PAUL J. ANDRE (State Bar No. 196585)
pandre@kslaw.com
LISA KOBIALKA (State Bar No. 191404)
lkobialka@kslaw.com
KING & SPALDING LLP
333 Twin Dolphin Drive
Suite 400
Redwood Shores, CA 94065
Telephone: (650) 590-0700
Facsimile: (650) 590-1900

BRUCE W. SLAYDEN II (TX Bar No. 18496695) (*pro hac vice* to be filed)
bslayden@kslaw.com
R. WILLIAM BEARD, JR. (TX Bar No. 00793318) (*pro hac vice* to be filed)
wbeard@kslaw.com
BRIAN C. BANNER (TX Bar No. 24059416) (*pro hac vice* to be filed)
bbanner@kslaw.com
KING & SPALDING LLP
401 Congress Avenue
Suite 3200
Austin, TX 78701
Telephone: (512) 457-2000
Facsimile: (512) 457-2100

Attorney for Plaintiffs
MICROCHIP TECHNOLOGY, INC., and
SILICON STORAGE TECHNOLOGY, INC.

E-filing
Filed
SEP 2 0 2010
RICHARD W. WIEKING
CLERK, U.S. DISTRICT COURT
NORTHERN DISTRICT OF CALIFORNIA
SAN JOSE
ADR
JCS

IN THE UNITED STATES DISTRICT COURT

FOR THE NORTHERN DISTRICT OF CALIFORNIA

SAN JOSE DIVISION

MICROCHIP TECHNOLOGY, INC. and SILICON STORAGE TECHNOLOGY, INC. Plaintiffs, v. UNITED MODULE CORP., and KERANOS, LLC, Defendants.	Case No. CV 10- 4241 **COMPLAINT** **Demand for Jury Trial**

Plaintiffs, Microchip Technology, Inc. ("Microchip") and Silicon Storage Technology, Inc. ("SST"), hereby demand a jury trial and seek a declaration that they do not infringe United States Patent Nos. 4,795,719 ("'719 Patent") attached hereto as Exhibit A, 4,868,629 ("'629 Patent") attached hereto as Exhibit B, and 5,042,009 ("'009 Patent") attached hereto as Exhibit C; and that the '719 Patent, '629 Patent, and '009 Patent are invalid and unenforceable.

PARTIES

1. Plaintiff Microchip is an Arizona corporation having a principal place of business at 2355 West Chandler Blvd., Chandler, Arizona, 85224. Microchip is a leading provider of microcontroller and analog semiconductors that are used in thousands of customer applications around the world.

2. Plaintiff SST is a California corporation having its principal place of business at 1020 Kifer Road, Sunnyvale, California, 94086. On April 8, 2010, Plaintiff SST became a wholly-owned subsidiary of Plaintiff Microchip.

3. Upon information and belief, Defendant United Module Corp. ("United Module") is a California corporation having its principal place of business located at 978 Highlands Circle, Los Altos, CA 94024.

4. Upon information and belief, Defendant Keranos, LLC ("Keranos") is a Texas limited liability company having its principal place of business located at 211 E. 7th Street, Suite 620, Austin, TX 78701.

JURISDICTION

5. This Complaint arises under the patent laws of the United States, Title 35 of the United States Code, and the Federal Declaratory Judgment Act, 28 U.S.C. §§ 2201–02. This Court has original jurisdiction over the subject matter of these claims made under 28 U.S.C. §§ 1331 & 1338(a).

6. On June 23, 2010, an action alleging infringement of the Patents-in-Suit was brought by Defendant Keranos against Plaintiff Microchip and other parties in the case captioned Keranos LLC v. Analog Devices, Inc., et al., Civil Action No. 2:10-cv-207 in the Eastern District of Texas ("Texas Case").

1 7. Defendant Keranos is not an owner of record of any of the Patents-in-Suit. In its
2 Complaint filed in the Texas Case, Keranos alleges that "Keranos currently holds all applicable
3 exclusive enforcement rights for infringement of the expired Patents-in-Suit through an
4 agreement with United Module, Inc., which owns all rights, title and interest in the Patents-in-
5 Suit."

6 8. Defendant United Module is listed with the U.S. Patent and Trademark Office
7 ("PTO") as the assignee of record of the '719 Patent, the '629 Patent, and the '009 Patent
8 (collectively, "Patents-in-Suit").

9 9. Defendant Keranos lacks constitutional standing to bring a patent infringement
10 suit on the Patents-in-Suit.

11 10. Upon information and belief, Defendant United Module resides and has
12 conducted business in this judicial district, and is subject to personal jurisdiction in this Court.

13 11. Upon information and belief, Defendant Keranos has conducted business in this
14 judicial district, and is subject to personal jurisdiction in this Court. For example, in its
15 Complaint filed in the Texas Case, Keranos admits to entering into an agreement with United
16 Module, who is a resident of this judicial district. As alleged in the Texas Case, that agreement
17 involves the enforcement rights of the Patents-in-Suit. Further, upon information and belief, J.
18 Nicholas Gross, a resident of this judicial district, is the sole governing member of Keranos.

19 12. In its Complaint filed in the Texas Case, Defendant Keranos alleges that
20 Microchip "infringed; induced others to infringe; and/or committed acts of contributory
21 infringement, literally or under the doctrine of equivalents, of one or more claims of the [Patents-
22 in-Suit] by importing, making using, offering to sell, and/or selling products and devices which
23 embody the patented invention, including, among other devices, integrated circuits using
24 embedded flash memory embodied in discrete form, wafer form, or incorporated within larger
25 systems on printed circuit boards." Microchip products that are specifically accused of
26 infringement in the Texas Case "include certain microcontrollers identified by [Microchip] in
27 press releases and other public literature as model numbers/series PIC10; PIC12; PIC16; PIC18;
28 PIC24; dsPIC DSCs; PIC32 and related family of products."

1 13. Because Defendant Keranos has accused Microchip's PIC10, PIC12, PIC16,
2 PIC18, PIC24, dsPIC DSCs, and PIC32 products of infringing the Patents-in-Suit in the Texas
3 Case, Keranos has taken a position that raises a substantial controversy, between parties having
4 adverse legal interests, that is of sufficient immediacy and reality to warrant the issuance of a
5 declaratory judgment. Accordingly, an actual controversy exists as to infringement and validity
6 of each of the Patents-in-Suit.

7 14. Since at least as early as its formation in 1989, Plaintiff SST has been in the
8 business of designing, manufacturing, and marketing a diversified range of semiconductor-based
9 memory and non-memory products for high volume applications. Included in those products are
10 SST's electrically erasable programmable read-only memory (EEPROM) products, which SST
11 has been designing, manufacturing, licensing, and marketing since at least 1993.

12 15. Since at least 1993, Plaintiff SST has licensed its EEPROM technology to various
13 third parties who, in turn, have used SST's EEPROM technology in their own products.

14 16. Plaintiff SST has licensed its EEPROM technology to parties who have been
15 named as defendants in the Texas Case, including Analog Devices, Inc.; Apple, Inc.; EM
16 Microelectronics-Marin SA; Freescale Semiconductor, Inc.; Qualcomm Global Trading, Inc.;
17 Samsung Electronics Co., Ltd.; Taiwan Semiconductor Manufacturing Co., Ltd.; and Winbond
18 Electronics Corporation (collectively, the "Third Parties").

19 17. Upon information and belief, some of the products that are manufactured and sold
20 by the Third Parties and that are accused of infringing the Patents-in-Suit in the Texas Case use
21 SST's EEPROM technology.

22 18. Because some of the third-party products Keranos has accused of infringing the
23 Patents-in-Suit in the Texas Case use SST's EEPROM technology, Defendant Keranos has taken
24 a position that raises a substantial controversy, between parties having adverse legal interests,
25 this is of sufficient immediacy and reality to warrant the issuance of a declaratory judgment.
26 Accordingly, an actual controversy exists as to infringement and validity of each of the Patents-
27 in-Suit.
28

VENUE

19. Venue is proper in this district under 28 U.S.C. §§ 1391 and 1400 because, on information and belief, Defendant United Module resides in this district. On information and belief, Defendant Keranos is subject to personal jurisdiction in this district. Plaintiffs Microchip and SST do business in this district.

INTRADISTRICT ASSIGNMENT

20. Under Civil Local Rules 3-2(c) and 3-5, this action, being a declaratory judgment action based on patent claims, is appropriate for assignment on a district-wide basis.

COUNT I
DECLARATORY JUDGMENT OF
NON-INFRINGEMENT OF THE '719 PATENT

21. Plaintiffs reallege the allegations in paragraphs 1–20 as though fully set forth herein.

22. An actual and justiciable controversy has arisen and exists between Plaintiffs Microchip, SST, and Defendants United Module and Keranos regarding the '719 Patent.

23. By making, using, selling, offering to sell, marketing, licensing, or importing its products, including the PIC10, PIC12, PIC16, PIC18, PIC24, dsPIC DSCs, PIC32, and related family of products, Microchip has not directly or indirectly infringed any claim of the '719 Patent, literally or under the doctrine of equivalents.

24. By making, using, selling, offering to sell, marketing, licensing, or importing its products, including the PIC10, PIC12, PIC16, PIC18, PIC24, dsPIC DSCs, PIC32, and related family of products, Microchip has not induced its customers to infringe any claim of the '719 Patent.

25. By making, using, selling, offering to sell, marketing, licensing, or importing its products, including the PIC10, PIC12, PIC16, PIC18, PIC24, dsPIC DSCs, PIC32, and related family of products, Microchip has not contributorily infringed any claim of the '719 Patent.

26. By making, using, selling, offering to sell, marketing, licensing, or importing its EEPROM technology, SST has not directly or indirectly infringed any claim of the '719 Patent, literally or under the doctrine of equivalents.

27. By making, using, selling, offering to sell, marketing, licensing, or importing its EEPROM technology, SST has not induced its customers to infringe any claim of the '719 Patent.

28. By making, using, selling, offering to sell, marketing, licensing, or importing its EEPROM technology, SST has not contributorily infringed any claim of the '719 Patent.

29. A judicial declaration concerning these matters is necessary and appropriate at this time so that Microchip can ascertain its rights, duties, and obligations with respect to the Defendants, the '719 Patent, and with regard to the design, development, manufacture, marketing, and sales of its products, including the PIC10, PIC12, PIC16, PIC18, PIC24, dsPIC DSCs, PIC32, and related family of products.

30. A judicial declaration concerning these matters is necessary and appropriate at this time so that SST can ascertain its rights, duties, and obligations with respect to the Defendants, the '719 Patent, and with regard to design, development, manufacture, marketing, and sales of its EEPROM technology.

COUNT II
DECLARATORY JUDGMENT OF INVALIDITY OF THE '719 PATENT

31. Plaintiffs reallege the allegations in paragraphs 1–30 as though fully set forth herein.

32. An actual and justiciable controversy has arisen regarding the validity of the '719 Patent.

33. The claims of the '719 Patent are invalid because of a failure to meet the conditions of patentability and/or otherwise comply with one or more of 35 U.S.C. §§ 100 *et seq.*, including §§ 102, 103, and 112.

34. A judicial declaration concerning these matters is necessary and appropriate at this time so that Microchip and SST can ascertain their rights, duties, and obligations with respect to the '719 Patent.

COUNT III
DECLARATORY JUDGMENT OF
NON-INFRINGEMENT OF THE '629 PATENT

35. Plaintiffs reallege the allegations in paragraphs 1–34 as though fully set forth herein.

36. An actual and justiciable controversy has arisen and exists between Plaintiffs Microchip, SST, and Defendants United Module and Keranos regarding the '629 Patent.

37. By making, using, selling, offering to sell, marketing, licensing, or importing its products, including the PIC10, PIC12, PIC16, PIC18, PIC24, dsPIC DSCs, PIC32, and related family of products, Microchip has not directly or indirectly infringed any claim of the '629 Patent, literally or under the doctrine of equivalents.

38. By making, using, selling, offering to sell, marketing, licensing, or importing its products, including the PIC10, PIC12, PIC16, PIC18, PIC24, dsPIC DSCs, PIC32, and related family of products, Microchip has not induced its customers to infringe any claim of the '629 Patent.

39. By making, using, selling, offering to sell, marketing, licensing, or importing its products, including the PIC10, PIC12, PIC16, PIC18, PIC24, dsPIC DSCs, PIC32, and related family of products, Microchip has not contributorily infringed any claim of the '629 Patent.

40. By making, using, selling, offering to sell, marketing, licensing, or importing its EEPROM technology, SST has not directly or indirectly infringed any claim of the '629 Patent, literally or under the doctrine of equivalents.

41. By making, using, selling, offering to sell, marketing, licensing, or importing its EEPROM technology, SST has not induced its customers to infringe any claim of the '629 Patent.

42. By making, using, selling, offering to sell, marketing, licensing, or importing its EEPROM technology, SST has not contributorily infringed any claim of the '629 Patent.

43. A judicial declaration concerning these matters is necessary and appropriate at this time so that Microchip can ascertain its rights, duties, and obligations with respect to the Defendants, the '629 Patent, and with regard to the design, development, manufacture,

marketing, and sales of its products, including the PIC10, PIC12, PIC16, PIC18, PIC24, dsPIC DSCs, PIC32, and related family of products.

44. A judicial declaration concerning these matters is necessary and appropriate at this time so that SST can ascertain its rights, duties, and obligations with respect to the Defendants, the '629 Patent, and with regard to the design, development, manufacture, marketing, and sales of its EEPROM technology.

COUNT IV
DECLARATORY JUDGMENT OF
INVALIDITY OF THE '629 PATENT

45. Plaintiffs reallege the allegations in paragraphs 1–44 as though fully set forth herein.

46. An actual and justiciable controversy has arisen regarding the validity of the '629 Patent.

47. The claims of the '629 Patent are invalid because of a failure to meet the conditions of patentability and/or otherwise comply with one or more of 35 U.S.C. §§ 100 *et seq.*, including § 102, 103, and 112.

48. A judicial declaration concerning these matters is necessary and appropriate at this time so that Microchip and SST can ascertain their rights, duties, and obligations with respect to the '629 Patent.

COUNT V
DECLARATORY JUDGMENT OF
NON-INFRINGEMENT OF THE '009 PATENT

49. Microchip realleges the allegations in paragraphs 1–48 as though fully set forth herein.

50. An actual and justiciable controversy has arisen and exists between Plaintiffs Microchip, SST, and Defendants United Module and Keranos regarding the '009 Patent.

51. By making, using, selling, offering to sell, marketing, licensing, or importing its products, including the PIC10, PIC12, PIC16, PIC18, PIC24, dsPIC DSCs, PIC32, and related family of products, Microchip has not directly or indirectly infringed any claim of the '009

Patent, literally or under the doctrine of equivalents.

52. By making, using, selling, offering to sell, marketing, licensing, or importing its products, including the PIC10, PIC12, PIC16, PIC18, PIC24, dsPIC DSCs, PIC32, and related family of products, Microchip has not induced its customers to infringe any claim of the '009 Patent.

53. By making, using, selling, offering to sell, marketing, licensing, or importing its products, including the PIC10, PIC12, PIC16, PIC18, PIC24, dsPIC DSCs, PIC32, and related family of products, Microchip has not contributorily infringed any claim of the '009 Patent.

54. By making, using, selling, offering to sell, marketing, licensing, or importing its EEPROM technology, SST has not directly or indirectly infringed any claim of the '009 Patent, literally or under the doctrine of equivalents.

55. By making, using, selling, offering to sell, marketing, licensing, or importing its EEPROM technology, SST has not induced its customers to infringe any claim of the '009 Patent.

56. By making, using, selling, offering to sell, marketing, licensing, or importing its EEPROM technology, SST has not contributorily infringed any claim of the '009 Patent.

57. A judicial declaration concerning these matters is necessary and appropriate at this time so that Microchip can ascertain its rights, duties, and obligations with respect to the Defendants, the '009 Patent, and with regard to the design, development, manufacture, marketing, and sales of its products, including the PIC10, PIC12, PIC16, PIC18, PIC24, dsPIC DSCs, PIC32, and related family of products.

58. A judicial declaration concerning these matters is necessary and appropriate at this time so that SST can ascertain its rights, duties, and obligations with respect to the Defendants, the '009 Patent, and with regard to the design, development, manufacture, marketing, and sales of its EEPROM technology.

COUNT VI
DECLARATORY JUDGMENT OF
INVALIDITY OF THE '009 PATENT

59. Microchip realleges the allegations in paragraphs 1–58 as though fully set forth

1 herein.

2 60. An actual and justiciable controversy has arisen regarding the validity of the '009
3 Patent.

4 61. The claims of the '009 Patent are invalid because of a failure to meet the
5 conditions of patentability and/or otherwise comply with one or more of 35 U.S.C. §§ 100 *et*
6 *seq.*, including § 102, 103, and 112.

7 62. A judicial declaration concerning these matters is necessary and appropriate at
8 this time so that Microchip and SST can ascertain their rights, duties, and obligations with
9 respect to the '009 Patent.

10 **REQUEST FOR RELIEF**

11 WHEREFORE, Plaintiffs Microchip and SST request judgment as follows:

12 A. For a declaration that the claims of U.S. Patent No. 4,795,719 are invalid;

13 B. For a declaration that neither Microchip nor any of its products infringe (directly,
14 indirectly, literally, and/or under the doctrine of equivalents) any valid claim of U.S. Patent No.
15 4,795,719;

16 C. For a declaration that Microchip has not contributorily infringed or induced
17 infringement of any valid claim of U.S. Patent No. 4,795,719;

18 D. For a declaration that no valid claim of U.S. Patent No. 4,795,719 has been
19 infringed (directly, indirectly, literally, and/or under the doctrine of equivalents) by any of
20 Microchip's customers by virtue of incorporating any Microchip product into any such
21 customer's products;

22 E. For a declaration that neither SST nor any of its products infringe (directly,
23 indirectly, literally, and/or under the doctrine of equivalents) any valid claim of U.S. Patent No.
24 4,795,719;

25 F. For a declaration that SST has not contributorily infringed or induced
26 infringement of any valid claim of U.S. Patent No. 4,795,719;

27 G. For a declaration that no valid claim of U.S. Patent No. 4,795,719 is infringed
28 (directly, indirectly, literally, and/or under the doctrine of equivalents) by any of SST's

customers/licensees by virtue of incorporating any SST technology into any such customer's/licensee's products;

H. For a declaration that the claims of U.S. Patent No. 4,868,629 are invalid;

I. For a declaration that neither Microchip nor any of its products infringe (directly, indirectly, literally, and/or under the doctrine of equivalents) any valid claim of U.S. Patent No. 4,868,629;

J. For a declaration that Microchip has not contributorily infringed or induced infringement of any valid claim of U.S. Patent No. 4,868,629;

K. For a declaration that no valid claim of U.S. Patent No. 4,868,629 is infringed (directly, indirectly, literally, and/or under the doctrine of equivalents) by any of Microchip's customers by virtue of incorporating any Microchip product into any such customer's products;

L. For a declaration that neither SST nor any of its products infringe (directly, indirectly, literally, and/or under the doctrine of equivalents) any valid claim of U.S. Patent No. 4,868,629;

M. For a declaration that SST has not contributorily infringed or induced infringement of any valid claim of U.S. Patent No. 4,868,629;

N. For a declaration that no valid claim of U.S. Patent No. 4,868,629 is infringed (directly, indirectly, literally, and/or under the doctrine of equivalents) by any of SST's customers/licensees by virtue of incorporating any SST technology into any such customer's/licensee's products;

O. For a declaration that the claims of U.S. Patent No. 5,042,009 are invalid;

P. For a declaration that neither Microchip nor any of its products infringe (directly, indirectly, literally, and/or under the doctrine of equivalents) any valid claim of U.S. Patent No. 5,042,009;

Q. For a declaration that Microchip has not contributorily infringed or induced infringement of any valid claim of U.S. Patent No. 5,042,009;

R. For a declaration that no valid claim of U.S. Patent No. 5,042,009 is infringed (directly, indirectly, literally, and/or under the doctrine of equivalents) by any of Microchip's

customers by virtue of incorporating any Microchip product into any such customer's products;

S. For a declaration that neither SST nor any of its products infringe (directly, indirectly, literally, and/or under the doctrine of equivalents) any valid claim of U.S. Patent No. 5,042,009;

T. For a declaration that SST has not contributorily infringed or induced infringement of any valid claim of U.S. Patent No. 5,042,009;

U. For a declaration that no valid claim of U.S. Patent No. 5,042,009 is infringed (directly, indirectly, literally, and/or under the doctrine of equivalents) by any of SST's customers/licensees by virtue of incorporating any SST technology into any such customer's/licensee's products;

V. For a determination that this case is exceptional under 35 U.S.C. § 285 and an award to Microchip and SST of their attorneys' fees, costs, and expenses in conjunction with this action; and

W. Such other and further relief as this Court or a jury may deem proper and just.

Dated: September 20, 2010

Respectfully submitted,

Paul J. Andre
Lisa Kobialka
KING & SPALDING LLP
333 Twin Dolphin Drive
Suite 400
Redwood Shores, CA 94065
Telephone: (650) 590-0700
Facsimile: (650) 590-1900

Attorney for Plaintiffs
MICROCHIP TECHNOLOGY, INC., and
SILICON STORAGE TECHNOLOGY, INC.

and

Bruce W. Slayden II (*pro hac vice* to be filed)
R. William Beard, Jr. (*pro hac vice* to be filed)
Brian C. Banner (*pro hac vice* to be filed)
KING & SPALDING LLP
401 Congress Avenue
Suite 3200
Austin, TX 78701
Telephone: (512) 457-2000
Facsimile: (512) 457-2100

Attorney for Plaintiffs
MICROCHIP TECHNOLOGY, INC., and
SILICON STORAGE TECHNOLOGY, INC.

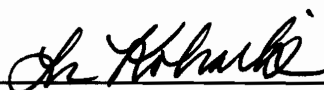
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28

DEMAND FOR JURY TRIAL

Microchip and SST hereby request a jury trial as to all issues triable to a jury.

Dated: September 20, 2010

Respectfully submitted,



Paul J. Andre
Lisa Kobialka
KING & SPALDING LLP
333 Twin Dolphin Drive
Suite 400
Redwood Shores, CA 94065
Telephone: (650) 590-0700
Facsimile: (650) 590-1900

Attorney for Plaintiffs
MICROCHIP TECHNOLOGY, INC., and
SILICON STORAGE TECHNOLOGY, INC.

and

Bruce W. Slayden II (*pro hac vice* to be filed)
R. William Beard, Jr. (*pro hac vice* to be filed)
Brian C. Banner (*pro hac vice* to be filed)
KING & SPALDING LLP
401 Congress Avenue
Suite 3200
Austin, TX 78701
Telephone: (512) 457-2000
Facsimile: (512) 457-2100

Attorney for Plaintiffs
MICROCHIP TECHNOLOGY, INC., and
SILICON STORAGE TECHNOLOGY, INC.

EXHIBIT A

United States Patent [19]

Eitan

[11] **Patent Number: 4,795,719**

[45] **Date of Patent: Jan. 3, 1989**

[54] **SELF-ALIGNED SPLIT GATE EPROM PROCESS**

[75] Inventor: **Boaz Eitan**, Sunnyvale, Calif.

[73] Assignee: **WaferScale Integration, Inc.**, Fremont, Calif.

[21] Appl. No.: **900,065**

[22] Filed: **Aug. 22, 1986**

Related U.S. Application Data

[62] Division of Ser. No. 610,369, May 15, 1984, Pat. No. 4,639,893.

[51] Int. Cl.⁴ H01L 27/10

[52] U.S. Cl. 437/43; 437/44; 437/48; 437/49; 437/52; 437/200; 437/984

[58] Field of Search .... 357/23.5, 54, 71; 148/DIG. 141, DIG. 109; 29/576 B, 571, 577 C, 578

[56] **References Cited**

U.S. PATENT DOCUMENTS

4,122,544	10/1978	McElroy .	
4,142,926	3/1979	Morgan	357/23.5
4,173,318	11/1979	Bassous et al.	29/571
4,173,791	11/1979	Bell	357/23.5
4,247,558	5/1981	Guterman	357/23.5
4,257,832	3/1981	Schwabe et al.	357/23.5
4,274,012	6/1981	Simko .	
4,297,719	10/1981	Hsu	357/23.5
4,318,216	3/1982	Hsu	437/29
4,334,292	6/1982	Eitan	357/23.5
4,336,603	6/1982	Kotecha et al.	357/23.5
4,380,866	4/1983	Countryman et al.	437/48
4,412,311	10/1983	Miccoli et al.	357/23.5
4,426,764	1/1984	Kosa et al.	29/571
4,462,090	7/1984	Iizuka .	
4,471,373	9/1984	Shimizu et al.	357/23.5
4,495,693	1/1985	Iwahashi et al.	29/576 B
4,561,004	12/1985	Kuo .	

FOREIGN PATENT DOCUMENTS

816931	7/1969	Canada .	
0045578	2/1982	European Pat. Off. .	
0158078	12/1982	Fed. Rep. of Germany	29/571
2437676	9/1979	France .	
0063684	5/1977	Japan	357/23.5
0089686	8/1978	Japan	357/23.5
0156369	12/1980	Japan	357/23.5
0071971	6/1981	Japan	29/571
0076878	5/1982	Japan	357/23.5
0206165	12/1983	Japan	357/23.5
2073484	10/1981	United Kingdom	357/23.5

OTHER PUBLICATIONS

IEEE Transactions, Electron Devices, vol. ED-29, No. 4, Apr. 1982, pp. 611-et seq., "Semiconductor MOSFET Structure for Minimizing Hot Carrier Generation".

IEEE Transactions, Electron Devices/vol. ED-32, No. 3, Mar. 1985, pp. 562-et seq., "Optimum Design of N+-N-Double-Diffused Drain MOSFET to Reduce Hot-Carrier Emission".

Electronics/Aug. 21, 1986, pp. 53-56.

Electronics/Sep. 4, 1986, p. 30.

Electronics/ Mar. 3, 1988, pp. 47-48.

IEE Transactions, Electron Devices/vol. ED-32, No. 5, May 1985, pp. 896-et seq., "Lightly Doped Drain Transistors for Advanced VLSI Circuits".

Primary Examiner—Brian E. Hearn
Assistant Examiner—Tom Thomas
Attorney, Agent, or Firm—Alan H. MacPherson; Gideon Gimlan; Forrest E. Gunnison

[57] **ABSTRACT**

A self-aligned split gate single transistor memory cell structure is formed by a process which self aligns the drain region to one edge of a floating gate. The portion of the channel underneath the floating gate is accurately defined by using one edge of the floating gate to align the drain region. The control gate formed over the floating gate controls the portion of the channel region between the floating gate and the source to provide split gate operation. The source region is formed sufficiently far from the floating gate so that the channel length between the source region and the closest edge of the floating gate is controlled by the control gate but does not have to be accurately defined.

