US011839086B2

(12) **United States Patent**
Petti et al.

(10) **Patent No.: US 11,839,086 B2**
(45) **Date of Patent: Dec. 5, 2023**

(54) **3-DIMENSIONAL MEMORY STRING ARRAY OF THIN-FILM FERROELECTRIC TRANSISTORS**

(71) Applicant: **SUNRISE MEMORY CORPORATION**, San Jose, CA (US)

(72) Inventors: **Christopher J. Petti**, Mountain View, CA (US); **Vinod Purayath**, Sedona, AZ (US); **George Samachisa**, Atherton, CA (US); **Wu-Yi Henry Chien**, San Jose, CA (US); **Eli Harari**, Saratoga, CA (US)

(73) Assignee: **SUNRISE MEMORY CORPORATION**, San Jose, CA (US)

(*) Notice: Subject to any disclaimer, the term of this patent is extended or adjusted under 35 U.S.C. 154(b) by 0 days.

(21) Appl. No.: **17/812,375**

(22) Filed: **Jul. 13, 2022**

(65) **Prior Publication Data**

US 2023/0027837 A1 Jan. 26, 2023

Related U.S. Application Data

(60) Provisional application No. 63/222,926, filed on Jul. 16, 2021.

(51) **Int. Cl.**

H10B 51/30	(2023.01)
G11C 11/22	(2006.01)
H10B 51/20	(2023.01)

(52) **U.S. Cl.**
CPC ***H10B 51/30*** (2023.02); ***G11C 11/223*** (2013.01); ***G11C 11/2273*** (2013.01); ***G11C 11/2275*** (2013.01); ***H10B 51/20*** (2023.02)

(58) **Field of Classification Search**
CPC H10B 51/30; H10B 51/20; H10B 51/50; H10B 51/10; G11C 11/223;
(Continued)

(56) **References Cited**

U.S. PATENT DOCUMENTS

4,213,139	A	7/1980	Rao
4,984,153	A	1/1991	Kregness et al.

(Continued)

FOREIGN PATENT DOCUMENTS

CN	111799263	A	*	10/2020
JP	1998-269789	A2		10/1998

(Continued)

OTHER PUBLICATIONS

"PCT Search Report and Written Opinion, PCT/US2022/039473", dated Dec. 6, 2022, 14 pages.

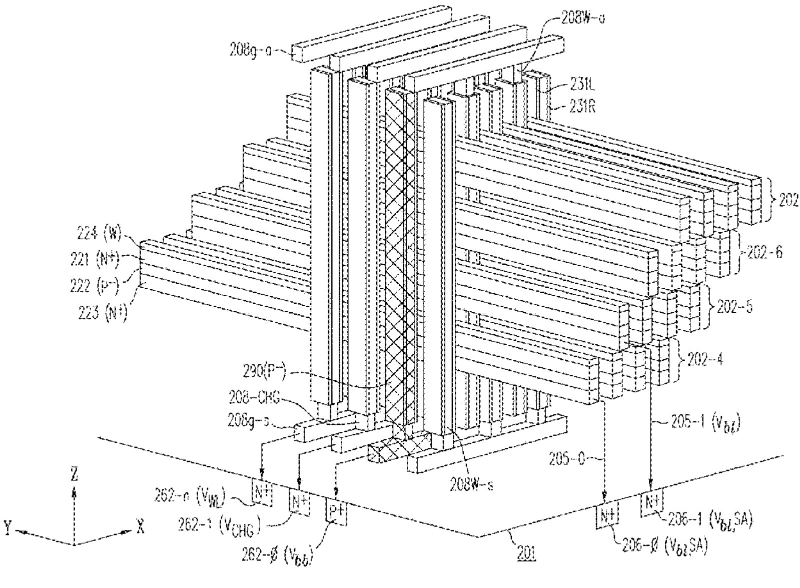
(Continued)

Primary Examiner — Fernando Hidalgo

(57) **ABSTRACT**

Thin-film Ferroelectric field-effect transistor (FeFET) may be organized as 3-dimensional NOR memory string arrays. Each 3-dimensional NOR memory string array includes a row of active stack each including a predetermined number of active strips each provided one on top of another and each being spaced apart from another by an isolation layer. Each active strip may include a shared source layer and a shared drain layer shared by the FeFETs provided along the active strip. Data storage in the active strip is provided by ferroelectric elements that can individually electrically set into one of two polarization states. FeFETs on separate active strips may be configured for read, programming or erase operations in parallel.

66 Claims, 49 Drawing Sheets



(58) **Field of Classification Search**
CPC G11C 11/2273; G11C 11/2275; G11C 11/2259; H01L 29/78391; H01L 29/40111; H01L 29/516
USPC 365/145
See application file for complete search history.

(56) References Cited

U.S. PATENT DOCUMENTS

5,388,246	A	2/1995	Kasai
5,583,808	A	12/1996	Brahmbhatt
5,646,886	A	7/1997	Brahmbhatt
5,656,842	A	8/1997	Iwamatsu
5,768,192	A	6/1998	Eitan
5,789,776	A	8/1998	Ancaster et al.
5,880,993	A	3/1999	Kramer et al.
5,915,167	A	6/1999	Leedy
6,040,605	A	3/2000	Sano et al.
6,049,497	A	4/2000	Yero
6,057,862	A	5/2000	Margulis
6,107,133	A	8/2000	Furukawa et al.
6,118,171	A	9/2000	Davies et al.
6,130,838	A	10/2000	Kim et al.
6,313,518	B1	11/2001	Ahn et al.
6,314,046	B1	11/2001	Kamiya et al.
6,362,508	B1	3/2002	Rasovaky et al.
6,396,744	B1	5/2002	Wong
6,434,053	B1	8/2002	Fujiwara
6,580,124	B1	6/2003	Cleeves et al.
6,587,365	B1	7/2003	Salling
6,627,503	B2	9/2003	Ma
6,744,094	B2	6/2004	Forbes
6,754,105	B1	6/2004	Chang et al.
6,774,458	B2	8/2004	Fricke et al.
6,873,004	B1	3/2005	Han et al.
6,881,994	B2	4/2005	Lee et al.
6,946,703	B2	9/2005	Ryu et al.
7,005,350	B2	2/2006	Walker et al.
7,177,977	B2	2/2007	Chen et al.
7,223,653	B2	5/2007	Cheng et al.
7,307,308	B2	12/2007	Lee
7,426,141	B2	9/2008	Takeuchi
7,465,980	B2	12/2008	Arimoto et al.
7,489,002	B2	2/2009	Forbes et al.
7,495,963	B2	2/2009	Edahiro et al.
7,512,012	B2	3/2009	Kuo
7,524,725	B2	4/2009	Chung
7,542,348	B1	6/2009	Kim
7,612,411	B2	11/2009	Walker
7,709,359	B2	5/2010	Boes et al.
7,804,145	B2	9/2010	Shimizu et al.
7,872,295	B2	1/2011	Park et al.
7,876,614	B2	1/2011	Kang et al.
7,898,009	B2	3/2011	Wilson et al.
8,026,521	B1	9/2011	Or-Bach et al.
8,139,418	B2	3/2012	Carman
8,178,396	B2	5/2012	Sinha et al.
8,237,213	B2	8/2012	Liu
8,242,504	B2	8/2012	Kim
8,278,183	B2	10/2012	Lerner
8,304,823	B2	11/2012	Boescke
8,383,482	B2	2/2013	Kim et al.
8,395,942	B2	3/2013	Samachisa et al.
8,604,618	B2	12/2013	Cooney et al.
8,630,114	B2	1/2014	Lue
8,743,612	B2	6/2014	Choi et al.
8,767,436	B2	7/2014	Scalia et al.
8,767,473	B2	7/2014	Shim et al.
8,848,425	B2	9/2014	Schloss
8,878,278	B2	11/2014	Alsmeier et al.
9,053,801	B2	6/2015	Sandhu et al.
9,053,802	B2	6/2015	Mueller
9,105,310	B2	8/2015	Li et al.
9,158,622	B2	10/2015	Lee et al.
9,190,293	B2	11/2015	Wang et al.
9,202,694	B2	12/2015	Konevecki et al.
9,219,225	B2	12/2015	Karda et al.
9,230,985	B1	1/2016	Wu et al.
9,231,206	B2	1/2016	Tao et al.
9,263,577	B2	2/2016	Ramaswamy et al.
9,281,044	B2 *	3/2016	Ramaswamy H01L 29/78391
9,299,580	B2	3/2016	Kong et al.
9,337,210	B2	5/2016	Karda et al.
9,362,487	B2	6/2016	Inumiya et al.
9,391,084	B2	7/2016	Lue
9,412,752	B1	8/2016	Yeh et al.
9,455,268	B2	9/2016	Oh et al.
9,472,560	B2	10/2016	Ramaswamy et al.
9,530,794	B2	12/2016	Ramaswamy et al.
9,558,804	B2	1/2017	Mue
9,620,605	B2	4/2017	Liang et al.
9,633,944	B2	4/2017	Kim
9,698,152	B2	7/2017	Peri
9,711,529	B2	7/2017	Hu et al.
9,748,172	B2	8/2017	Takaki
9,786,684	B2	10/2017	Ramaswamy et al.
9,799,761	B2	10/2017	Or-Bach et al.
9,818,468	B2	11/2017	Müller
9,818,848	B2	11/2017	Sun et al.
9,830,969	B2	11/2017	Slesazeck et al.
9,842,651	B2	12/2017	Harari
9,865,680	B2	1/2018	Okumura et al.
9,875,784	B1	1/2018	Li et al.
9,876,018	B2	1/2018	Chavan et al.
9,892,800	B2	2/2018	Harari
9,911,497	B1	3/2018	Harari
9,941,299	B1	4/2018	Chen et al.
10,014,317	B2	7/2018	Peng
10,038,092	B1	7/2018	Chen et al.
10,043,567	B2	8/2018	Slesazeck et al.
10,056,393	B2	8/2018	Schroeder et al.
10,074,667	B1	9/2018	Higashi
10,090,036	B2	10/2018	Van Houdt
10,096,364	B2	10/2018	Harari
10,121,553	B2	11/2018	Harari et al.
10,157,780	B2	12/2018	Wu et al.
10,211,223	B2	2/2019	Van Houdt et al.
10,211,312	B2	2/2019	Van Houdt et al.
10,217,667	B2	2/2019	Or-Bach et al.
10,249,370	B2	4/2019	Harari
10,254,968	B1	4/2019	Gazit et al.
10,283,493	B1	5/2019	Nishida
10,319,696	B1	6/2019	Nakano
10,355,121	B2	7/2019	Or-Bach et al.
10,373,956	B2	8/2019	Gupta et al.
10,381,370	B2	8/2019	Shin et al.
10,381,378	B1	8/2019	Harari
10,395,737	B2	8/2019	Harari
10,403,627	B2	9/2019	Van Houdt et al.
10,418,377	B2	9/2019	Van Houdt et al.
10,424,379	B2	9/2019	Slesazeck et al.
10,431,596	B2	10/2019	Herner et al.
10,438,645	B2	10/2019	Muell et al.
10,460,788	B2	10/2019	Mueller
10,475,812	B2	11/2019	Harari
10,510,773	B2	12/2019	Ramaswamy et al.
10,600,808	B2	3/2020	Schröder et al.
10,608,008	B2	3/2020	Harari et al.
10,608,011	B2	3/2020	Harari et al.
10,622,051	B2	4/2020	Mueller et al.
10,622,377	B2	4/2020	Harari et al.
10,636,471	B2	4/2020	Ramaswamy et al.
10,650,892	B2	5/2020	Noack
10,651,153	B2	5/2020	Fastow et al.
10,651,182	B2	5/2020	Morris et al.
10,651,196	B1	5/2020	Sharangpani et al.
10,692,837	B1	6/2020	Nguyen et al.
10,692,874	B2	6/2020	Harari et al.
10,700,093	B1	6/2020	Kalitsov et al.
10,720,437	B2	7/2020	Yoo
10,825,834	B1	11/2020	Chen
10,872,905	B2	12/2020	Muel
10,879,269	B1	12/2020	Zhang et al.
10,896,711	B2	1/2021	Lee et al.
10,937,482	B2	3/2021	Sharma et al.

