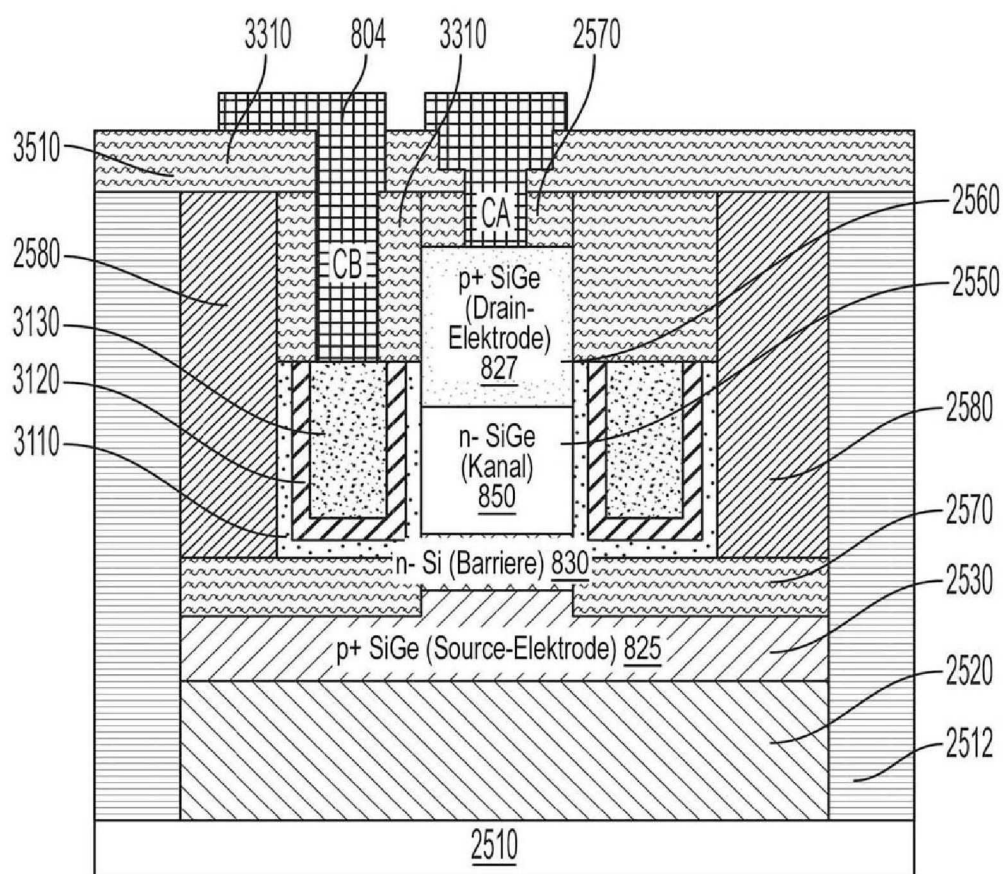


FIG. 36

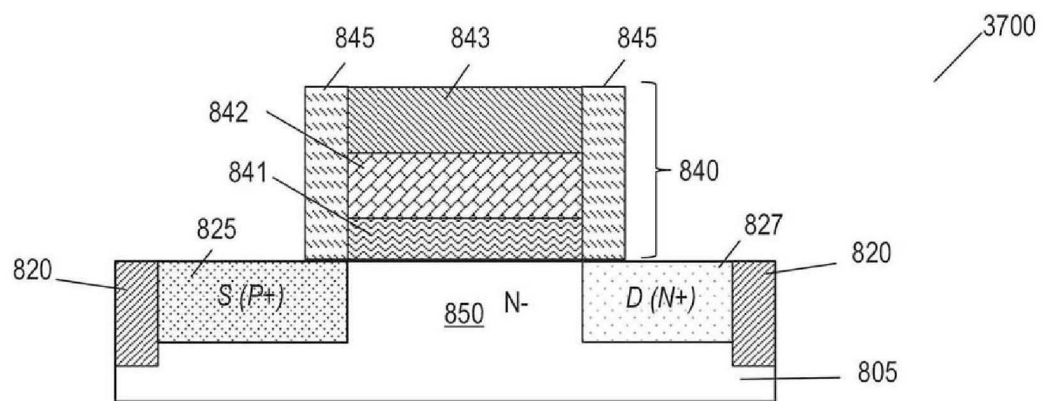


FIG. 37