8 Claims, 4 Drawing Sheets

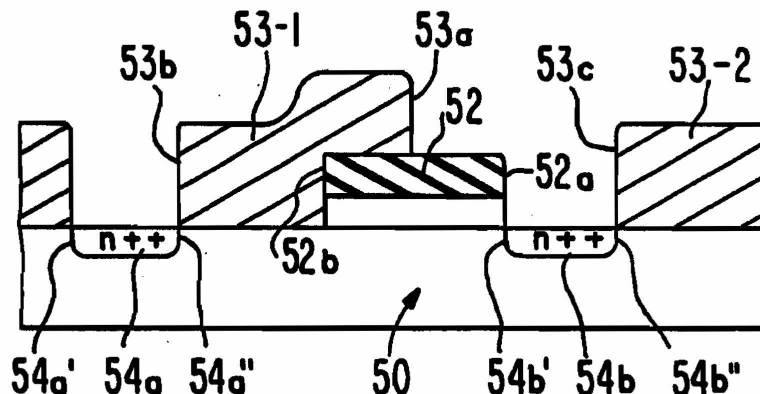


FIG. 1
PRIOR ART

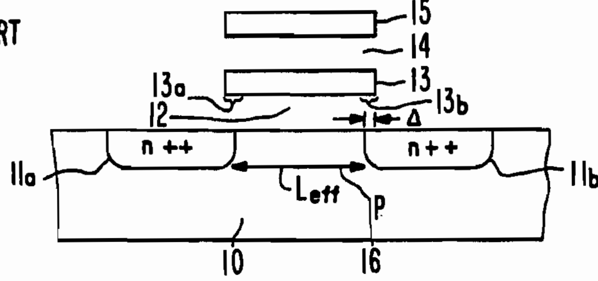


FIG. 2
PRIOR ART

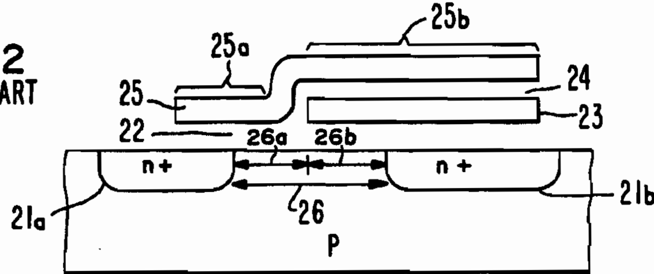


FIG. 3

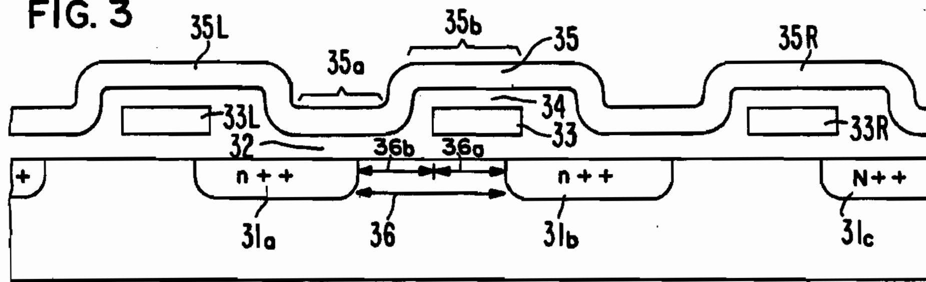


FIG. 4
PRIOR ART

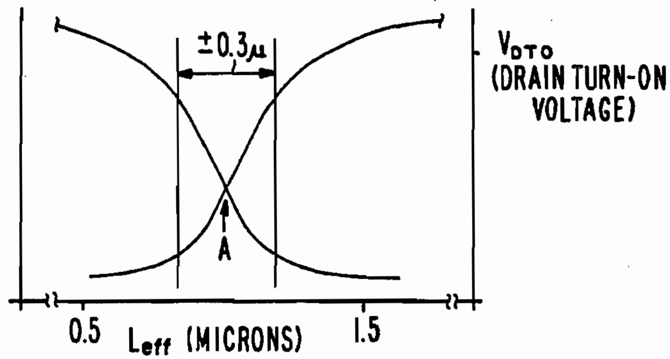


FIG. 5a

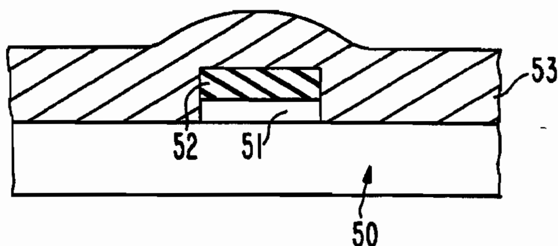
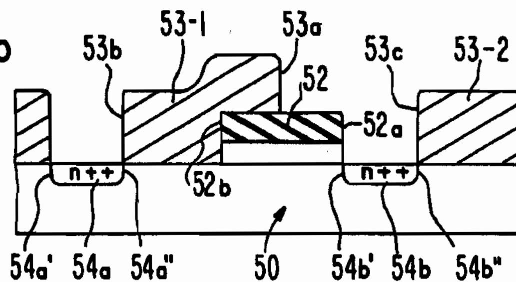


FIG. 5b



4,795,719

FIG. 6a

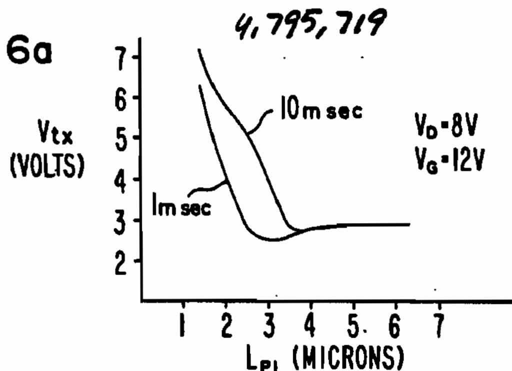


FIG. 6b

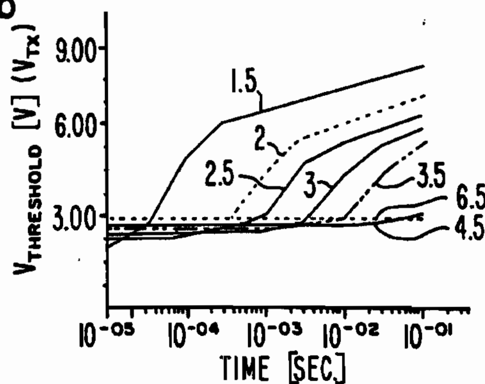


FIG. 6c

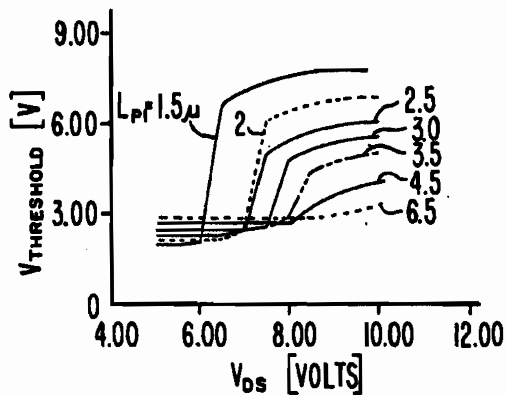


FIG. 6d

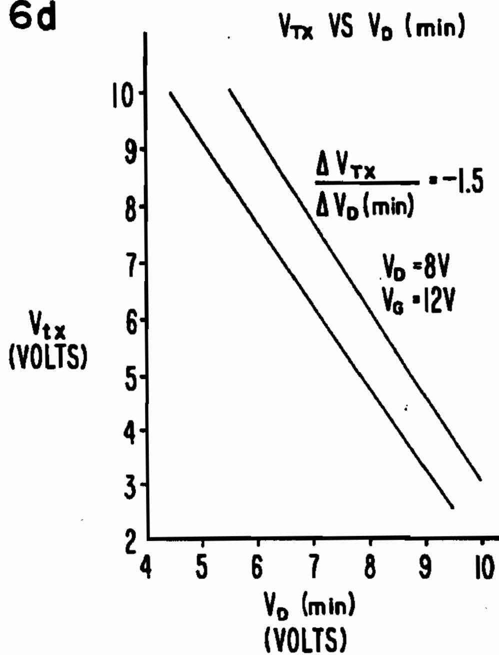


FIG. 7a

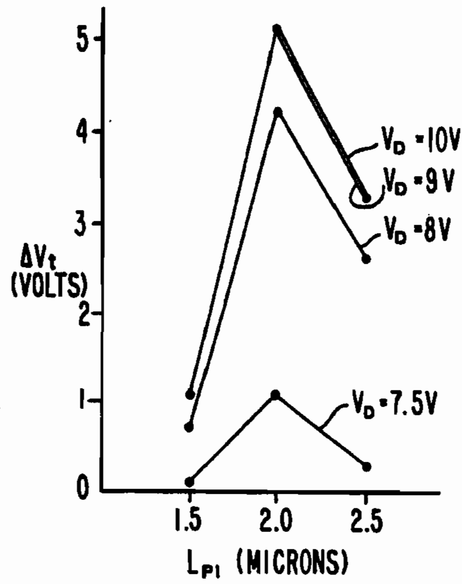


FIG. 7b

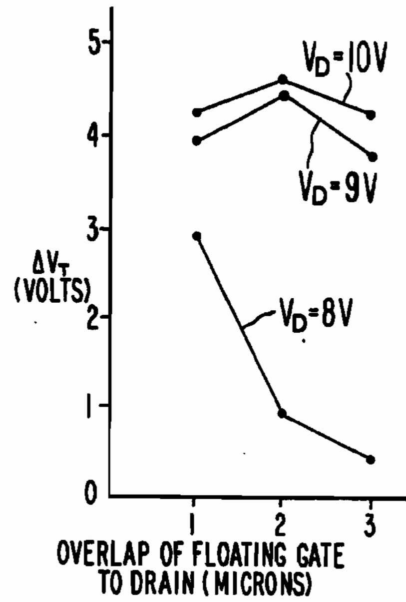
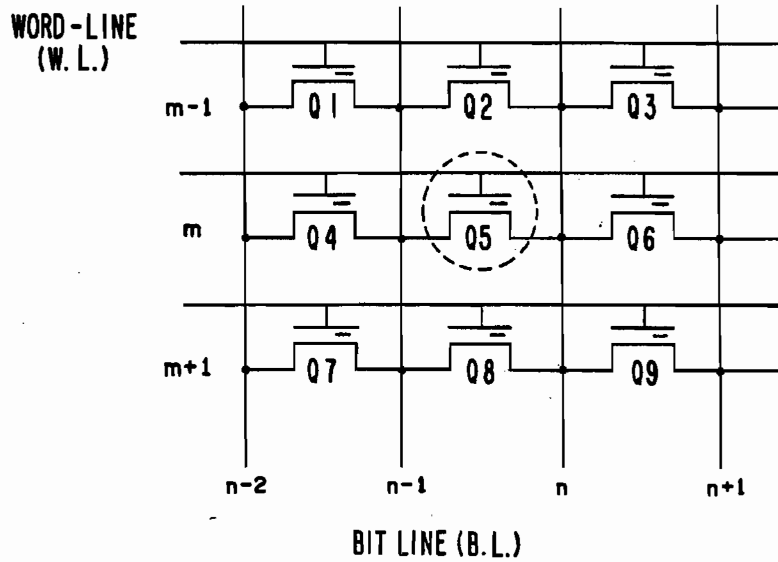


FIG. 8



Case5:10-cv-04241-LHK Document1 Filed09/20/10 Page21 of 75

SELF-ALIGNED SPLIT GATE EPROM PROCESS

This application is a division of application Ser. No. 610,369 filed May 15, 1984, now U.S. Pat. No. 4,639,893.

BACKGROUND OF THE INVENTION

1. Field of the Invention

This invention relates to a nonvolatile EPROM and more particularly to such an EPROM having a split gate (i.e., both a floating gate and a control gate) for controlling the writing and reading of each cell wherein the floating gate is self-aligned with the drain and the channel underlying the floating gate and the control gate is not self-aligned.

2. Prior Art

A split gate nonvolatile EPROM with increased efficiency is disclosed in U.S. Pat. No. 4,328,565 issued May 4, 1982 on an application of Harari, filed Apr. 7, 1980. As disclosed by Harari, the floating gate in an n channel EPOM cell extends over the drain diffusion and over a portion of the channel thereby to form a "drain" capacitance between the drain and the floating gate and a "channel" capacitance between the channel and the floating gate. A control gate then overlaps the floating gate and extends over the remainder of the channel near the source diffusion thereby to form a "control" capacitance between the floating gate and the control gate. These three capacitances form the coupling for driving each cell. The inversion region in the channel directly under the control gate is established directly by a "write or read access" voltage applied to the control gate. The inversion region in the channel directly under the floating gate is established indirectly through the drain and control capacitances and the channel capacitance by the control gate voltage and by another write access voltage applied to the drain. A cell is erased either by ultraviolet illumination or by electrons from the floating gate tunneling through a region of thinned oxide. The non-symmetrical arrangement of the control gate and floating gate with respect to source and drain allows a very dense array implementation. Other split gate structures are disclosed in an article by Barnes, et al. entitled "Operation and Characterization of N-Channel EPROM Cells", published in Solid State Electronics, Vol. 21, pages 521–529 (1978) and an article by Guterman, et al. entitled "An Electrically Alterable Nonvolatile Memory Cell Using a Floating-Gate Structure", published in the IEEE Journal of Solid-State Circuits, Vol. SC-14, No. 2, April 1979.

FIG. 1 illustrates a typical EPROM of the prior art. In FIG. 1 a memory cell comprises n++ source region 11*a* and n++ drain region 11*b* separated by channel region 16. Channel region 16 has an effective length L_{eff} as shown. Overlying channel region 16 is gate dielectric 12 on which is formed a floating gate 13. Typically floating gate 13 is formed of polycrystalline silicon. Overlying floating gate 13 is insulation 14, typically thermally grown silicon dioxide. Control gate 15 is formed above floating gate 13 on insulation 14. The state of the transistor in FIG. 1 is determined by the charge placed on floating gate 13. When electrons are placed on floating gate 13, the threshold voltage V_{tx} required on gate 15 to turn on the transistor (i.e., to form an n channel between source 11*a* and drain 11*b* thereby allowing current to flow from one to the other) is much greater than when no electrons are placed on floating gate 13. As shown in FIG. 1, regions 13*a* and 13*b* of floating gate 13 overlie the source 11*a* and drain 11*b*, respectively, by a small amount "Δ". Consequently, a capacitance is formed between the source 11*a* and floating gate region 13*a* and between the drain 11*b* and floating gate region 13*b*. If the overlap by gate 13 of the source 11*a* and the drain 11*b* is the amount "Δ", then the capacitance C_{pp} between the floating gate 13 and the control gate 15 (both made of polycrystalline silicon) is given by the following equation:

$$C_{pp} \alpha A_{pp} \alpha W(L_{eff}+2\Delta_{FG,D}) \quad (1)$$

In equation 1, C_{pp} is the capacitance between the floating gate 13 and the overlying control gate 15 (this capacitance is proportional to A_{pp}) and A_{pp}, the area of the floating gate 13, is just the width W of the floating gate 13 (perpendicular to the sheet of the drawing) times the length of the floating gate 13 which is (L_{eff} $+2\alpha_{FG,D}$).

The capacitance C_{PROM} between the floating gate 13 and the substrate 10 is proportional to the effective width W_{eff} (i.e. the width perpendicular to the sheet of the paper of the active area underneath the floating gate 13) of the floating gate 13 times L_{eff}. Thus the capacitance C_{PROM} is

$$C_{PROM} \alpha A_{PROM} \alpha W_{eff}(L_{eff}) \quad (2)$$

The capacitive coupling $C_{FG,D}$ of the floating gate 13 to the drain 11*b* is given by

$$C_{FG,D} \alpha A_{FG,D} \alpha W_{eff}(\Delta_{FG,d}) \quad (3)$$

The coupling ratio $CR_{FG,D}$ of the capacitive coupling $C_{FG,D}$ of the floating gate 13 to drain 11*b* to the capacitive coupling C_{pp} of the floating gate 13 to the control gate 15 and the capacitive coupling C_{PROM} of the floating gate 13 to the substrate 10 is

$$CR_{FG,D} \alpha W_{eff}(\Delta_{FG,D})/[W_{eff}(L_{eff})+W(L_{eff}+2\Delta_{FG,D})] \quad (4)$$

As L_{eff} becomes smaller and smaller the impact of the coupling of the drain on the performance of the PROM cell becomes greater and greater until in the limit, as L_{eff} becomes very, very small, this coupling approaches 0.3 (taking into account different oxide thicknesses and the difference between W and W_{eff}, for example). The overlay "Δ" depends on the process and is substantially fixed.

FIG. 2 shows the prior art split gate structure as illustrated by Harari in U.S. Pat. No. 4,328,565 issued May 4, 1982. The major concern in this structure relates to the length of portion 26*b* of channel 26 beneath floating gate 23. The structure of FIG. 2 is a nonself-aligned split gate structure. The total effective channel length 26 is defined by one mask and therefore is constant. Unfortunately, the length of the portion 26*b* of channel 26 beneath the floating gate 23 varies with mask alignment tolerances. Thus the effective channel length 26*b* depends strongly on the alignment process. As a result the best technology available today yields an effective tolerance of channel length 26*b* no better than ±0.5 to ±0.6 microns. For a typical nominal one micron effective channel length 26*b* the actual channel length will vary, due to manufacturing tolerances, over the range of about 1±0.6 micron. The result is a very wide variation in performance from one transistor memory cell to

the next. Programming and read current are both very sensitive to channel length. Good cells will be perfect but bad cells will not work. A good device has an effective channel *26b* (in one embodiment 0.8 microns) which lies between a too-short channel length (for example 0.2 microns or less, so that considering manufacturing variations there may not be any overlap at all of floating gate 23 over the channel 26 and thus there will be no programming of the cell) and a too-long channel length (for example greater than 1.4. microns) with unacceptably slow programming. The major issue in this prior art structure thus is the length of channel portion *26b* (L_{eff}) rather than the coupling. Therefore in a structure such as that shown in FIG. 2 there can be coupling between drain *21b* and floating gate 23 but if the channel length *26b* is not carefully controlled, the memory cell is not going to perform as expected.

A major problem in the prior art EPROM of FIG. 1 relates to the relationship between the program threshold voltage V_{tx} and the drain turn on voltage V_{DTO} of the device. V_{DTO} is the voltage on the drain which, when capacitively coupled to the floating gate 13, turns on the transistor. As shown in FIG. 4, for L_{eff} as shown in FIG. 1 increasing from about 0.5 to 1.2 microns, the program threshold V_{tx} drops below the acceptable program threshold. On the other hand the drain turn-on voltage V_{DTO} becomes as high as the junction breakdown voltage for L_{eff} greater than about one micron. Below one micron, V_{DTO} is very low and may go as low as three to five volts which causes the array of EPROMS to fail. The crossover point is shown as "A" in FIG. 4. In designing a regular EPROM, the crossover point A should be such that V_{tx} is high enough (i.e. greater than five volts) while V_{DTO} is not too low (i.e. not lower than eight volts). However, both curves V_{DTO} and V_{tx} are quite steep at the crossover point A and thus the characteristics of the device are very sensitive to L_{eff}. So if the tolerance on L_{eff} is even ±0.3 microns, which is very good, then the characteristics of the device are still relatively unpredictable. Obviously the desired solution is to eliminate the effect of V_{DTO} and optimize L_{eff} for V_{tx}.

SUMMARY OF THE INVENTION

In accordance with my invention, I overcome the problems of the prior art by providing a memory cell using a split gate structure containing both a control gate and a floating gate in which the floating gate is self-aligned to the drain region. The control gate is not self-aligned. By "self-aligned" I mean that the portion of the transistor channel length under the floating gate will be defined by the floating gate itself regardless of any processing misalignments thereby insuring a constant channel length under the floating gate. To do this, a special process is employed wherein the floating gate is used to define one edge of the drain region. The source region is defined at the same time as the drain region but the alignment of the source region relative to the floating gate is not critical so long as the source region does not underlie and is spaced from the floating gate.

In a process in accordance with this invention, the diffused drain region (which also functions as a bit line and which corresponds to an elongated drain region of the type shown in the above-mentioned '565 patent) is formed using the floating gate to define one edge of the drain region. In the preferred embodiment, the drain and source regions are formed by ion implantation and one edge of the floating gate defines the lateral limit of one side of the drain region. A photoresist material partially extends over the floating gate in one direction and beyond the floating gate in the other direction and the source region is defined by an opening in the portion of this photoresist extending beyond the floating gate in the other direction. The result is to form a precisely defined channel portion L_{eff} of the channel region beneath the floating gate and a remaining relatively imprecisely defined portion of the channel region (to be controlled by a to-be-formed control gate electrode which is part of the word line) underneath the photoresist between the other edge of the floating gate and the source region.

In accordance with my invention, any misalignment between the floating gate and the source region is covered by a to-be-formed control gate and has little effect on the operation of the memory cell while the floating gate is self-aligned to the drain region.

This invention will be understood in more detail in conjunction with the following drawings:

DESCRIPTION OF THE DRAWINGS

FIG. 1 illustrates a prior art EPROM using a single floating gate beneath the control gate;

FIG. 2 illustrates the split gate structure of the prior art wherein the floating gate is not self-aligned to the drain region and the control gate is formed over part of the channel region;

FIG. 3 illustrates the split gate structure of this invention wherein the floating gate is self-aligned to the drain region and overlies but is insulated from an accurately defined portion L_{eff} of the channel region between the source and drain and the control gate overlies the floating gate and that portion of the channel region not overlain by the floating gate but is insulated therefrom;

FIG. 4 illustrates the relationship between threshold voltage V_{tx} and drain turn-on voltage V_{DTO} for the structure of FIG. 1;

FIGS. *5a* and *5b* illustrate the novel process which I use to manufacture the novel self-aligned split gate structure of my invention;

FIGS. *6a* through *6d* illustrate the effect of the channel length L_{P1} under the floating gate on programming;

FIGS. *7a* and *7b* show the tight envelope of operation for the nonself-aligned structure and illustrate graphically the advantages of my self-aligned split gate structure; and

FIG. 8 shows in schematic form a memory array formed using the self-aligned split gate structure of my invention.

DETAILED DESCRIPTION

The following detailed description is meant to be illustrative only and not limiting. Other embodiments of this invention will be obvious to those skilled in the art in view of the following description. In FIGS. *5a* and *5b*, only the cross-section of a single memory cell or a portion thereof is shown while in FIG. 3 two cells and part of a third are shown in cross-section. It should be understood that a semiconductor integrated circuit memory made in accordance with this invention employs a plurality of such cells together with peripheral circuits for writing data into memory and for accessing the data stored in the memory. For simplicity these circuits are not shown.

The starting point for the process of my invention to yield my novel self-aligned split gate structure is the

same as in the nonself-aligned split gate structure of the prior art and in particular of the '565 patent. Thus as shown in FIG. 5*a* a silicon substrate 50 typically having a resistivity of 10–50 ohm-centimeters has formed thereon in a standard manner a layer of gate oxide 51. Gate oxide 51, typically 300 angstroms thick, then has formed on it a first layer of polycrystalline silicon (often called "poly 1") which is patterned as shown in FIG. 5*a* to form a floating gate 52. The oxide 51 beneath the portions of polycrystalline silicon removed to form floating gate 52 is then removed by an etching process (typically a plasma etch) and a photoresist layer 53 is then formed over the top surface of the structure.

As shown in FIG. 5*b* photoresist layer 53 is then patterned so that a particular segment 53-1 of photoresist is formed to partially overlie floating gate 52. Photoresist 53-1 has a right edge 53*a* which is formed to overlie the floating gate 52 somewhere near its middle and a left edge 53*b* which is formed to the left of left edge 52*b* of floating gate 52. The width of floating gate 52 is typically 1.5 to 2 microns and thus it is not difficult to ensure with sufficient certainty given typical tolerances in the manufacturing process that edge 53*a* is to the left of right edge 52*a* of floating gate 52 even for a reasonably expected worst case mask misalignment during the manufacturing process. It is also quite simple to insure that left edge 53*b* is sufficiently to the left of left edge 52*b* of floating gate 52 so that left edge 52*b* of floating gate 52 is never exposed, even in a worst case alignment mismatch of masks during manufacture. Thus the to-be-formed source 54*a* will always be laterally spaced from the left edge 52*b* of floating gate 52.

Following the formation of patterned photoresist 53-1 the structure is subjected to an ion implantation at a selected well-known dosage (typically 4×10^{12} per cm^2) to form n++ drain region 54*b* and n++ source region 54*a* in the top surface of the semiconductor material 50. The region 54*b* has its left edge 54*b*' defined by the right edge 52*a* of floating gate 52 and its right edge 54*b*" defined by the left edge 53*c* of patterned photoresist 53-2. The source region 54*a* has its right edge 54*a*" defined by the left edge 54*b* of patterned photoresist 53-1. Thus the drain region 54*b* is self-aligned to the right edge 52*a* of floating gate 52. However, the right edge 54*a*" of source region 54*a* is self-aligned to the left edge 53*b* of photoresist 53-1. The uncertainty in the location of left edge 53*b* of patterned photoresist 53-1 relative to left edge 52*b* of floating gate 52 represents an uncertainty in the length of the control gate channel (corresponding to channel portion 36*b* in FIG. 3) and not of the floating gate channel L_{eff} (corresponding to channel portion 36*a* in the center cell of FIG. 3). By placing a proper voltage on the to-be-formed control gate, (corresponding to control gate 35 in FIG. 3), the channel length under the control gate becomes irrelevant and the conduction or nonconduction of the total channel is determined by the voltage placed on the floating gate 52 (corresponding to floating gate 33 in FIG. 3). Because floating gate 52 is uniformly coupled to drain 54*b* in all transistors in a memory array made in accordance with this invention and further because the effective channel length L_{eff} (corresponding to channel 36*a* in FIG. 3) is substantially the same underneath all floating gates 52 in all transistors in a memory array formed in accordance with this invention, the structure of this invention yields a split gate programmable EPROM capable of being manufactured with much higher yield than the prior art EPROMs.

The remaining steps in the process are standard well known steps in the silicon gate EPROM technology. Insulation (not shown) is formed over floating gate 52. A control gate (often called "poly 2") corresponding to control gate 35 in FIG. 3 is formed, usually as part of a word line. The resulting structure appears as shown in FIG. 3. FIG. 3 shows floating gate 33 and floating gates 33L and 33R formed to the left and right of floating gate 33. All three floating gates have their right edges self-aligned to the left edges of the underlying drain regions in accordance with this invention. The drain region for a given cell doubles as the source region for the cell to the right.

The finished structure made by the process of this invention as illustrated in FIGS. 5*a* and 5*b* is shown in FIG. 3. In FIG. 3 the floating gate 33 has been formed prior to the formation of the source and drain regions 31*a* and 31*b*. The floating gate 33 is formed on a thin layer of insulation overlying a portion of the to-be-formed channel region between the source and drain. The right edge of the floating gate 33 has been used to define one edge of the drain region 31*b*. Overlying floating gate 33 is insulation 34 (typically silicon oxide) and overlying oxide 34 is the control gate 35. A portion 35*a* of control gate 35 overlies a second portion of the channel region between the left end of the floating gate 33 and the source region 31*a*. As described herein, the channel region 36*b* beneath portion 35*a* of control gate 35 can have a length 36*b* which varies substantially without affecting the performance of the device.

The structure shown in the central part of the cross-section in FIG. 3 is but one cell of a plurality of such cells. In a typical virtual ground structure the drain 36*b* for the cell shown in cross-section in FIG. 3 serves as the source for another cell located just to the right. Likewise the source 36*b* serves as the drain for a second cell located just to the left. The portions of the floating gates 33L and 33R associated with these adjacent cells are shown in FIG. 3.

Note that the floating gate 52 (FIG. 5*a* and 5*b*) becomes capacitively coupled to drain region 54*b* by the lateral diffusion of left edge 54*b*' beneath floating gate 52 during further processing of the structure. This lateral diffusion is typically around 0.3 microns. However contrary to the prior art, the floating gate 52 is formed before the formation of the drain region 54*b*, rather than after, and is precisely self-aligned to one edge of the drain region 54*b*.

FIG. 6*a* illustrates the variation in threshold voltage versus the drain channel length (L_{P1}) of the floating gate ("poly 1"). In FIG. 6*a* the ordinate is the program threshold and the abscissa is the length of the floating gate channel L_{P1} in microns. (Of importance, FIGS. 6*a* through 6*d* and 7*a* and 7*b* use drawn dimensions. However, the channel lengths 36*a* and 36*b* shown in FIG. 3 are the effective dimensions after processing. Thus channel length 36*a* is denoted L_{eff} to represent the effective length of this channel after processing, while before processing this channel length is a drawn dimension and as such is denoted by the symbol L_{P1}. Accordingly, each of the dimensions L_{P1} shown in FIGS. 6*a* through 6*d* and 7*a* and 7*b* must be corrected (i.e., reduced) by a given amount (approximately 0.5 microns), to reflect the effect of processing. Naturally the amount of the correction will vary with the processing.) The threshold voltage V_{tx} obtained or programmed in a given time for a given drain voltage and gate voltage (corresponding in FIG. 6*a* to a drain voltage of 8 volts and a gate

voltage of 12 volts) drops rapidly as the length of the channel L_{P1} under the floating gate 52 (FIG. 5b) increases to a minimum V_{tx} of about 2.5 volts for L_{P1} of somewhere between 3 to 4 microns and then increases slightly. This minimum V_{tx} corresponds to the initial device threshold before programming. The threshold V_{tx} represents the voltage which must be applied to the control gate (such as gate 35 in FIG. 3) to turn on the transistor beneath the control gate as shown in FIG. 3 when the cell containing that transistor has been programmed. Thus as the length of the channel 36a underneath the floating gate 33 increases (FIG. 3) the threshold voltage necessary to turn on the transistor and create a channel from the source region 31b to the drain 31a decreases. As is shown in FIG. 6a, both 1 millisecond and 10 millisecond programming times yield substantially the same shaped curve.

FIG. 6b illustrates the effect of the length of the channel 36a underneath floating gate 33 on the threshold voltage (ordinate) versus programming time (abscissa). The various curves reflect different lengths L_{P1} of the channel 36a (FIG. 3) beneath floating gate 33 in microns. As these channel lengths increase, the threshold voltage for a given programming time drops. Thus for a programming time of 10^{-2} seconds, the threshold voltage for a 1.5 micron channel length L_{P1} is approximately 7 volts whereas the threshold voltage for a 3.0 micron channel L_{P1} is about 4 volts. These curves were obtained for a voltage V_{DS} from the drain to the source of 8 volts and a voltage on the control gate 35 of 12 volts. The curves of FIG. 6b illustrate that the shorter the floating gate the stronger the field which is formed and therefore the greater the number of electrons which are placed on the floating gate thereby resulting in a larger threshold voltage V_{tx} to turn on the transistor.

FIG. 6c is a plot of threshold voltage V_{tx} (ordinate) versus the voltage on the drain 31b (FIG. 3) with the length of channel 36a beneath floating gate 33 as the parameter on the various curves. For a given drain voltage V_D (for example 8 volts) the threshold voltage V_{tx} goes up as the length L_{P1} of the channel 36a beneath floating gate 33 goes down. The curves of FIG. 6c were taken with a control channel L_{P2} (corresponding to the drawn dimension of channel 36b in FIG. 3) beneath the control gate 35 of 2.5 microns, a gate voltage on control gate 35 of 12 volts and a programming time of 10 milliseconds (10^{-2} seconds). These curves illustrate that once a given drain voltage difference V_{DS} is achieved between the drain and the source, increasing the drain voltage beyond a given amount has substantially little effect on the threshold voltage V_{tx} of the transistor. In other words, $\Delta V_{tx}/\Delta V_{DS}$ becomes substantially zero thereby showing that increasing the drain voltage coupled to the floating gate has little effect on the programming of the transistor. Thus after the program threshold voltage V_{tx} is reached, increasing the drain to source voltage V_{DS} does not achieve any significant improvement in performance.

As L_{P1} increases, the threshold voltage V_{tx} at which ΔV_{tx} over ΔV_{DS} becomes very small decreases. So increasing V_{DS} does even less for structures with longer floating gates.

In FIG. 6c each consecutive point on a given line for a given L_{P1} represents an additional 10 milliseconds of programming time rather than just 10 milliseconds of programming time. Accordingly the curves for V_{tx} versus V_{DS} in FIG. 6c would be even flatter than shown in FIG. 6c if a constant programming time was applied to program the cell from different V_{DS} start points.

FIG. 6d illustrates the very tight predictability of threshold voltage V_{tx} versus V_D (min) for the structure of this invention. V_D (min) is defined as the minimum V_{DS} needed to start programming (i.e., to start efficient electron flow onto the floating gate). In FIG. 6c V_D(min) is the V_{DS} at which the curve shows a break point sharply to the right. This break point or "knee" corresponds to the V_D(min) plotted in FIG. 6d.

The relationship of FIG. 6d to FIG. 6c illustrates a basic point of my invention. In a 256K EPROM the time to program the cells in the EPROM theoretically equals 256K times the time to program each cell divided by 8 (ROMs are programmed one byte at a time). Therefore, if the programming time of each cell can be significantly reduced, the efficiency of programming a large number of EPROMs can be proportionally increased. I have discovered that to program to a given threshold voltage V_{tx} in a given programming time, the key is to control the length of L_{P1} and in particular to make this length (which is related to the channel 36a in FIG. 3) as small as practical without generating punch-through from the source to the drain. As shown by analysis of FIG. 6d, the threshold voltage V_{tx} is increased for a given programming time by decreasing V_D(min). As shown in FIG. 6c V_D(min) decreases as L_{P1} decreases in length. Accordingly, decreasing L_{P1} is the key to programming to a given threshold voltage V_{tx} in a given time. My invention not only allows a small effective channel length L_{eff} to be achieved beneath the floating gate but allows this channel length to be achieved in a controllable and reproducible manner throughout an EPROM array thereby to obtain repeatable and consistent results throughout the array.

FIG. 7a illustrates change of threshold voltage, ΔV_T for three different L_{P1} (i.e., three different drawn channel lengths beneath the floating gate) for the structure shown in FIG. 2. In a nonself-aligned structure, the proper length of the channel under the floating gate is crucial to achieve maximum threshold voltage V_{tx}. As shown in FIG. 7a if the channel length 36a becomes too short (for example, 1.5 microns), then punch-through occurs between the source 31a and drain 31b during programming resulting in a failure to program the device. The proper alignment of a floating gate in the nonselfaligned structure to optimize the length of the channel 36a beneath the floating gate 33 and the overlap of the floating gate to drain is crucial. The very sharp peak in FIG. 7a reflects the variation in V_{tx} with channel length L_{P1}. FIG. 7a shows that to optimize the device for the minimum channel length L_{P1} in terms of programming efficiency results in a lower initial threshold before programming and higher final threshold after programming so as to obtain a higher read current. This means a lower impedance in the circuit which in turn means that during read a capacitor in the sense amplifier in the peripheral circuitry of the memory discharges faster through a programmed transistor than otherwise would be the case resulting in shorter access time.