(56) References Cited

U.S. PATENT DOCUMENTS

10,950,616	B2	3/2021	Harari et al.
11,043,280	B1	6/2021	Prakash et al.
11,171,157	B1	11/2021	Lai et al.
11,411,025	B2	8/2022	Lai et al.
2001/0030340	A1	10/2001	Fujiwara
2001/0053092	A1	12/2001	Kosaka et al.
2002/0012271	A1	1/2002	Forbes
2002/0028541	A1	3/2002	Lee et al.
2002/0051378	A1	5/2002	Ohsawa
2002/0193484	A1	12/2002	Albee
2003/0038318	A1	2/2003	Forbes
2004/0214387	A1	10/2004	Madurawe et al.
2004/0246807	A1	12/2004	Lee
2004/0262681	A1	12/2004	Masuoka et al.
2004/0262772	A1	12/2004	Ramanathan et al.
2004/0264247	A1	12/2004	Kim
2005/0128815	A1	6/2005	Ishikawa et al.
2005/0236625	A1	10/2005	Schuele et al.
2005/0280061	A1	12/2005	Lee
2006/0001083	A1	1/2006	Bhattacharyya
2006/0080457	A1	4/2006	Hiramatsu et al.
2006/0155921	A1	7/2006	Gorobets et al.
2006/0212651	A1	9/2006	Ashmore
2006/0261404	A1	11/2006	Forbes
2007/0012987	A1	1/2007	McTeer
2007/0023817	A1	2/2007	Dao
2007/0045711	A1	3/2007	Bhattacharyya
2007/0134876	A1	6/2007	Lai et al.
2007/0192518	A1	8/2007	Rupanagunta et al.
2008/0022026	A1	1/2008	Yang et al.
2008/0054346	A1	3/2008	Saitoh et al.
2008/0160765	A1	7/2008	Lee et al.
2008/0173930	A1	7/2008	Watanabe
2008/0178794	A1	7/2008	Cho et al.
2008/0212358	A1*	9/2008	Mitsui H01L 21/76849 438/3
2008/0239812	A1	10/2008	Naofumi et al.
2008/0266960	A1	10/2008	Kuo
2008/0291723	A1	11/2008	Wang et al.
2008/0301359	A1	12/2008	Smith
2009/0057722	A1	3/2009	Masuoka et al.
2009/0140318	A1	6/2009	Dong
2009/0157946	A1	6/2009	Arya
2009/0237996	A1	9/2009	Kirsch et al.
2009/0268519	A1	10/2009	Ishii
2009/0279360	A1	11/2009	Peter et al.
2009/0290442	A1	11/2009	Rajan
2009/0316487	A1	12/2009	Lee et al.
2010/0013001	A1	1/2010	Cho et al.
2010/0121994	A1	5/2010	Kim et al.
2010/0124116	A1	5/2010	Takashi et al.
2010/0128509	A1	5/2010	Kim et al.
2010/0148215	A1	6/2010	Schulze et al.
2010/0207185	A1	8/2010	Lee et al.
2010/0219392	A1	9/2010	Awaya et al.
2010/0254191	A1	10/2010	Son et al.
2010/0327413	A1	12/2010	Lee et al.
2011/0003418	A1	1/2011	Sakata et al.
2011/0044113	A1	2/2011	Kim
2011/0047325	A1	2/2011	Mishima
2011/0115011	A1	5/2011	Masuoka et al.
2011/0134705	A1	6/2011	Jones et al.
2011/0143519	A1	6/2011	Lerner
2011/0170266	A1	7/2011	Haensh et al.
2011/0208905	A1	8/2011	Shaeffer et al.
2011/0298013	A1	12/2011	Hwang et al.
2011/0310683	A1	12/2011	Gorobets
2012/0063223	A1	3/2012	Lee
2012/0146126	A1	6/2012	Lai et al.
2012/0182801	A1	7/2012	Lue
2012/0208347	A1	8/2012	Hiwang et al.
2012/0223380	A1	9/2012	Lee et al.
2012/0243314	A1	9/2012	Takashi
2012/0307568	A1	12/2012	Banna et al.
2012/0327714	A1	12/2012	Lue
2013/0031325	A1	1/2013	Nakamoto et al.
2013/0126957	A1	5/2013	Higashitani et al.
2013/0256780	A1	10/2013	Kai et al.
2014/0015036	A1	1/2014	Fursin et al.
2014/0040698	A1	2/2014	Loh et al.
2014/0070289	A1	3/2014	Tanaka et al.
2014/0070290	A1	3/2014	Inumiya et al.
2014/0075135	A1	3/2014	Choi et al.
2014/0112075	A1	4/2014	Dunga et al.
2014/0117366	A1	5/2014	Saitoh
2014/0151774	A1	6/2014	Rhie
2014/0173017	A1	6/2014	Takagi et al.
2014/0213032	A1	7/2014	Kai et al.
2014/0229131	A1	8/2014	Cohen et al.
2014/0247674	A1	9/2014	Karda et al.
2014/0328128	A1	11/2014	Louie et al.
2014/0340952	A1	11/2014	Ramaswamy et al.
2014/0355328	A1	12/2014	Muller et al.
2015/0054507	A1	2/2015	Gulaka et al.
2015/0098272	A1	4/2015	Kasolra et al.
2015/0113214	A1	4/2015	Sutardja
2015/0155876	A1	6/2015	Jayasena et al.
2015/0194440	A1	7/2015	Noh et al.
2015/0220463	A1	8/2015	Fluman et al.
2015/0249143	A1	9/2015	Sano
2015/0263005	A1	9/2015	Zhao et al.
2015/0340371	A1	11/2015	Lui
2015/0372099	A1	12/2015	Chen et al.
2016/0013156	A1	1/2016	Zhai et al.
2016/0019951	A1	1/2016	Park et al.
2016/0035711	A1	2/2016	Hu
2016/0086970	A1	3/2016	Peng
2016/0118404	A1	4/2016	Peng
2016/0141294	A1	5/2016	Peri et al.
2016/0225860	A1	8/2016	Karda et al.
2016/0276360	A1	9/2016	Doda et al.
2016/0300724	A1	10/2016	Levy et al.
2016/0314042	A1	10/2016	Plants
2016/0321002	A1	11/2016	Jung et al.
2016/0358934	A1	12/2016	Lin et al.
2017/0053906	A1	2/2017	Or-Bach et al.
2017/0062456	A1	3/2017	Sugino et al.
2017/0092370	A1	3/2017	Harari
2017/0092371	A1	3/2017	Harari
2017/0140807	A1	5/2017	Sun et al.
2017/0148517	A1	5/2017	Harari
2017/0148810	A1	5/2017	Kai et al.
2017/0194341	A1	7/2017	Yamada
2017/0213731	A1	7/2017	Yoon et al.
2017/0213821	A1	7/2017	Or-Bach et al.
2017/0358594	A1	12/2017	Lu et al.
2018/0006044	A1	1/2018	Chavan et al.
2018/0095127	A1	4/2018	Pappu et al.
2018/0108416	A1	4/2018	Harari
2018/0151419	A1	5/2018	Wu et al.
2018/0261613	A1	9/2018	Ariyoshi et al.
2018/0269229	A1	9/2018	Or-Bach et al.
2018/0286918	A1	10/2018	Bandyopadhyay et al.
2018/0330791	A1	11/2018	Li et al.
2018/0331042	A1	11/2018	Manusharow et al.
2018/0342455	A1	11/2018	Nosho et al.
2018/0342544	A1	11/2018	Lee et al.
2018/0366471	A1	12/2018	Harari et al.
2018/0366485	A1	12/2018	Harari
2018/0366489	A1	12/2018	Harari et al.
2018/0374929	A1	12/2018	Yoo
2019/0006009	A1	1/2019	Harari
2019/0006015	A1	1/2019	Norman et al.
2019/0019564	A1	1/2019	Li et al.
2019/0067327	A1	2/2019	Herner et al.
2019/0148286	A1	5/2019	Or-Bach et al.
2019/0157296	A1	5/2019	Harari et al.
2019/0180821	A1	6/2019	Harari
2019/0206890	A1	7/2019	Harari et al.
2019/0214077	A1	7/2019	Oh et al.
2019/0237470	A1	8/2019	Mine et al.
2019/0238134	A1	8/2019	Lee et al.
2019/0244971	A1	8/2019	Harari
2019/0259769	A1	8/2019	Karda et al.

(56) **References Cited**

U.S. PATENT DOCUMENTS

2019/0303042	A1	10/2019	Kim et al.	
2019/0304988	A1	10/2019	Dong et al.	
2019/0325945	A1	10/2019	Linus	
2019/0325964	A1	10/2019	Harari	
2019/0319044	A1	11/2019	Harari	
2019/0348424	A1	11/2019	Karda et al.	
2019/0355747	A1	11/2019	Herner et al.	
2019/0370117	A1	12/2019	Fruchtman et al.	
2020/0020718	A1	1/2020	Harari et al.	
2020/0051990	A1	2/2020	Harari et al.	
2020/0063263	A1	2/2020	Yang et al.	
2020/0065647	A1	2/2020	Mulaosmanovic et al.	
2020/0075631	A1	3/2020	Dong	
2020/0098738	A1	3/2020	Herner et al.	
2020/0098779	A1	3/2020	Cernea et al.	
2020/0098881	A1*	3/2020	Vellianitis	H01L 29/66628
2020/0176468	A1	6/2020	Herner et al.	
2020/0201718	A1	6/2020	Richter et al.	
2020/0203378	A1	6/2020	Harari et al.	
2020/0219572	A1	7/2020	Harari	
2020/0243486	A1	7/2020	Quader et al.	
2020/0258897	A1	8/2020	Yan et al.	
2020/0350324	A1*	11/2020	Hoffmann	H10B 53/30
2020/0357453	A1	11/2020	Slesazeck et al.	
2020/0357455	A1	11/2020	Noack et al.	
2020/0357470	A1	11/2020	Noack	
2020/0357822	A1*	11/2020	Chen	G11C 11/223
2020/0365609	A1	11/2020	Harari	
2020/0388313	A1	12/2020	Cho et al.	
2020/0388711	A1	12/2020	Doyle et al.	
2020/0403002	A1	12/2020	Harari	
2020/0411533	A1	12/2020	Alsmeier et al.	
2021/0005238	A1	1/2021	Mueller	
2021/0013224	A1	1/2021	Purayath et al.	
2021/0020659	A1	1/2021	Chen	
2021/0066502	A1	3/2021	Karda et al.	
2021/0074725	A1	3/2021	Lue	
2021/0074726	A1	3/2021	Lue	
2021/0111179	A1	4/2021	Shivaraman et al.	
2021/0175251	A1	6/2021	Zhang et al.	
2021/0247910	A1	8/2021	Kim et al.	
2021/0248094	A1	8/2021	Norman et al.	
2021/0265308	A1	8/2021	Norman et al.	
2021/0375933	A1*	12/2021	Lu	H01L 29/40111
2022/0028876	A1	1/2022	Purayath et al.	
2022/0028886	A1	1/2022	Pur et al.	
2022/0139933	A1*	5/2022	Noack	G11C 11/221 365/145
2022/0231049	A1*	7/2022	Lin	H10B 51/20
2022/0246777	A1*	8/2022	Kim	B32B 17/10
2022/0254390	A1*	8/2022	Gans	G11C 5/04
2022/0351776	A1*	11/2022	Nam	G11C 11/223
2022/0384459	A1	12/2022	Lu et al.	
2022/0393031	A1*	12/2022	Ando	H01L 29/66484
2023/0052477	A1*	2/2023	Ha	H01L 29/0847

FOREIGN PATENT DOCUMENTS

JP	2000-243972	A2	9/2000
JP	2000339978	A	12/2000
JP	2004079606	A	3/2004
JP	2006099827	A1	4/2006
JP	2009206451	A1	9/2009
JP	2010108522	A1	5/2010
JP	2010251572	A	11/2010
JP	2011028540	A1	2/2011
JP	2012-150876	A2	8/2012
KR	20120085591	A	8/2012
KR	20120085591	A1	8/2012
KR	20120085603	A	8/2012
WO	2015025357	A1	2/2015
WO	2018236937	A1	12/2018
WO	2019066948	A1	4/2019

OTHER PUBLICATIONS

Lee, Sang Yeol, "Comprehensive Review on Amorphous Oxide Semiconductor Thin Film Transistor", Trans. Electr. Electron. Mater. 21, https://doi.org/10.1007/s42341-020-00197-w, Mar. 28, 2020, pp. 235-248.

Lee, Chea-Young, et al., "Etching Characteristics and Changes in Surface Properties of IGZO Thin Films by O2 Addition in CF4/Ar Plasma", Coatings. 2021, 11(8):906. https://doi.org/10.3390/coatings11080906, Jul. 29, 2021.

Liao, P.J., et al., "Characterization of Fatigue and Its Recovery Behavior in Ferroelectric HfZrO", 2021 Symposium on VLSI Technology, Jun. 2021, 2 p.

Liao, C.Y., et al., "Multibit Ferroelectric FET Based on Nonidentical Double HfZrO2 for High-Density Nonvolatile Memory", IEEE Electron Device Letters, vol. 42, No. 4, doi: 10.1109/LED.2021.3060589., Apr. 2021, pp. 617-620.

Lue, H.T., et al., "A Highly Scalable 8-Layer 3D Vertical-gate {VG) TFT NANO Flash Using Junction-Free Buried Channel BE-SONOS Device", Symposium on VLSI: Tech. Dig. Of Technical Papers, 2010, pp. 131-132.

Materano, Monica, et al., "Interplay between oxygen defects and dopants: effect on structure and performance of HfO2-based ferroelectrics", Inorg. Chem. Front. (8), Apr. 9, 2021, pp. 2650-2672.

Mo, Fei, et al., "Experimental Demonstration of Ferroelectric HfO2 FET with Ultrathin-body IGZO for High-Density and Low-Power Memory Application", 2019 Symposium on VLSI Technology, 3, doi: 10.23919/VLSIT.2019.8776553., 2019, pp. T42-T4.

Mo, Fei, et al., "Experimental Demonstration of Ferroelectric HfO2 FET with Ultrathin-body IGZO for High-Density and Low-Power Memory Application", Symposium on VLSI Technology, 2019, pp. T42-T43, doi: 10.23919/VLSIT.2019.8776553., Jul. 13, 2020, 7 pages.

Mo, Fei, et al., "Low-Voltage Operating Ferroelectric FET with Ultrathin IGZO Channel for High-Density Memory Application", IEEE Journal of the Electron Devices Society, vol. 8, doi: 10.1109/JEDS.2020.3008789., Jul. 13, 2020, pp. 717-723.

Mueller, S., et al., "Development Status of Gate-First FeFET Technology", 2021 Symposium on VLSI Technology, Jun. 2021, 2 pages.

Mueller, J., et al., "Ferroelectric Hafnium Oxide Based Materials and Devices: Assessment of Current Status and Future Prospects", ECS Journal of Solid State Science and Technology, 4 (5) N30-N35 (2015), Feb. 21, 2015, pp. N29-N35.

Mueller, J., et al., "Ferroelectric hafnium oxide: A CMOS-compatible and highly scalable approach to future ferroelectric memories", IEEE International Electron Devices Meeting, Washington, DC, USA, doi: 10.1109/IEDM.2013.6724605., 2013, pp. 10.8.1-10.8.4.

Mueller, Stephen, et al., "From MFM Capacitors Toward Ferroelectric Transistors: Endurance and Disturb Characteristics of HfO2-Based FeFET Devices", IEEE Transactions on Electron Devices, vol. 60, No. 12, doi: 10.1109/TED.2013.2283465., Dec. 2013, pp. 4199-4205.

Müller, J., et al., "Ferroelectric hafnium oxide: A CMOS-compatible and highly scalable approach to future ferroelectric memories", 2013 IEEE International Electron Devices Meeting, Washington, DC, USA, doi: 10.1109/IEDM.2013.6724605., 2013, pp. 10.8.1-10.8.4.

Nguyen, Manh-Cuong, et al., "Wakeup-Free and Endurance-Robust Ferroelectric Field-Effect Transistor Memory Using High Pressure Annealing", IEEE Electron Device Letters, vol. 42, No. 9, doi: 10.1109/LED.2021.3096248., Sep. 2021, pp. 1295-1298.

On, Nuri , et al., "Boosting carrier mobility and stability in indium-zinc-tin oxide thin-film transistors through controlled crystallization", Sci Rep 10, 18868, https://doi.org/10.1038/s41598-020-76046-w, 2020, 16 pages.

Onuki, Tatsuya , et al., "Embedded memory and ARM Cortex-MO core using 60-nm C-axis aligned crystalline indium-gallium-zinc oxide FET integrated with 65-nm Si CMOS", IEEE Journal of Solid-State Circuits, vol. 52, No. 4,, Apr. 2017, pp. 925-932.

(56) References Cited

OTHER PUBLICATIONS

Park, Goon-Ho , et al., "Electrical Characteristics of SiO2/High-k Dielectric Stacked Tunnel Barriers for Nonvolatile Memory Applications", Journal of the Korean Physical Society, vol. 55, No. 1, Jul. 2009, pp. 116-119.

Park, Min Hyuk, et al., "Review of defect chemistry in fluorite-structure ferroelectrics for future electronic devices", J. Mater. Chem. C, vol. 8, No. 31 10526-10550, Jun. 9, 2020., Aug. 21, 2020, pp. 10526-10550.