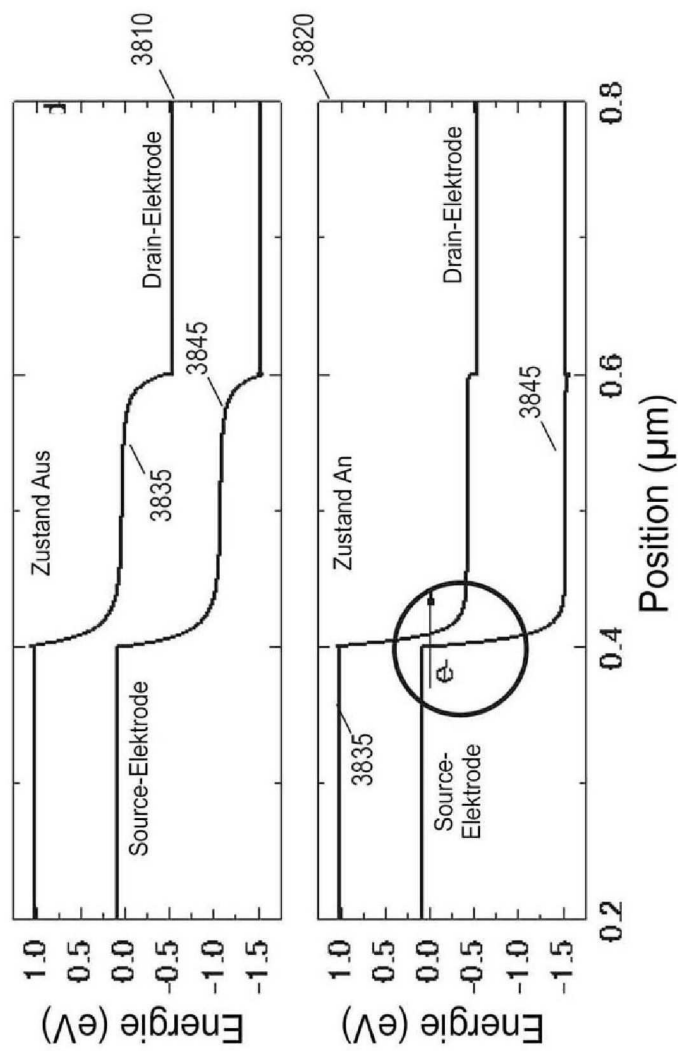


FIG. 38

FIG. 39

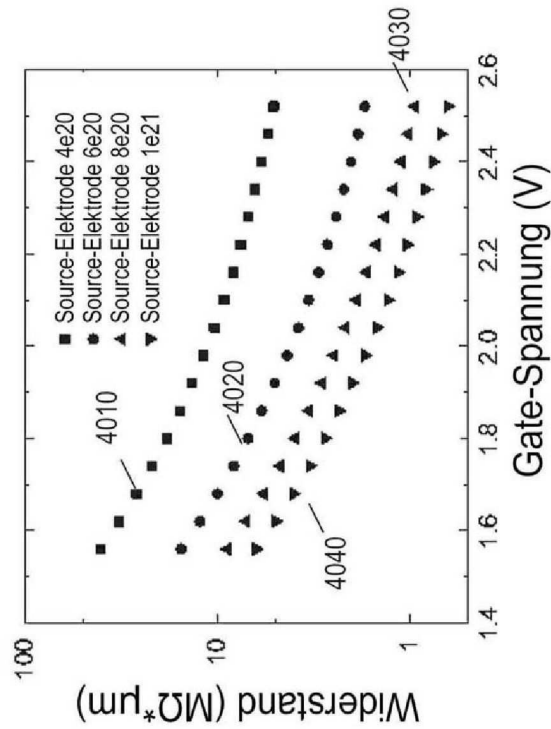
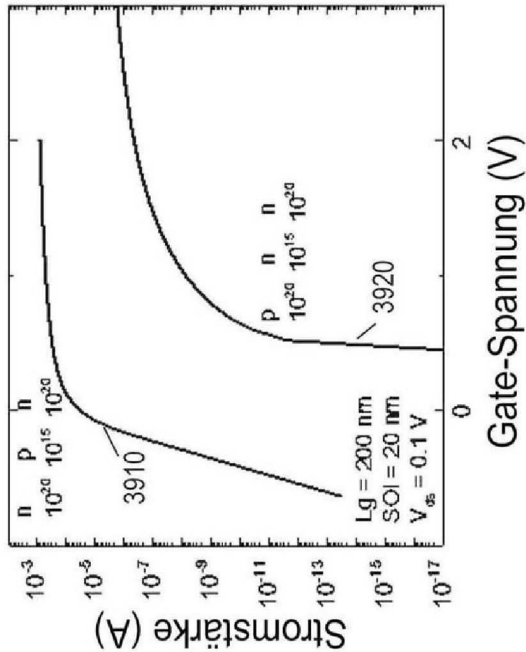