Three effective channels beneath the floating gate (1.5 micron, 2.0 micron and 2.5 micron) are shown in FIG. 7a. The parameter ΔV_T (representing the change in threshold voltage as a function of different channel length) is illustrated by the curves. This change in voltage is particularly pronounced as one goes from 1.5 to 2 to 2.5 micron length for L_{P1}. The change in V_{tx} as a function of channel length is similar to that shown in

FIG. 6*a* for the self-aligned structure of my invention. However, as one goes from a 2 micron L_{P1} to 1.5 micron L_{P1} and shorter, a new phenomenon appears reflecting possible punch through from the source to the drain and V_{tx} thus is lower than would be expected. The nonselfaligned curve shows that a proper L_{P1} is critical to obtaining a predicted threshold voltage. However, with nonself-aligned floating gate technology L_{P1} can vary even across a given chip causing a variation in V_{tx} from cell to cell within a given memory. Often this variation is unacceptable. As can be seen by the curves of FIG. 7*a*, a given memory can have L_{P1} from cell to cell varying for example from 1.5 microns all the way to 2.5 microns or greater because of misalignment in the masking during the processing of the wafer. Accordingly, V_{tx} is unpredictably variable across the wafer often resulting in unacceptable performance.

FIG. 7*b* shows the effect of overlap and V_D on threshold voltage. For the nonself-aligned device the structure must be aligned so that the 3 sigma worst case of alignment gives a satisfactory channel length 36*a* beneath floating gate 33. Increasing the coupling between the floating gate and the drain does not improve the threshold voltage of the device for given programming conditions so overlapping the drain with the floating gate does not help. The more overlap of the floating gate to the drain means the more electrons required to charge the floating gate for a given channel length 36*a* beneath the floating gate. So instead of improving the efficiency of the device, increasing the overlap of the floating gate to the drain actually decreases this efficiency. A minimum overlap of the floating gate to the drain is needed to insure that accelerated electrons hit and lodge in the floating gate rather than in the control gate or the word line.

FIG. 7*b* shows that as the overlap of the nonselfaligned structure increases, the ΔV_T actually declines for a given V_D. Again, this shows that the coupling between the drain and the floating gate is not helpful to achieving a desired V_{tx} and indeed can even be harmful.

The circuit of this invention is highly scaleable and retains its self-aligned character as it is scaled.

An important effect of this invention is that by choosing the correct L_{P1} the programming time for a memory array can be substantially reduced. For example, a prior art 256K EPROM takes approximately 150 seconds or 2½ minutes to program. A 256K EPROM using the structure of this invention can be programmed in approximately 30 seconds. This is a substantial improvement resulting in lower programming costs and lower test costs.

An additional advantage flowing from this invention is that the uncertainty in the location of the floating gate due to mask alignment tolerances is substantially reduced compared to the uncertainty in the location of the floating gate in the prior art nonself-aligned structure and in the standard prior art EPROM (nonsplit gate with self-aligned). Table 1 illustrates this improvement with respect to the self-aligned split gate structure of this invention compared to the standard non-split gate self-aligned structure of the prior art.

TABLE I

	Standard EPROM (Non-split but self-aligned gate)	Self-aligned split gate structure of this invention
STEP 1	Poly 1 (Floating gate) Critical dimension not defined but non-critical dimensions are defined	1 (Floating gate) Critical dimension defined
STEP 2	Poly 2 (Control gate) Define critical dimensions of control gate - Accuracy degraded because of rough, non-planar topology associated with two layers of polycrystaline silicon	
STEP 3	Poly 1 critical dimension defined using Poly 2 as a mask	

Table I compares only the critical steps in the two processes used to define the floating gate and thus the crucial channel length L_{eff}. L_{eff} is the important channel length in the self-aligned split gate structure of this invention and in any EPROM structure. Note that in a standard non-split gate self-aligned structure L_{eff} is the total channel length between the source and drain.

As shown in Table I three steps are required to define the critical dimension of the floating gate in the standard non-split gate self-aligned structure. In the first step only the noncritical dimensions corresponding to the width (but not the length) of the channel beneath the floating gate are defined. The critical dimensions of the floating gate corresponding to the channel length beneath the floating gate are not defined. In step 2 the second layer polycrystalline silicon from which the control gate will be fabricated is deposited. The critical dimension of this second layer (known as "poly 2") is defined in step 2. This dimension corresponds to the channel length between the to-be-formed source and drain regions. However, the accuracy with which the critical dimension of the control gate is fabricated is degraded because of the rough nonplanar topology associated with the two layers of polycrystalline silicon deposited on the wafer. In the third step the first layer of polycrystalline silicon (poly 1) has its critical dimension (corresponding to channel length L_{P1}) defined using the second layer of polycrystalline silicon as a mask. Again, the accuracy with which the critical dimension of the first layer of polycrystalline silicon is defined is degraded due to the uneven topology of the structure.

In contrast, the self-aligned split gate structure of my invention defines the critical dimension of the poly 1 floating gate layer in step 1.

As the above comparison shows, the channel length L_{P1} for the standard nonsplit gate self-aligned structure is equal to the drawn length of the channel plus or minus the uncertainty in the critical dimension associated with the poly 2 definition step plus or minus the uncertainty introduced in the critical dimension of the channel length associated with poly 1 using poly 2 as a mask. Thus the uncertainty in the effective channel length in the standard nonsplit gate self-aligned structure has two components introduced by two critical dimensions. On the other hand, using the self-aligned split gate structure of my invention, only one uncertainty in a critical dimension occurs and that occurs in the first step where the poly 1 critical dimension is defined and the topology is smooth. Accordingly my

invention yields a double processing advantage over the process by which the standard non-split gate self-aligned structure of the prior art is made by eliminating one critical dimension in defining L_{eff} and by introducing a much smoother topology during the formation of the critical channel length L_{eff}.

Table 2 compares the critical steps required to define the poly 1 floating gate in the nonself-aligned split gate structure of the prior art compared to the single step required to define the floating gate in the self-aligned split gate structure of my invention.

TABLE II

	Nonself-aligned split gate	Self-aligned split gate of this invention
STEP 1	Source and Drain Implanted	Poly 1 (Floating Gate) Define critical dimension
STEP 2	Poly 1 (Floating Gate) Define critical dimension	

Step 1 in fabricating the prior art nonself-aligned split gate structure is to implant the source and drain regions in the device. Step 2 is then to deposit the poly 1 layer and then form the floating gate from this layer. The critical dimension L_{P1} is defined by this step. Unfortunately, uncertainty in the length of LP_1 results from the uncertainty in the critical dimension of the poly 1 plus or minus the misalignment of the mask used to define the critical dimension of the floating gate relative to the underlying drain region. Typically the uncertainty in the critical dimension is ±0.3 microns while the uncertainty due to the mask misalignment is ±0.6 microns. When combined in a statistical sense (root mean square) the total uncertainty in L_{P1} can be ±0.6 or ±0.7 microns. To the contrary, using the self-aligned split gate structure of my invention, the critical dimension of the poly 1 floating gate is defined with an uncertainty at most of about ±0.3 microns. Accordingly, my invention achieves a substantial improvement in manufacturing accuracy over the prior art nonself-aligned split gate structure.

FIG. 8 illustrates an EPROM array fabricated using the self-aligned split gate structure of my invention. For simplicity, an array of nine (9) transistors or cells is shown. The programming and reading of cell or transistor Q5 will be described. Note that the array comprises of word line rows m-1, m and m+1 and bit line columns n-2, n-1, n and n+1. Column n-2 is the source of transistors Q1, Q4 and Q7 while column n-1 is the drain of transistors Q1, Q4, and Q7 and the source of transistors Q2, Q5 and Q8. Similarly, column n is the drain of transistors Q2, Q5 and Q8 and the source of transistors Q3, Q6 and Q9. Column n+1 is the drain of transistors Q3, Q6 and Q9.

In operation, to read device m,n (i.e. cell Q5) all bit lines except n-1 are set at 2 volts. Bit line n-1 is set at ground. Word line m is set at 5 volts while all other word lines except m are set at ground.

To program device m,n (i.e., cell Q5) all bit lines except n are set at ground while bit line n is set at 8 or 9 volts. All word lines except m are set at ground while word line m is set at 12 volts. During programming, device m,n+1 (i.e., cell Q6) is also in programming condition but in the reverse configuration (i.e., the high voltage is applied away from the floating gate). In this configuration there is no programming of m, n 1. This asymmetry in the split gate EPROM is what enables one to utilize the virtual ground approach.

While one embodiment of this invention has been described, other embodiments of this invention will be obvious to those skilled in the semiconductor arts in view of this disclosure.

What is claimed is:

1. A method of manufacturing a memory cell containing a split gate transistor comprising:
   - forming first polycrystalline silicon on, but separated from a semiconductor substrate by first insulation, said first polycrystalline silicon defining a floating gate having a first edge and a second edge opposite said first edge;
   - forming a photoresist pattern over said substrate and over a surface of said first polycrystalline silicon, said surface extending laterally between the first and second edges, a first opening being formed in said photoresist pattern to expose both the first edge of said floating gate and a first portion of the semiconductor substrate extending laterally from said first edge and a second opening being formed in said photoresist pattern to expose a second portion of the semiconductor substrate laterally spaced apart from said floating gate;
   - implanting selected impurities into those portions of the semiconductor substrate exposed by the openings of said photoresist thereby to form a source region laterally spaced apart from said floating gate and a drain region extending from but self-aligned to the first edge of said floating gate.

2. The method of claim 1 wherein said drain region has a selected edge self-aligned to the first edge of said floating gate.

3. The method of claim 1 wherein the first opening in the photoresist pattern is patterned to expose a laterally extending surface portion of the floating gate, said surface portion extending from the first edge to a point near the middle of the floating gate.

4. A method according to claim 1 further comprising:
   - forming second polycrystalline silicon to insulatively overlap the floating gate and a channel portion of the substrate located between a portion of the substrate overlapped by the floating gate and the second portion of the substrate, said second polycrystalline silicon defining a control gate of the split gate transistor.

5. A method for manufacturing a split gate transistor having an insulated floating gate overlying a channel region of the transistor and a control gate extending over the floating gate, the method comprising:
   - forming a first insulative layer extending laterally on a semiconductor substrate;
   - forming a first poly layer made of polycrystalline silicon on the first insulative layer;
   - patterning the first poly layer to include opposed first and second edges defining opposed ends of the floating gate;
   - forming a photoresist layer over the first poly layer;
   - patterning the photoresist layer so that the first edge of the first poly layer and a surface portion of the first poly layer extending laterally from the first edge are exposed to define a peripheral portion of a drain implantation window;
   - further patterning the photoresist layer to cover the second edge of the first poly layer and to extend laterally beyond the second edge of the first poly layer to terminate at an edge of the photoresist

layer to thereby define a peripheral portion of a source implantation window, the edge of the photoresist layer being positioned such that the source implantation window is spaced apart from the floating gate; and

implanting impurities through the source and drain implantation windows to form respective source and drain regions of the transistor, the drain region being self aligned thereby to the first edge of the floating gate and the source region being spaced apart from the floating gate.

6. A method according to claim 5 further comprising:

forming a second insulative layer extending laterally over the first poly layer; and

forming a second poly layer made of polycrystalline silicon on the second insulative layer, the second poly layer extending to insulatively overlay a channel portion of the transistor between the source and drain regions, said second poly layer defining a control gate of the split gate transistor.

7. A manufacturing method for assuring consistency over process variations in the effective channel length of a plurality of split gate transistors which are to be formed each to have a floating gate laterally spaced apart from a source region of the transistor, the method comprising:

insulatively disposing the floating gate of each transistor on a semiconductive substrate;

forming a photoresistive coating on the floating gate of each transistor, the coating extending laterally beyond the floating gate to cover the substrate;

creating a first opening in the coating to expose an edge portion of the floating gate of each transistor and a first portion of the substrate directly adjacent to the edge portion;

creating a second opening in the coating, laterally spaced apart from the floating gate, to expose a second portion of the substrate; and

implanting doping impurities through the first and second openings to create for each of the plurality of transistors a drain region which is self-aligned to the edge portion of the floating gate of the transistor and a source region which is spaced apart from the floating gate of the transistor.

8. A method according to claim 7 further comprising:

forming a control line to insulatively overlap the floating gates of each of the transistors and to further insulatively overlap channel portions of each of the transistors between the source and drain regions of the transistors, said control line defining a control gate for each of the transistors.

* * * * *

EXHIBIT B

United States Patent [19]

Eitan

[11] **Patent Number: 4,868,629**

[45] **Date of Patent: Sep. 19, 1989**

[54] **SELF-ALIGNED SPLIT GATE EPROM**

[75] Inventor: **Boaz Eitan, Sunnyvale, Calif.**

[73] Assignee: **WaferScale Integration, Inc., Fremont, Calif.**

[21] Appl. No.: **762,582**

[22] Filed: **Aug. 2, 1985**

Related U.S. Application Data

[63] Continuation-in-part of Ser. No. 610,369, May 15, 1984.

[51] Int. Cl.[4] H01L 27/10

[52] U.S. Cl. 357/45; 357/41; 357/23.5; 357/23.9; 365/185

[58] Field of Search 357/23.5, 23.9, 41, 357/45; 365/185

[56] **References Cited**

U.S. PATENT DOCUMENTS

4,122,544	10/1978	McElroy	357/23.5
4,142,926	3/1979	Morgan	357/23.5
4,173,791	11/1979	Bell	357/23.5
4,173,818	11/1979	Bassous et al.	357/23.5
4,257,832	3/1981	Schwabe et al.	357/23.5
4,267,558	5/1981	Guterman	357/23.5
4,274,012	6/1981	Simko	357/23.5
4,297,719	10/1981	Hsu .	
4,300,212	11/1981	Simko	357/23.5
4,318,216	3/1982	Hsu	357/23.5
4,328,565	5/1982	Harari	357/23.5
4,334,292	6/1982	Kotecha	357/23.5
4,336,603	6/1982	Kotecha et al.	357/23.5
4,412,311	10/1983	Miccoli et al.	357/23.5
4,426,764	1/1984	Kosa et al.	357/23.5
4,462,090	7/1984	Iizuka	365/185
4,471,373	9/1984	Shimizu et al.	357/23.5
4,495,693	1/1985	Iwahashi et al.	357/23.5
4,561,004	12/1985	Kuo et al.	357/41

FOREIGN PATENT DOCUMENTS

816931	7/1969	Canada	357/23.5
0045578	2/1982	European Pat. Off.	357/23.5
1647781	5/1985	European Pat. Off. .	
0158078	12/1982	Fed. Rep. of Germany	357/23.5
2437676	9/1979	France	357/23.5
0063684	4/1977	Japan	357/23.5
0089686	7/1978	Japan	357/23.5
54-156484	12/1979	Japan	357/23.5
0156369	12/1980	Japan	357/23.5
0071971	6/1981	Japan	357/23.5
0076878	5/1982	Japan	357/23.5
57-96572	6/1982	Japan	357/23.5
0206165	12/1983	Japan	357/23.5
2073484	10/1981	United Kingdom	357/23.5

OTHER PUBLICATIONS

Article entitled "High Density Flash EEPROMs Are About To Burst On The Market", pp. 47 and 48, *Electronics*, Mar, 3, 1988.

Shirota, Paper entitled "A New NAND Cell for Ultra High Density 5V only EPROM".

IEEE Transactions on Electron Devices, vol. ED-32, No. 5, 5/85 "Lightly Doped Drain Transistors For Advanced VLSI Circuits", pp. 896 et seq.

IEEE Trans. on EL.DV., vol. ED-29, No. 4, 4/82, By takeda et al, pp. 611 et seq., "Semiconductor MOSFET Structure For Minimizing Hot-Carrier Generation".

IEEE Trans. on EL. DV., vol. ED-32, No. 3, 3/85, By Koyanagi et al, pp. 562 et seq., "Optimum Design of n+-n−Double-Diffused Drain MOSFET To Reduce Hot-Carrier Emission".

Primary Examiner—Martin H. Edlow
Attorney, Agent, or Firm—Skjerven, Morrill, MacPherson, Franklin & Friel

[57] **ABSTRACT**

A self-aligned split gate single transistor memory cell structure is formed by a process which self aligns the drain region to one edge of a floating gate. The portion of the channel underneath the floating gate is accurately defined by using one edge of the floating gate to align the drain region. The control gate formed over the floating gate controls the portion of the channel region between the floating gate and the source to provide split gate operation. The source region is formed sufficiently far from the floating gate so that the channel length between the source region and the closest edge of the floating gate is controlled by the control gate but does not have to be accurately defined.

8 Claims, 7 Drawing Sheets

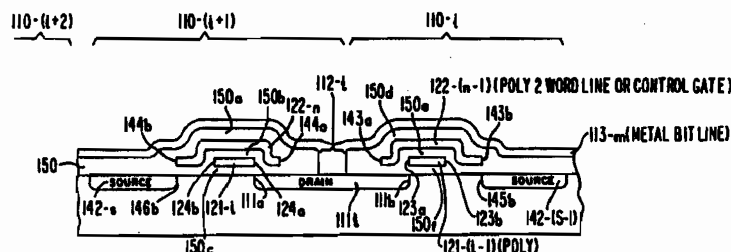


FIG. 1
PRIOR ART

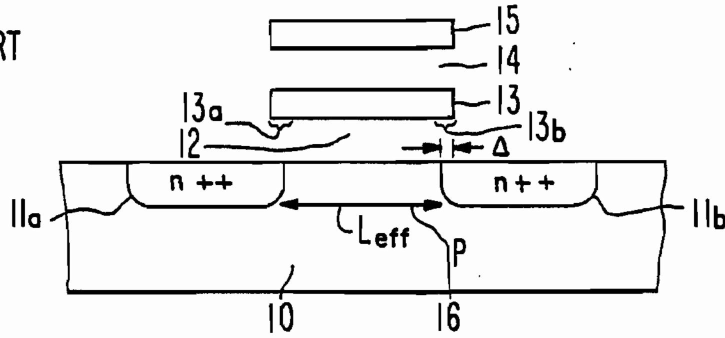


FIG. 2
PRIOR ART

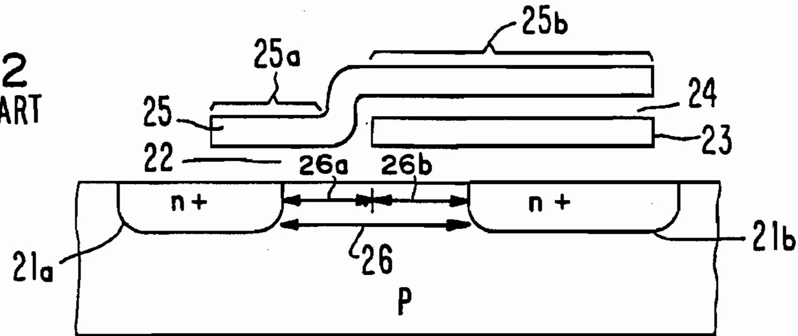


FIG. 3

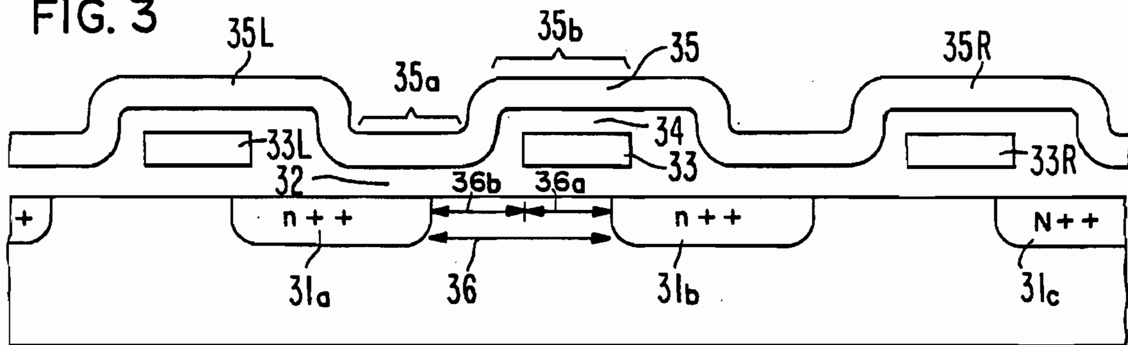


FIG. 4
PRIOR ART

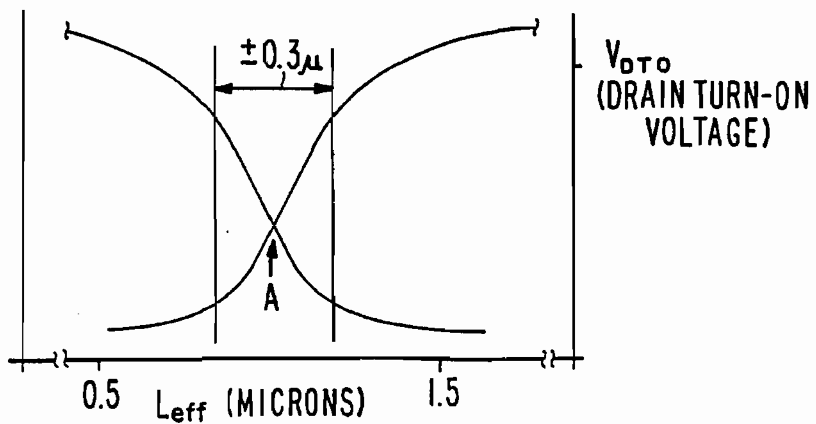


FIG. 5a

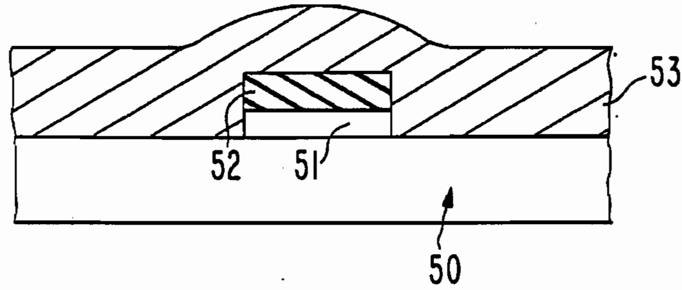


FIG. 5b

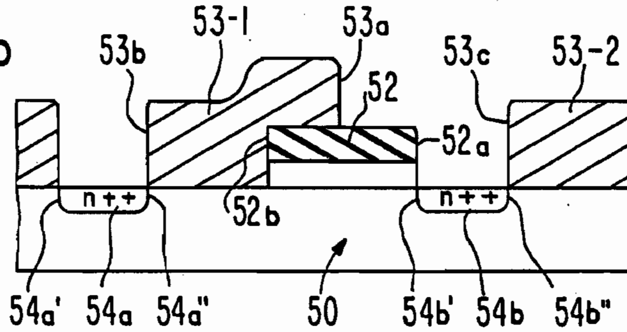


FIG. 6a

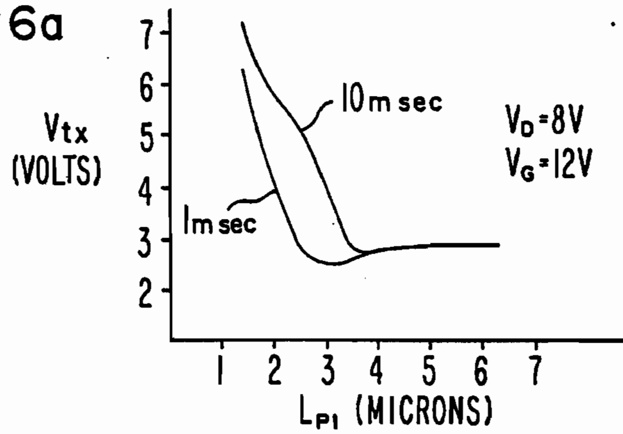


FIG. 6b

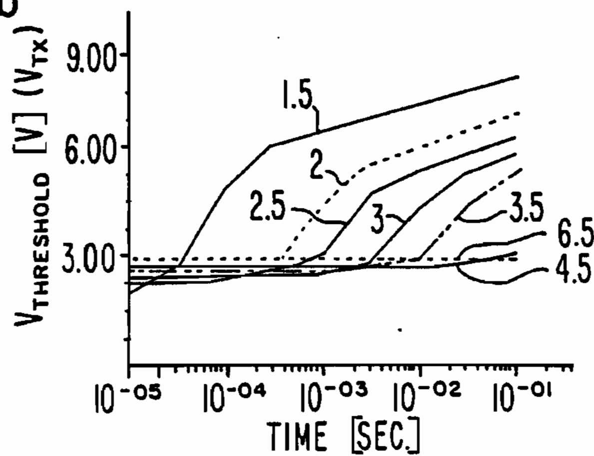


FIG. 6c

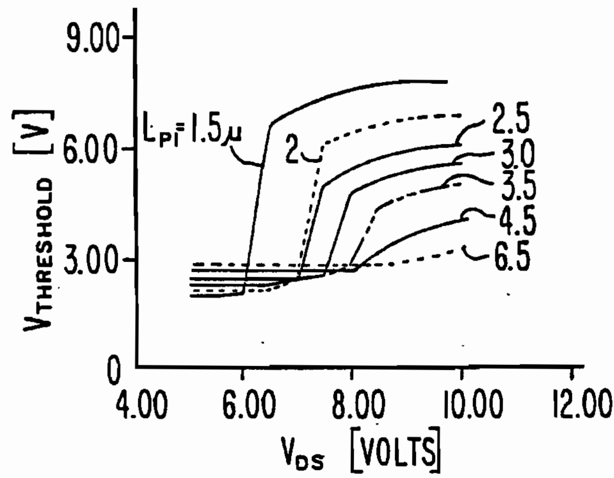


FIG. 6d

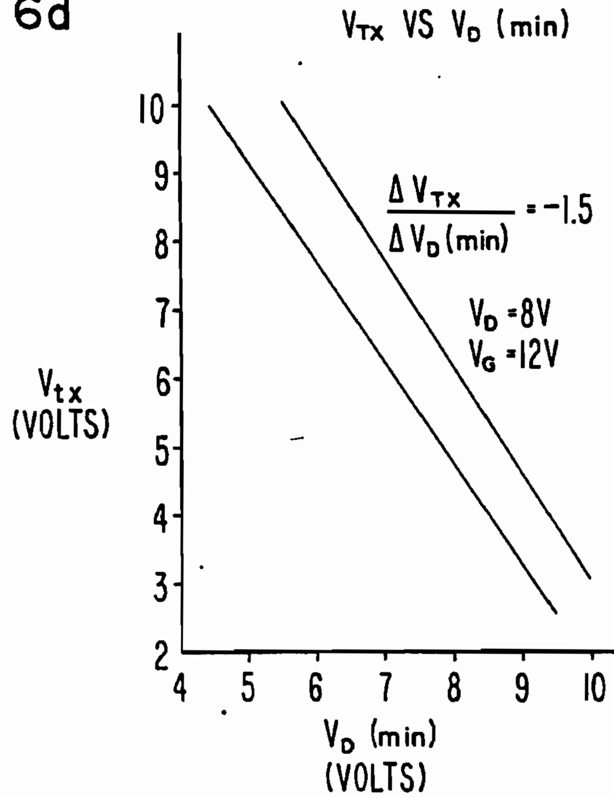


FIG. 7a

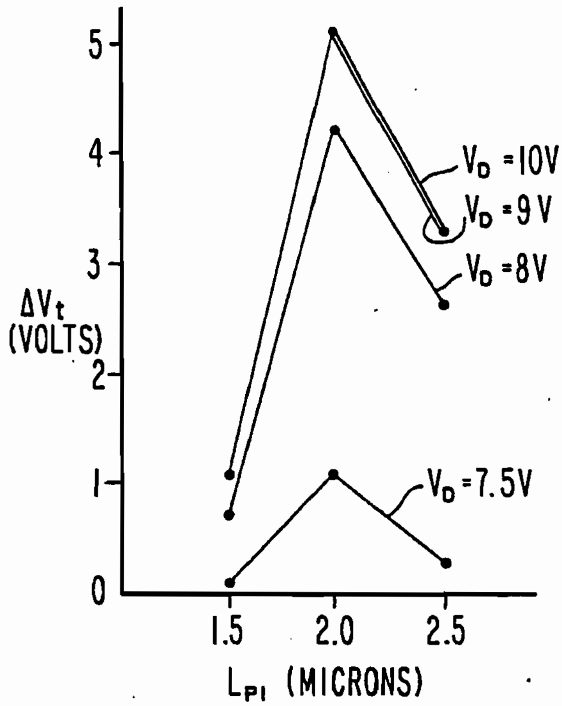


FIG. 7b

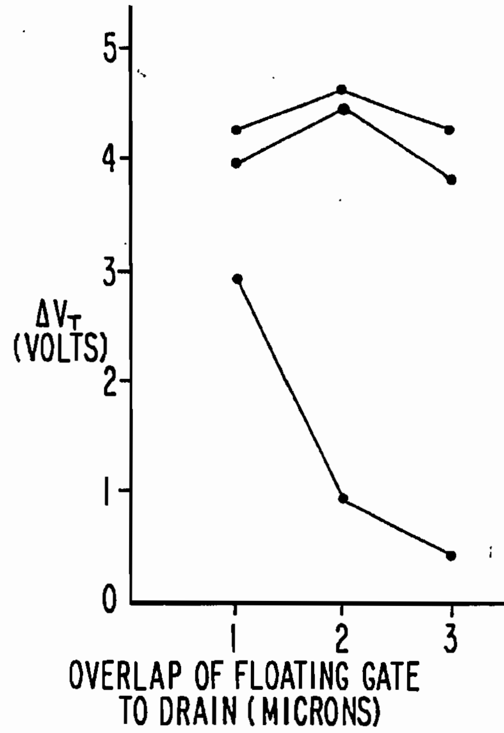
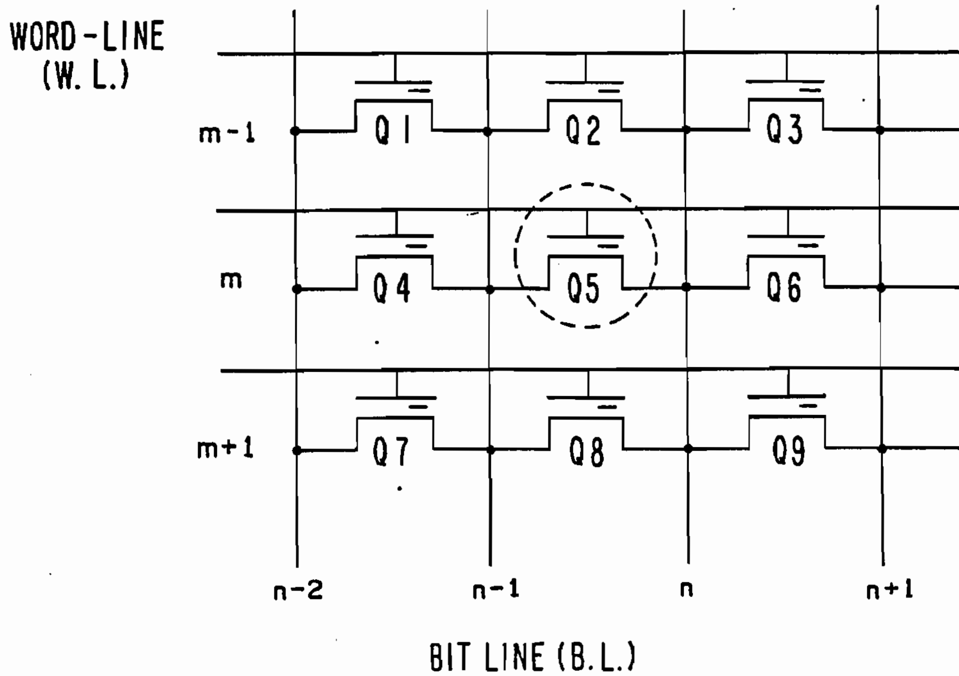