Rios, Rafael , et al., "A Physically Based Compact Model for IGZO Transistors", IEEE Transactions on Electron Devices, vol. 68, No. 4, doi: 10.1109/TED.2021.3059387., Apr. 2021, pp. 1664-1669.

Rzehak, Volker , "Low-Power FRAM Microcontrollers and Their Applications", Texas Instruments White Paper, SLAA502, Jul. 2019, 7 pages.

Saitoh, Masumi , et al., "HfO2-based FeFET and FTJ for Ferroelectric-Memory Centric 3D LSI towards Low-Power and High-Density Storage and AI Applications", 2020 IEEE International Electron Devices Meeting (IEDM), doi: 10.1109/IEDM13553.2020.9372106., 2020, pp. 18.1.1-18.1.4.

Saitoh, Masumi , et al., "HfO2-based FeFET and FTJ for Ferroelectric-Memory Centric 3D LSI towards Low-Power and High-Density Storage and AI Applications", IEEE International Electron Devices Meeting (IEDM), doi: 10.1109/IEDM13553.2020.9372106., 2020, pp. 18.1.1-18.1.4.

Salahuddin, Sayeef , et al., "FeFETs for Near-Memory and In-Memory Compute", 2021 IEEE International Electron Devices Meeting (IEDM), Department of Electrical Engineering and Computer Sciences, Dec. 2021, 4 pages.

Sharma, Abhishek A., et al., "High Speed Memory Operation in Channel-Last, Back-gated Ferroelectric Transistors", IEEE International Electron Devices Meeting (IEDM), doi: 10.1109/IEDM13553.2020.9371940., 2020, pp. 18.5.1-18.5.4.

Sheng, Jiazhen , et al., "Review Article: Atomic layer deposition for oxide semiconductor thin film transistors: Advances in research and development", J. Vac. Sci. Technol. A 36, 060801, https://doi.org/10.1116/1.5047237., Nov. 2, 2018, 14 pages.

Si, Mengwei , et al., "Why In2O3 Can Make 0.7 nm Atomic Layer Thin Transistors?", https://arxiv.org/ftp/arxiv/papers/2012/2012.12433.pdf., 2012, 26 pages.

Si, Mengwei , et al., "Why In2O3 Can Make 0.7 nm Atomic Layer Thin Transistors?", available at https://arxiv.org/ftp/arxiv/papers/2012/2012.12433.pdf, School of Electrical and Computer Engineering and Birck Nanotechnology Center, Purdue University, West Lafayette, IN 47907, United States, 2012, pp. 1-26.

Su, Nai-Chao , et al., "A Nonvolatile InGaZnO Charge-Trapping-Engineered Flash Memory With Good Retention Characteristics", IEEE Electron Device Letters, vol. 31, No. 3, Mar. 2010, pp. 201-203.

Sun, Chen , "First Demonstration of BEOL-Compatible Ferroelectric TCAM Featuring a-IGZO Fe-TFTs with Large Memory Window of 2.9 V, Scaled Channel Length of 40 nm, and High Endurance of 108 Cycles", 2021 Symposium on VLSI Technology Digest of Technical Papers, Kyoto, doi: 978-4-86348-779-6., Jun. 2021, 2 pages.

Tan, Ava J., et al., "A Nitrided Interfacial Oxide for Interface State Improvement in Hafnium Zirconium Oxide-Based Ferroelectric Transistor Technology", IEEE Electron Device Letters, vol. 39, No. 1, doi: 10.1109/LED.2017.2772791., Jan. 2018, pp. 95-98.

Tan, Ava J., et al., "Experimental Demonstration of a Ferroelectric HfO2-Based Content Addressable Memory Cell", IEEE Electron Device Letters, vol. 41, No. 2, doi: 10.1109/LED.2019.2963300., Feb. 2020, pp. 240-243.

Tan, Ava Jiang, et al., "Ferroelectric HfO2 Memory Transistors with High-κ Interfacial Layer and Write Endurance Exceeding 1010 Cycles", arXiv:2103.08806 [physics.app-ph], available at https://arxiv.org/abs/2103.08806., Mar. 16, 2021.

Tan, Ava J., et al., "Hot Electrons as the Dominant Source of Degradation for Sub-5nm HZO FeFETs", IEEE Symposium on VLSI Technology, Honolulu, HI, USA, doi: 10.1109/VLSITechnology18217.2020.9265067., 2020, pp. 1-2.

Tan, Yan-Ny , et al., "Over-Erase Phenomenon in SONOS-Type Flash Memory and its Minimization Using a Hafnium Oxide Charge Storage Layer", IEEE Transactions on Electron Devices, vol. 51, No. 7, Jul. 2004, pp. 1143-1147.

Tanaka, T. , et al., "A 768 GB 3b/cell 3D-Floaling-Gate NANO Flash Memory", Digest of Technical Papers, the 2016 EEE International Solid-Slate Circuits Conference, 2016, pp. 142-144.

Van Houdt, Jan , "The 3D FeFET: contender for 3D-NAND Flash memory and machine learning", available at https://www.imec-int.com/en/imec-magazine/imec-magazine-october-2017/the-vertical-ferroelectric-fet-a-new-contender-for-3d-nand-flash-memory-and-machine-learning, Sep. 30, 2019, 8 pages.

Wann, H.C. , et al., "High-Endurance Ultra-Thin Tunnel Oxide in Monos Device Structure for Dynamic Memory Application", IEEE Electron Device letters, vol. 16, No. 11, Nov. 1995, pp. 491-493.

Yang, Jin , et al., "Highly Optimized Complementary Inverters Based on p-SnO and n-InGaZnO With High Uniformity", IEEE Electron Device Letters, vol. 39, No. 4, doi: 10.1109/LED.2018.2809796., Apr. 2018, pp. 516-519.

Wu, Jixuan , et al., "A Monolithic 3D Integration of RRAM Array with Oxide Semiconductor FET for In-Memory Computing in Quantized Neural Network AI Applications", 2020 IEEE Symposium on VLSI Technology Digest of Technical Papers, Honolulu, HI, USA, Jun. 2020, 4 pages.

"EP Extended Search Report EP168690149.3", dated Oct. 18, 2019.

"European Search Report, EP 16852238.1", dated Mar. 28, 2019.

"European Search Report, EP17844550.8", dated Aug. 12, 2020, 11 pages.

"Imec Demonstrates Capacitor-less IGZO-Based DRAM Cell With >400s Retention Time", IMEC, Press release, available at https://www.imec-int.com/en/press/imec-demonstrates-capacitor-less-igzo-based-dram-cell-400s-retention-time, Dec. 15, 2020, 15 pages.

"Invitation to Pay Additional Fees (PCT/ISA/206), PCT/US2020/015710", dated Mar. 20, 2020, 2 pages.

"Invitation to Pay Additional Fees, PCT/US2019/065256", dated Feb. 13, 2020, 2 pages.

"Notification of Reasons for Refusal, Japanese Patent Application 2018-527740", (English translation), dated Nov. 4, 2020, 8 pages.

"Partial European Search Report EP 16869049.3", dated Jul. 1, 2019, pp. 1-12.

"PCT Search Report and Written Opinion, PCT/US2018/038373", dated Sep. 10, 2018.

"PCT Search Report and Written Opinion, PCT/US2018/067338", dated May 8, 2019.

"PCT Search Report and Written Opinion, PCT/US2019/014319", dated Apr. 15, 2019.

"PCT Search Report and Written Opinion, PCT/US2019/041678", dated Oct. 9, 2019.

"PCT Search Report and Written Opinion, PCT/US2019/052164", dated Feb. 27, 2020.

"PCT Search Report and Written Opinion, PCT/US2019/052446", dated Dec. 11, 2019.

"PCT Search Report and Written Opinion, PCT/US2019/065256", dated Apr. 14, 2020.

"PCT Search Report and Written Opinion, PCT/US2020/015710", dated Jun. 9, 2020.

"PCT Search Report and Written Opinion, PCT/US2020/017494", dated Jul. 20, 2020, 13 pages.

"PCT Search Report and Written Opinion, PCT/US2020/065374", dated Mar. 15, 2021, 17 pages.

"PCT Search Report and Written Opinion, PCT/US2020/065670", dated Apr. 5, 2021, 12 pages.

"PCT Search Report and Written Opinion, PCT/US2021/016964", dated Jun. 15, 2021, 19 pages.

"PCT Search Report and Written Opinion, PCT/US2021/025722", dated Jun. 15, 2021, 10 pages.

"PCT Search Report and Written Opinion, PCT/US2021/042607", dated Nov. 4, 2021, 17 pages.

"PCT Search Report and Written Opinion, PCT/US2021/064844", dated Mar. 8, 2022, 15 paged.

(56) **References Cited**

OTHER PUBLICATIONS

"PCT Search Report and Written Opinion, PCT/US2021/42620", dated Oct. 28, 2021, 18 pages.
"PCT Search Report and Written Opinion, PCT/US2022/016729", Applicant: SunRise Memory Corporation, dated May 17, 2022, 20 pages.
Ahn, Min-Ju , et al., "Transparent multi-level-cell nonvolatile memory with dual-gate amorphous indiumgallium-zinc oxide thin-film transistors", Appl. Phys. Lett. 109, 252106; doi: 10.1063/1.4972961., 2016, 6 pages.
Alessandri, Cristobal , et al., "Monte Carlo Simulation of Switching Dynamics in Polycrystalline Ferroelectric Capacitors", IEEE Transactions on Electron Devices, vol. 66, No. 8, doi: 10.1109/TED.2019.2922268., Aug. 2019, pp. 3527-3534.
Ali, T. , et al., "A Multilevel FeFET Memory Device based on Laminated HSO and HZO Ferroelectric Layers for High-Density Storage", IEEE International Electron Devices Meeting (IEDM), 2019, doi: 10.1109/ EDM19573.2019.8993642., Dec. 2019, pp. 28.7.1-28.7.4.
Ali, T. , et al., "High Endurance Ferroelectric Hafnium Oxide-Based FeFET Memory Without Retention Penalty,", IEEE Transactions on Electron Devices, vol. 65, No. 9, doi: 10.1109/TED.2018.2856818.—paper cited in Bae (Berkeley) paper, Sep. 2018, pp. 3769-3774.
Bae, Jong-Ho , et al., "Highly Scaled, High Endurance, Q-Gate, Nanowire Ferroelectric FET Memory Transistors", IEEE Electron Device Letters, vol. 41, No. 11, doi: 10.1109/LED.2020.3028339.—Sayeef-Berkeley paper on FeFET memory, Nov. 2020, pp. 1637-1640.
Beyer, Sven , et al., "FeFET: A versatile CMOS compatible device with game-changing potential", IEEE International Memory Workshop (IMW), doi: 10.1109/IMW48823.2020.9108150., 2020, pp. 1-4.
Böscke, T.S. , et al., "Ferroelectricity in hafnium oxide: CMOS compatible ferroelectric field effect transistors", 2011 International Electron Devices Meeting, Washington, DC, USA, doi: 10.1109/IEDM.2011.6131606., 2011, pp. 24.5.1-24.5.4.
Chang, Sou-Chi , et al., "Anti-ferroelectric HfxZr1-xO2 Capacitors for High-density 3-D Embedded-DRAM", IEEE International Electron Devices Meeting (IEDM), doi: 10.1109/IEDM13553.2020.9372011., 2020, pp. 28.1.1-28.1.4.
Cho, Min Hoe, et al., "Achieving a Low-Voltage, High-Mobility IGZO Transistor through an ALD-Derived Bilayer Channel and a Hafnia-Based Gate Dielectric Stack", ACS Applied Materials & Interfaces, Apr. 1, 2021, 13 (14), , DOI: 10.1021/acsami.0c22677, 2021, pp. 16628-16640.
Cho, Min Hoe, et al., "Comparative Study on Performance of IGZO Transistors With Sputtered and Atomic Layer Deposited Channel Layer", IEEE Transactions on Electron Devices, vol. 66, No. 4, doi: 10.1109/TED.2019.2899586., Apr. 2019, pp. 1783-1788.
Choi, Seonjun , et al., "A novel three-dimensional NAND flash structure for improving the erase performance", IEICE Electronics Express, 2019 vol. 16 Issue 3, 2019, 6 pages.
De, Sourav , et al., "Ultra-Low Power Robust 3bit/cell Hf0.5Zr0.5O2 Ferroelectric FinFET with High Endurance for Advanced Computing-In-Memory Technology", 2021 Symposium on VLSI Technology, Kyoto, Japan, Jun. 13-19, 2021, Jun. 2021, 2 pages.
Dünkel , "A FeFET based super-low-power ultra-fast embedded NVM technology for 22nm FDSOI and beyond", IEEE International Electron Devices Meeting (IEDM), San Francisco, CA, USA, 2017, doi: 10.1109/IEDM.2017.8268425., 2017, pp. 19.7.1-19.7.4.
Dutta, Sourav , et al., "Logic Compatible High-Performance Ferroelectric Transistor Memory", available at https://arxiv.org/abs/2105.11078, Department of Electrical Engineering, University of Notre Dame, Notre Dame, IN 46556, USA, May 24, 2021, 28 pages.
Florent, K. , "First demonstration of vertically stacked ferroelectric AI doped HfO2 devices for NAND applications", Symposium on VLSI Technology, doi: 10.23919/VLSIT.2017.7998162., 2017, pp. T158-T159.
Florent, K. , et al., "Vertical Ferroelectric HfO2 FET based on 3-D NAND Architecture: Towards Dense Low-Power Memory", IEEE International Electron Devices Meeting (IEDM), San Francisco, CA, 2018, doi: 10.1109/IEDM.2018.8614710., 2018, pp. 2.5.1-2.5.4.
Hendy, Ian , "The Role of Increased Semiconductor Mobility—an IGZO Case Study", Display Daily, available at https://www.displaydaily.com/article/display-daily/the-role-of-increased-semiconductor-mobility-an-igzo-case-study, Apr. 2021, 11 pages.
Hou, S. Y., et al., "Wafer-Leval Integration of an Advanced Logic-Memory System Through the Second-Generation CoWoS Technology", IEEE Transactions on Electron Devices, vol. 64, No. 10, Oct. 2017, 4071-4077.
Hsain, Hanan Alexandra, et al., "Many routes to ferroelectric HfO2: A review of current deposition methods", J. Vac. Sci. Technol. A 40, 010803 (2022), published Dec. 1, 2021., Dec. 1, 2021, pp. 010803-1-010803-36.
Jeewandara, Thamarasee , et al., "High-speed 3-D memory with ferroelectric NAND flash memory", available at https://techxplore.com/news/2021-01-high-speed-d-memoryferroelectric-nand.html, Jan. 26, 2021, 5 pages.
Kawai, H. , et al., "Design Principle of Channel Material for Oxide-Semiconductor Field-Effect Transistor with High Thermal Stability and High On-current by Fluorine Doping", IEEE International Electron Devices Meeting (IEDM), doi: 10.1109/IEDM13553.2020.9372121., 2020, pp. 22.2.1-22.2.4.
Kim, Min-Kyu , et al., "CMOS-compatible ferroelectric NAND flash memory for high-density, low-power, and high-speed three-dimensional memory", Science Advances, vol. 7, No. 3, eabe 1341, DOI: 10.1126/sciadv.abe1341. (Replacement of polysilicon channel material—use IZO for channel), Jan. 13, 2021, 10.
Kim, Taeho , et al., "Effects of high pressure nitrogen annealing on ferroelectric Hf0.5Zr0.5O2 films", Appl. Phys. Lett. 112, 092906 https://doi.org/10.1063/1.5003369, Mar. 2, 2018, 5 pages.
Kim, N. , et al., "Multi-layered Vertical gate NANO Flash Overcoming Stacking Limit for Terabit Density Storage", Symposium on VLSI Tech. Dig. of Technical Papers, 2009, pp. 188-189.
Kunitake, Hitoshi , et al., "High-temperature Electrical Characteristics of 60nm CAAC-IGZO FET : Comparison with Si FET", International Conference on Solid State Devices and Materials, https://doi.org/10.7567/SSDM.2018.N-3-04, Sep. 2018, pp. 787-788.
Chen, Chi-Yu , et al., "FeFET Memory Featuring Large Memory Window and Robust Endurance of Long-Pulse Cycling by Interface Engineering using High-k AION", 2020 IEEE Symposium on VLSI Technology, doi: 10.1109/VLSITechnology18217.2020.9265103., 2020, pp. 1-2.
Kim, Hyungwoo , et al., "Two-step deposition of TiN capping electrodes to prevent degradation of ferroelectric properties in an in-situ crystallized TiN/Hf0.5Zr0.5O2/TIN device", Nano Express 3 015004., 2022, 12 pages.
McBriarty, Martin E., et al., "Crystal Phase Distribution and Ferroelectricity in Ultrathin HfO2—ZrO2 Bilayers", Phys. Status Solidi B, 257: 1900285. https://doi.org/10.1002/pssb.201900285, Aug. 21, 2019, pp. 1-25.
Mittmann, T. , et al., "Impact of Oxygen Vacancy Content in Ferroelectric HZO films on the Device Performance", 2020 IEEE International Electron Devices Meeting (IEDM), doi: 10.1109/IEDM13553.2020.9372097., 2020, pp. 18.4.1-18.4.4.
Onaya, Takashi , et al., "Improvement in ferroelectricity of HfxZr1—xO2 thin films using top- and bottom-ZrO2 nucleation layers", APL Materials 7, 061107; https://doi.org/10.1063/1.5096626, 2019, pp. 1-8.
Sato, Yuta , et al., "Source/Drain Contact Engineering of InGaZnO Channel BEOL Transistor for Low Contact Resistance and Suppressing Channel Shortening Effect,", 20th International Workshop on Junction Technology (IWJT), doi: 10.23919/IWJT52818.2021.9609366., 2021, 3 pages.
Wang, Chin-I , et al., "Atomic layer deposited TIN capping layer for sub-10 nm ferroelectric Hf0.5Zr0.5O2 with large remnant polarization and low thermal budget", Applied Surface Science, vol. 570, 2021, 151152, ISSN 0169-4332, https://doi.org/10.1016/j.apsusc.2021.151152., Aug. 9, 2021, pp. 1-8.