FIG. 40

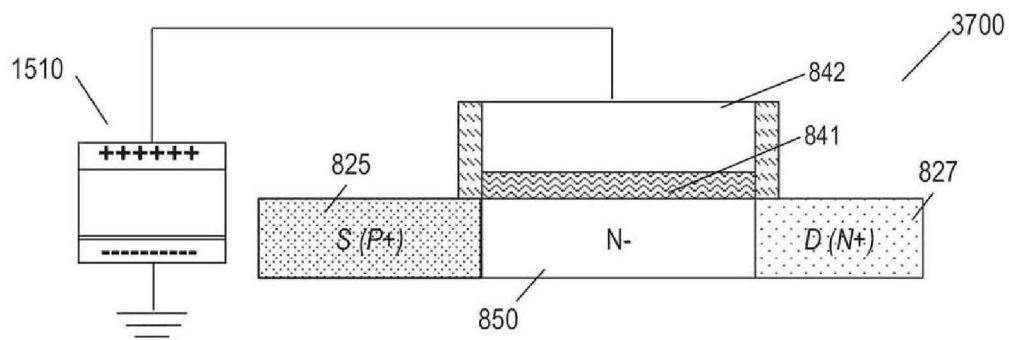


FIG. 41

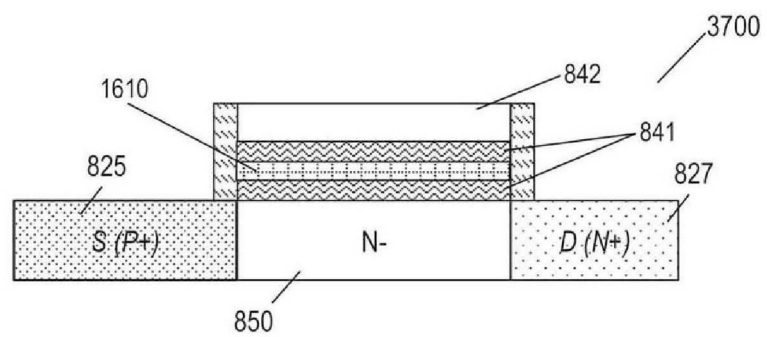


FIG. 42