FIG. 8



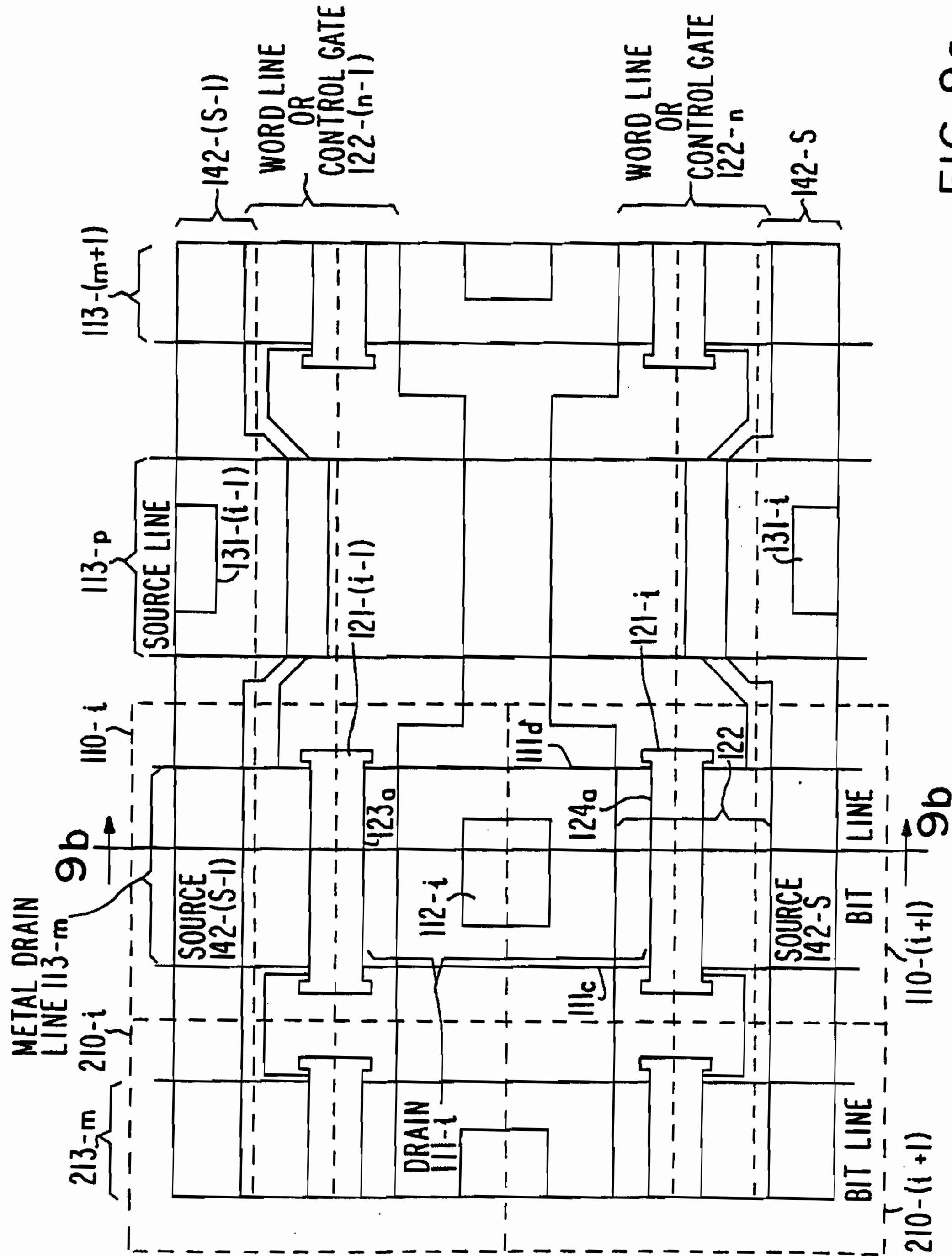


FIG. 9a

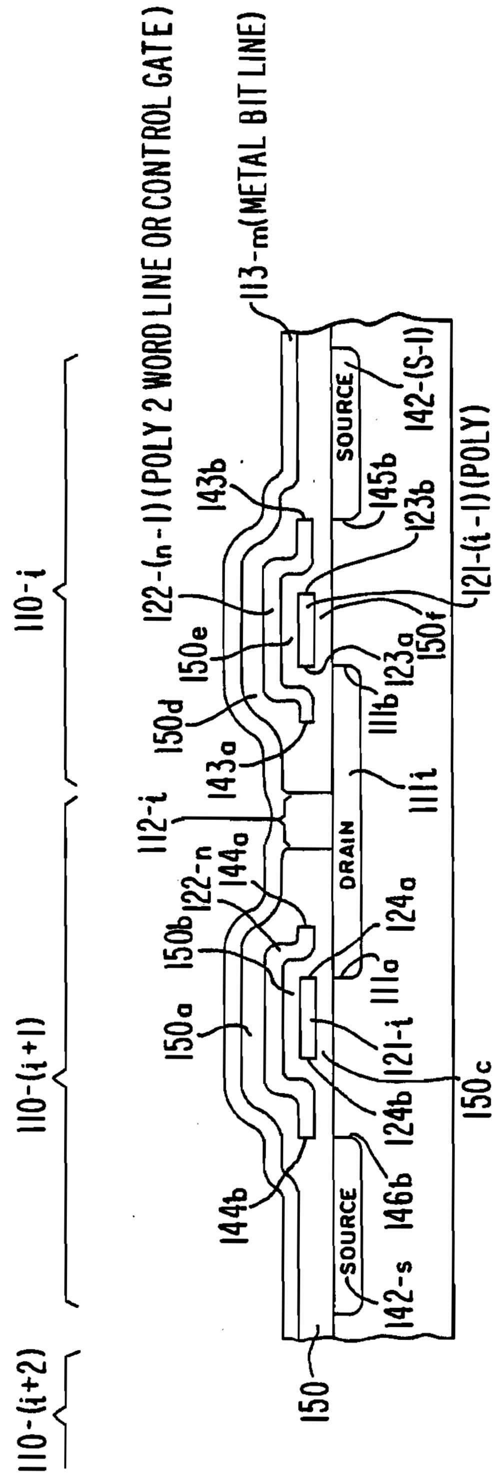


FIG. 9b

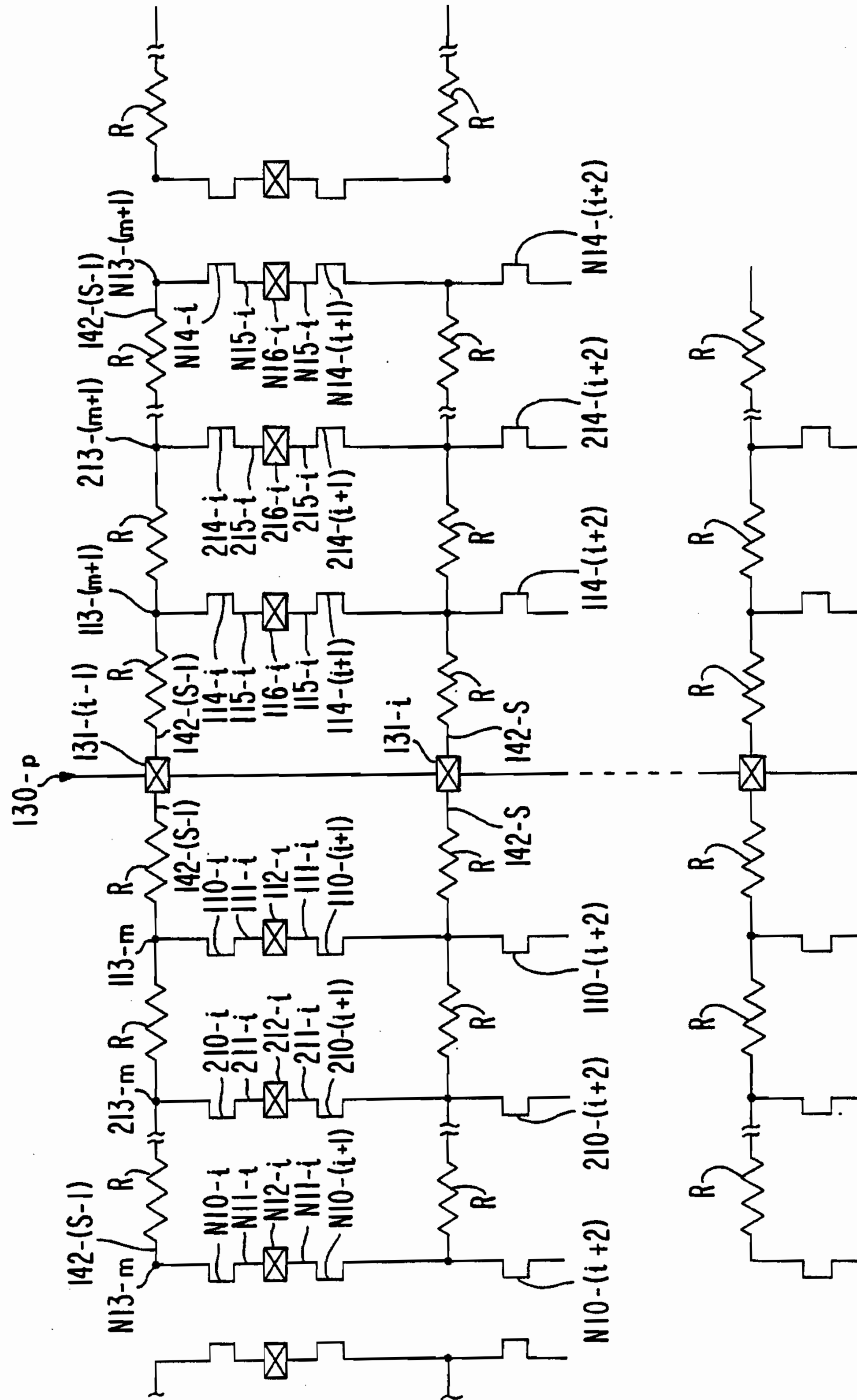


FIG. 9c

SELF-ALIGNED SPLIT GATE EPROM

RELATED APPLICATION

This application is a continuation-in-part of application Ser. No. 06/610,369, filed May 15, 1984, entitled "A SELF-ALIGNED SPLIT GATE EPROM", which application is assigned to Wafer Scale Integration, Inc. the assignee of this case.

BACKGROUND OF THE INVENTION

1. Field of the Invention

This invention relates to a nonvolatile EPROM and more particularly to such an EPROM having a split gate (i.e., both a floating gate and a control gate) for controlling the writing and reading of each cell wherein the floating gate is self-aligned with the drain and the channel underlying the floating gate and the control gate is not self-aligned.

2. Prior Art

A split gate nonvolatile EPROM with increased efficiency 1B is disclosed in U.S. Pat No. 4,328,565 issued May 4, 1982 on an application of Harari, filed Apr. 7, 1980. As disclosed by Harari, the floating gate in an n channel EPROM cell extends over the drain diffusion and over a portion of the channel thereby to form a "drain" capacitance between 23 the drain and the floating gate and a "channel" capacitance between the channel and the floating gate. A control gate then overlaps the floating gate and extends over the remainder of the channel near the source diffusion thereby to form a "control" capacitance between the floating gate and the control gate. These three capacitances form the coupling for driving each cell. The inversion region in the channel directly under the control gate is established directly by a "write or read access" voltage applied to the control gate. The inversion region in the channel directly under the floating gate is established indirectly through the drain and control capacitances and the channel capacitance by the control gate voltage and by another write access voltage applied to the drain. A cell is erased either by ultraviolet illumination or by electrons from the floating gate tunneling through a region of thinned oxide. The nonsymmetrical arrangement of the control gate and floating gate with respect to source and drain allows a very dense array implementation. Other split gate structures are disclosed in an article by Barnes, et al. entitled "Operation and Characterization of N-Channel EPROM Cells", published in Solid State Electronics, Vol. 21, pages 521-529 B (1978) and an article by Guterman, et al. entitled "An Electrically Alterable Nonvolatile Memory Cell Using a Floating-Gate Structure", published in the IEEE Journal of Solid-State Circuits, Vol. SC-14, No. 2, April 1979.

FIG. 1 illustrates a typical EPROM of the prior art. In FIG. 1 a memory cell comprises n source region 11*a* and n++ drain region 11*b* separated by channel region 16. Channel region 16 has an effective length L_{eff} as shown. Overlying channel region 16 is gate dielectric 12 on which is formed a floating gate 13. Typically floating gate 13 is formed of polycrystalline silicon. Overlying floating gate 13 is insulation 14, typically thermally grown silicon dioxide. Control gate 15 is formed above floating gate 13 on insulation 14. The state of the transistor in FIG. 1 is determined by the charge placed on floating gate 13 When electrons are placed on floating gate 13, the threshold voltage V_{tx} required on gate 15 to turn on the transistor (i.e., to form an n channel between

source 11*a* and drain 11*b* thereby allowing current to flow from one to the other) is much greater than when no electrons are placed on floating gate 13. As shown in FIG. 1, regions 13*a* and 13*b* on floating gate 13 overlie the source 11*a* and drain 11*b*, respectively, by a small amount "Δ". Consequently, a capacitance is formed between the source 11*a* and floating gate region 13*a* and between the drain 11*b* and floating gate region 13*b*. If the overlap by gate 13 of the source 11*a* drain llb is the amount "Δ", then the capacitance C_{pp} between the floating gate 13 and the control gate 15 (both made of polycrystalline silicon) is given by the following equation:

$$C_{pp} \alpha A_{pp} \alpha W(L_{eff}+2\Delta_{FG,D}) \quad (1)$$

In equation 1, Cpp is the capacitance between the floating gate 13 and the overlying control gate 15 (this capacitance is proportional to A_{pp}) and A_{pp}, the area of the floating gate 13, is just the width W of the floating gate B 13 (perpendicular to the sheet of the drawing) times the length of the floating gate 13 which is (L_{eff} $2\Delta_{FG,D}$)

The capacitance C_{PROM} between the floating gate 13 and the substrate 10 is proportional to the effective width W_{eff} (i.e. the width perpendicular to the sheet of the paper of the active area underneath the floating gate 13) of the floating gate 13 times L_{eff}. Thus the capacitance C_{PROM} is

$$C_{PROM} \alpha A_{PROM} \alpha W_{eff}(L_{eff}) \quad (2)$$

The capacitive coupling $C_{FG,D}$ of the floating gate 13 to the drain 11*b* is given by

$$C_{FG,D} \alpha A_{FG,D} \alpha W_{eff}(\Delta_{FG,D}) \quad (3)$$

The coupling ratio $CR_{FG,D}$ of the capacitive coupling $C_{FG,D}$ of the floating gate 13 to drain 11*b* to the capacitive coupling C_{pp} of the floating gate 13 to the control gate 15 and the capacitive coupling C_{PROM} of the floating gate 13 to the substrate 10 is

$$CR_{FG,D} \alpha W_{eff}(\Delta_{FG,D})/[W_{eff}(L_{eff})+W(L_{eff}+2\Delta_{FG,D})] \quad (4)$$

As L_{eff} becomes smaller and smaller the impact of the coupling of the drain on the performance of the PROM cell becomes greater and greater until in the limit, as L_{eff} becomes very, very small, this coupling approaches 0.3 (taking into account different oxide thicknesses and the difference between W and W_{eff}, for example). The overlay "Δ" depends on the process and is substantially fixed.

FIG. 2 shows the prior art split gate structure as illustrated by Harari in U.S. Pat. No. 4,328,565 issued May 4, 1982. The major concern in this structure relates to the length of portion 26*b* of channel 26 beneath floating gate 23 The structure of FIG. 2 is a nonself-aligned split gate structure. The total effective channel length 26 is defined by one mask and therefore is constant. Unfortunately, the length of the portion 26*b* of channel 26 beneath the floating gate 23 varies with mask alignment tolerances. Thus the effective channel length 26*b* depends strongly on the alignment process. As a result the best technology available today yields an effective

channel length tolerance 26*b* no better than ±0.5 to ±0.6 microns. For a typical nominal one micron effective channel length 26*b* the actual channel length will vary, due to manufacturing tolerances, over the range of about 1±0.6 micron. The result is a very wide variation in performance from one transistor memory cell to the next. Programming and read current are both very sensitive to channel length. Good cells will be perfect but bad cells will not work. A good device has an effective channel 26*b* (in one embodiment 0.8 microns) which lies between a too-short channel length (for example 0.2 microns or less, so that considering manufacturing variations there may not be any overlap at all of floating gate 23 over the channel 26 and thus there will be no programming of the cell) and a too-long channel length (for example greater than 1.4 microns) with unacceptably slow programming. The major issue in this prior art structure thus is the length of channel portion 26*b* (L_{eff}) rather than the coupling. Therefore in a structure such as that shown in FIG. 2 there can be coupling between drain 21*b* and floating gate 23 but if the channel length 26*b* is not carefully controlled, the memory cell is not going to perform as expected.

A major problem in the prior art EPROM of FIG. 1 relates to the relationship between the program threshold voltage V_{tx} and the drain turn on voltage V_{DTO} of the device. V_{DTO} is the voltage on the drain which, when capacitively coupled to the floating gate 13, turns on the transistor. As shown in FIG. 4, for L_{eff} as shown in FIG. 1 increasing from about 0.5 to 1.2 microns, the program threshold V_{tx} drops below the acceptable program threshold. On the other hand the drain turn-on voltage V_{DTO} becomes as high as the junction breakdown voltage for L_{eff} greater than about one micron. Below one micron, V_{DTO} is very low and may go as low as three to five volts which causes the array of EPROMS to fail. The crossover point is shown as "A" in FIG. 4. In designing a regular EPROM, the crossover point A should be such that V_{tx} is high enough (i.e. greater than five volts) while VDTO is not too low (i.e. not lower than eight volts). However, both curves V_{DTO} and V_{tx} are quite steep at the crossover point A and thus the characteristics of the device are very sensitive to L_{eff}. So if the tolerance on Leff is even ±0.3 microns, which is very good, then the characteristics of the device are still relatively unpredictable. Obviously the desired solution is to eliminate the effect of V_{DTO} and optimize L_{eff} for V_{tx}.

SUMMARY OF THE INVENTION

In accordance with my invention, I overcome the problems of the prior art by providing a memory cell using a split gate structure containing both a control gate and a floating gate in which the floating gate is self-aligned to the drain region. The control gate is not self-aligned. "Self-aligned" means here that the portion of the transistor channel length under the floating gate will be defined by the floating gate itself regardless of any processing misalignments thereby insuring a constant channel length under the floating gate. To do this, a special process is employed wherein the floating gate is used to define one edge of the drain region. The source region is defined at the same time as the drain region but the alignment of the source region relative to the floating gate is not critical so long as the source region does not underlie and is spaced from the floating gate.

In a process in accordance with this invention, the diffused drain region (which also functions as a bit line and which corresponds to an elongated drain region of the type shown in the above-mentioned '565 patent) is formed using the floating gate to define one edge of the drain region. In the preferred embodiment, the drain and source regions are formed by ion implantation and one edge of the floating gate defines the lateral limit of one side of the drain region. A photoresist material partially extends over the floating gate in one direction and beyond the floating gate in the other direction and the source region is defined by an opening in the portion of this photoresist extending beyond the floating gate in the other direction. The result is to form a precisely defined channel portion L_{eff} of the channel region beneath the floating gate and a remaining relatively imprecisely defined portion of the channel region (to be controlled by a to-be-formed control gate electrode which is part of the word line) underneath the photoresist between the other edge of the floating gate and the source region.

In accordance with my invention, any misalignment between the floating gate and the source region is covered by a to-be-formed control gate and has little effect on the operation of the memory cell while the floating gate is self-aligned to the drain region.

This invention will be understood in more detail in conjunction with he following drawings:

DESCRIPTION OF THE DRAWINGS

FIG. 1 illustrates a prior art EPROM using a single floating gate beneath the control gate:

FIG. 2 illustrates the split gate structure of the prior art wherein the floating gate is not self-aligned to the drain region and the control gate is formed over part of the channel region:

FIG. 3 illustrates the split gate structure of this invention wherein the floating gate is self-aligned to the drain region and overlies but is insulated from an accurately defined portion L_{eff} of the channel region between the source and drain and the control gate overlies the floating gate and that portion of the channel region not overlain by the floating gate but is insulated therefrom;

FIG. 4 illustrates the relationship between threshold voltage V_{tx} and drain turn-on voltage V_{DTO} for the structure of FIG. 1;

FIGS. 5*a* and 5*b* illustrate the novel process which I use to manufacture the novel self-aligned split gate structure of my invention;

FIGS. 6*a* through 6*d* illustrate the effect of the channel length L_{P1} under the floating gate on programming:

FIGS. 7*a* and 7*b* show the tight envelope of operation for the nonself-aligned structure and illustrate graphically the advantages of my self-aligned split gate structure: and

FIG. 8 shows in schematic form a memory array formed using the self-aligned split gate structure of my invention.

FIG. 9*a* shows the layout of a portion of a novel, high-speed EPROM incorporating the self-aligned split-gate structure of this invention;

FIG. 9*b* shows in cross-section two transistors from the structure of FIG. 9*a;* and

FIG. 9*c* shows schematically the architecture of one embodiment of the EPROM array portions of which are shown in FIGS. 9*a* and 9*b*.

DETAILED DESCRIPTION

The following detailed description is meant to be illustrative only and not limiting. Other embodiments of this invention will be obvious to those skilled in the art in view of the following description. In FIGS. 5a and 5b, only the cross-section of a single memory cell or a portion thereof is shown while in FIG. 3 two cells and part of a third are shown in cross-section. It should be understood that a semiconductor integrated circuit memory made in accordance with this invention employs a plurality of such cells together with peripheral circuits for writing data into memory and for accessing the data stored in the memory. For simplicity these circuits are not shown

The starting point for the process of my invention to yield my novel self-aligned split gate structure is the same as in the nonself-aligned split gate structure of the prior art and in particular of the '565 patent. Thus as shown in FIG. 5a a silicon substrate 50 typically having a resistivity of 10-50 ohm-centimeters has formed thereon in a standard manner a layer of gate oxide 51. Gate oxide 51, typically 300 angstroms thick, then has formed on it a first layer of polycrystalline silicon (often called "poly 1") which is patterned as shown in FIG. 5a to to form floating gate 52 a floating gate 52. The oxide 51 beneath the portions of polycrystalline silicon removed to form floating gate 52 is then removed by an etching process (typically a plasma etch) and a photoresist layer 53 is then formed over the top surface of the structure.

As shown in FIG. 5b, photoresistor layer 53 is then patterned so that a particular segment 53-1 of photoresist is formed to partially overlie floating gate 52. Photoresist segment 53-1 has a right edge 53a which is formed to overlie the floating gate 52 somewhere near its middle and a left edge 53b which is formed to the left of left edge 52b of floating gate 52. The length of floating gate 52 is typically 1.5 to 2 microns and thus it is not difficult to ensure with sufficient certainty given typical tolerances in the manufacturing process that edge 53a is to the left of right edge 52a of floating gate 52 even for a reasonably expected worst case mask misalignment during the manufacturing process. It is also quite simple to insure that left edge 53b is sufficiently to the left of left edge 52b of floating gate 52 so that left edge 52b of floating gate 52 is never exposed, even in a worst case alignment mismatch of masks during manufacture. Thus the to-be-formed source 54a will always be laterally spaced from the left edge 52b for floating gate 52.

Following the formation of patterned photoresist segment 53-1 the structure is subjected to an ion implantation at a selected well-known dosage (typically 4×10^{12} per cm^2) to form n++ drain region 54b and n. source region 54a in the top surface of the semiconductor material 50. The region 54b has its left edge 54b' defined by the right edge 52a of floating gate 52 and its right edge 54b" defined by the left edge 53c of patterned photoresist 53-2. The source region 54a has its right edge 54a" defined by the left edge 53b of patterned photoresist segment 53-1 . Thus the drain region 54b is self-aligned to the right edge 52a of floating gate 52. However, the right edge 54a" of source region 54a is self-aligned to the left edge 53b of photoresist segment 53-1. The uncertainty in the location of left edge 53b of patterned photoresist segment 53-1 relative to left edge 52b of floating gate 52 represents an uncertainty in the length of the control gate channel (corresponding to channel portion 36b in FIG. 3) and not of the floating gate channel L_{eff} (corresponding to channel portion 36a in the center cell of FIG. 3). By placing a proper voltage on the to-be-formed control gate, (corresponding to control gate 35 in FIG. 3), the channel length under the control gate becomes irrelevant and the conduction or nonconduction of the total channel is determined by the voltage placed on the floating gate 52 (corresponding to floating gate 33 in FIG. 3). Because floating gate 52 is uniformly coupled to drain 54b in all transistors in a memory array made in accordance with this invention and further because the effective channel length L_{eff} (corresponding to channel 36a in FIG. 3) is substantially the same underneath all floating gates 52 in all transistors in a memory array formed in accordance with this invention, the structure of this invention yields a split gate programmable EPROM capable of being manufactured with much higher yield than the prior art EPROMs.

The remaining steps in the process are standard well known steps in the silicon gate EPROM technology. Insulation (not shown) is formed over floating gate 52. A control gate (often called "poly 2") corresponding to control gate 35 in FIG. 3 is formed, usually as part of a word line. The resulting structure appears as shown in FIG. 3. FIG. 3 shows floating gate 33 and floating gates 33L and 33R formed to the left and right of floating gate 33. All three floating gates have their right edges self-aligned to the left edges of the underlying drain regions in accordance with this invention. The drain region for a given cell doubles as the source region for the cell to the right.

The finished structure made by the process of this invention as illustrated in FIGS. 5a and 5b is shown in FIG. 3. In FIG. 3 the floating gate 33 has been formed prior to the formation of the source and drain regions 31a and 31b. The floating gate 33 is formed on a thin layer of insulation overlying a portion of the to-be-formed channel region between the source and drain. The right edge of the floating gate 33 has been used to define one edge of the drain region 31b. Overlying floating gate 33 is insulation 34 (typically silicon oxide) and overlying oxide 34 is the control gate 35. A portion 35a of control gate 35 overlies a second portion of the channel region between the left end of the floating gate 33 and the source region 31a. As described herein, the channel region 36b beneath portion 35a of control gate 35 can have a length 36b which varies substantially without affecting the performance of the device.

The structure shown in the central part of the cross-section in FIG. 3 is but one cell of a plurality of such cells. In a typical virtual ground structure the drain 31b for the cell shown in cross-section in FIG. 3 serves as the source for another cell located just to the right. Likewise the source 31a serves as the drain for a second cell located just to the left. The portions of the floating gates 33L and 33R associated with these adjacent cells are shown in FIG. 3.

Note that the floating gate 52 (FIGS. 5a and 5b) becomes capacitively coupled to drain region 54b by the lateral diffusion of left edge 54b' beneath floating gate 52 during further processing of the structure. This lateral diffusion is typically around 0.3 microns. However contrary to the prior art, the floating gate 52 is formed before the formation of the drain region 54b, rather than after, and is precisely self-aligned to one edge of the drain region 54b.

FIG. 6a illustrates the variation in threshold voltage versus the drawn channel length (L_{P1}) of the floating gate ("poly 1"). In FIG. 6a the ordinate is the program threshold and the abscissa is the length of the floating gate channel L_{P1} in microns. (Of importance, FIGS. 6a through 6d and 7a and 7b use drawn dimensions. However, the channel lengths 36a and 36b shown in FIG. 3 are the effective dimensions after processing. Thus channel length 36a is denoted Leff to represent the effective length of this channel after processing, while before processing this channel length is a drawn dimension and as such is denoted by the symbol L_{P1}. Accordingly, each of the dimensions L_{Pf} shown in FIGS. 6a through 6d and 7a and 7b must be corrected (i.e., reduced) by a given amount (approximately 0.5 microns), to reflect the effect of processing. Naturally the amount of the correction will vary with the processing.) The threshold voltage V_{tx} obtained or programmed in a given time for a given drain voltage and gate voltage (corresponding in FIG. 6a to a drain voltage of 8 volts and a gate voltage of 12 volts) drops rapidly as the length of the channel LP1 under the floating gate 52 (FIG. 5b) increases to a minimum V_{tx} of about 2.5 volts for L_{P1} of somewhere between 3 to 4 microns and then increases slightly. This minimum V_{tx} corresponds to the initial device threshold before programming. The threshold V_{tx} represents the voltage which must be applied to the control gate (such as gate 35 in FIG. 3) to turn on the transistor beneath the control gate as shown in FIG. 3 when the cell containing that transistor has been programmed Thus as the length of the channel 36a underneath the floating gate 33 increases (FIG. 3) the threshold voltage necessary to turn on the transistor and create a channel from the source region 31b to the drain 31a decreases. As is shown in FIG. 6a, both 1 millisecond and 10 millisecond programming times yield substantially the same shaped curve.