* cited by examiner

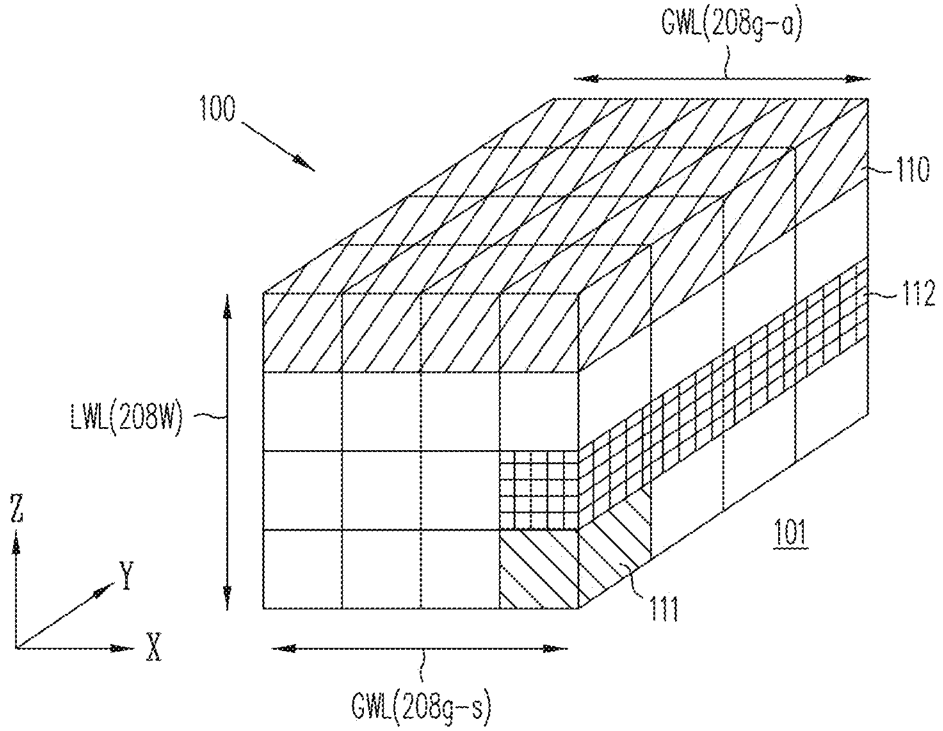

GWL(208g-a)
100
110
112
LWL(208W)
Z
Y
X
101
111
GWL(208g-s)

FIG. 1a-1

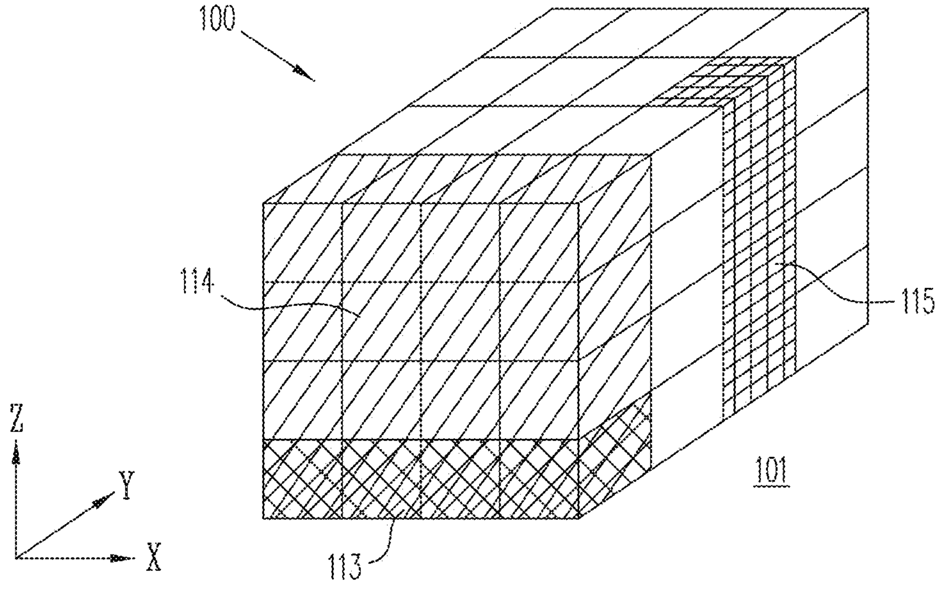

100
114
115
Z
Y
X
101
113

FIG. 1a-2

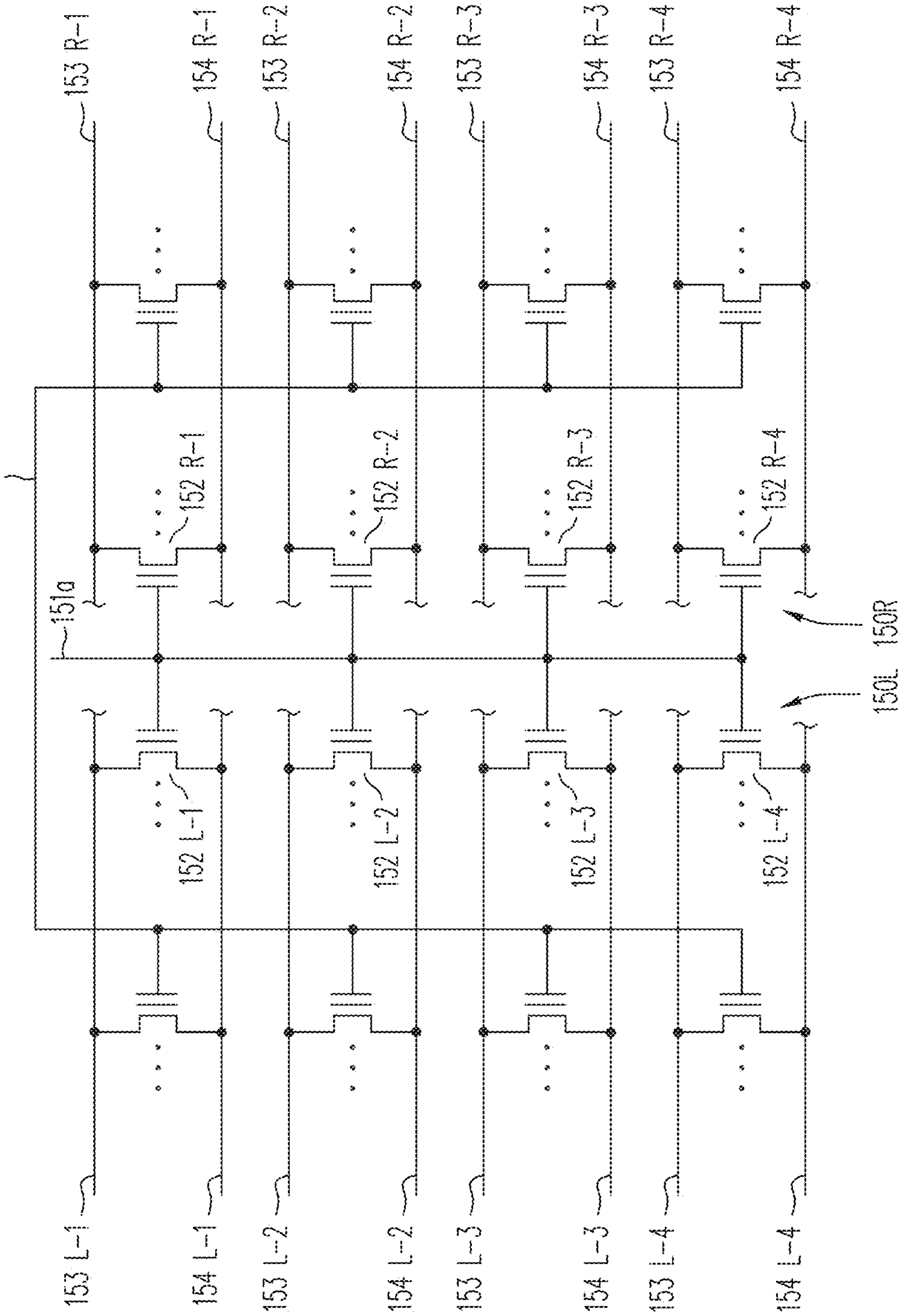

153 R-1
154 R-1
153 R-2
154 R-2
153 R-3
154 R-3
153 R-4
154 R-4
151n
152 R-1
152 R-2
152 R-3
152 R-4
151a
150L
150R
152 L-1
152 L-2
152 L-3
152 L-4
153 L-1
154 L-1
153 L-2
154 L-2
153 L-3
154 L-3
153 L-4
154 L-4

FIG. 1b

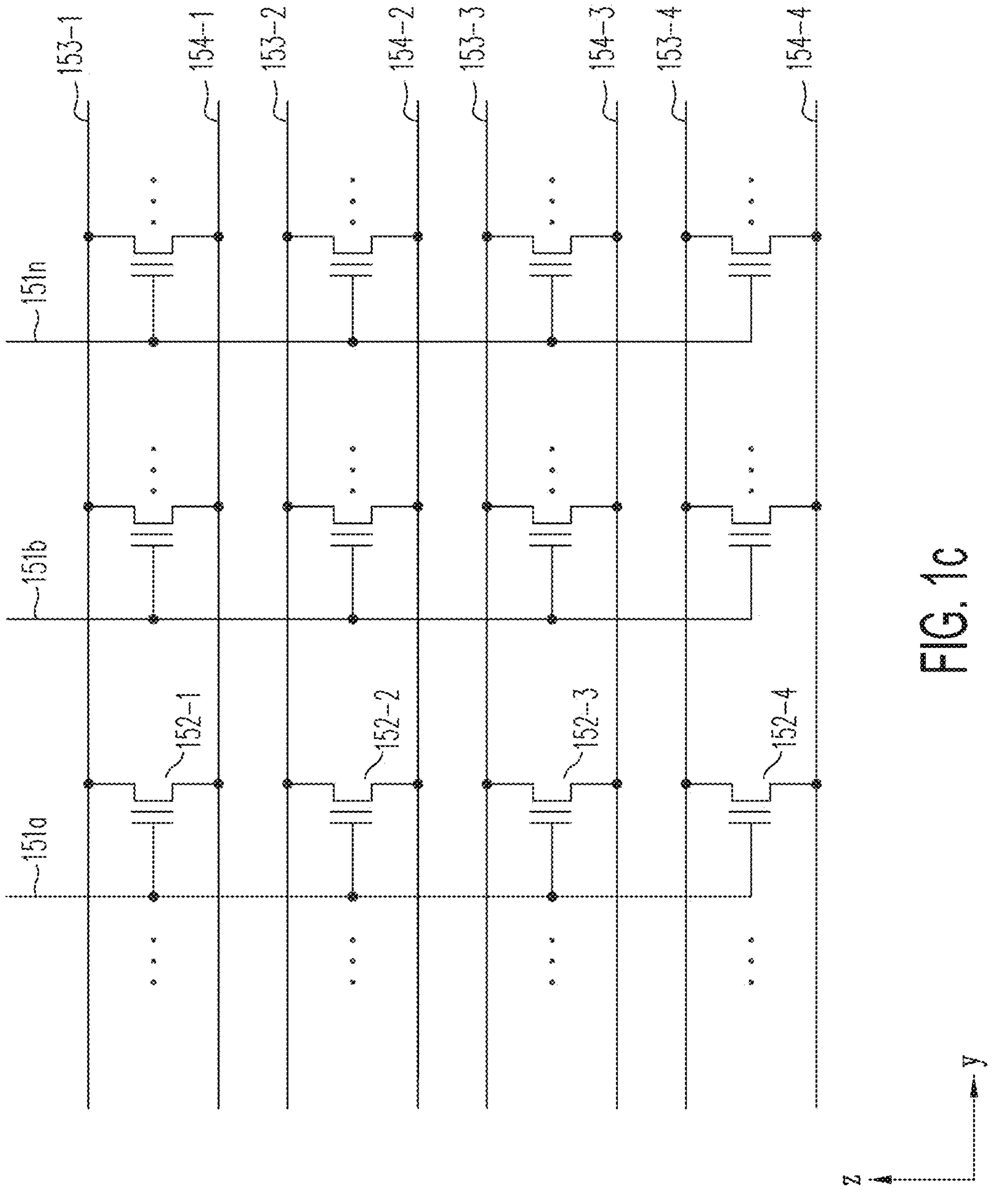

153-1
154-1
153-2
154-2
153-3
154-3
153-4
154-4
151n
151b
151a
152-1
152-2
152-3
152-4
y
z