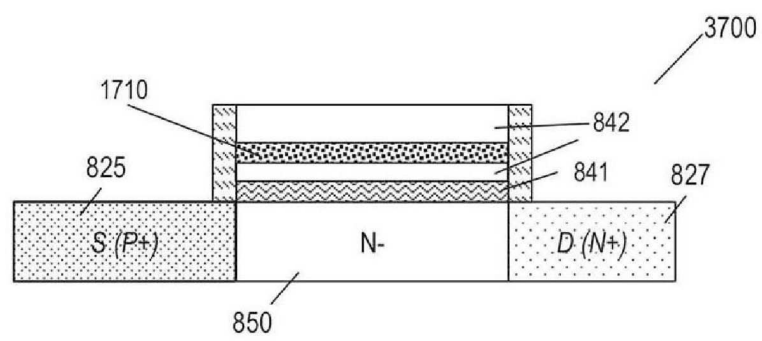


FIG. 43

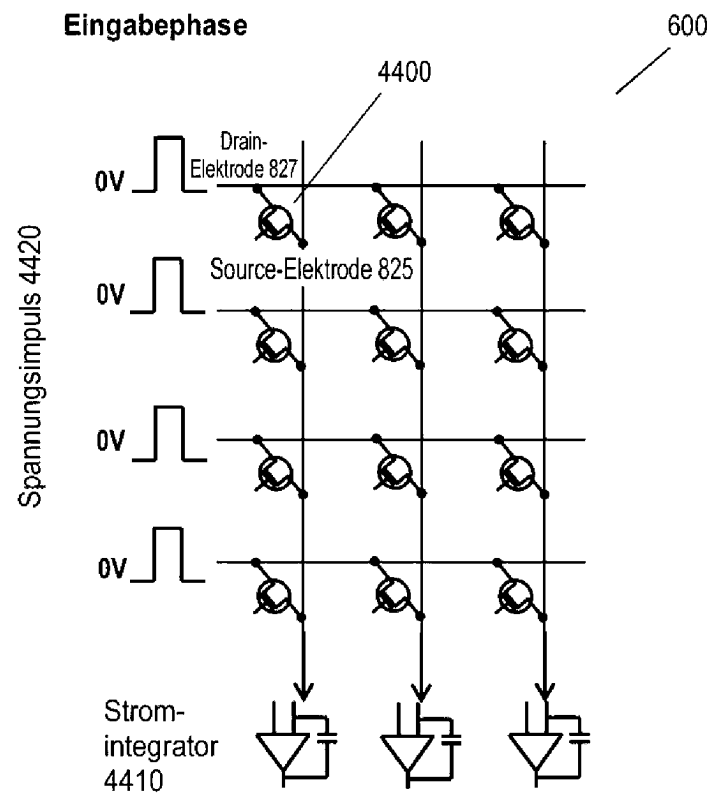


FIG. 44

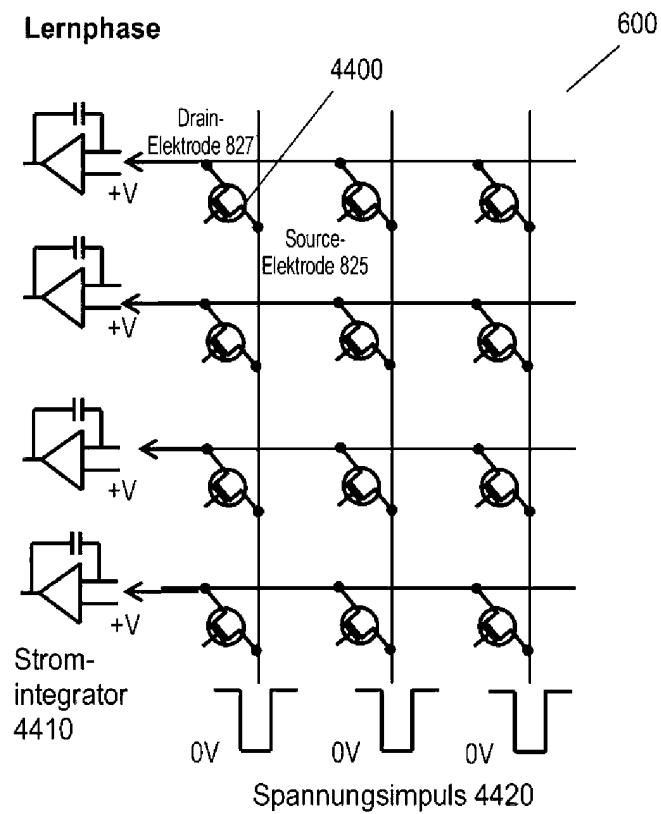