FIG. 6b illustrates the effect of the length of the channel 36a underneath floating gate 33 on the threshold voltage (ordinate) versus programming time (abscissa). The various curves reflect different lengths L_{P1} of the channel 36a (FIG. 3) beneath floating gate 33 in microns. As these channel lengths increase, the threshold voltage for a given programming time drops. Thus for a programming time of 10^{-2} seconds, the threshold voltage for a 1.5 micron channel length L_{P1} is approximately 7 volts whereas the threshold voltage for a 3.0 micron channel L_{P1} is about 4 volts. These curves were obtained for a voltage VDS from the drain to the source of 8 volts and a voltage on the control gate 35 of 12 volts. The curves of FIG. 6b illustrate that the shorter the floating gate the stronger the field which is formed and therefore the greater the number of electrons which are placed on the floating gate thereby resulting in a larger threshold voltage V_{tx} to turn on the transistor.

FIG. 6c is a plot of threshold voltage V_{tx} (ordinate) versus the voltage on the drain 31b (FIG. 3) with the length of channel 36a beneath floating gate 33 as the parameter on the various curves. For a given drain voltage V_{DS} (for example 8 volts) the threshold voltage V_{tx} goes up as the length L_{P1} of the channel 36a beneath floating gate 33 goes down. The curves of FIG. 6c were taken with a control channel L_{P2} (corresponding to the drawn dimension of channel 36b in FIG. 3) beneath the control gate 35 of 2.5 microns, a gate voltage on control gate 35 of 12 volts and a programming time of 10 milliseconds (10^{-2} seconds). These curves illustrate that once a given drain voltage difference V_{DS} is achieved between the drain and the source, increasing the drain voltage beyond a given amount has substantially little effect on the threshold voltage V_{tx} of the transistor. In other words, $\Delta V_{tx}/\Delta V_{DS}$ becomes substantially zero thereby showing that increasing the drain voltage coupled to the floating gate has little effect on the programming of the transistor. Thus after the program threshold voltage V_{tx} is reached increasing the drain to source voltage V_{DS} does not achieve any significant improvement in performance.

As L_{P1} increases, the threshold voltage V_{tx} at which ΔV_{tx} over ΔV_{DS} becomes very small decreases. So increasing V_{DS} does even less for structures with longer floating gates.

In FIG. 6c each consecutive point on a given line for a given L_{P1} represents an additional 10 milliseconds of programming time rather than just 10 milliseconds of programming time. Accordingly the curves for V_{tx} versus V_{DS} in FIG. 6c would be even flatter than shown in FIG. 6c if a constant programming time was applied to program the cell from different V_{DS} start points.

FIG. 6d illustrates the very tight predictability of threshold voltage V_{tx} versus V_D (min) for the structure of this invention. V_D (min) is defined as the minimum VDS needed to start programming (i.e., to start efficient electron flow onto the floating gate). In FIG. 6c V_D(min) is the V_{DS} at which the curve shows a break point sharply to the right. This break point or "knee" corresponds to the V_D(min) plotted in FIG. 6d.

The relationship of FIG. 6d to FIG. 6c illustrates a basic print of my invention. In a 256 K EPROM the time to program the cells in the EPROM theoretically equals 256 K times the time to program each cell divided by 8 (ROMs are programmed one byte at a time). Therefore, if the programming time of each cell can be significantly reduced, the efficiency of programming a large number of EPROMs can be proportionally increased. I have discovered that to program to a given threshold voltage V_{tx} in a given programming time, the key is to control the length of L_{P1} and in particular to make this length (which is related to the channel 36a in FIG. 3) as small as practical without generating punch through from the source to the drain. As shown by analysis of FIG. 6d, the threshold voltage V_{tx} is increased for a given programming time by decreasing V_D(min). As shown in FIG. 6c V_D(min) decreases as L_{P1} decreases in length. Accordingly, decreasing L_{P1} is the key. to programming to a given threshold voltage V_{tx} in a given time. My invention not only allows a small effective channel length L_{eff} to be achieved beneath the floating gate but allows this channel length to be achieved in a controllable and reproducible manner throughout an EPROM array thereby to obtain repeatable and consistent results throughout the array.

FIG. 7a illustrates change of threshold voltage, ΔV_T for three different valves of L_{P1} (i.e., three different drawn channel lengths beneath the floating gate) for the structure shown in FIG. 2. In a nonself-aligned structure, the proper length of the channel under the floating gate is crucial to achieve maximum threshold voltage V_{tx}. As shown in FIG. 7a if the channel length 36a becomes too short (for example, 1.5 microns), then punch-through occurs between the source 31a and drain 31b during programming resulting in a failure to program the device. The proper alignment of a floating gate in the nonself-aligned structure to optimize the length of the channel 36a beneath the floating gate 33 and the overlap of the floating gate to drain is crucial.

The very sharp peak in FIG. 7a reflects the variation in V_{tx} with channel length L_{P1}. FIG. 7a shows that to optimize the device for the minimum channel length L_{P1} in terms of programming efficiency results in a lower initial threshold before programming and higher final threshold after programming so as to obtain a higher read current. This means a lower impedance in the circuit which in turn means that during read a capacitor in the sense amplifier in the peripheral circuitry of the memory discharges faster through a programmed transistor than otherwise would be the case resulting in shorter access time. Three effective channels beneath the floating gate (1.5 micron, 2.0 micron and 2.5 micron) are shown in FIG. 7a. The parameter ΔV_T(representing the change in threshold voltage as a function of different channel length) is illustrated by the curves. This change in voltage is particularly pronounced as one goes from 1.5 to 2 to 2.5 micron length for L_{P1}. The change in V_{tx} as a function of channel length is similar to that shown in FIG. 6a for the self-aligned structure of my invention. However, as one goes from a 2 micron L_{P1} to 1.5 micron L_{P1} and shorter, a new phenomenon appears reflecting possible punch through from the source to the drain and Vtx thus is lower than would be expected. The nonself-aligned curve shows that a proper L_{P1} is critical to obtaining a predicted threshold voltage. However, with nonself-aligned floating gate technology L_{P1} can vary even across a given chip causing a variation in V_{tx} from cell to cell within a given memory. Often this variation is unacceptable. As can be seen by the curves of FIG. 7a, a given memory can have L_{P1} from cell to cell varying for example from 1.5 microns all the way to 2.5 microns or greater because of misalignment in the masking during the processing of the wafer. Accordingly, V_{tx} is unpredictably variable across the wafer often resulting in unacceptable performance.

FIG. 7b shows the effect of overlap and V_D on threshold voltage. For the nonself-aligned device the structure must be aligned so that the 3 sigma worst case of alignment gives a satisfactory channel length 36a beneath floating gate 33. Increasing the coupling between the floating gate and the drain does not improve the threshold voltage of the device for given programming conditions so overlapping the drain with the floating gate does not help. The more overlap of the floating gate to the drain means the more electrons required to charge the floating gate for a given channel length 36a beneath the floating gate. So instead of improving the efficiency of the device, increasing the overlap of the floating gate to the drain actually decreases this efficiency. A minimum overlap of the floating gate to the drain is needed to insure that accelerated electrons hit and lodge in the floating gate rather than in the control gate or the word line.

FIG. 7b shows that as the overlap of the nonself-aligned structure increases, the ΔV_T actually declines for a given V_D. Again, this shows that the coupling between the drain and the floating gate is not helpful to achieving a desired V_{tx} and indeed can even be harmful.

The circuit of this invention is highly scaleable and retains its self-aligned character as it is scaled.

An important effect of this invention is that by choosing the correct L_{P1} the programming time for a memory array can be substantially reduced. For example, a prior art 256 K EPROM takes approximately 150 seconds or 2½ minutes to program. A 256 K EPROM using the structure of this invention can be programmed in ap-

proximately 30 seconds. This is a substantial improvement resulting in lower programming costs and lower test costs.

An additional advantage flowing from this invention is that the uncertainty in the location of the floating gate due to mask alignment tolerances is substantially reduced compared to the uncertainty in the location of the floating gate in the prior art nonself-aligned structure and in the standard prior art EPROM (nonsplit gate but self-aligned). Table 1 illustrates this improvement with respect to the self-aligned split gate structure of this invention compared to the standard non-split gate self-aligned structure of the prior art.

TABLE I

	Standard EPROM (Non-split but self-aligned gate)	Self-aligned split gate structure of this invention
STEP 1	Poly 1 (Floating gate) Critical dimension not defined but non-critical dimensions are defined	Poly 1 (Floating gate) Critical dimension defined
STEP 2	Poly 2 (Control gate) Define critical dimensions of control gate Structure - Accuracy degraded because of rough, non-planar topology associated with two layers of polycrystalline silicon	
STEP 3	Poly 1 critical dimension defined using Poly 2 as a mask	

Table I compares only the critical steps in the two processes used to define the floating gate and thus the crucial channel length L_{eff}. L_{eff} is the important channel length in the self-aligned split gate structure of this invention and in any EPROM structure. Note that in a standard non-split gate self-aligned structure L_{eff} is the total channel length between the source and drain.

As shown in Table I three steps are required to define the critical dimension of the floating gate in the standard non-split gate self-aligned structure. In the first step only the noncritical dimensions corresponding to the width (but not the length) of the channel beneath the floating gate are defined. The critical dimensions of the floating gate corresponding to the channel length beneath the floating gate are not defined. In step 2 the second layer polycrystalline silicon from which the control gate will be fabricated is deposited. The critical dimension of this second layer (known as "poly 2") is defined in step 2. This dimension corresponds to the channel length between the to-be-formed source and drain regions. However, the accuracy with which the critical dimension of the control gate is fabricated is degraded because of the rough nonplanar topology associated with the two layers of polycrystalline silicon deposited on the wafer. In the third step the first layer of polycrystalline silicon (poly 1) has its critical dimension (corresponding to channel length L_{P1}) defined using the second layer of polycrystalline silicon as a mask. Again, the accuracy with which the critical dimension of the first layer of polycrystalline silicon is defined is degraded due to the uneven topology of the structure.

In contrast, the self-aligned split gate structure of my invention defines the critical dimension of the poly 1 floating gate layer in step 1.

As the above comparison shows, the channel length L_{P1} for the standard nonsplit gate self-aligned structure is equal to the drawn length of the channel plus or minus the uncertainty in the critical dimension associated with the poly 2 definition step plus or minus the uncertainty introduced in the critical dimension of the channel length associated with poly 1 using poly 2 as a mask. Thus the uncertainty in the effective channel length in the standard nonsplit gate self-aligned structure has two components introduced by two critical dimensions. On the other hand, using the self-aligned split gate structure of my invention, only one uncertainty in a critical dimension occurs and that occurs in the first step where the poly 1 critical dimension is defined and the topology is smooth. Accordingly my invention yields a double processing advantage over the process by which the standard non-split gate self-aligned structure of the prior art is made by eliminating one critical dimension in defining L_{eff} and by introducing a much smoother topology during the formation of the critical channel length L_{eff}.

Table 2 compares the critical steps required to define the poly 1 floating gate in the nonself-aligned split gate structure of the prior art compared to the single step required to define the floating gate in the self-aligned split gate structure of my invention.

TABLE II

	Nonself-aligned split gate structure	Self-aligned split gate structure of this invention
STEP 1	Source and Drain Implanted	Poly 1 (Floating Gate) Define critical dimension
STEP 2	Poly 1 (Floating Gate) Define critical dimension	

Step 1 in fabricating the prior art nonself-aligned split gate structure is to implant the source and drain regions in the device. Step 2 is then to deposit the poly 1 layer and then form the floating gate from this layer. The critical dimension L_{P1} is defined by this step. Unfortunately, uncertainty in the length of L_{P1} results from the uncertainty in the critical dimension of the poly 1 plus or minus the misalignment of the mask used to define the critical dimension of the floating gate relative to the underlying drain region. Typically the uncertainty in the critical dimension is +0.3 microns while the uncertainty due to the mask misalignment is +0.6 microns. When combined in a statistical sense (root means square) the total uncertainty in L can be +0.6 or +0.7 microns. To the contrary, using the self-aligned split gate structure of my invention, the critical dimension of the poly 1 floating gate is defined with an uncertainty at most of about +0.3 microns. Accordingly, my invention achieves a substantial improvement in manufacturing accuracy over the prior art nonself-aligned split gate structure.

FIG. 8 illustrates an EPROM array fabricated using the self-aligned split gate structure of my invention. For simplicity, an array of nine (9) transistors or cells is shown. The programming and reading of cell or transistor Q5 will be described. Note that the array comprises word line rows m−1, m and m+1 and bit line columns n−2, n−1, n and n+1. Column n−2 is the source of transistors Q1, Q4 and Q7 while column n−1 is the drain of transistors Q1, Q4, and Q7 and the source of transistors Q2, Q5 and Q8. Similarly, column n is the drain of transistors Q2, Q5 and Q8 and the source of transistors Q3, Q6 and Q9. Column n+1 is the drain of transistors Q3, Q6 and Q9.

In operation, to read device m,n (i.e. cell Q5) all bit lines except n−1 are set at 2 volts. Bit line n−1 is set at ground. Word line m is set at 5 volts while all other word lines except m are set at ground.

To program device m,n (i.e., cell Q5) all bit lines except n are set at ground while bit line n is set at 8 or 9 volts. All word lines except m are set at ground while word line m is set at 12 volts. During programming, device m, n+1 (i.e., cell Q6) is also in programming condition but in the reverse configuration (i.e., the high voltage is applied away from the floating gate). In this configuration there is no programming of m, n+1. This asymmetry in the split gate EPROM is what enables one to utilize the virtual ground approach.

An additional embodiment of this invention is illustrated in FIGS. 9*a*, 9*b* and 9*c*. FIG. 9*a* illustrates in top view the layout of an embodiment of this invention which decreases the switching time necessary to read the state of a cell. Naturally, to increase the speed of a memory the time necessary to read the state of each cell in the memory should be decreased. The smaller the cell current or the larger the capacitance associated with the bit line connected to a given cell, the longer it takes to read the state of the cell. In the previously described embodiments of this invention, the drain region for one cell may serve as the source region for another cell. Thus when a cell is being read the source may serve as virtual ground. To read one cell in a memory and then read a second cell, the drain region of the second cell which has previously served as virtual ground must be switched to a higher voltage. The time necessary to do this depends upon the capacitance of the drain region. To do this more rapidly, the drain capacitance (also called the bit line capacitance) must be reduced. The structure of FIG. 9*a* does this by using a novel array architecture utilizing the self-aligned split gate EPROM of this invention. Instead of a virtual ground which can function as both a source and a drain region and which achieves the high cell density as in the embodiment described above in conjunction with FIGS. 3, 5*a* and 5*b*, the structure of FIG. 9*a* uses a solid, dedicated source line 130-p as in a standard EPROM and dedicated bit lines (such as bit lines 213-m, 113-m and 113-(m+1). The source line is not switched from virtual ground to a high voltage but rather is always kept at a voltage near the level at which the cell is to be switched. By doing this the switching time can be decreased. The source line 130-p comprises a metal line formed on insulation over the array in a direction orthogonal to source regions 142-(s-1) and 142-s. Vias 131-(i-1) and 131-i connect metal source line 130-p to source regions 142-(s-1) and 142-s, respectively. Source regions 142-(s-1) and 142-s are typically formed by ion implantation. Formed orthogonal to source line 130-p are word lines 122-n and 122-(n-1). Word lines 122-n and 122-(n-1) function as control gates and are formed over but separated on insulation from floating gates 121-i and 121S3 (i-1). Floating gates (of which gates 121-i and 121-(i-1) are shown in top view in their entirety) are formed of polycrystalline silicon overlying but insulated from the channel region between an underlying drain and source. In accordance with the description given above in conjunction with the structure of FIGS. 3, 5*a* and 5*b*, drain 111-i is formed by ion-implantation using edges 123-a and 124-a of floating

Case5:10-cv-04241-LHK Document1 Filed09/20/10 Page43 of 75

gates 121-(i-1) and 121-i, respectively to define the top and bottom edges 111*b* and 111*a*, respectively (FIG. 9*b*) of drain 111-i. Sides 111*c* and 111*d* of drain 111-i are bounded by oxide isolation. Thus each drain region such as region 111-i is disconnected from the other similar drain regions.

As will be shown shortly, the width of the source regions 142-(s-1) and 142-s can be reduced substantially by forming each source region (such as source regions 142-s and 142-(s-1)) by ion implantation using one edge of each word line (such as word line 122-n or 122-(n-1) respectively) as a mask to define the edge of the corresponding source region during the n+ ion implantation used to form the n+ regions of the peripheral access and logic transistors on the EPROM.

The main difference in the structure shown in FIG. 9*a* using the self-aligned split gate embodiments shown in FIGS. 3, 5*a* and 5*b*, and prior art EPROM arrays is the location of the bit lines 213-m, 113-m and 113-(m+1) and the word lines 122-(n-1) and 122-n in terms of layout. The advantage of the structure of FIG. 9*a* is that the bit line 113 associated with a given cell 110 does not have to be switched all the way from ground to a voltage necessary to detect the state of this cell but is always held at a voltage close to the read voltage. The selection of the particular cell to be read is done by the word line 122. A second advantage is that a bit line 113-m, 213-m . . . has much less capacitance than a typical prior art bit line. The reason for this is that while the bit line (such as bit line 113-m) is connected to a plurality of drain regions (such as drain 111-i of cell 110-i+1) arranged in a column, each drain such as drain 111-i functions as the drain in only two adjacent transistors (transistors 110-(i+1) and 110-i as shown in FIG. 9*a*) and each drain is not connected as part of a continuous diffused (or in this case, ion-implanted) region to the other drain regions. Thus the capacitance associated with each drain region 111 is reduced compared to the capacitance associated with a continuous drain diffusion of the prior art.

As shown in FIG. 9*a* and in cross-section in FIG. 9*b*, the ion-implanted drain region 111-i serving memory cell 110-i+1 is contacted by a via 112-i formed in insulation 150 over the drain region 111-i. Metal bit line 113-m (also called a metal drain line) electrically contacts drain 111-i through via 112-i. Formed directly adjacent to drain 111-i and self-aligned with drain 111-i as described above in conjunction with FIGS. 3, 5*a* and 5*b*, are two floating gates 121-(i-1) and 121-i formed of a first layer of polycrystalline silicon ("poly 1"). Overlying floating gate 121-i is a control gate 122-n (also called a word line) formed of a second layer of polycrystalline silicon ("poly 2"). The second layer of polycrystalline silicon 122-n extends the length of 2N transistors in the array to form a word line (also called a control gate) and is orthogonal, in the embodiment shown, to both metal drain line 113-m and metal source contact line 130-p. N is an integer which in accordance with this invention is preferrably 2, 4 or 8.

The cell 110-(i+1) includes part of a source diffusion denoted as 142-s in FIG. 9*a*. The source diffusion 142-s for cell 110-(i+1) is also the source diffusion for cell 110-(i+2) (not shown in FIG. 9*a* but shown schematically in FIG. 9*b* and 9*c*) just as the drain diffusion 111-i is the drain diffusion-for cell 110-i as well as for cell 110-(i+1). Source diffusion 142-s also serves as the source region for other cells in a given row. Thus each source diffusion except the first actually serves as the source region for 4N memory cells. The polycrystalline silicon word line 122-n is formed with substantial overlap over the source and drain to prevent misalignment from affecting the ability of a given cell to turn on when read. Edges 123*a* and 124*a* of the poly-1 floating gates 121-(i-1) and 121-i serve, as described above in conjunction with FIGS. 3, 5*a* and 5*b*, to define the edges 111*b* and 111*a*, respectively, of drain region 111-i as illustrated in FIGS. 9*a* and 9*b*.

FIG. 9*b* also illustrates the symmetrical structure memory cells 110-i+1 and 110-i of this invention. These two cells share a common drain region 111-i as shown. Control gate 122-n (typically formed of polycrystalline silicon) has a left edge 144*b* which extends over source region 142-s. Source 142-s serves as a source not only for cell 110-(i+1) but also for cell 110-(i+2) (adjacent and below cell 110(i+1) in FIGS. 9*a* but not shown in FIG. 9*a*). Overlying source region 142-s is a metal contact layer 130-p (FIG. 9*a*) which is connected to the source line 142-s by a contact 131-i.

In fabricating the structure of FIGS. 9*a*, 9*b* and 9*c* it is clear that the source region 142-s, for example, can be fabricated using left edge 144*b* of polycrystalline word line 122-n as a guide to define right edge 146*b* of source region 142-s. When this is done, the width of control line 122-n and all similar control lines can be reduced by the tolerance otherwise placed on this width to insure that it is properly aligned over source region 142-s. Thus a saving in space can be achieved using this technique of at least half a micron in width of control line 122-n. When similar savings are made in conjunction with control line 122-(n-1) by self-aligning left edge 145*b* of source 142-(s-1) with the right edge 143*b* of control line 122-(n-1), a substantially smaller array can be achieved. Advantageously, sources 142 are formed using the n+ ion implant used to form the MOS transistors in the logic and access circuitry in the peripheral regions of the memory array.

FIG. 9*c* illustrates schematically the layout of an array utilizing the structure shown in FIGS. 9*a* and 9*b*. As shown in FIG. 9*c*, metal source line 130-p is shown extending vertically down the center of the array. Contacts 131-i and 131-(i-1) are shown schematically to illustrate the vias through the underlying insulation through which metal line 130-p electrically contacts the laterally extending source regions 142-s and 142-(s-1). As described above, source regions 142-s and 142-(s-1) are preferrably formed by ion implantation. Each lateral source region serves as the source for up to 2N transistors on each side of metal source line 130-p where N typically is an integer and can be 2, 4 or 8. Metal bit lines 113-m, 213-m . . . to N13-m are shown to the left of metal source line 130-p and extending parallel to metal source line 130-p while metal bit lines 113-m+1, 213-(m+1), . . . to N13-(m+1) are shown to the right of metal source line 130-p but extending parallel thereto. Each metal bit line such as bit line 113-m contacts underlying drain regions such as drain region 111-i through a via and contact region such as 112-i. As explained above, drain region 111-i and comparable drain regions 211-i through N11-i and 115-i through N15-i each serve as the drain regions for two self-aligned transistors such as transistors 110-i and 110-(i+1). Each source line 142 serves as the source for each of the 2N or 4N transistors connected thereto. However, each incremental section of the source line between a given pair of adjacent transistors (such as transistors 110-i and 210-i) has an incremental resistance R associated there-

with. Thus, when a read current I_r passes through each incremental resistance there is a drop in voltage by the amount I_rR with the result that the drain to source voltage difference at transistor N10-i is reduced by the amount NI_rR. Accordingly, the voltage drop generated by the read current passing through the source regions to metal contact line 130-p places a practical limit on the maximum size of the number N.

As shown schematically in FIG. 9c source region 142-s serves as the source not only for transistors 110-i+1 through N10-(i+1) and 114-(i+1) through N14-(i+1), but also for transistors 110-(i+2), 210-(i+2) through N10-(i+2) and transistors 114-(i+2) through N14-(i+2). Thus the second through subsequent rows of source lines 142 each are connected to two lines of transistors.

The metal bit lines 113-m through N13-m and 113-(m+1) through N13-(m+1) are each connected in rows to a plurality of pairs of transistor cells in the same manner as shown in conjunction with line 113-m, the contact region 112-i and transistor cells 110-i and 110-(i+1) (FIGS. 9a and 9b). Typically a row of the array will include 256, 512 or more metal contact regions such as contact region 112-i contacting the drains of 512, 1024 or more transistor cells such as cell 110-i and cell 110-(i+1).

Among the advantages of the structure shown in FIGS. 9a, 9b and 9c is that the capacitance of a typical word line 142 to a typical bit line 113 is substantially reduced. The reason for this can be seen in FIG. 9b. In the prior art each time a word line 122 passes over a bit line 113, in the prior art a fairly high capacitance exists because the bit line has a high donor concentration (typically n+) and the oxide between the word line and the bit line is thin. However, with this invention, the word line 122-n is separated from the overlying bit line 113 by fairly thick oxide 150a (typically around one micron thick) and thus the capacitance is very low. Moreover, the word line 122-n has very little overlap associated with edge 144a overlying the drain region 111-i. Accordingly, the bit line to word line capacitance is substantially reduced.

Another advantage of this invention lies in the fact that the read current I_r is quite high because of the use of a split gate. With a split gate the floating gate 121-i can be made shorter but wider because high coupling to the gate is not required. Since the current is proportional to the width of the floating gate over the length of the floating gate, the current is increased by having a short floating gate Finally, the bit line capacitance itself is smaller than in the prior art because the bit line 113-m contacts 256 discrete drains rather than a continuous drain line. The reduction in size of the drain regions by using unconnected drain regions such as drain 111-i substantially reduces the capacitance associated with the bit line.

While several embodiments of this invention have been described, other embodiments of this invention will be obvious to those skilled in the semiconductor arts in view of this disclosure.

What is claimed is:

1. An EPROM array comprising:
- a substrate composed of semiconductor material;
- a plurality of memory cells formed on the substrate, each memory cell including a split gate transistor;
- a metal source contact line running in a first direction across the array;
- a source diffusion line having a multiplicity of portions serving as source regions of the split gate transistors, the source diffusion line being integrally formed in the substrate and running a selected distance across said array orthogonal to said metal source contact line;
- a contact between said metal source contact line and the source diffusion line for coupling a potential on said metal source contact line to each of said source regions;
- a plurality of metal drain lines running across said array substantially parallel to said metal source contact line each metal drain line contacting drain region of a selected number of the split gate transistors in said array;
- a plurality of control lines formed over said array running orthogonal to said metal source contact line and said plurality of metal drain lines;
- wherein each split gate transistor comprises: a channel region; a floating gate formed over but insulated from a first portion of the channel region, a first edge of the floating gate being aligned with and used to define one edge of the drain region of the split gate transistor, a second edge of said floating gate being over said channel region, the second edge being positioned away from the first edge of the floating gate by a predetermined distance and separated from the closest edge of the source region of the transistor by a second portion of said channel region, and a control gate formed over but insulated from said floating gate and formed over but insulated from said second portion of said channel region; and
- wherein said control gate of each split gate transistor comprises part of one of said plurality of control lines.

2. An EPROM array as in claim 1 wherein the source diffusion line has no more than 2N portions serving as source regions and N is an integer selected from the group consisting of 2, 4 and 8.

3. A memory device comprising the structure of claim 1 replicated M times to form an EPROM device with M times the memory cells of the structure of claim 1.

4. Structure as in claim 1 wherein said control gate extends over the second portion of said channel region to the source region and the edge of said control gate furthest from said floating gate is used to define, and is aligned with, the edge of the source region adjacent the channel region.

5. An EPROM array containing a plurality of memory cells wherein each cell in the array includes a transistor containing a source region, a drain region and a channel region therebetween, a floating gate formed over but insulated from a first portion of the channel region, a first edge of the floating gate being aligned with and used to define one edge of said drain region and a second edge of said floating gate being over said channel region, separated from the first edge of the floating gate by a predetermined distance, and separated from the closest edge of said source region by a second portion of said channel region, and a control gate formed over but insulated from said floating gate and formed over but insulated from said second portion of said channel region.

6. Structure as in claim 5 wherein said control gate extends over the second portion of said channel region to the source region and the edge of said control gate

furthest from said floating gate is used to define, and is aligned with the edge of the source region adjacent the second portion of the channel region and wherein the control gate further extends beyond the first edge of the floating gate.

7. A memory array comprising:

a semiconductor substrate; and

a plurality of split gate transistors formed in the substrate, each having: a source region; a drain region spaced apart from the source region; a first channel portion interposed between the source and drain regions; a second channel portion interposed between the first channel portion and the source region; a floating gate insulatively disposed over the first channel portion, the floating gate having opposed first and second edges spaced apart by a predetermined distance, the first edge of the floating gate being self-aligned with and used to define an edge of the drain region; and a control gate overlapping the second channel portion and the floating gate, the control gate having a portion extending over and beyond the first edge of the floating gate.

8. The memory array of claim 7 wherein the control gate of each transistor has an edge self-aligned with and used to define an edge of the source region.

* * * * *

5 10 15 20 25 30 35 40 45 50 55 60 65

UNITED STATES PATENT AND TRADEMARK OFFICE

CERTIFICATE OF CORRECTION

PATENT NO. : 4,868,629

DATED : September 19, 1989

INVENTOR(S) : Boaz Eitan

It is certified that error appears in the above-identified patent and that said Letters Patent is hereby corrected as shown below:

Sheets 1-7 of the drawing should be deleted to be replaced with sheets 1-7 of the drawing as shown on the attached sheets.