FIG. 1c

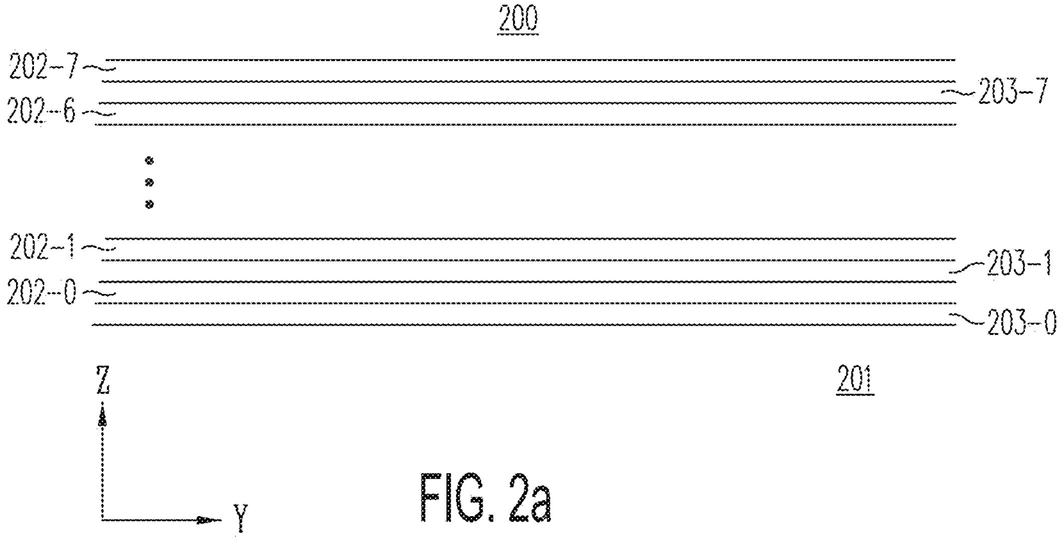

200
202-7
202-6
202-1
202-0
203-7
203-1
203-0
201
Z
Y

FIG. 2a

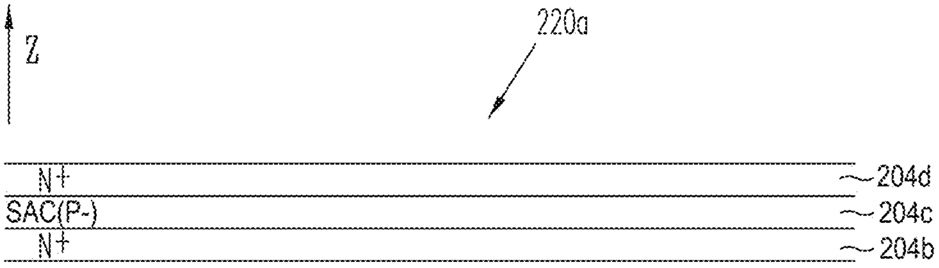

Z
220a
N+
SAC(P-)
N+
204d
204c
204b

FIG. 2b-1

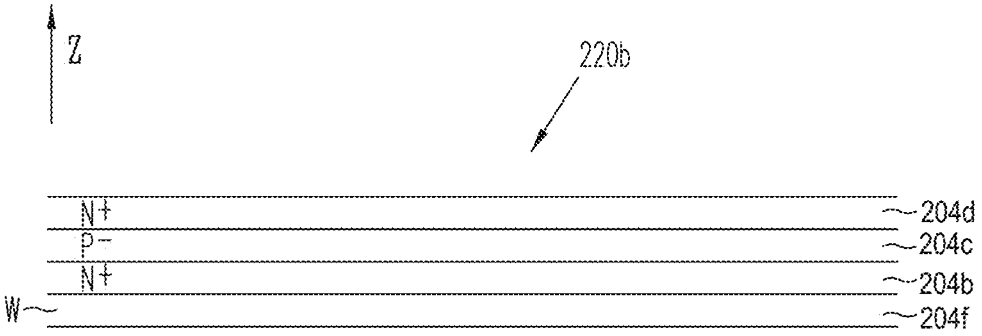

Z
220b
N+
P-
N+
W
204d
204c
204b
204f

FIG. 2b-2

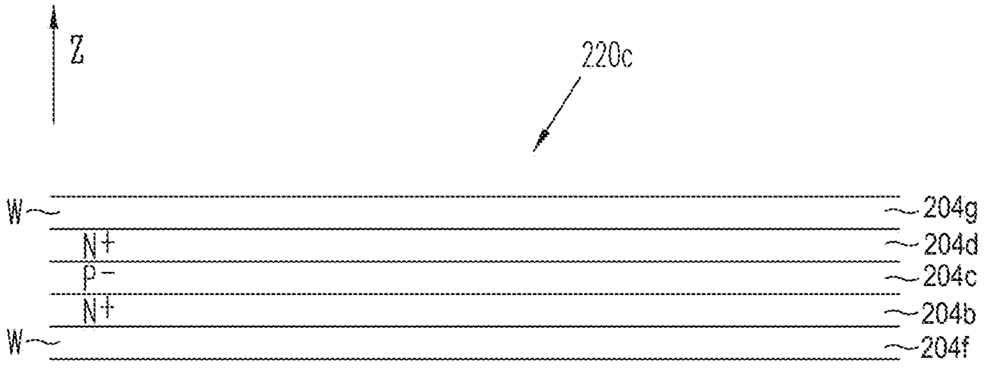

Z
220c
W
N+
P−
N+
W
204g
204d
204c
204b
204f

FIG. 2b-3

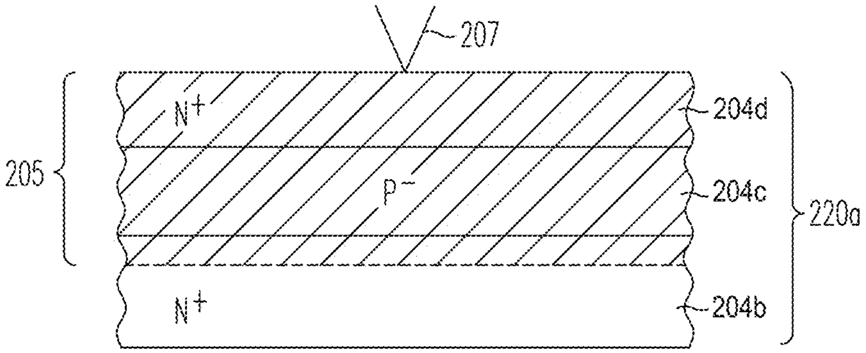

207
205
N+
P−
N+
204d
204c
204b
220a

FIG. 2b-4

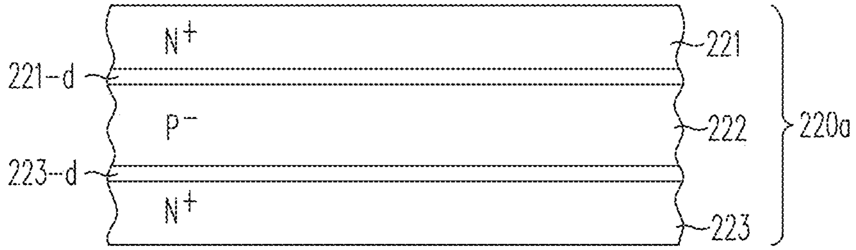

N+
221-d
P−
223-d
N+
221
222
223
220a

FIG. 2b-5

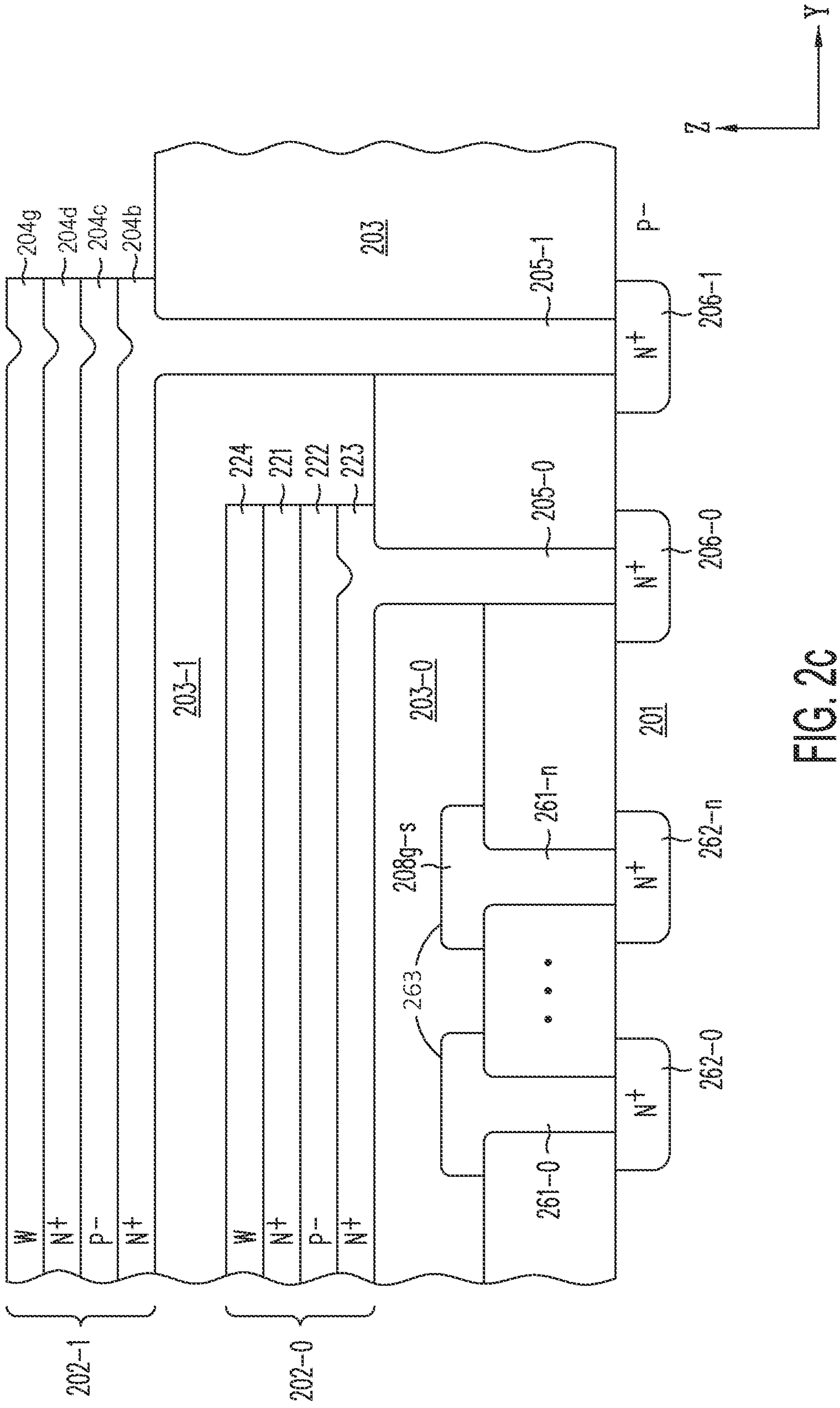

Y
Z
203
P−
205−1
206−1
N+
204g
204d
204c
204b
W
N+
P−
N+
202−1
203−1
224
221
222
223
W
N+
P−
N+
202−0
205−0
206−0
N+
203−0
201
261−n
262−n
N+
208g−s
263
262−0
N+
261−0

FIG. 2C

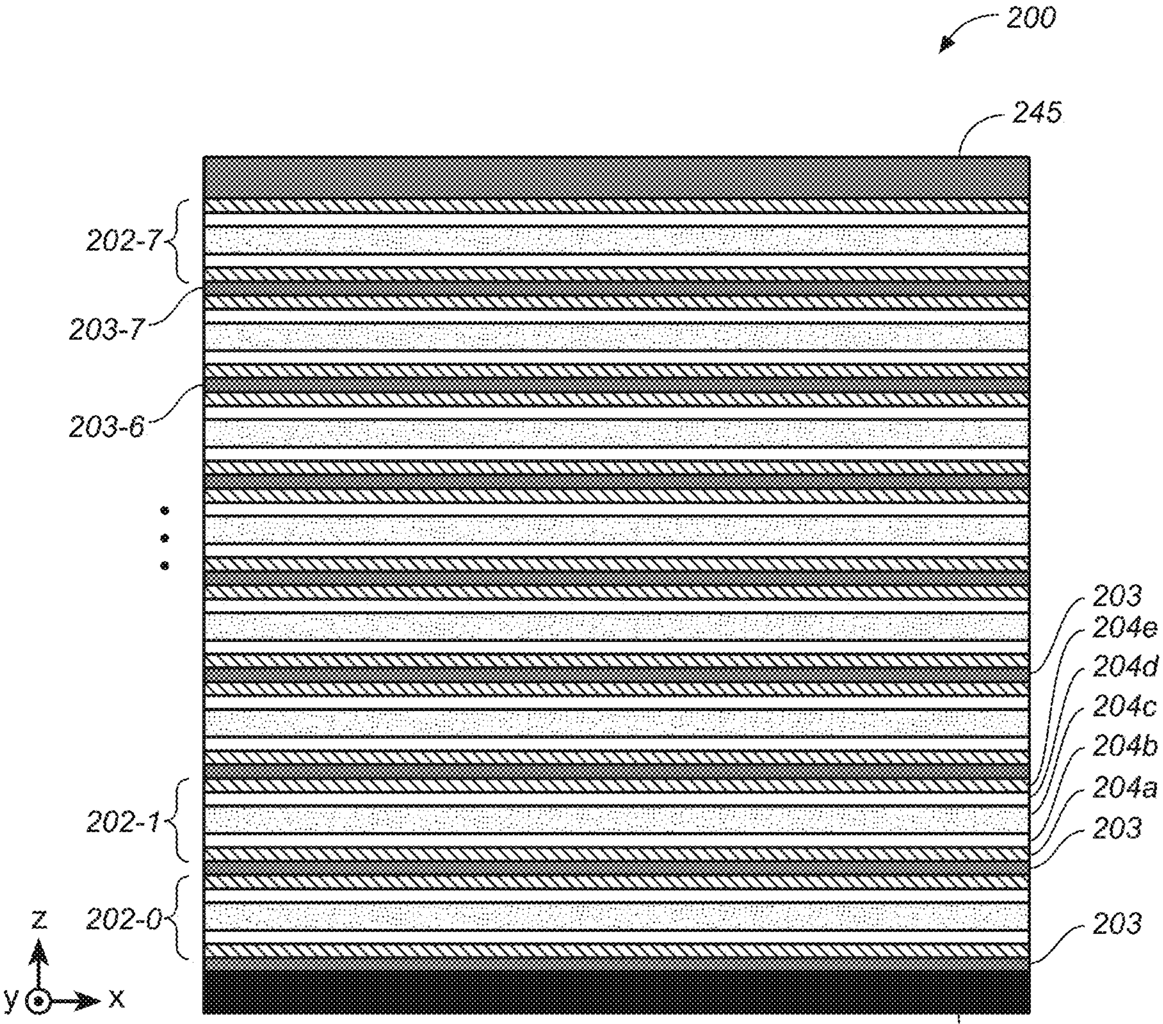

200
245
202-7
203-7
203-6
203
204e
204d
204c
204b
204a
203
202-1
202-0
203
z
y
x
241