FIG. 45

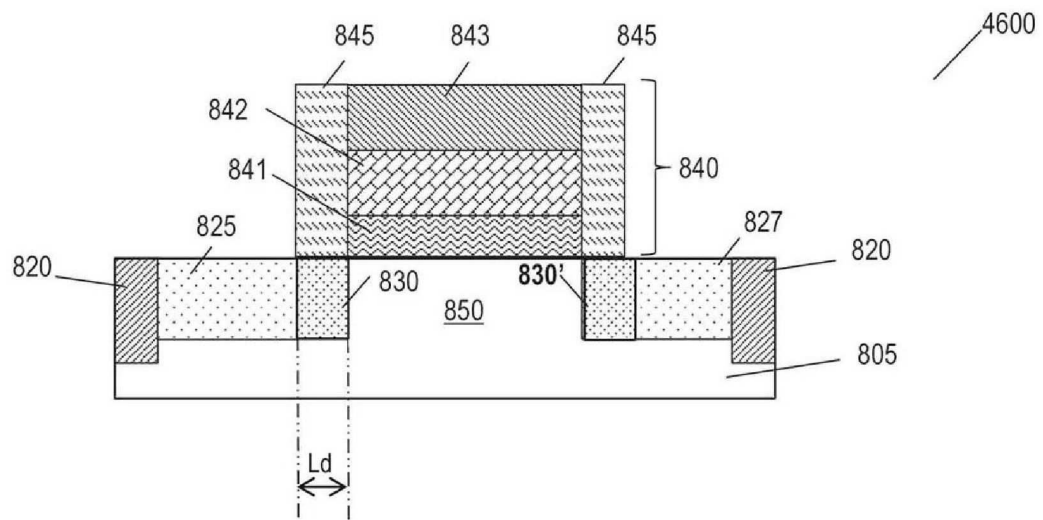


FIG. 46

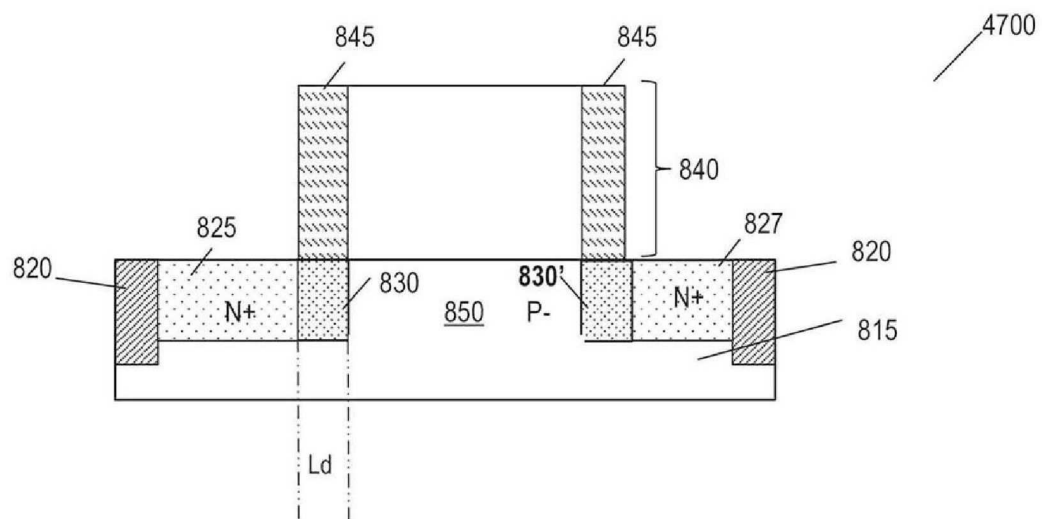


FIG. 47