Signed and Sealed this

Thirteenth Day of March, 1990

Attest:

JEFFREY M. SAMUELS

Attesting Officer

Acting Commissioner of Patents and Trademarks

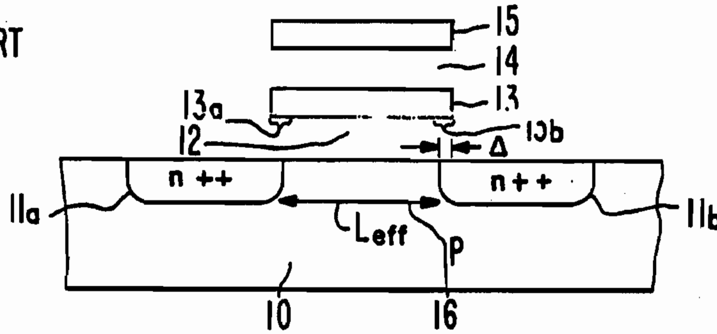


FIG. 1
PRIOR ART
15
14
13a
13
12
13b
Δ
n ++
n ++
11a
11b
Leff
p
10
16

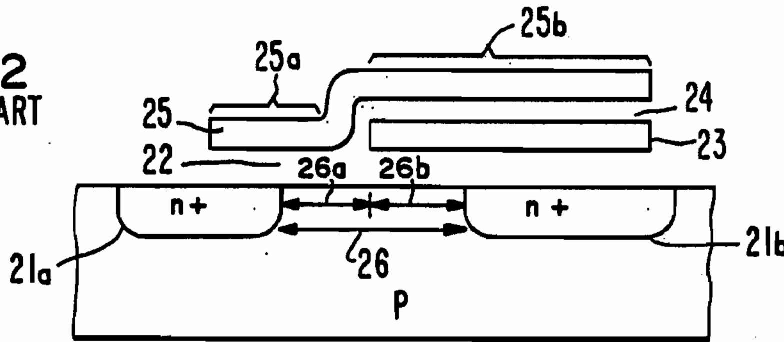


FIG. 2
PRIOR ART
25b
25a
24
25
23
22
26a
26b
n+
n +
21a
21b
26
p

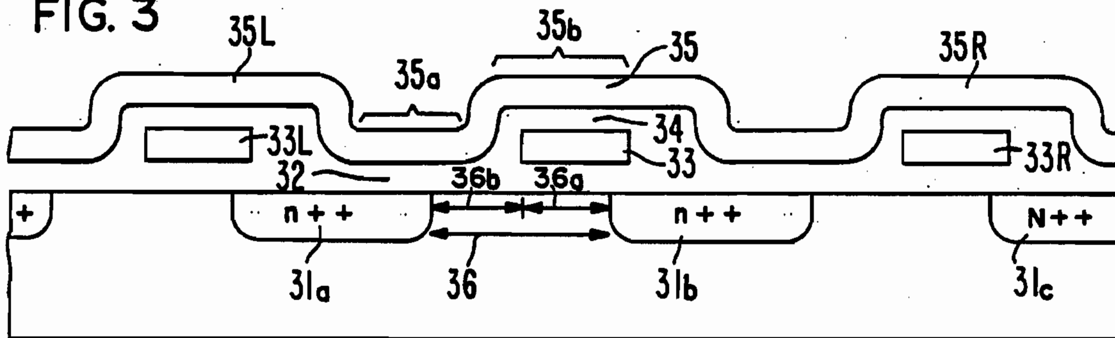


FIG. 3
35L
35b
35
35R
35a
34
33L
33
33R
32
36b
36a
+
n + +
n + +
N++
31a
36
31b
31c

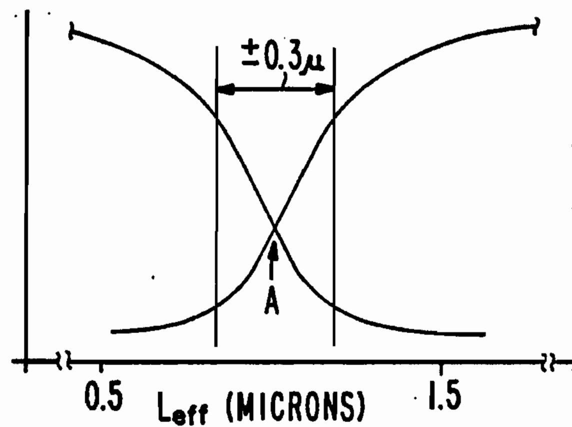


FIG. 4
PRIOR ART
Vtx
(THRESHOLD
VOLTAGE)
±0.3μ
VDTO
(DRAIN TURN-ON
VOLTAGE)
A
0.5
Leff (MICRONS)
1.5

FIG. 5a

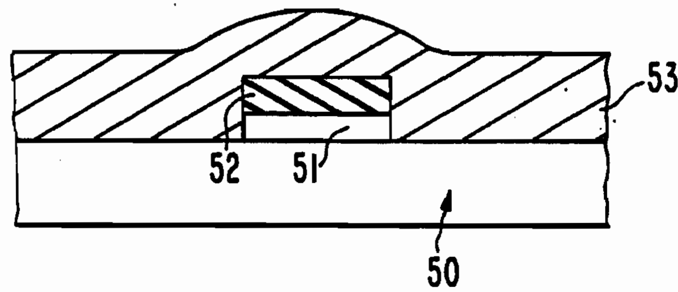


FIG. 5b

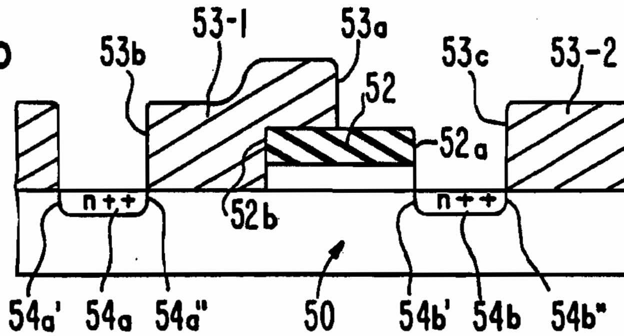


FIG. 6a

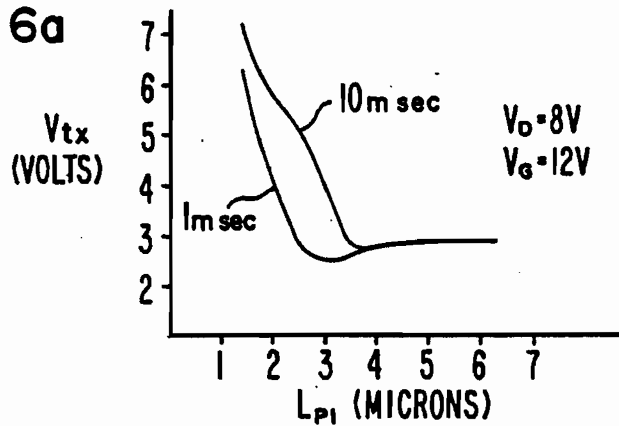


FIG. 6b

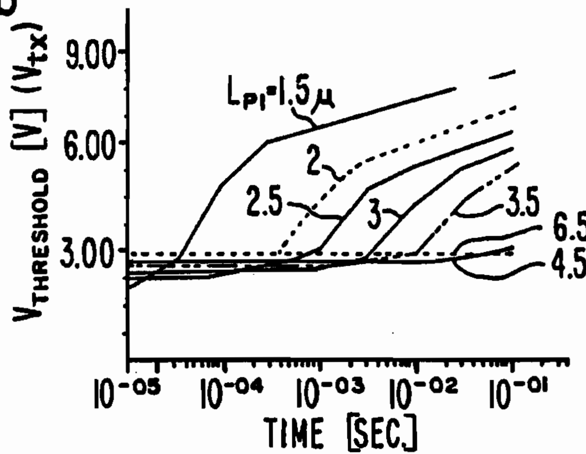


FIG. 6c

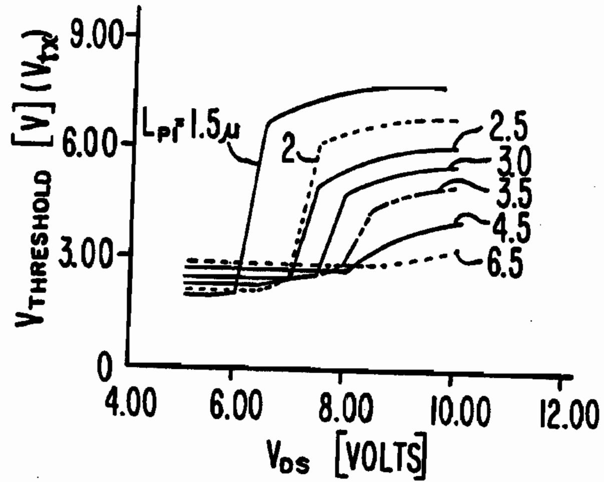


FIG. 6d

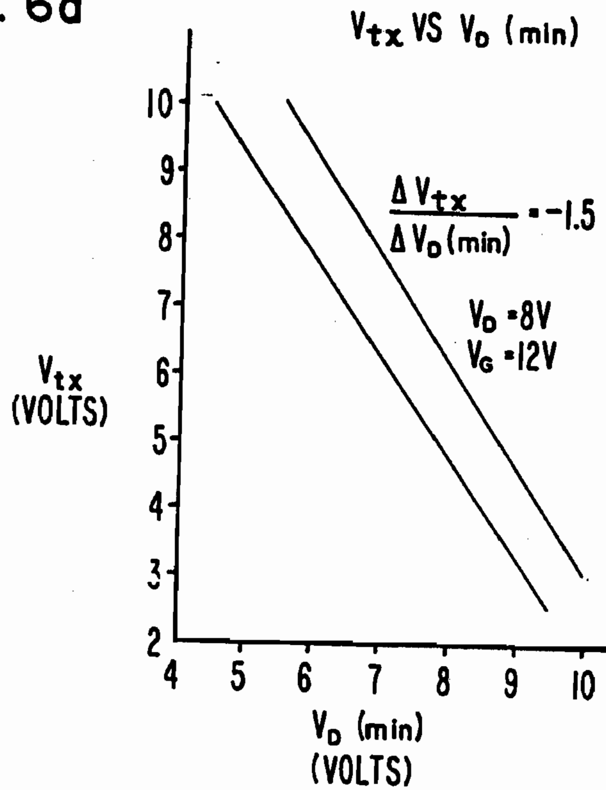


FIG. 7a

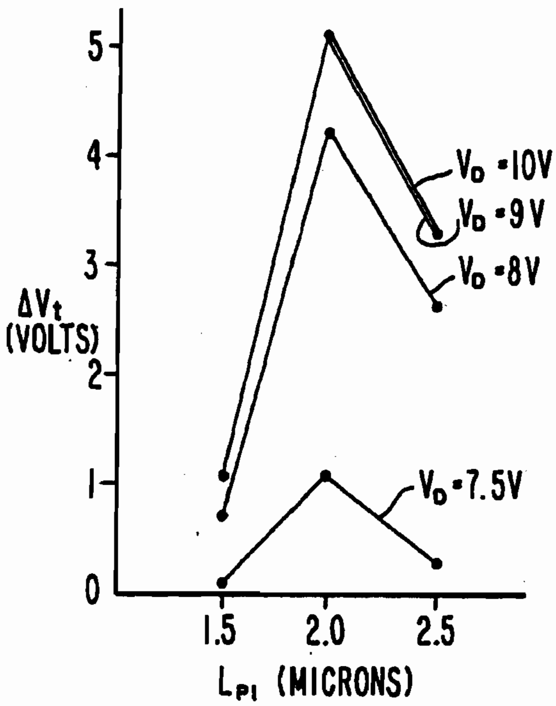


FIG. 7b

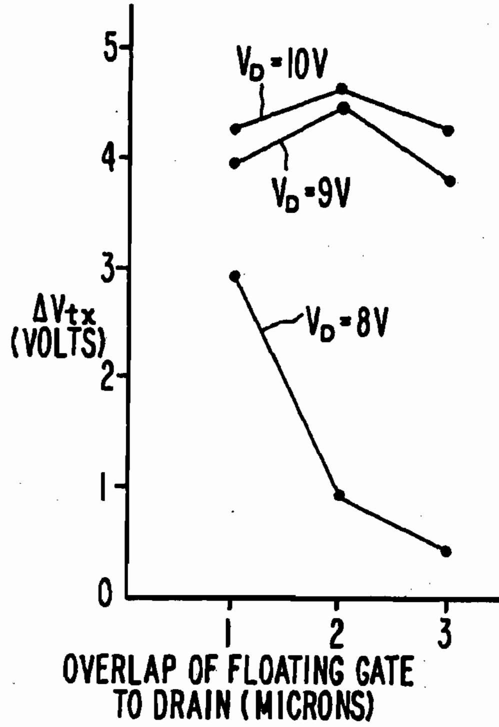
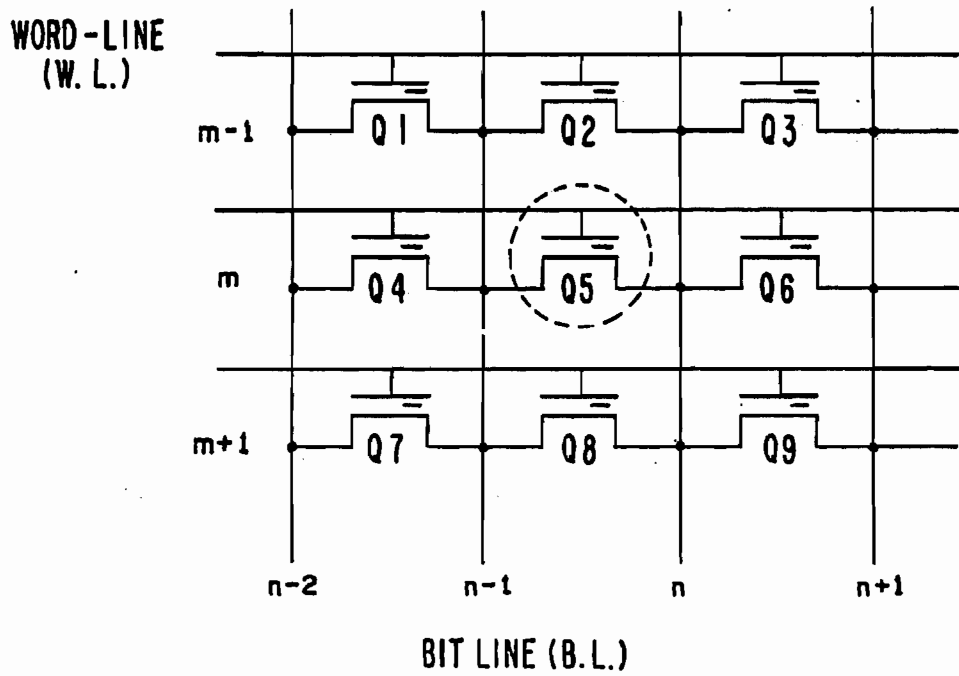


FIG. 8



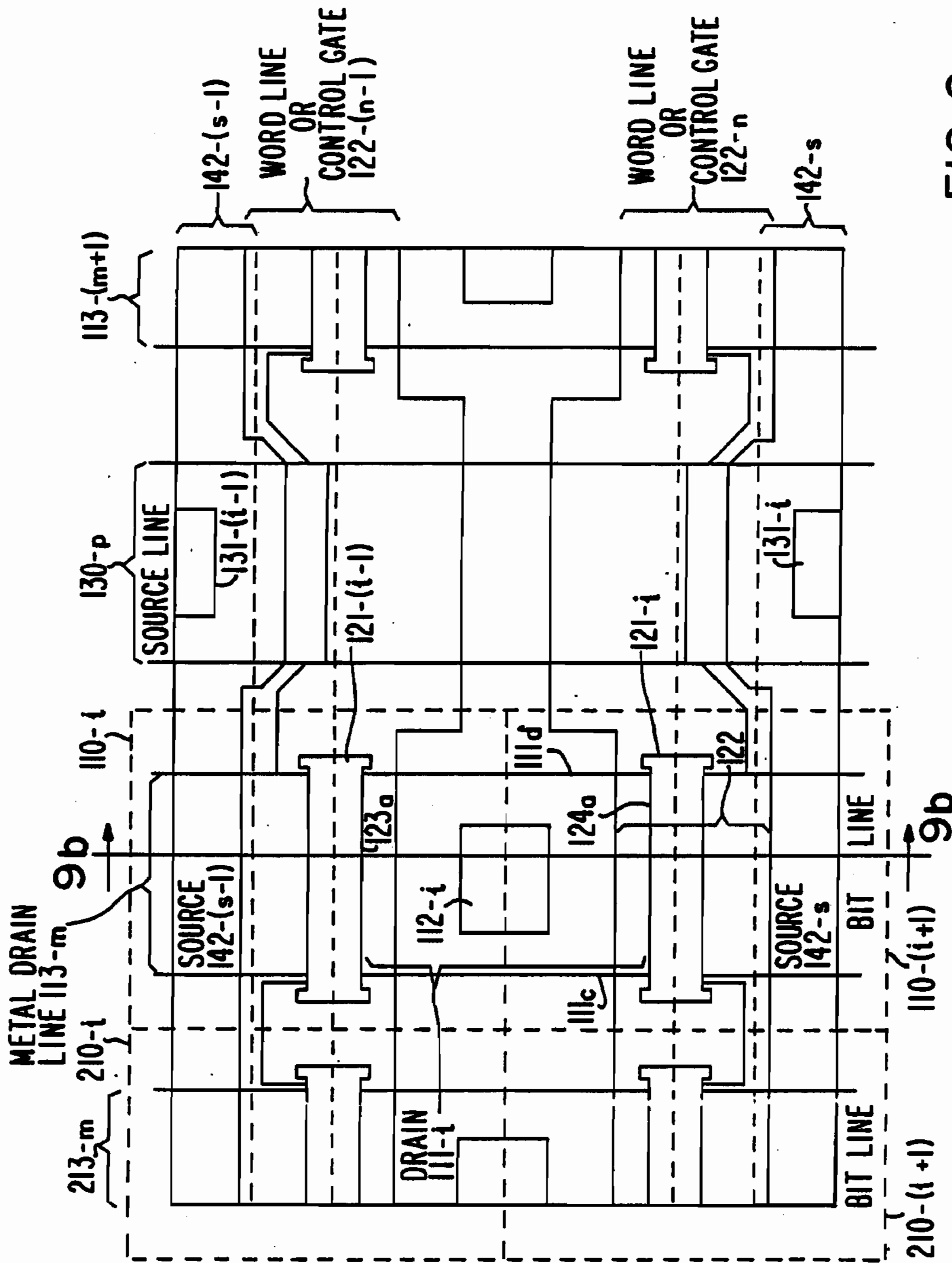


FIG. 9a

Case5:10-cv-04241-LHK Document1 Filed09/20/10 Page52 of 75

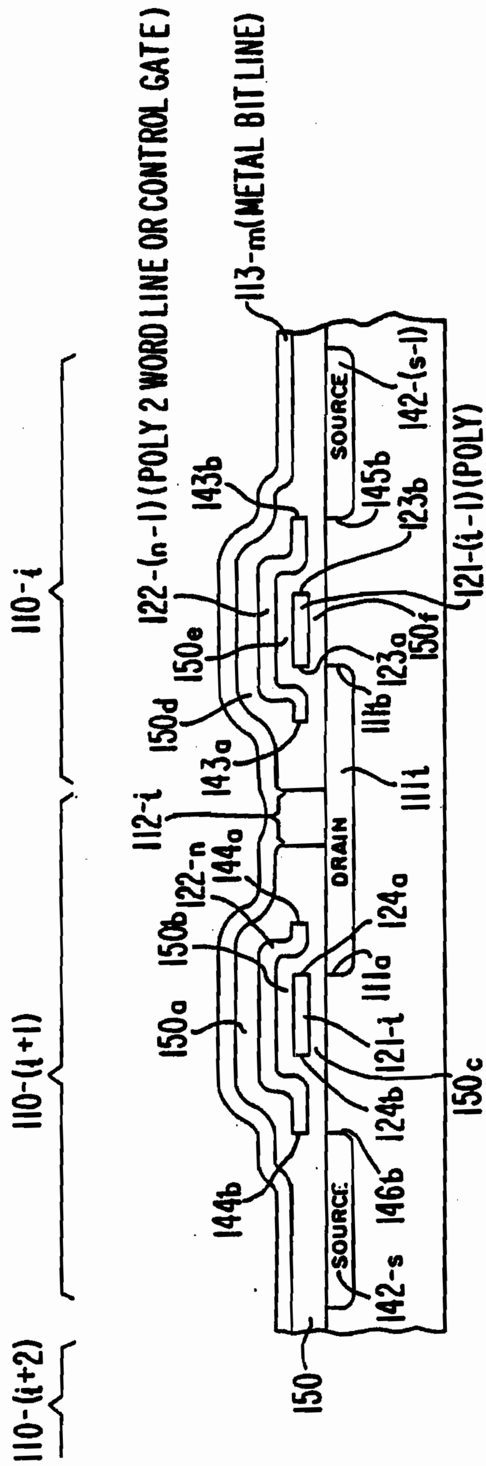


FIG. 9b

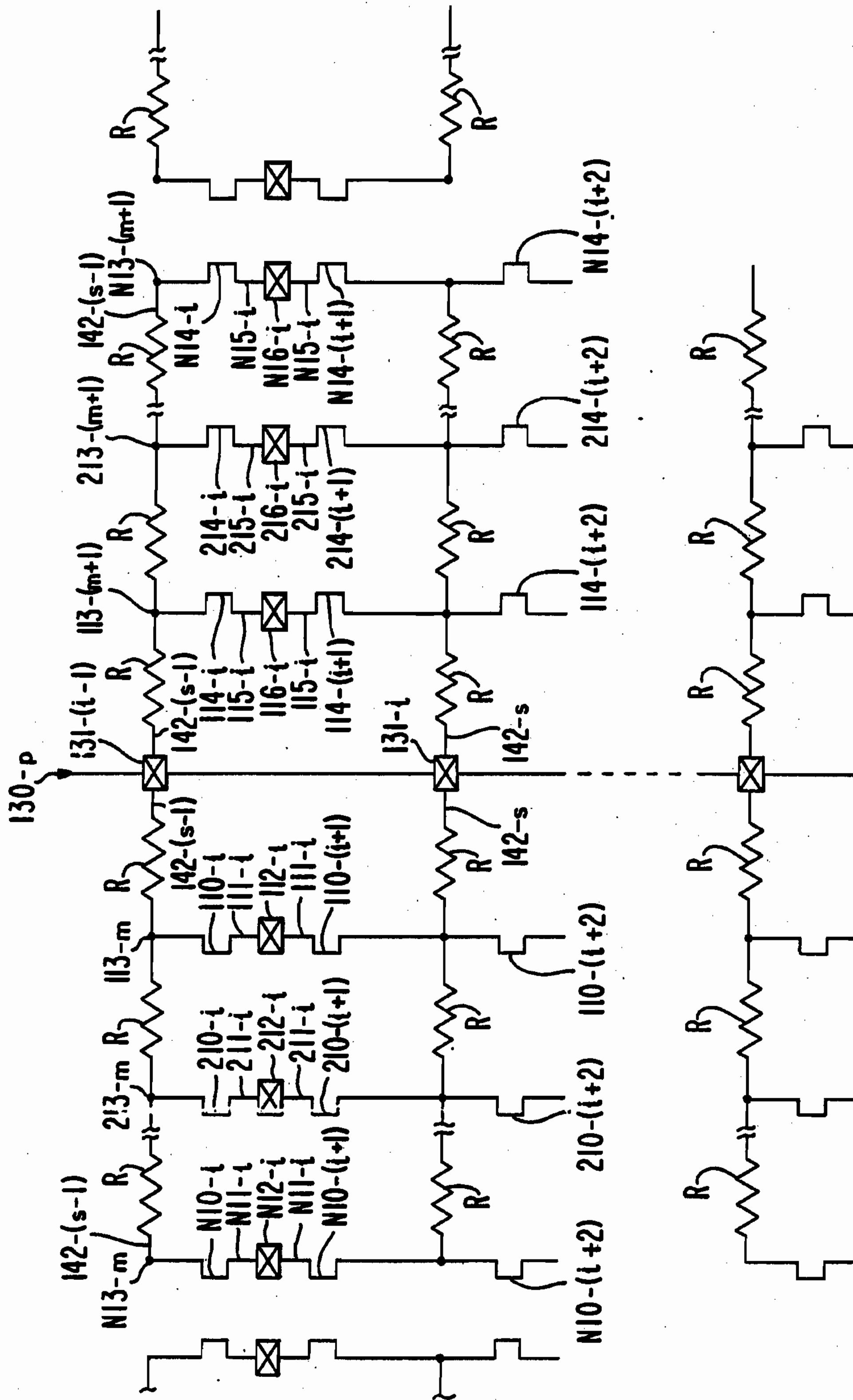


FIG. 9c

UNITED STATES PATENT AND TRADEMARK OFFICE

CERTIFICATE OF CORRECTION

PATENT NO. : 4,868,629 Page 1 of 3

DATED : September 19, 1989

INVENTOR(S) : Boaz Eitan

It is certified that error appears in the above-identified patent and that said Letters Patent is hereby corrected as shown below:

Col. 1, line 21, insert --structure-- after "gate".

Col. 1, line 22, delete "1B".

Col. 1, line 27, delete "23".

Col. 1, line 50, delete "B".

Col. 1, line 56, insert --++-- after "n".

Col. 1, line 66, insert --.-- after "13".

Col. 2, line 9, "drain 11b" should read --and the drain 11b-- with number "11b" in bold.

Col. 2, line 22, "B" should be deleted.

Col. 5, line 26, the line should read --to form a floating gate 52. The oxide--.

Col. 5, line 54, change "n." to --n++--.

Col. 5, line 59, insert --segment-- after "photoresist".

Col. 7, line 9, "Leff should read --L_{eff}--.

Col. 7, line 21, insert --,-- after "rapidly".

Col. 7, line 23, insert --,-- after "increases".

Col. 7, line 49, "VDS" should read --V_{DS}--.

UNITED STATES PATENT AND TRADEMARK OFFICE

CERTIFICATE OF CORRECTION

PATENT NO. : 4,868,629

DATED : September 19, 1989

INVENTOR(S) : Boaz Eitan

It is certified that error appears in the above-identified patent and that said Letters Patent is hereby corrected as shown below:

Col. 8, line 8, insert --,-- after "reached".

Col. 8, line 24, "V_D" should read --V_{DS}--.

Col. 8, line 31, "print" should read --point--.

Col. 8, line 55, "ΔV_T" should read --ΔV_{tx}--.

Col. 9, line 12, "Three effective ..." should begin a new paragraph.

Col. 9, line 14, "ΔV_T" should read --ΔV_{tx}--.

Col. 9, line 24, "Vtx" should read --V_{tx}--.

Col. 9, line 33, delete "." after "microns".

Col. 9, line 57, "ΔV_T" should read --ΔV_{tx}--.

Col. 10, Table 1, insert --structure-- after "gate" in left column heading of table;
delete "structure" after "gate" in left column step 2, line 3, of table.

Col. 11, line 47, "+0.3" should read --$\pm$0.3--.

Col. 11, line 48, "+0.6" should read --$\pm$0.6--.

Col. 11, line 50, "L" should read --L_{P1}--.

Col. 11, line 50, "+0.6 or +0.7" should read --$\pm$0.6 or $\pm$0.7--.

UNITED STATES PATENT AND TRADEMARK OFFICE

CERTIFICATE OF CORRECTION

PATENT NO. : 4,868,629 Page 3 of 3

DATED : September 19, 1989

INVENTOR(S) : Boaz Eitan

It is certified that error appears in the above-identified patent and that said Letters Patent is hereby corrected as shown below:

Col. 11, line 54, "+0.3" should read --±0.3--.

Col. 12, line 11, insert --a-- after "is also in".

Col. 12, line 61, "121S3 (i-1)" should read --121-(i-1)--

Col. 15, line 50, insert "." after "gate".

Signed and Sealed this

Fourth Day of June, 1991

Attest:

HARRY F. MANBECK, JR.

Attesting Officer *Commissioner of Patents and Trademarks*

EXHIBIT C

United States Patent [19]

Kazerounian et al.

[11] **Patent Number: 5,042,009**

[45] **Date of Patent: Aug. 20, 1991**

[54] METHOD FOR PROGRAMMING A FLOATING GATE MEMORY DEVICE

[75] Inventors: Reza Kazerounian, Alameda; Boaz Eitan, Sunnyvale, both of Calif.

[73] Assignee: WaferScale Integration, Inc., Fremont, Calif.

[21] Appl. No.: 282,788

[22] Filed: Dec. 9, 1988

[51] Int. Cl.[5] G11C 13/00

[52] U.S. Cl. 365/185; 365/189.01; 365/230.01

[58] Field of Search 365/185, 189.01, 230.01

[56] References Cited

U.S. PATENT DOCUMENTS

4,580,067 4/1986 Proebsting et al. .... 307/296 R
4,628,487 12/1986 Smayling 365/185
4,667,312 5/1987 Doung et al. 365/189
4,723,225 2/1988 Kaszubinski et al. 365/185
4,763,299 8/1989 Hazane 365/51

OTHER PUBLICATIONS

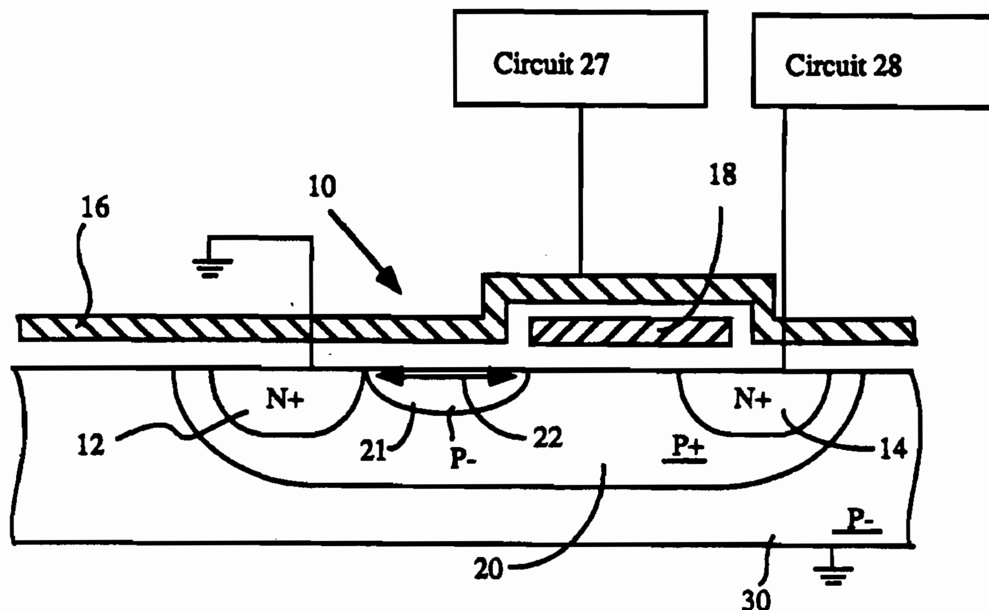
R. Kazerounian, et al., "A 5 Volt Only High Density Poly-Poly Erase Flash Eprom Cell", IEDM. Dec. 1988, pp. 436-439.

Primary Examiner—Terrell W. Fears
Attorney, Agent, or Firm—Skjerven, Morrill, MacPherson, Franklin & Friel

[57] ABSTRACT

A method of programming a floating gate transistor permits the use of a charge pump to provide drain programming current. The programming drain current is typically held below about 1 μA. This programming drain current can be provided by a conventional charge pump. In the first embodiment, the drain current can be limited by connecting a resistor between the source and ground. In a second embodiment, the drain current is limited by limiting the transistor control gate voltage. In a third embodiment, a charge pump is coupled to the drain while the control gate is repetitively pulsed. Each time the control gate is pulsed, the transistor turns on, and although the drain is initially discharged through the transistor, some hot electrons are accelerated onto the floating gate, and eventually the floating gate is programmed. In these embodiments the erase gate voltage may be raised to enhance programming efficiency.