FIG. 2d(i)

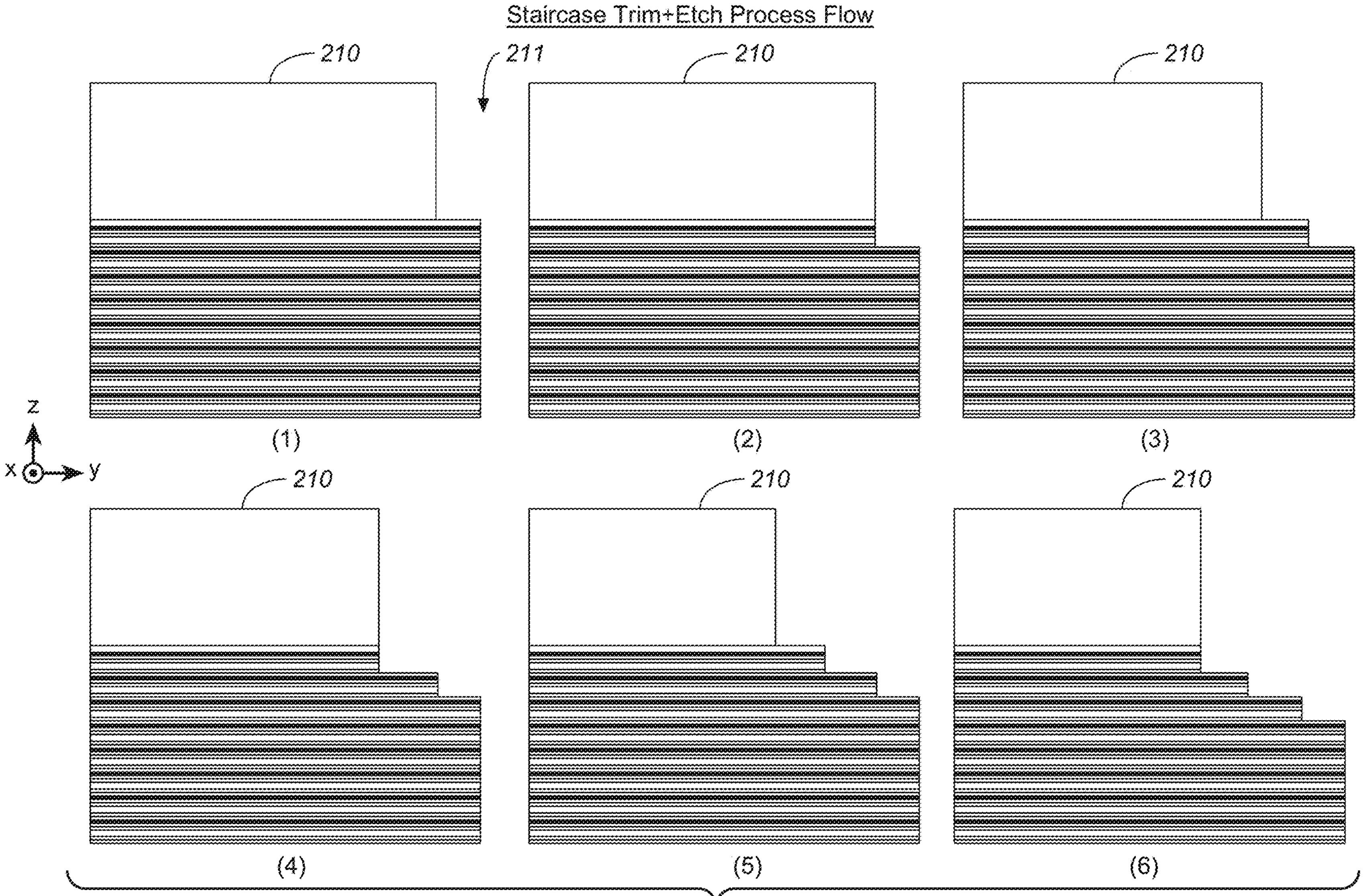

Staircase Trim+Etch Process Flow
211
210
210
210
(1)
(2)
(3)
210
210
210
(4)
(5)
(6)
z
x
y

FIG. 2d(ii)

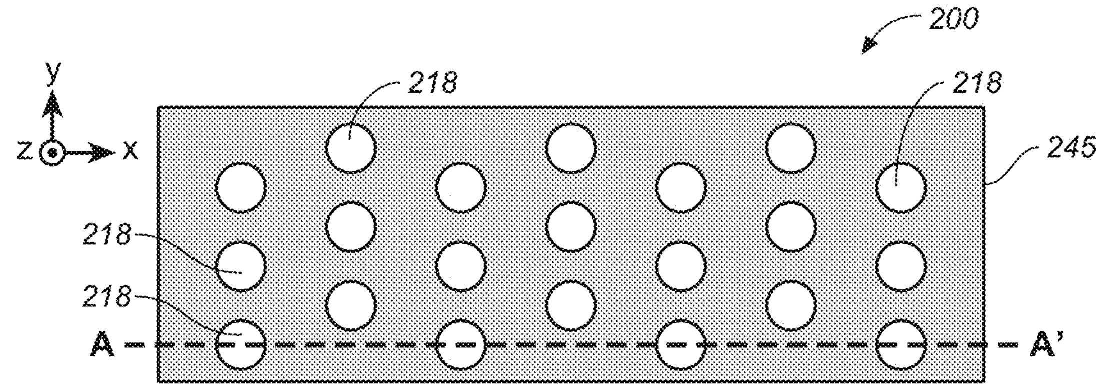

200
218
218
245
218
218
y
z
x
A
A'

FIG. 2e(i)

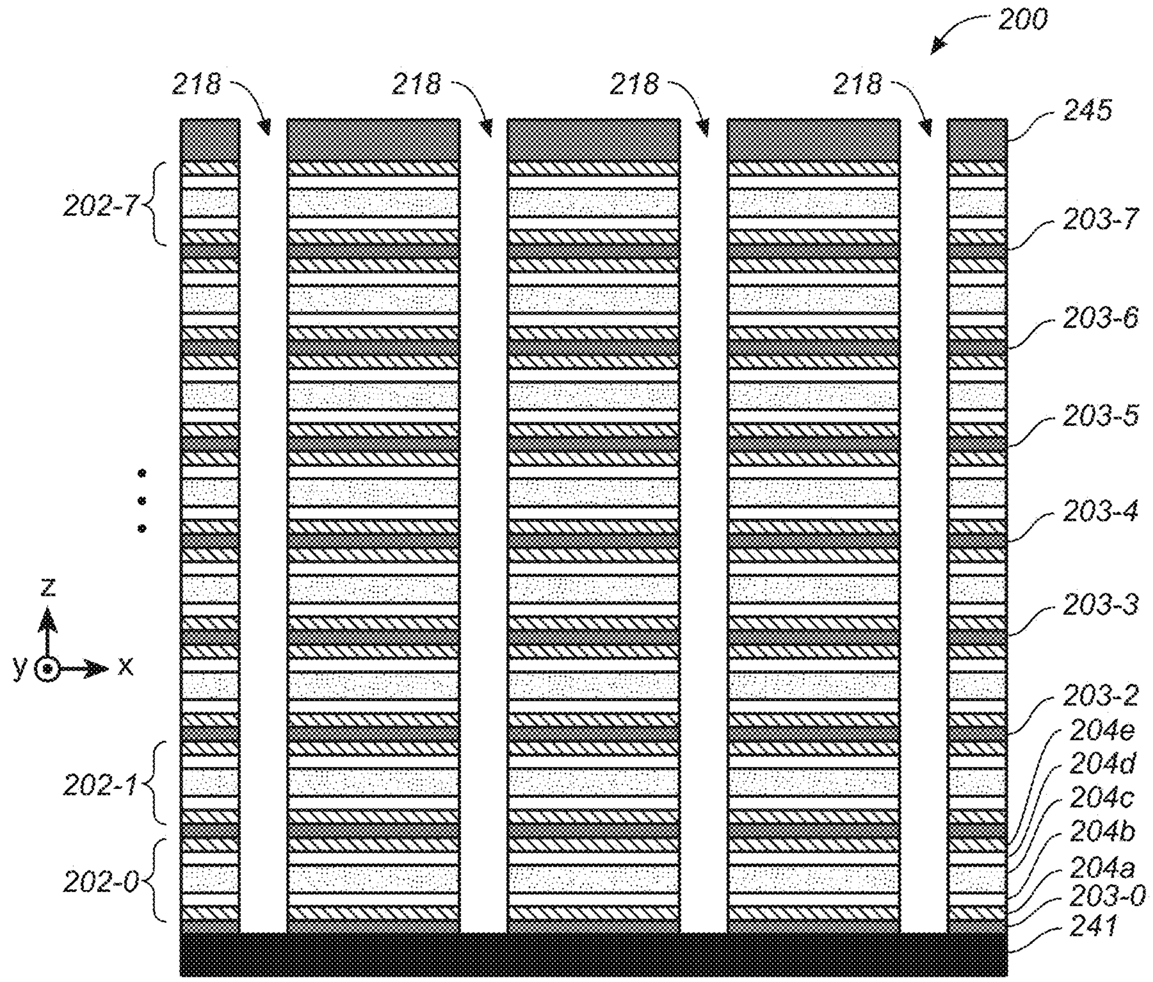
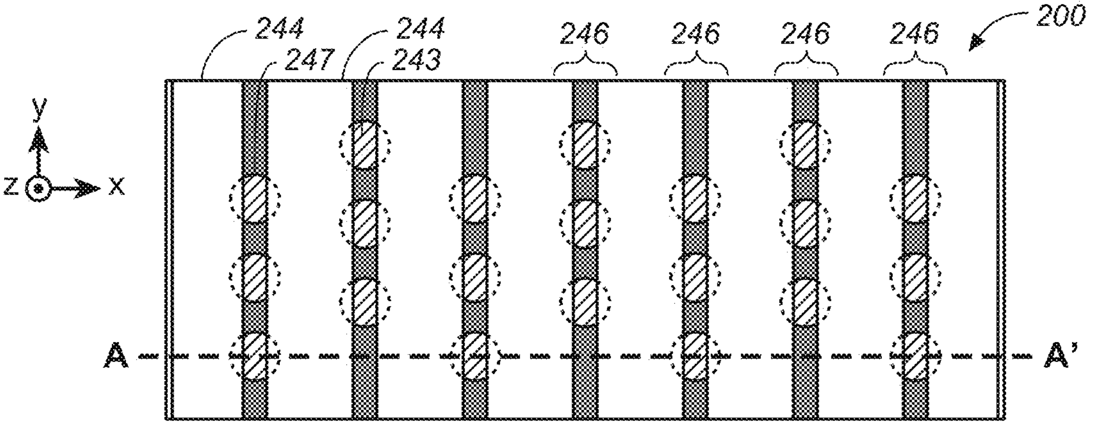
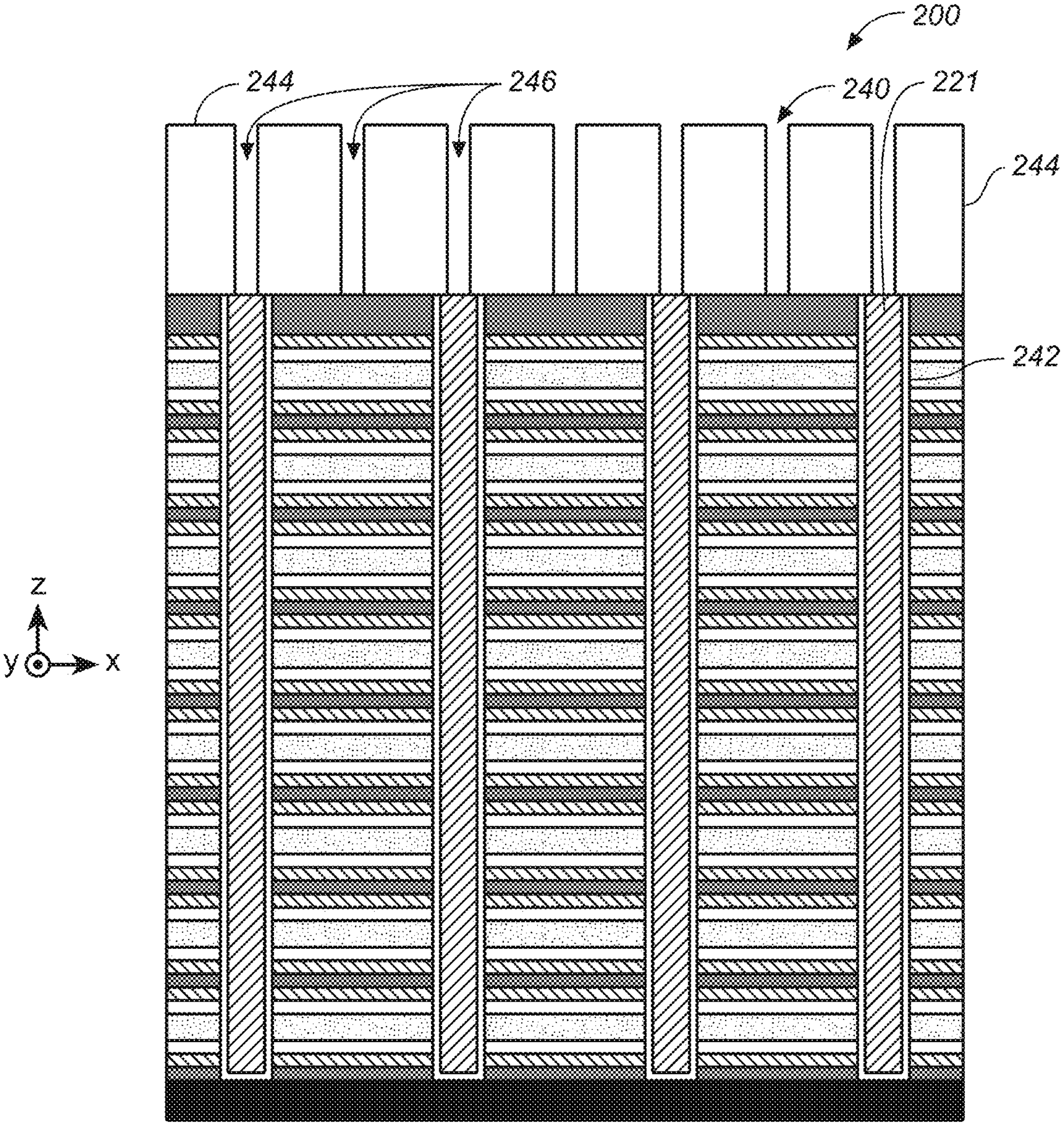

200
218
218
218
218
245
202-7
203-7
203-6
203-5
203-4
203-3
z
y
x
203-2
204e
204d
204c
204b
204a
203-0
241
202-1
202-0

FIG. 2e(ii)



244
247
244
243
246
246
246
246
200
y
z
x
A
A'

FIG. 2f(i)



200
244
246
240
221
244
242
z
y
x

FIG. 2f(ii)

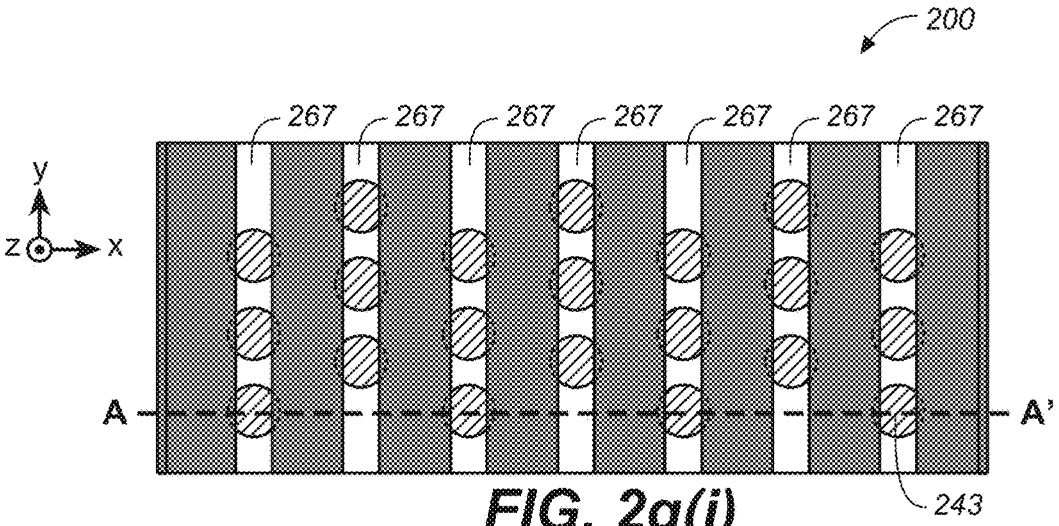

200
267
267
267
267
267
267
267
y
z
x
A
A'
243

FIG. 2g(i)

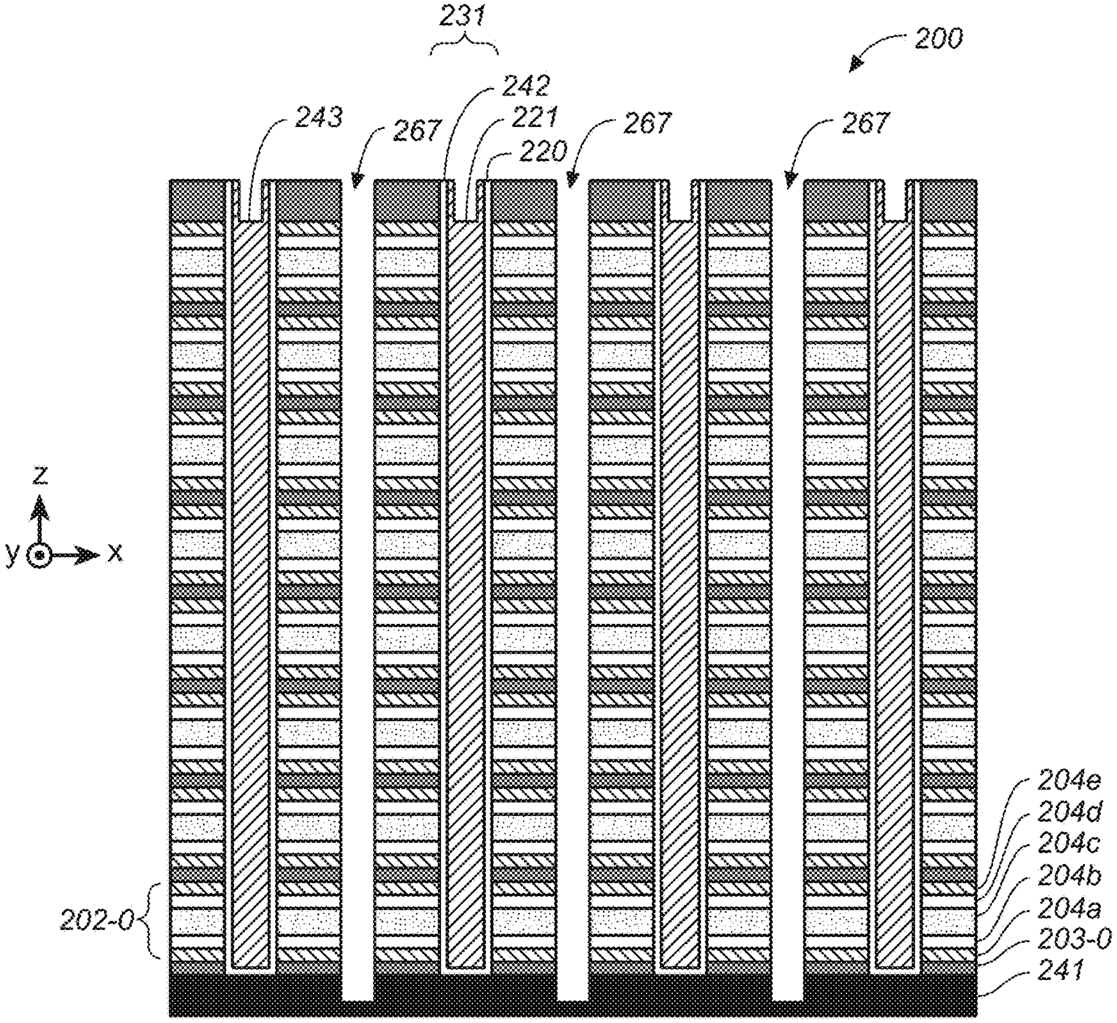

231
200
242
243
267
221
267
267
220
z
y
x
204e
204d
204c
204b
204a
202-0
203-0
241

FIG. 2g(ii)

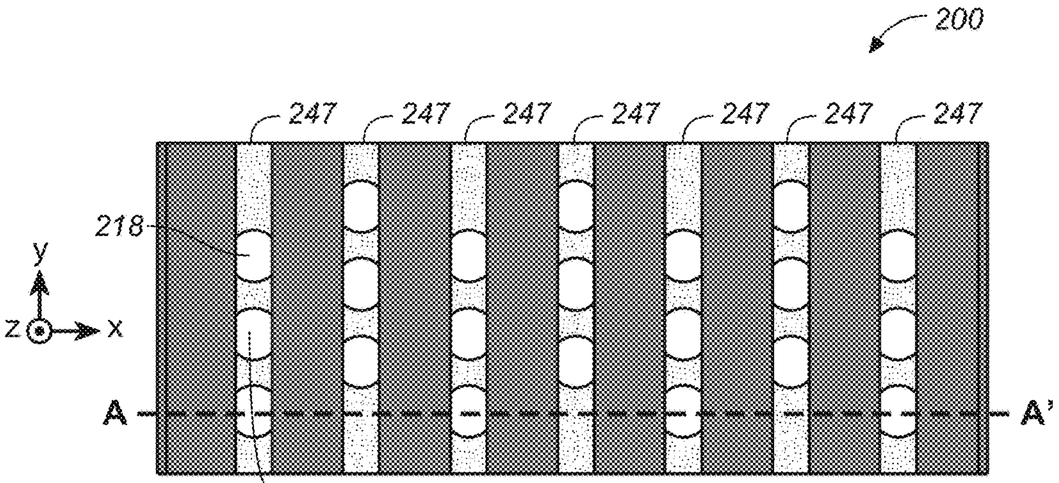

200
247
247
247
247
247
247
247
218
218
y
z
x
A
A'

FIG. 2h(i)

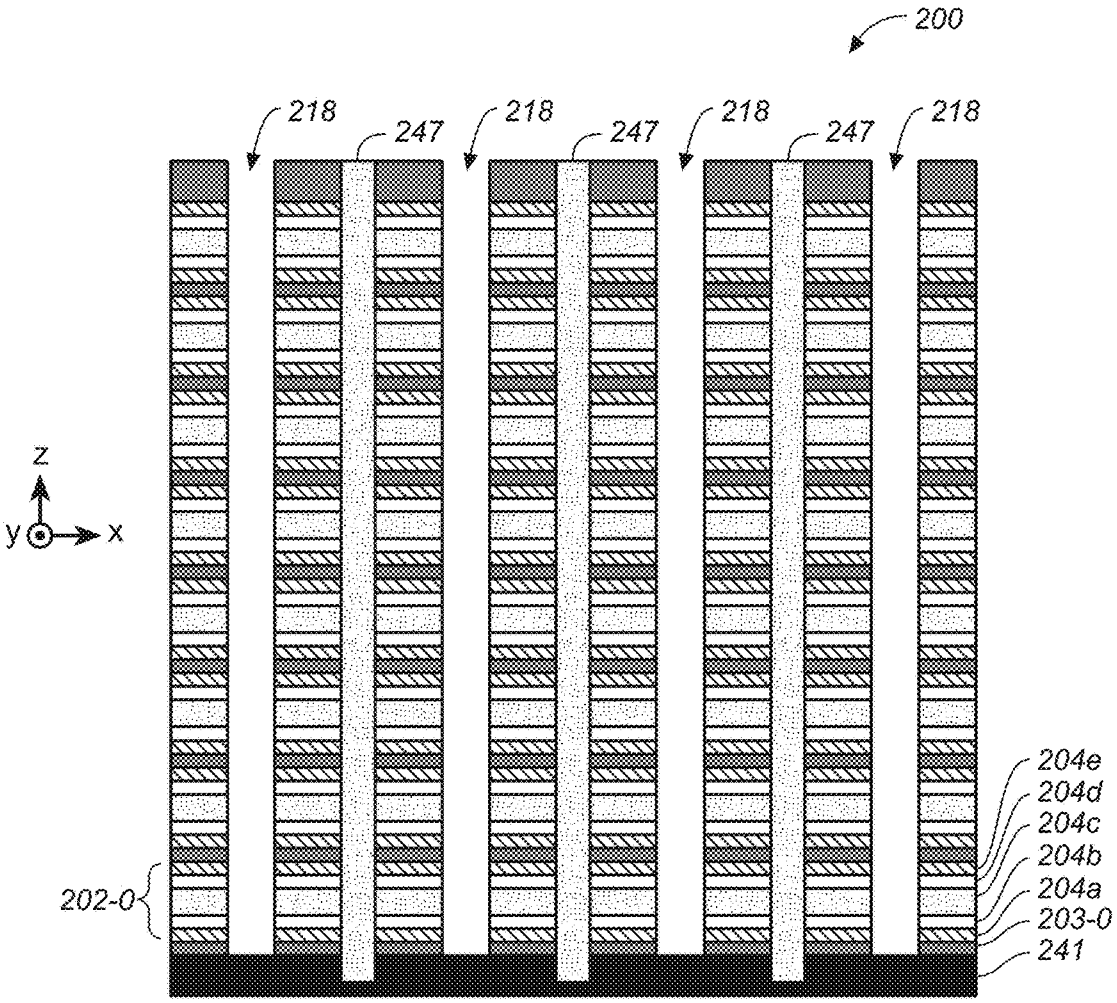

200
218
247
218
247
218
247
218
z
y
x
204e
204d
204c
204b
204a
203-0
241
202-0

FIG. 2h(ii)

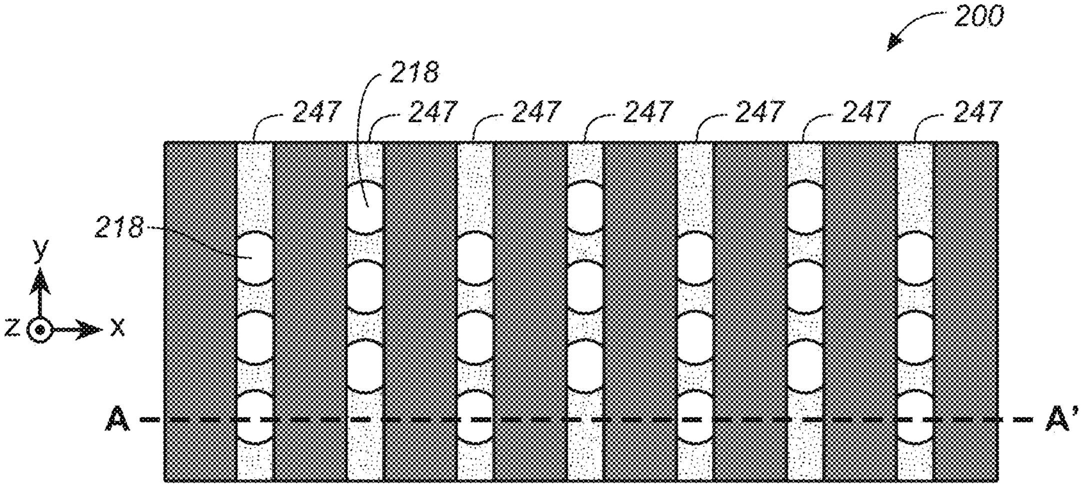

200
218
247
247
247
247
247
247
247
218
y
z
x
A
A'

FIG. 2i(i)