30 Claims, 10 Drawing Sheets



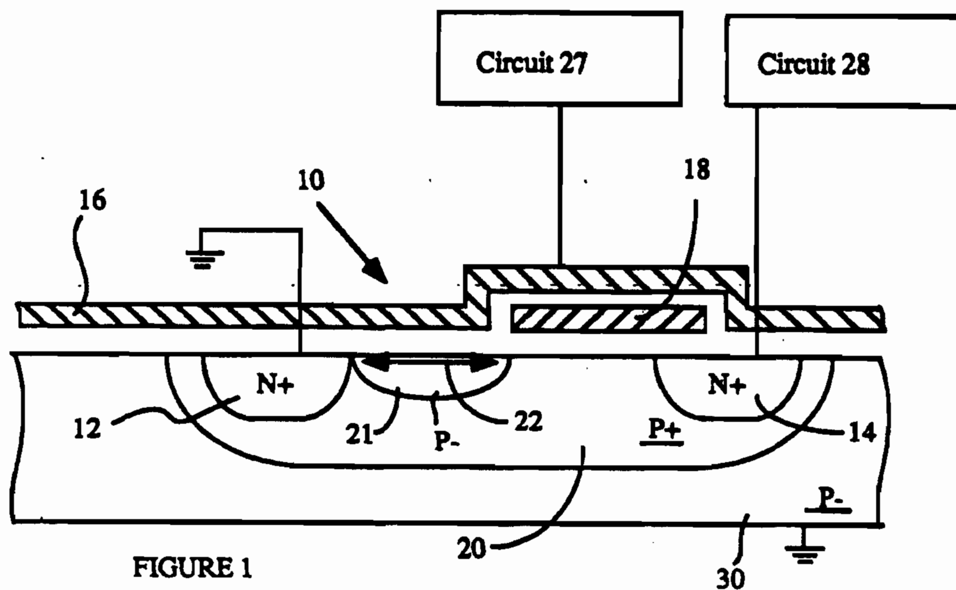


FIGURE 1

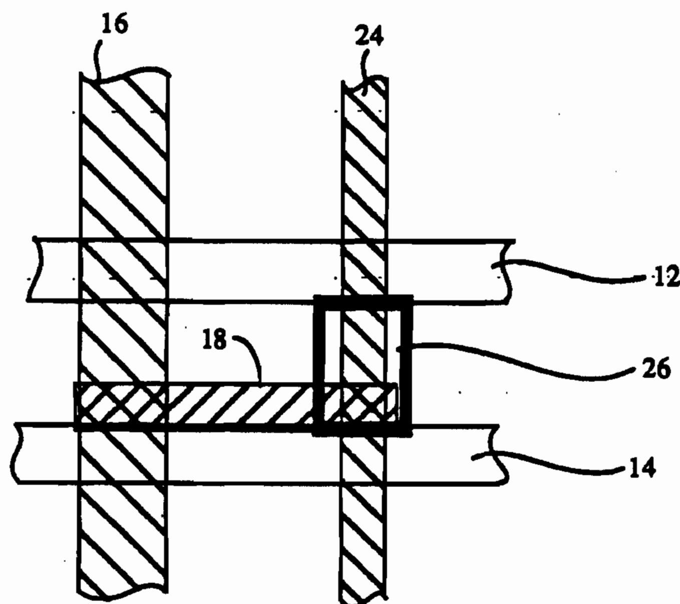


FIGURE 1a

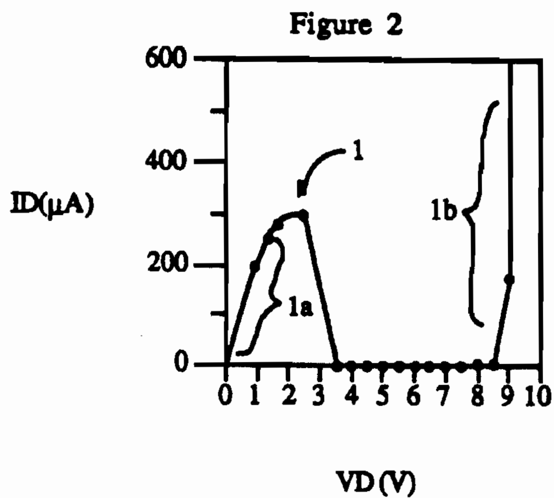


Figure 2
600
400
200
0
ID(μA)
1
1a
1b
0 1 2 3 4 5 6 7 8 9 10
VD (V)

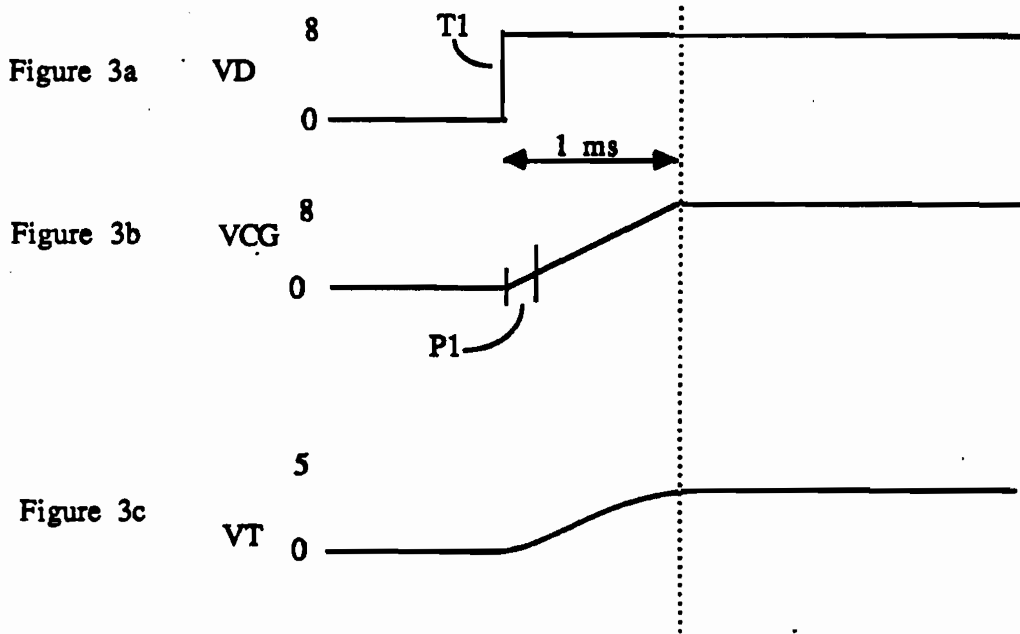


Figure 3a
VD
8
0
T1
1 ms
Figure 3b
VCG
8
0
P1
Figure 3c
VT
5
0

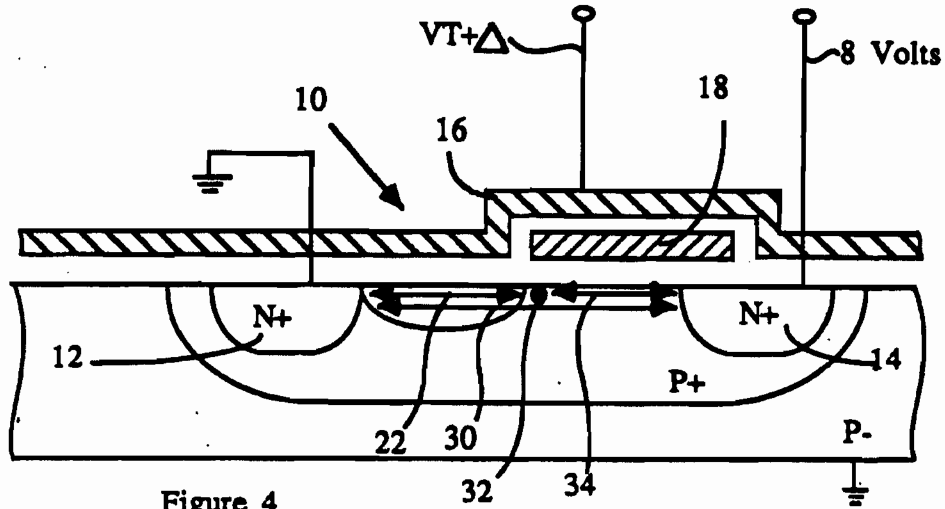


Figure 4

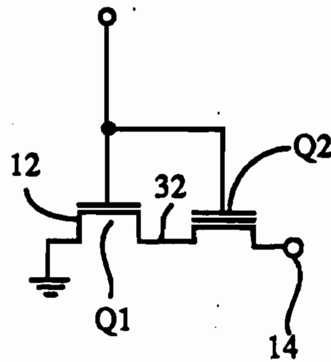


Figure 4a

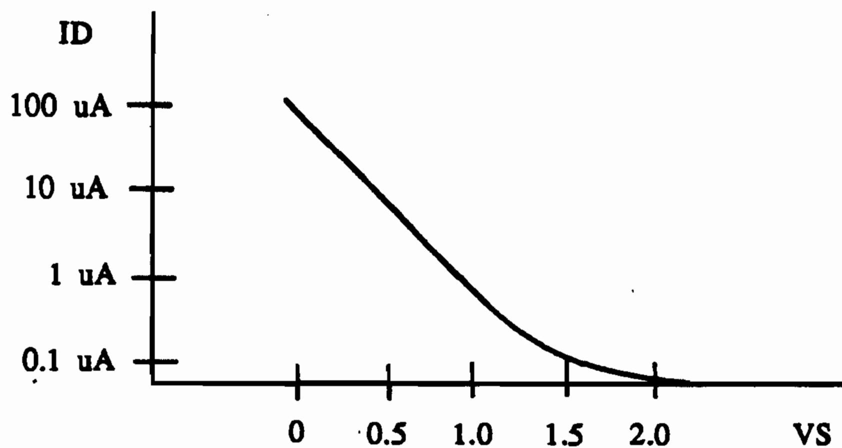


Figure 5

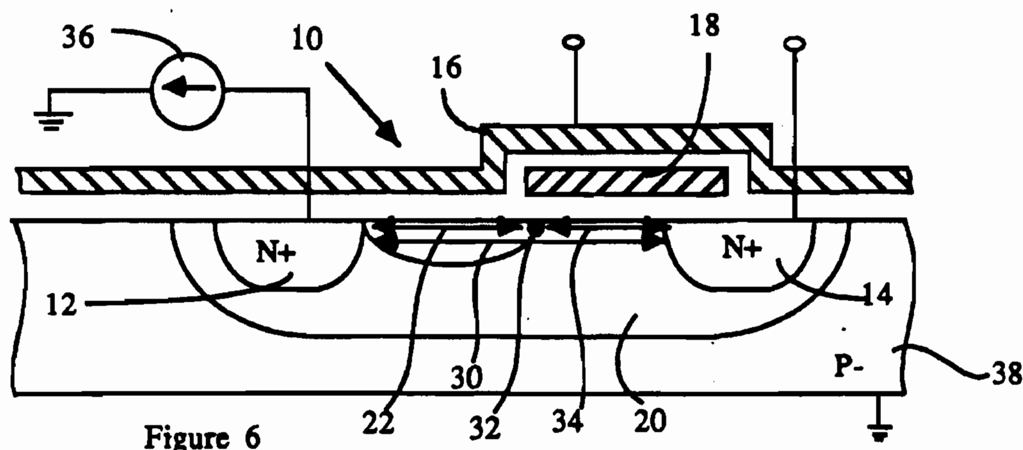


Figure 6

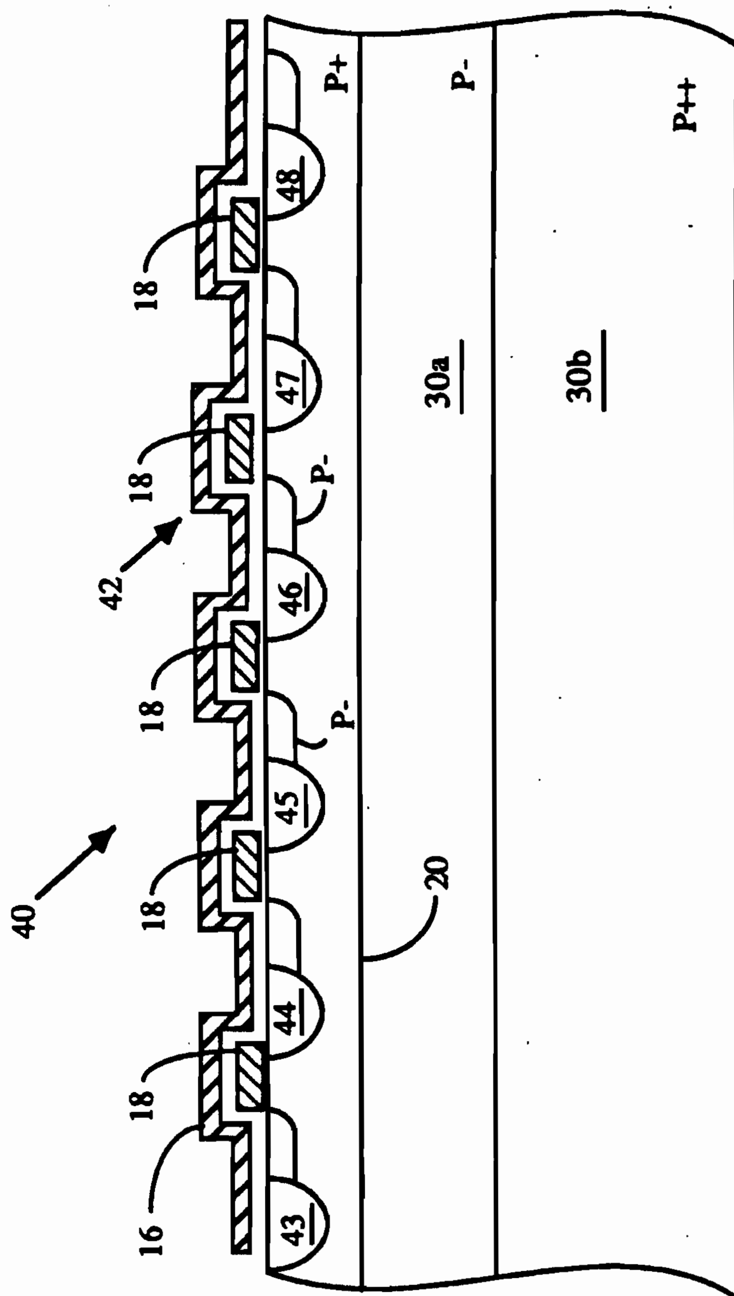


Figure 6a

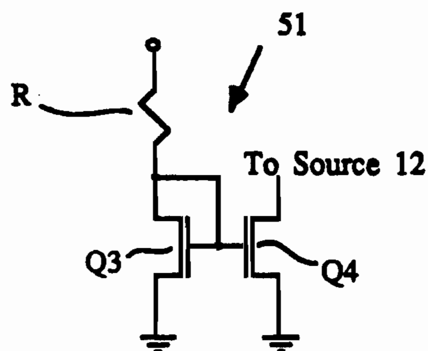


Figure 7

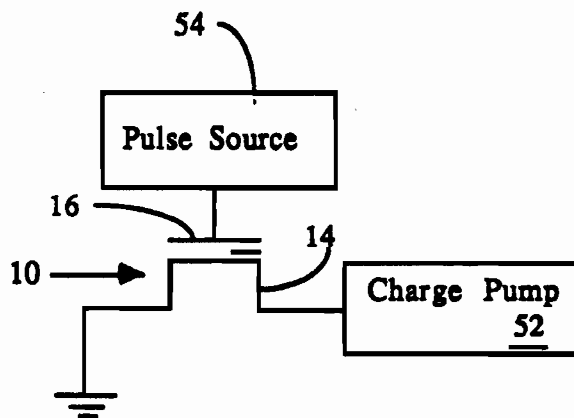


Figure 8

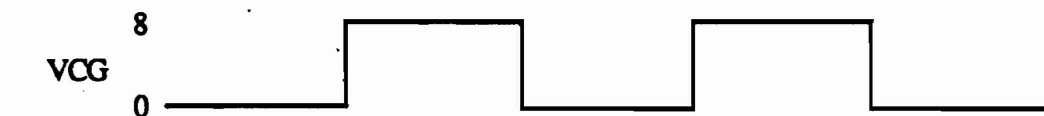


Figure 9a

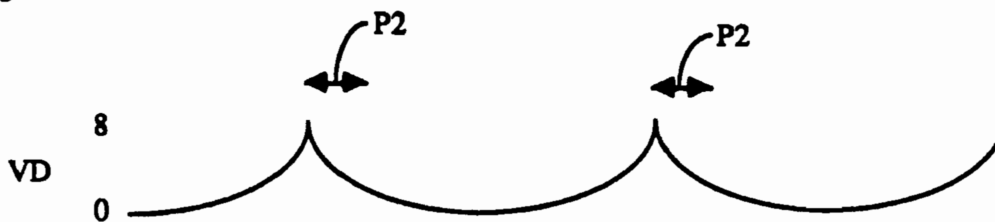


Figure 9b

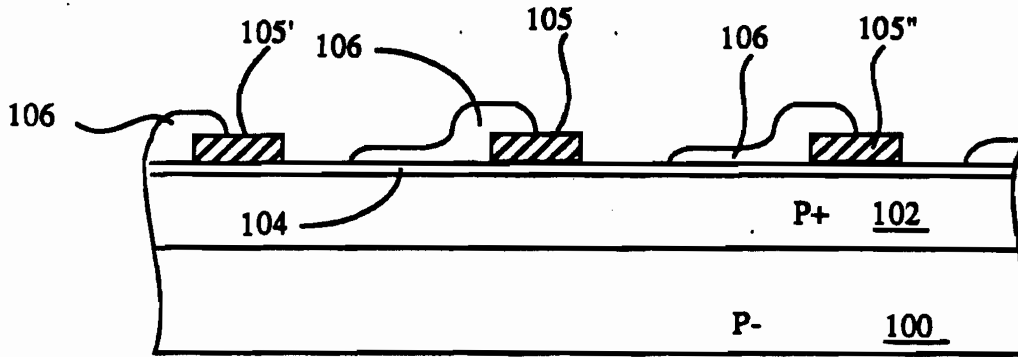


Figure 10a

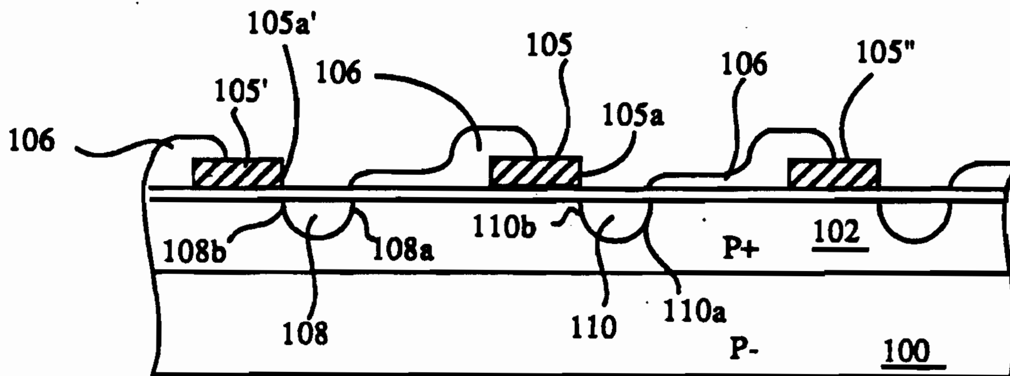


Figure 10b

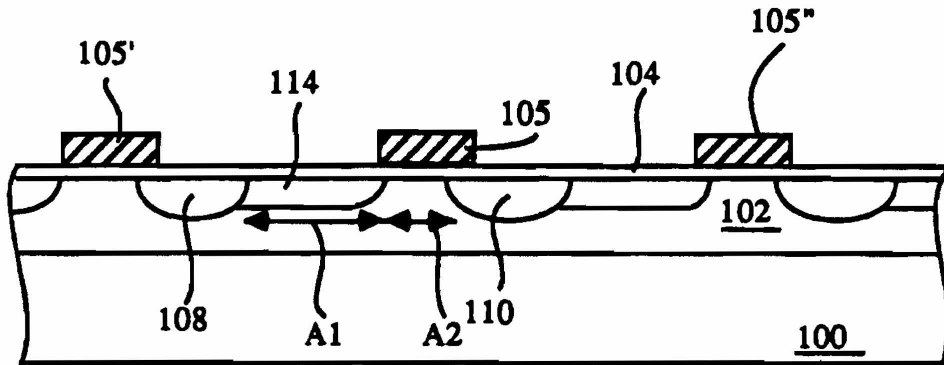


Figure 10c

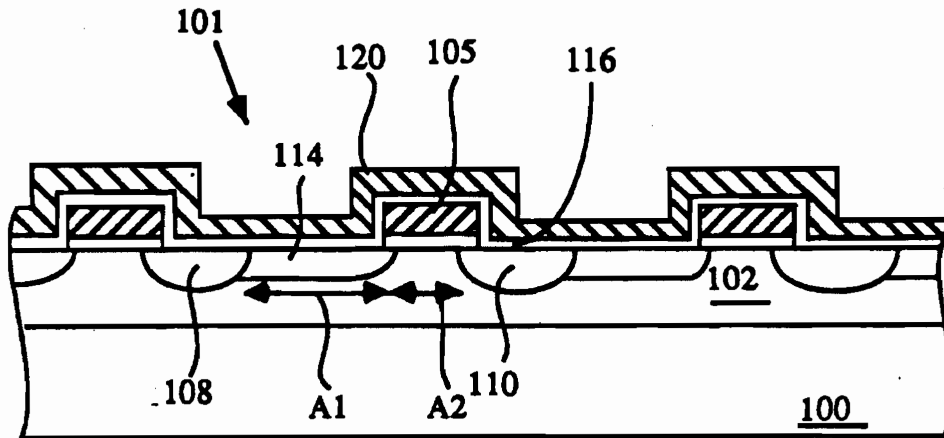


Figure 10d

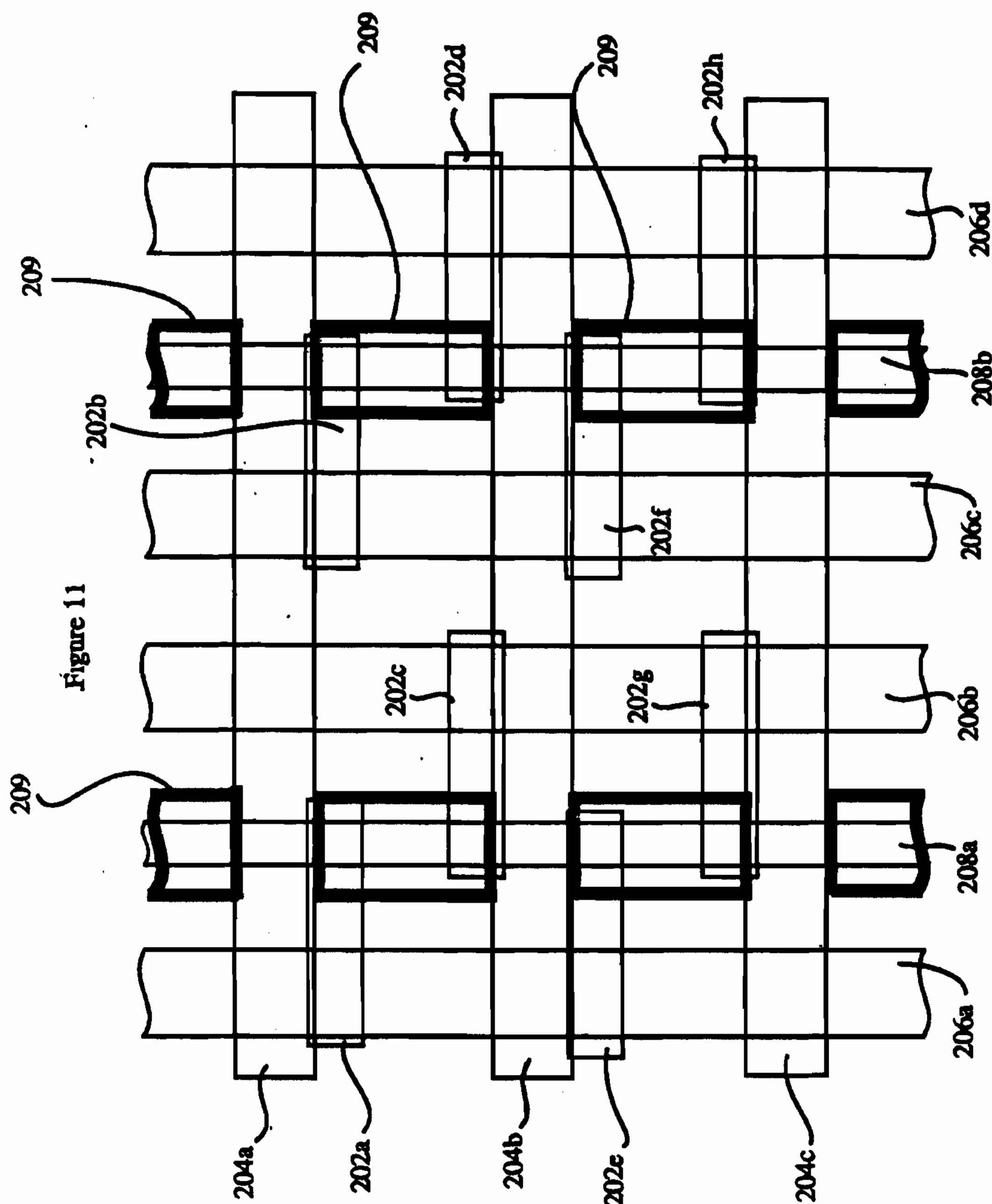


Figure 11

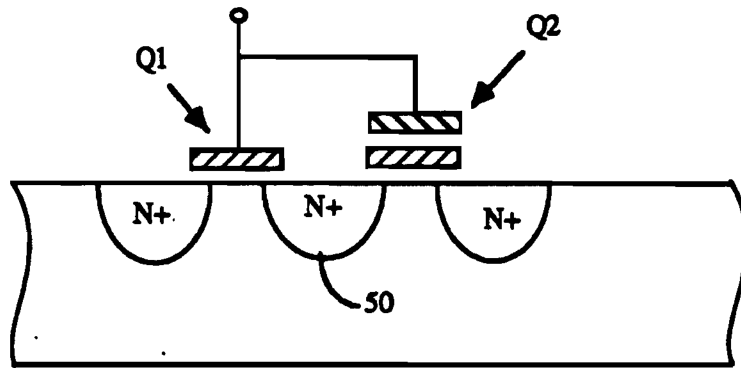


Figure 12

METHOD FOR PROGRAMMING A FLOATING GATE MEMORY DEVICE

BACKGROUND OF THE INVENTION

This invention relates to floating gate memory devices and methods for programming such devices.

There are a number of floating gate memory devices known in the art. One type of floating gate memory comprises an array of floating gate transistors which are programmed and erased by an electron tunneling mechanism. An example of such a device is discussed by Johnson et al. in "A 16 Kb Electrically Erasable Nonvolatile Memory," published at the IEEE International Solid State Circuits Conference in 1980, page 152-153, incorporated herein by reference. Johnson's device uses programming and erase voltages of about 25 volts. Although most digital electronic systems include a 5 volt power supply but do not include a 25 volt power supply, 25 volts can be generated on-chip from a 5 volt power supply with a conventional charge pump, since the amount of current required for tunneling is on the order of 1 nA. Unfortunately, memory cells which are programmed and erased by tunneling tend to be large, and thus expensive.

Another type of floating gate memory is the EPROM, which is programmed by hot electron injection and erased by exposure to UV light. EPROM cells are small, and are less expensive to build than EEPROM cells, but the data stored in the EPROM cannot be reprogrammed unless the EPROM is removed from a system and exposed to UV light prior to reprogramming. Further, such devices are programmed by hot electron injection, which requires a voltage in excess of 5 volts (e.g. about 12 volts) and a high programming current. Such programming currents are too large to generate with a charge pump. Thus, if one wanted to program an EPROM in-system, one would have to include an extra power supply, which would entail an undesirable expense.

Another type of floating gate memory is the flash EPROM, which is programmed by hot electron injection and erased by tunneling. Such a device is discussed by Kynett et al. in "An In-System Reprogrammable 256K CMOS Flash Memory", published at the IEEE International Solid State Circuits Conference in 1988, pages 132 to 133, incorporated herein by reference. Advantageously, flash EPROMs have small memory cells, and are thus relatively inexpensive. However, since flash EPROMs of the type discussed by Kynett are erased by electron tunneling either between the floating gate and drain or between the floating gate and source, they draw a large current during electrical erase due to band to band tunneling across the drain/substrate or source/substrate junction. Flash EPROMs also have a number of other disadvantages. For example, they are hot electron programmed, and thus require a programming voltage in excess of 5 volts (typically 8 to 12 volts) with about 1 mA of programming current per cell. This combination of high current and high programming voltage cannot be economically generated from an on-chip charge pump. (Flash EPROMs cannot be efficiently programmed merely by connecting a 5 volt power supply to the drain, especially at high operating temperatures, e.g. 125° C. Also, since the output voltage of a nominally 5 volt power supply may vary by plus or minus 10%, and thus be as low as 4.5 volts, programming cannot be efficiently accomplished by connecting

the 5 volt power supply to the drain for this reason as well.) Another limitation of the above flash EPROM is the need for a tightly regulated erase voltage to prevent over-erase, i.e. to prevent the erase circuitry from leaving the floating gate with a large positive charge. (Since Kynett's floating gate extends from the source to the drain, a positively charged floating gate would leave Kynett's transistor on regardless of the state of his control gate.)

It would be desirable to provide a floating gate memory device which combines the following features:

(1) The small cell size of a flash EPROM;

(2) The erasability of an EEPROM, i.e. a device which can be erased in-system, wherein the erase voltage is generated by a charge pump from a single 5 volt power supply; and

(3) In-system programmability from a single 5 volt power supply.

These goals could be achieved if a method were found for programming a flash EPROM without requiring more than a few microamps of drain current.

SUMMARY

A erasable floating gate memory device constructed in accordance with an embodiment of the invention has the small cell size of a flash EPROM, but can be programmed and erased using a single 5 volt power supply. Of importance, the programming and erase voltages are generated on-chip from the 5 volt power supply, e.g. using a charge pump.

One embodiment of the invention includes means for limiting the amount of current permitted to flow through the drain during programming. Because of this, the programming drain voltage can be generated by a charge pump and it is not necessary to provide an additional power supply for programming the memory device.

In a first embodiment, during programming, the control gate voltage of the floating gate memory device is ramped from a first voltage (e.g. ground) to a programming voltage (e.g. between 5 and 8 volts) over a time period such as 1 millisecond. Because of this, the programming drain current ramps up slowly during the 1 millisecond period, hot electrons are continuously injected onto the floating gate during the 1 millisecond period, the threshold voltage of the transistor is constantly increasing, and there is no period of time during which the drain current exceeds a value greater than that which the charge pump can provide.

In a second embodiment of the invention, during programming, the control gate is raised to a value just slightly greater than the threshold voltage of the transistor while a programming drain voltage is applied to the drain region. This ensures that the drain current through the transistor is of a magnitude which can be provided by a charge pump. The memory device typically incorporates a split gate architecture, i.e. the floating gate covers a first portion of the channel but not a second portion. The control gate covers the second portion of the channel and part of the floating gate. Thus, the control gate controls the amount of current permitted to flow through the channel, even if the floating gate is positively charged.

In this embodiment, the memory device includes an erase gate which is capacitively coupled to the floating gate. During programming, the erase gate voltage is raised, e.g. to about 10 volts, to thereby increase the

electrical potential at the floating gate in order to enhance the programming efficiency of the memory device. It is thus seen that the control gate is used to control the amount of programming current, while the erase gate enhances programming efficiency.

In another embodiment of the invention, during programming, the source is coupled to ground via a current limiting element. The current limiting element limits the source current to a value between 1 and 5 μA. An example of such an element is a 1 megaohm resistor. This raises the source voltage during programming, thereby increasing the threshold voltage of the transistor due to the back bias effect, thus reducing the amount of drain current permitted to flow during programming. Because of this increase in threshold voltage, the programming current permitted to flow between the source and drain is limited to a value which can be generated by the charge pump. In this embodiment, in addition to coupling a current limiting element between the source and ground, the transistor erase gate voltage is raised, e.g. to about 10 volts. Since the floating gate is capacitively coupled to the erase gate, this has the effect of increasing the floating gate voltage and enhancing programming efficiency. However, in other embodiments, the erase gate is grounded during programming.

In yet another embodiment, the charge pump is coupled to the transistor drain region while the control gate is periodically pulsed. When the control gate voltage is low, the drain voltage rises to about 8 volts. When the control gate is pulsed, the drain is discharged through the floating gate transistor, and when the control gate voltage is low again, the drain region is permitted to charge to 8 volts. As described in greater detail below, repeatedly pulsing the control gate permits one to program the floating gate transistor with a charge pump, even though the charge pump cannot provide more than a few microamps of current. This programming technique can be used while raising the erase gate voltage to enhance programming efficiency or in conjunction with a grounded erase gate.

In one embodiment, the memory device comprises a staggered virtual ground array of split gate floating gate memory cells. The array comprises a set of elongated source/drain regions and a plurality of rows of floating gates, each row of floating gates formed between a pair of source/drain regions. The floating gates are arranged so that in a given row, every other floating gate is adjacent to a first one of the source/drain regions within the pair, and the remaining floating gates within the row are adjacent a second one of the source/drain regions within the pair. Because of this, the array can be constructed in a smaller surface area than would be possible if all of the floating gates in a given row were adjacent the same source/drain region.

In one embodiment, each cell comprises a channel region between a pair of associated source/drain regions. The channel region includes a first portion under the floating gate (and adjacent to one of the source/drain regions within the pair) which is heavily doped, and a second portion adjacent the other source/drain region which is more lightly doped. The first portion of the channel enhances the programming efficiency of the cell, while the low dopant concentration of the second portion of the channel causes the second portion of the channel to exhibit a low threshold voltage. We have discovered a novel method for doping the channel so that the first and second portions of the channel are self-aligned with the edges of the floating gate. This is done by (1) heavily doping the entire channel region, (2) forming the floating gate, and (3) partially counterdoping the portion of the channel that is not under the floating gate, using the floating gate as a mask. This technique improves yields because it is impossible to misalign the first and second portions of the channel with respect to the rest of the transistor.

These and other advantages of the present invention are better understood with reference to the detailed description below.

BRIEF DESCRIPTION OF THE DRAWINGS

FIG. 1 illustrates a circuit for programming a flash EPROM in accordance with a first embodiment of our invention.

FIG. 1*a* illustrates in plan view the flash EPROM of FIG. 1.

FIG. 2 illustrates the drain current versus drain voltage characteristic curve of the transistor of FIG. 1 with a constant control gate voltage.

FIGS. 3*a* to 3*c* are waveform diagrams illustrating the drain voltage, control gate voltage and threshold voltage during programming of the transistor of FIG. 1.

FIG. 4 illustrates a floating gate transistor which is programmed using a method in accordance with a second embodiment of our invention.

FIG. 4*a* schematically illustrates a circuit equivalent to that shown in FIG. 4.

FIG. 5 illustrates the effect of a source bias voltage on drain current.

FIG. 6 illustrates a floating gate transistor in which the source is biased with respect to the substrate during programming by a current limiting circuit element.

FIG. 6*a* illustrates a row of transistors constructed in accordance with the embodiment of FIG. 6.

FIG. 7 illustrates a circuit which provides an MOS transistor equivalent of a resistor which is used as the current limiting circuit element of FIG. 6.

FIG. 8 illustrates a transistor, the drain of which is coupled directly to a charge pump and the control gate of which is coupled to a pulse source.

FIGS. 9*a* and 9*b* illustrate the control gate and drain voltage waveform applied to the floating gate transistor of FIG. 8.

FIGS. 10*a* to 10*d* illustrate a floating gate transistor during a manufacturing process in accordance with our invention.

FIG. 11 illustrates in plan view a flash EPROM array constructed in accordance with our invention.

FIG. 12 illustrates a floating gate transistor coupled to a single gate transistor.

DETAILED DESCRIPTION

FIG. 1 illustrates in cross section a flash EPROM transistor 10 coupled to programming circuitry. Referring to FIG. 1, transistor 10 includes an N+source 12, an N+drain 14, a control gate 16 and a floating gate 18. Transistor 10 is formed within a P+region 20 to enhance the programming efficiency of transistor 10. A P−region 21 is formed in a portion 22 of the channel region to reduce the effective threshold voltage of portion 22. Transistor 10 also includes an erase gate 24 which extends over but is insulated from floating gate 18. Erase gate 24 is outside of the cross section of FIG. 1, but is illustrated in plan view in FIG. 1*a*. The portion of floating gate 18 which extends under erase gate 24 is formed over a field oxide region 26.

If one were to attempt to program transistor 10 by raising the control gate and drain voltages to a programming voltage (e.g., about 8 to 12 volts), the drain current would initially rise to several hundred microamps as the drain voltage increased. The "one-shot" drain current versus drain voltage characteristic curve of transistor 10 is illustrated in FIG. 2. Curve 1 in FIG. 2 indicates that current initially increases (portion 1*a*), but thereafter drops as floating gate 18 is programmed. If the drain voltage keeps increasing, current rises again (portion 1*b*) due to injection induced breakdown between drain 14 and P+region 20. A charge pump cannot economically provide the several hundred microamps required to get past portion 1a of curve 1. However, we have discovered a method for programming transistor 10 without providing such a large drain current.

In accordance with one embodiment of our invention, a control gate voltage waveform as illustrated in FIG. 3*b* is applied to control gate 16 by a circuit 27 while the voltage 12 waveform of FIG. 3*a* is applied to drain 14 by a circuit 28. As can be seen, at or after a time Tl when a programming drain voltage of about 8 volts is applied to drain 14, the voltage at control gate 16 ramps up from 0 to 8 volts over a time period (typically 0.1 to 10 ms, and preferably 1 ms). During a first portion P1 of this 1 ms period, control gate voltage VCG never exceeds a few volts, and the conditions required to draw a drain current of more than 1 μA never exist. However, during portion Pl, electrons are slowly injected onto floating gate 18, and threshold voltage VT slowly starts to rise (FIG. 3*c*).

After portion Pl, control gate voltage VCG continues to increase to 8 volts. However, transistor 10 still does not draw more than 1 μA because threshold voltage VT also continues to increase, and conditions are never created which would permit a large drain current to flow. By the time voltage VCG reaches 8 volts, threshold voltage VT reaches about 8 volts, and transistor 10 is programmed without ever requiring more than 1 μA of drain current.

It will be apparent to those skilled in the art how to build circuits 27 and 28 capable of generating the voltage waveforms of FIGS. 3*a* and 3*b*. Thus circuits 27 and 28 will not be described in further detail herein except to note that the voltage applied to control gate 16 and drain 14 by circuits 27, 28 is derived from a charge pump. Although 8 volts are applied to the transistor of FIG. 1, this value is merely exemplary, and other voltages can also be applied to transistor 10.

Transistor 10 is read in a conventional manner, e.g., by raising the voltage at control gate 16 to about 5 volts, raising the voltage at drain 14 to 1.5 volts, grounding source 12 and erase gate 24, and sensing whether current flows through transistor 10. Transistor 10 is erased by grounding control gate 16, drain 14 and source 12 and raising the erase gate voltage to about 25 volts, thereby causing electrons to tunnel from floating gate 18 to erase gate 24. This leaves floating gate 18 positively charged.

Although the embodiment discussed above functions adequately and comes within the scope of the invention, it does have some drawbacks. For example, different transistors in the array may be programmed at different rates. Assume, for example, that hot electrons reach the floating gate of one of the transistors in the array at a low rate. If the control gate voltage of that transistor increases too rapidly, the transistor will start to draw a large current (e.g. in excess of 100 μA) before charge is injected into its floating gate. Thus, the drain voltage V_D will start to drop, and programming will cease. Accordingly, the ramp rate must be selected to rise as slowly as the programming rate of the slowest transistor in the array allows. (If the ramp rate is too slow, programming will take too long.)

FIG. 4 illustrates another embodiment of my invention. In FIG. 4, during programming a voltage of about 8 volts is applied to drain 14 while a voltage $V_T+\Delta$ is applied to control gate 16, where V_T is the threshold voltage which, if applied to control gate 16, will permit up to 1 μA to flow through portion 22 of transistor channel 30. Δ is an incremental voltage, such that if VT +Δ is applied to control gate 16, several microamps will be 34 permitted to flow through portion 22. (V_T is typically about 1.0 volt, while Δ is about 0.2 volts.) It is thus seen that as long as the control gate voltage is less than or equal to $V_T+\Delta$, the drain current will be less than several microamps, and thus the transistor of FIG. 4 can be programmed using a conventional charge pump.

As is known in the art, the higher the electrical potential at floating gate 18, the greater the programming efficiency of transistor 10. In one embodiment, the electrical potential of floating gate 18 is enhanced by raising the voltage at erase gate 24. Because of capacitive coupling between erase gate 24 and floating gate 18, the increase in erase gate voltage, e.g. to about 10 volts, enhances programming of transistor 10. Of course, the erase gate voltage cannot be raised too high, e.g., greater than 20 volts, or electrons will tunnel off of floating gate 18 and onto erase gate 24.

It should be noted that although control gate 16 is biased such that portion 22 of channel 30 limits current below a few microamps, the voltage drop across portion 22 is only between 2 and 3 volts, even when 8 volts are applied to drain 14. The reason for this is that transistor 10 can be envisioned as two transistors, i.e. a first transistor Q1 (FIG. 4*a*) whose source is source 12, whose channel is channel portion 22, and whose drain is point 32 between channel portions 22 and 34. The drain, channel and source of the second transistor Q2 comprise drain 14, channel portion 34, and point 32, respectively. As point 32 is biased with respect to P+region 20, the back bias effect (also known as the body effect) of second transistor Q2 increases the effective threshold voltage of the second transistor, thus ensuring a large voltage drop between drain 14 and point 32. (The relation between the source-substrate voltage and drain current for a transistor is illustrated in FIG. 5). It is this voltage drop which accelerates hot electrons onto floating gate 18. The enhanced dopant concentration at P+region 20 increases the back bias effect exhibited by second transistor Q2. (The back bias effect is discussed at pages 32 to 43 of "MOS Field-Effect Transistors and Integrated Circuits" by Paul Richman, published by John Wiley and Sons in 1973, incorporated herein by reference.) It is thus seen that the transistor of FIG. 4 is programmed without requiring a large drain current.

Although the embodiment of FIG. 4 functions adequately and comes within the scope of the present invention, it too has several drawbacks. For example, because dopant concentrations, oxide thicknesses and other parameters vary over the wafer surface area, the threshold voltages of the various transistors in the array may vary, and it may be difficult to generate a control gate voltage $V_T+\Delta$ which will permit programming of

the various flash EPROM transistors at an acceptable rate without permitting too much drain current to flow, and thus cause the drain voltage to drop.

In another embodiment, instead of applying about 8 volts to drain 14 and $V_T+\Delta$ to control gate 16, a voltage of about 6 volts is applied to drain 14 and a voltage between about 2.5 and 3.5 volts is applied to control gate 16. This will permit a programming drain current between about 50 μA and 100 μA. It is noted that while it is difficult to economically generate 100 μA from a charge pump which generates 8 volts from a 5 volt. 10% supply, 100 μA can be economically generated from a charge pump which generates 6 volts from a 5 volt±10% supply.

Since this embodiment permits between 50 and 100 μA to flow through the transistor, the voltage at point 32 will be lower in this embodiment than in the embodiment in which $V_T+\Delta$ is applied to control gate 16. Since it is the voltage difference between drain 14 and point 32 which provides electrons with enough energy to reach floating gate 18, the smaller drain voltage in this embodiment is offset by the lower voltage at point 32.

It should be noted that in this embodiment, the control gate voltage need not be regulated as tightly as the embodiment in which $V_T+\Delta$ is applied to control gate 16. Also, in this embodiment, the erase gate voltage is raised, e.g. to a voltage generally less than 15 volts and preferably about 10 volts.

FIG. 6 illustrates another embodiment of our invention in which programming drain current is held below 1 μA automatically without requiring the generation of a control gate voltage within very tight constraints. Referring to FIG. 6, 8 volts are applied to drain 14, about 4 volts are applied to control gate 16, and a current limiter 36 is coupled between source 12 and ground during programming. (During reading and erasing, source 12 is connected directly to ground.) Current limiter 36 is typically a 1 MΩ resistor which limits the amount of current permitted to flow through transistor 10. As current flows through transistor 10, current limiter 36 has the effect of biasing source 12 relative to substrate 38 to generate the above-mentioned back bias effect by virtue of the ohmic voltage drop across the resistor. As the voltage at source 12 reaches about one volt, the back bias effect of first transistor Q1 causes the programming current to drop to about 1 μA. This causes the voltage at point 32 to rise, e.g. to a value between 2 and 3 volts, thereby increasing the back bias effect of transistor Q2. Because of this, transistor Q2 limits the drain current flowing through drain 14, thereby ensuring that point 32 is at a voltage such that the voltage drop between drain 14 and point 32 is sufficient to accelerate hot electrons onto floating gate 18. (Because of the enhanced dopant concentration of channel portion 34, the drain current of transistor Q2 is more sensitive to its source voltage than transistor Q1. P+region 20 is grounded via its electrical connection to grounded substrate 38.)

As electrons are accelerated onto floating gate 18, the threshold voltage of transistor Q2 starts to increase, and the voltage at point 32 starts to decrease, so that the voltage across the source and drain of transistor Q2 increases. This increase in voltage facilitates further injection of hot electrons onto floating gate 18.

As in the embodiment described above in relation to FIG. 4, the erase gate voltage is typically raised during programming, e.g. to about 10 volts, to enhance programming efficiency.

After electrical erase, floating gate 18 is typically positively charged. This positive charge also effectively raises the floating gate electrical potential to further enhance programming efficiency.

As mentioned above, only about 4 volts are applied to control gate 16 during programming. The reason for this is that a transistor in accordance with the embodiment of FIG. 6 is typically part of a row of transistors such as row 40 of FIG. 6a. This row comprises a plurality of source/drain regions 43 to 48, and the junction between each source/drain region and P+region 20 forms a capacitor. If it were desired to program a transistor 42 in row 40 and control gate 16 were raised to a voltage in excess of 6 volts, and all of floating gates 18 in row 40 were positively charged, all of source/drain regions 43 to 45 would be effectively connected to source/drain region 46 (source/drain 46 serves as the source of transistor 42). That would be the equivalent of connecting a very large parasitic capacitance to source/drain region 46, and it would take an unacceptably long amount of time to raise the voltage at region 46 and to program transistor 42. By only raising the control gate voltage to only 4 volts, a resistance between source/drain region 46 and the other source/drain regions to the left of transistor 42 is created to reduce the effect of the above-mentioned parasitic capacitance. (In the embodiment of FIG. 6a, P− region 30a can be an epitaxial layer on a P++substrate 30b.)

Transistor 10 of FIG. 6 is a split gate flash EPROM, meaning that floating gate 18 covers portion 34 of channel 30 but not portion 22. As mentioned above, this is the equivalent of the pair of transistors Q1, Q2 in FIG. 4a. However, split gate transistor 10 has two advantages over an embodiment in which the EPROM cell was actually constructed as two transistors (FIG. 12). First, the transistor of FIG. 6 (and FIGS. 1 and 4) is smaller than transistors Q1 and Q2 of FIG. 12. Second, in FIG. 6, electrons gain energy while travelling from source 12 to point 32, and for at least some of these electrons, this energy can be added to the energy gained by the electrons as they travel through channel portion 34, to enhance programming efficiency. In the embodiment of FIG. 12, any energy gained by electrons moving through the channel of transistor Q1 is completely lost as the electrons move through N+region 50, and this lost energy cannot be used to enhance programming efficiency.

One of the major advantages of the transistor of FIG. 6 is the fact that the transistor is programmed (1) without drawing more than a few microamps of drain current, and (2) without requiring precise regulation of control gate, erase gate or drain voltages. As mentioned above, in the embodiment in which a ramp voltage is applied to the control gate and the embodiment in which $V_T+\Delta$ is applied to the control gate, the control gate voltage has to be precisely controlled to permit programming without drawing too much drain current. In FIG. 6, programming is achieved without having to tightly regulate the control gate voltage.

Instead of using a resistor as current limiter 36, in one embodiment, a circuit 51, comprising a first MOSFET Q3 and a second MOSFET Q4, coupled in a current mirror configuration, provides an MOS equivalent of a resistor (FIG. 7). A resister R is coupled between VCC and the drain of transistor Q3. The effective resistance

R_{EQ} between the source and drain of transistor Q4 is as follows:

$$R_{EQ}=R_1\times(W3/L3)/(W4/L4)$$

where W3, L3, W4 and L4 are the channel width of transistor Q3, the channel length of transistor Q3, the channel width of transistor Q4 and the channel length of transistor Q4, respectively, and R1 is the resistance of resistor R. W3, L3, W4 and L4 are selected so that transistor Q4 exhibits a desired amount of resistance. This effective resistance R_{EQ} is typically within the range of 100 KΩ to 2 MΩ, and preferably about 1 MΩ, to permit a drain current less than about 10 μA to flow through floating gate transistor 10.

In accordance with another embodiment of our invention, drain 14 is coupled to a charge pump 52 while control gate 6 is coupled to a pulse source 54 (FIG. 8). Pulse source 54 provides a stream of pulses having an amplitude of about 5 volts (VCC), an on-time of 0.1 microseconds and an off-time of 0.9 microseconds. (The waveform provided by pulse source 54 is illustrated in FIG. 9*a*.) Of importance, when control gate 16 is at ground, transistor 10 is off, and drain 14 charges to about 8 volts. When control gate 16 is pulsed, drain 14 is discharged through transistor 10, and when control gate 16 is again grounded, drain 14 charges back to about 8 volts. FIG. 9*b* illustrates the drain voltage waveform resulting from coupling charge pump 52 to drain 14 and pulsing control gate 16. The repetitive application of the waveform of FIGS. 9*a* and 9*b* to control gate 16 and drain 14 over about a 1 ms time period is sufficient to program transistor 10, because at least during time periods P2, the voltage conditions are appropriate for accelerating hot electrons onto floating gate 18. (If control gate 16 were not pulsed, drain 14 would remain at a low voltage because the charge pump coupled to drain 14 cannot provide a large output current, and the conditions required for hot electron injection would not exist.)

A novel technique for constructing a flash EPROM transistor 101 (FIG. 10*d*) for use with the above described programming technique is described below.

First, a P− silicon substrate 100 is implanted with P type impurities to form a P+layer 102 approximately 0.8 microns thick and having a dopant concentration of between 10^{17} and $10^{18}/cm^3$ (FIG. 10*a*). An insulating layer 104 (typically thermally grown SiO_2) is formed on the wafer, and a heavily doped polysilicon floating gate 105 is formed on insulating layer 104 in a conventional manner. (During formation of floating gate 105, other floating gates such as floating gates 105' and 105" are formed elsewhere on the surface. The description herein only refers to structures within transistor 101, it being understood that similar structures constituting the rest of a flash EPROM array are formed elsewhere on the wafer.) A photoresist layer 106 is then formed on the wafer and patterned.

Referring to FIG. 10*b*, the wafer is then subjected to an N type ion implantation step to form N+source 108 and drain 110. One edge 108*a* of source 108 and one edge 110*a* of drain 110 are defined by photoresist 106, while the other edge 108*b* of source 108 and edge 110*b* of drain 110 are defined by edges 105*a*' and 105*a* of floating gates 105' and 105, respectively. This is done for reasons described in U.S. Pat. No. 4,639,893, issued to Boaz Eitan, and incorporated herein by reference.

Photoresist layer 106 is removed, and the wafer is subjected to a diffusion step. The wafer is then subjected to a blanket N type ion implantation step to partially counter-dope a portion 114 of P+layer 102, so that portion 114 becomes P− material (see FIG. 10*c*). It will be appreciated that at the conclusion of this process step, the transistor channel will include a first area Al which comprises P− material and a second area A2 which comprises P+material. P+area A2 serves to enhance the transistor programming efficiency, while area Al is P- material so that the effective threshold voltage of area Al is about one volt. Of importance, the lateral extent of areas Al and A2 are self-aligned with the other transistor structures. Thus, it is impossible to misalign the lateral extent of areas Al, A2 and degrade manufacturing yields.

The wafer is then subjected to an oxide etching step (e.g. using HF acid) to remove the exposed portions of insulating layer 104. An additional insulation layer 116 is then formed on the wafer (e.g. by thermal oxidation). Transistor 101 is completed by forming control gate 120 on the wafer using conventional techniques. (See FIG. 10*d*).

The threshold voltage of area A2 when floating gate 105 is electrically neutral is approximately 3 to 5 volts because of the enhanced channel doping concentration. However, transistor 101 is a flash EPROM. Prior to use, charge is removed from floating gate 105 with an erase gate (not shown in FIGS. 10*a* to 10*d*, but described below) prior to use. This reduces the threshold voltage of area A2 below zero volts. (Although this may result in an inversion region forming under floating gate 105 independently of the voltage at control gate 120, this will not create a problem since area Al will only conduct when a high voltage is applied to control gate 120.) Under these circumstances, transistor 101 stores a zero. Floating gate 105 can then be programmed to raise the threshold voltage of area A2 and to thereby store a one in transistor 101.

During the process of constructing control gate 120, an erase gate (not shown in FIGS. 10*a* to 10*d*) is also formed over floating gate 105, typically outside of the cross section of FIGS. 10*a* to 10*d*. The resulting cell may have a layout as illustrated and described in U.S. Pat. application Ser. No. 07/189,874, entitled "EEPROM WITH IMPROVED ERASE STRUCTURE" filed by Eitan et al. on May 3, 1988, incorporated herein by reference. FIG. 11 illustrates a portion of the layout of an array 200 of flash EPROM cells constructed in accordance with an alternative embodiment of the invention. As can be seen, array 200 includes an array of floating gates 202*a* to 202*h*, source/drain regions 204*a* to 204*c*, control gates 206*a* to 206*d*, tunneling erase gates 208*a*, 208*b*, and field oxide regions 209. Array 200 is constructed using a staggered virtual grounded architecture. When it is desired to read or program floating gates 202*a* or 202*b*, source/drain region 204*a* serves as a drain while source/drain region 204*b* serves as a source. When it is desired to read or program floating gates 202*c* or 202*d*, source/drain region 204*b* serves as a drain while source/drain region 204*a* serves as a source. Source/drain regions 204*b*, 204*c* similarly serve as a source or a drain to read or program one of floating gates 202*e* to 202*h*. Control gate 206*a* is used to read or program the floating gates within the column comprising floating gates 202*a* and 202*e*. The other control gates are used to read or program the floating gates within other associated columns of floating gates. Erase gate 208*a* is used to erase floating gates 202*a*, 202*c*, 202*e*

and 202*g*, while erase gate 208*b* is used to erase floating gates 202*b*, 202*d*, 202*f* and 202*h*.

In the array of FIG. 11, the floating gates are staggered relative to one another, i.e. the floating gates 202*a* is formed against source/drain region 204*a* while adjacent floating gate 202*c* is formed against source/drain region 202*b*. If floating gates 202*a* and 202*c* were both formed against source/drain region 204*a*, the cell size would have to be increased to permit both floating gates 202*a* and 202*c* to extend underneath erase gate 208*a*. Thus, staggering the floating gates permits the flash EPROM array to be constructed on a small surface area.

An address decoder appropriate for use with the array of FIG. 11 is discussed in U.S. patent application Ser. No. 07/258,926, filed on Oct. 17, 1988 by Syed Ali and incorporated herein by reference. Also see U.S. patent application Ser. No. 07/258,952, filed by Eitan et al. on Oct. 17, 1988.

While the invention has been described with regard to specific embodiments, those skilled in the art will recognize that changes can be made in form and detail without departing from the spirit and scope of the invention. Accordingly, all such changes come within the present invention.

I claim:

1. A method for programming a floating gate transistor, said floating gate transistor comprising a source, a drain spaced apart from said source, said source and drain being of a first conductivity type and formed in a semiconductor region of a second conductivity type, a channel extending between said source and drain, a floating gate extending over at least a portion of said channel, and a control gate extending over at least a portion of said floating gate, said method comprising the steps of:

applying a programming voltage to said drain and control gate sufficient to cause hot electron injection programming of said transistor; and

ensuring that the programming drain current for said transistor is less than a predetermined value.

2. Method of claim 1 wherein said predetermined value is less than or equal to about 150 μA.

3. Method of claim 1 wherein said predetermined value is less than or equal to about 10 μA.

4. Method of claim 1 wherein said programming drain voltage is provided by a charge pump and said programming drain current is held to a value sufficiently low so that said charge pump can provide said programming drain current.

5. A method for programming a floating gate transistor, said floating gate transistor comprising a source, a drain spaced apart from said source, said source and drain being of a first conductivity type and formed in a semiconductor region of a second conductivity type, a channel extending between said source and drain, a floating gate extending over at least a portion of said channel, and a control gate extending over at least a portion of said floating gate, said method comprising the steps of:

applying a programming voltage to said drain; and

applying to said control gate a voltage which rises from a first value to a second value such that during the time said voltage at said control gate is rising, electrons are being injected into said floating gate so that the threshold voltage of said transistor increases at a rate which ensures that said transistor does not draw a drain current over a predetermined value during programming.

6. Method of claim 5 wherein the voltage applied to said control gate rises from said first value to said second value over a 0.1 ms time period.

7. Method of claim 1 wherein said step of ensuring comprises the step of applying a voltage to said control gate to keep the drain current below said predetermined value.

8. Method of claim 7 wherein said transistor comprises an erase gate capacitively coupled to said floating gate, said method further comprising the step of raising the voltage at said erase gate.

9. A method for programming a floating gate transistor, said floating gate transistor comprising a source, a drain spaced apart from said source, said source and drain being of a first conductivity type and formed in a semiconductor region of a second conductivity type, a channel extending between said source and drain, a floating gate extending over at least a portion of said channel, and a control gate extending over at least a portion of said floating gate, said method comprising the steps of:

applying a programming voltage to said drain and control gate;

ensuring that the programming drain current is less than a predetermined value, wherein said step of ensuring comprises the step of providing an electrical resistance between said source and ground, said semiconductor region being grounded.

10. Method of claim 9 wherein said transistor comprises an erase gate capacitively coupled to said floating gate, said method further comprising the step of raising the voltage at said erase gate.

11. Method of claim 10 wherein the voltage at said erase gate is greater than 5 volts during programming.

12. Method of claim 10 wherein the voltage at said erase gate is less than 20 volts during programming.

13. Method of claim 9 wherein said transistor comprises an erase gate which is grounded during programming.

14. A method for programming a floating gate transistor, said transistor comprising a source, a drain, a channel extending between said source and drain, a floating gate extending over at least a portion of said channel, and a control gate extending over at least a portion of said floating gate, said method comprising the steps of:

coupling a programming voltage generator to said drain; and

repetitively applying pulses to said control gate to thereby cause hot electron injection programming of said transistor.

15. Method of claim 14 wherein said programming voltage generator is a charge pump.

16. Method of claim 14 wherein said transistor comprises an erase gate capacitively coupled to said floating gate, said method further comprising the step of raising the voltage at said erase gate.

17. Method of claim 14 wherein said transistor comprises an erase gate which is grounded during programming.

18. A method for programming a floating gate transistor, said transistor comprising a source, a drain, a channel extending between said source and drain, a floating gate extending over at least a portion of said channel, and a control gate extending over at least a

portion of said floating gate, said method comprising the steps of:

coupling a programming voltage generator to said drain; and

repetitively applying pulses to said control gate,

wherein said programming voltage generator is incapable of generating an output current which said transistor would normally conduct if (1) said floating gate were unprogrammed, (2) said programming drain voltage was applied to said drain, and (3) a programming control gate voltage equal to the amplitude of said pulses was applied to said control gate.

19. Structure comprising:

a floating transistor including a source, a drain spaced apart from said source, said source and drain being of a first conductivity type and formed in a semiconductor region of a second conductivity type, a channel extending between said source and drain, a floating gate extending over at least a portion of said channel, and a control gate extending over at least a portion of said floating gate; and

means for applying a programming voltage to said drain and control gate to thereby program said transistor by hot electron injection and ensuring that the programming drain current of said transistor is less than a predetermined value.

20. Structure of claim 19 wherein said predetermined value is less than or equal to about 150 μA.

21. Structure comprising:

a floating gate transistor including a source, a drain spaced apart from said source, said source and drain being of a first conductivity type and formed in a semiconductor region of a second conductivity type, a channel extending between said source and drain, a floating gate extending over at least a portion of said channel, and control gate extending over at least a portion of said floating gate; and

means for applying a programming voltage to said drain and control gate and ensuring that the programming drain current is less than a predetermined value, wherein said means for applying applies to said control gate a voltage which rises from a first value to a second value such that during the time said voltage at said control gate is rising, electrons are being injected into said floating gate so that the threshold voltage of said transistor increases at a rate which ensures that said transistor does not draw a drain current over said predetermined value during programming.

22. Structure of claim 21 wherein said means for applying causes the voltage at said control gate to rise from said first value to said second value over a 0.1 ms time period.

23. Structure of claim 19 wherein an electrical resistance is provided between said source and ground to prevent said programming drain current from exceeding said predetermined value, said semiconductor region being grounded.

24. Structure of claim 23 wherein said transistor comprises an erase gate capacitively coupled to said floating gate, and said means for applying also raises the voltage at said erase gate during programming.

25. Structure comprising:

a floating gate transistor including a source, a drain, a channel extending between said source and drain, a floating gate extending over at least a portion of said channel, and a control gate extending over at least a portion of said floating gate; and

means for repetitively applying pulses to said control gate to thereby program said floating gate transistor by hot electron injection.

26. Structure of claim 25 wherein said transistor includes an erase gate capacitively coupled to said floating gate, said structure further comprising means for raising the voltage at said erase gate during programming.

27. Method of claim 1 wherein said step of ensuring comprises the step of coupling the source of said floating gate transistor to an additional transistor, said additional transistor limiting the current of said floating gate transistor during programming.

28. Method of claim 1 wherein said step of ensuring comprises the step of raising the source voltage of said floating gate transistor during programming.

29. Structure of claim 19 further comprising means for increasing the voltage at said source during programming.

30. Structure of claim 19 further comprising an additional transistor coupled to the source of said floating gate transistor, said additional transistor limiting the programming current of said floating gate transistor.

* * * * *