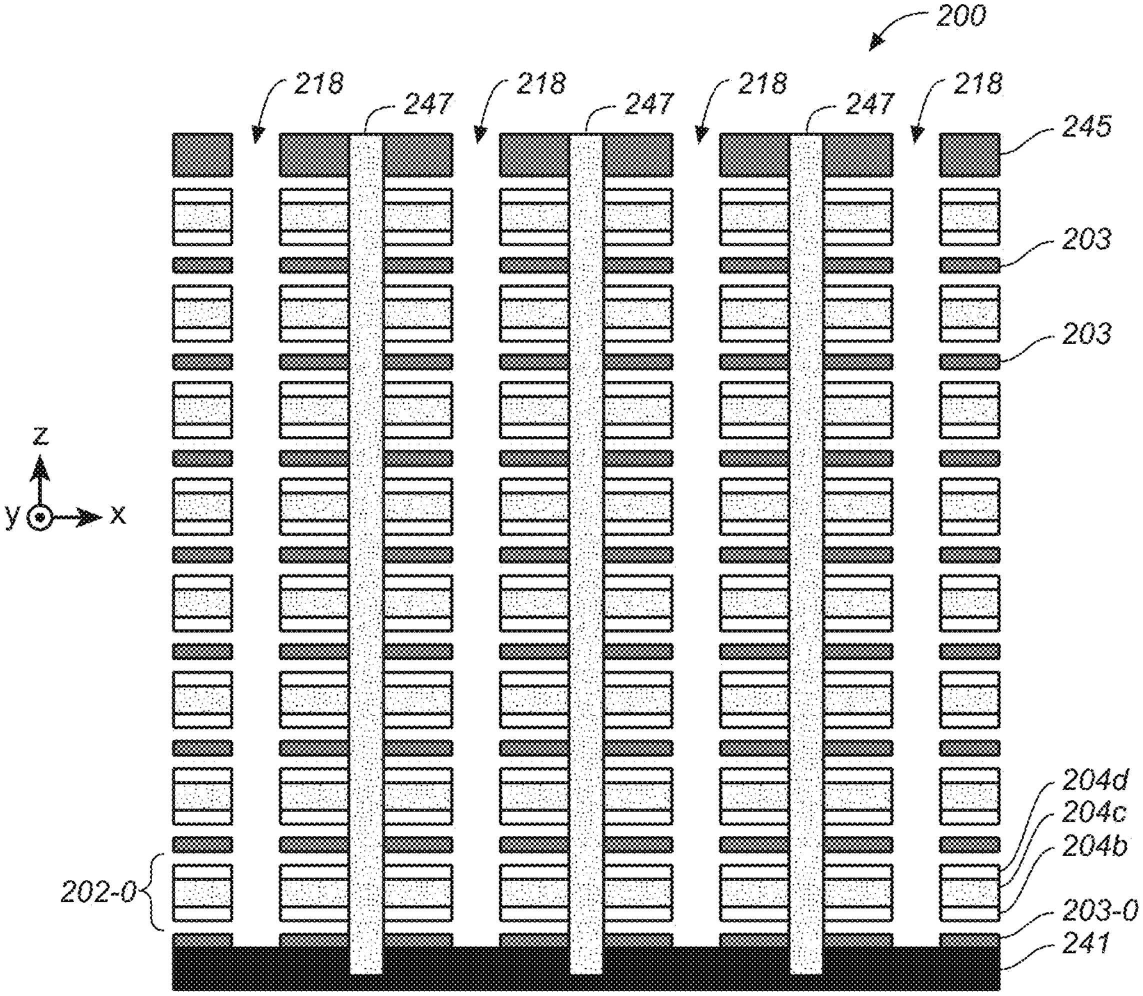

200
218
247
218
247
218
247
218
245
203
203
z
y
x
204d
204c
204b
202-0
203-0
241

FIG. 2i(ii)

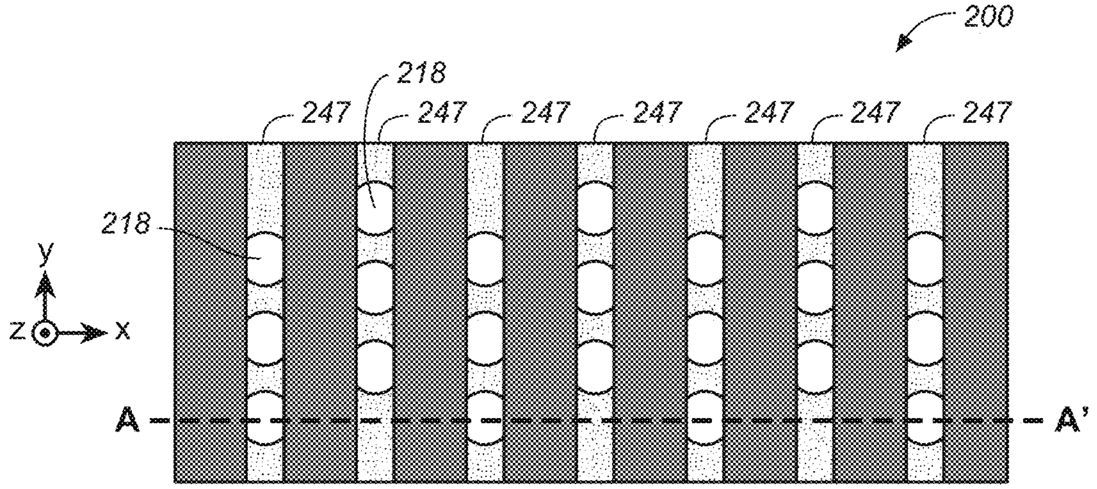

200
218
247
247
247
247
247
247
247
218
y
z
x
A
A'

FIG. 2j(i)

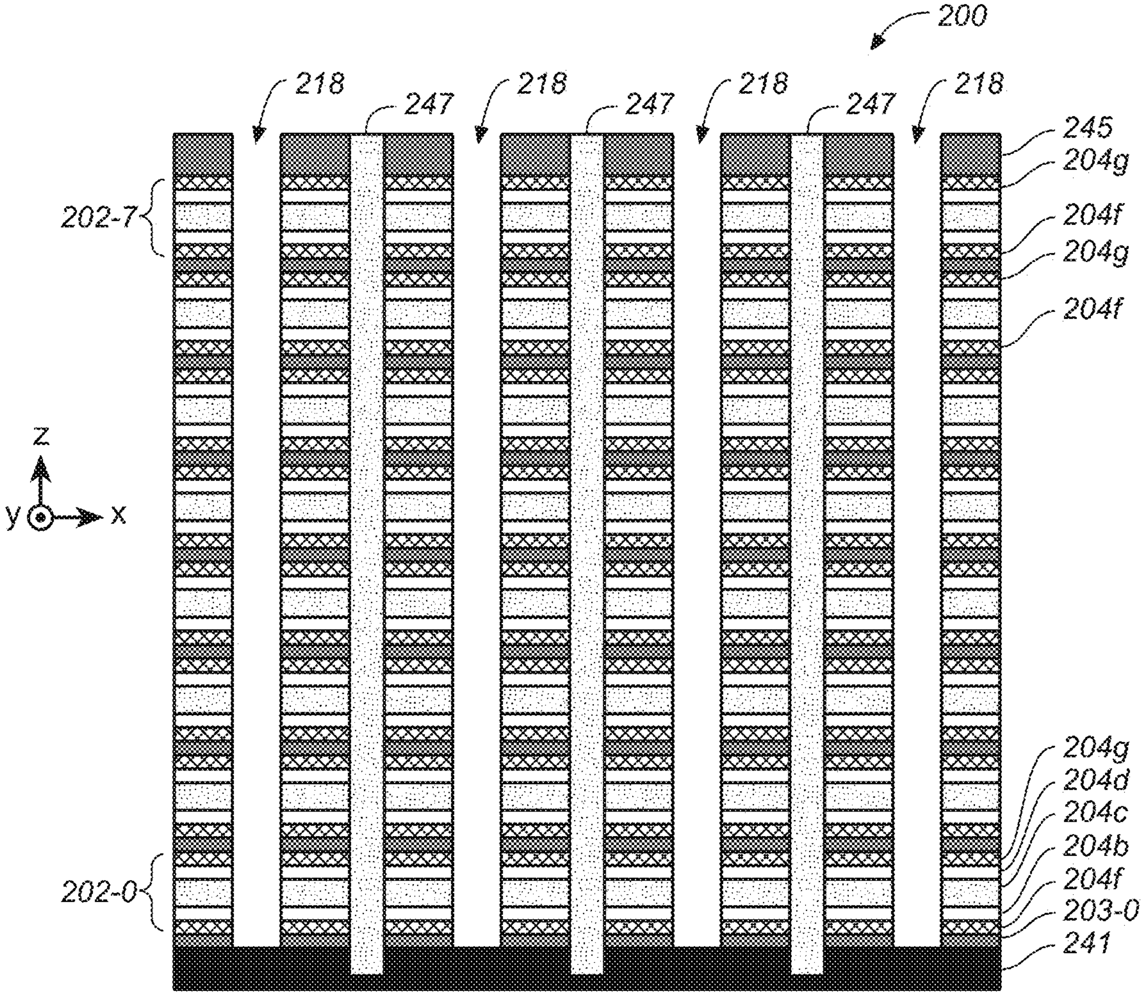

200
218
247
218
247
218
247
218
245
204g
202-7
204f
204g
204f
z
y
x
204g
204d
204c
204b
202-0
204f
203-0
241

FIG. 2j(ii)

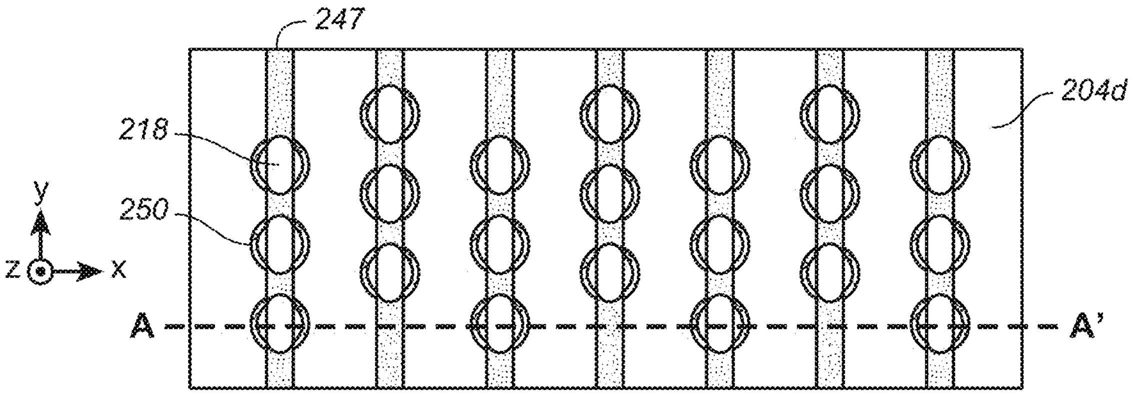

247
204d
218
250
y
z
x
A
A'

FIG. 2k(i)

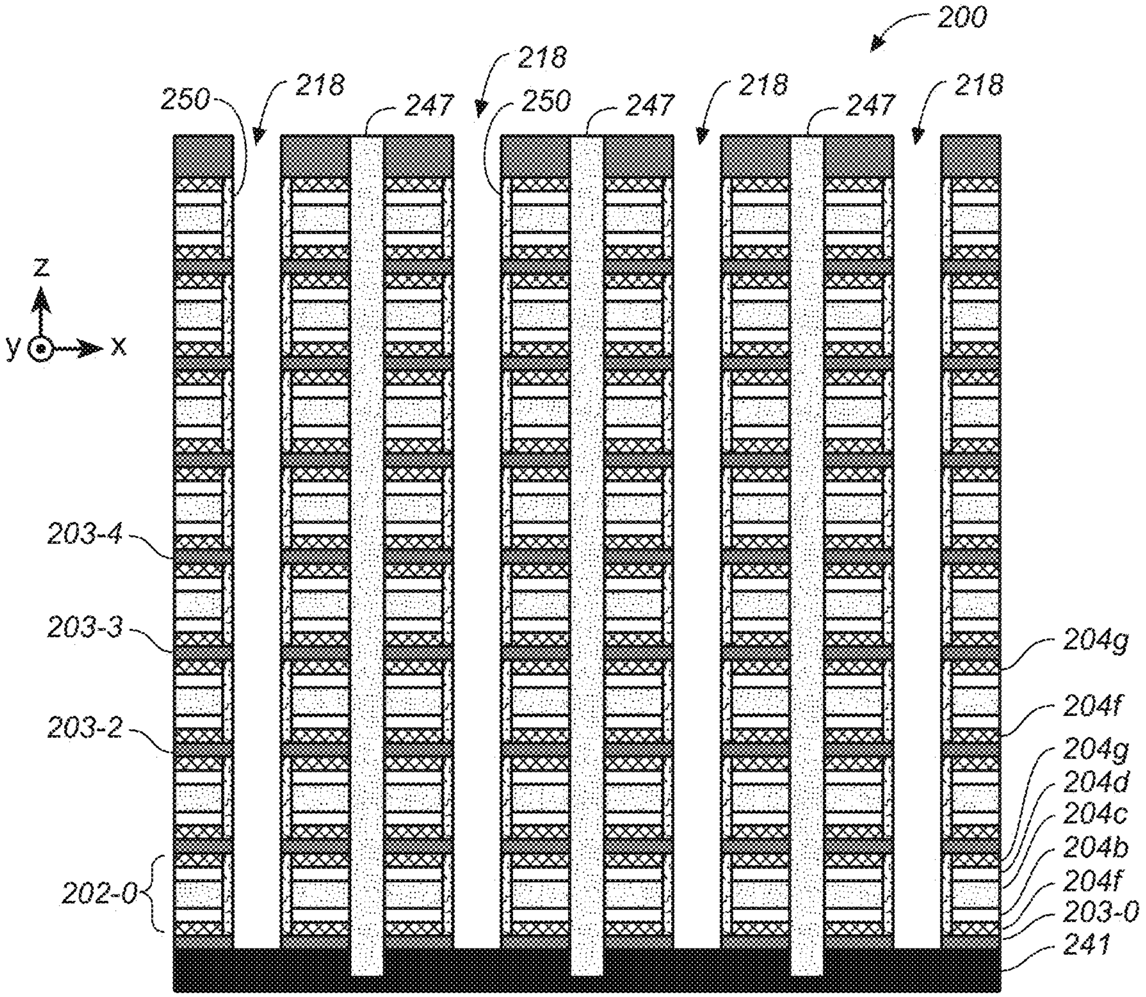

200
218
250
218
247
250
247
218
247
218
z
y
x
203-4
203-3
203-2
202-0
204g
204f
204g
204d
204c
204b
204f
203-0
241

FIG. 2k(ii)

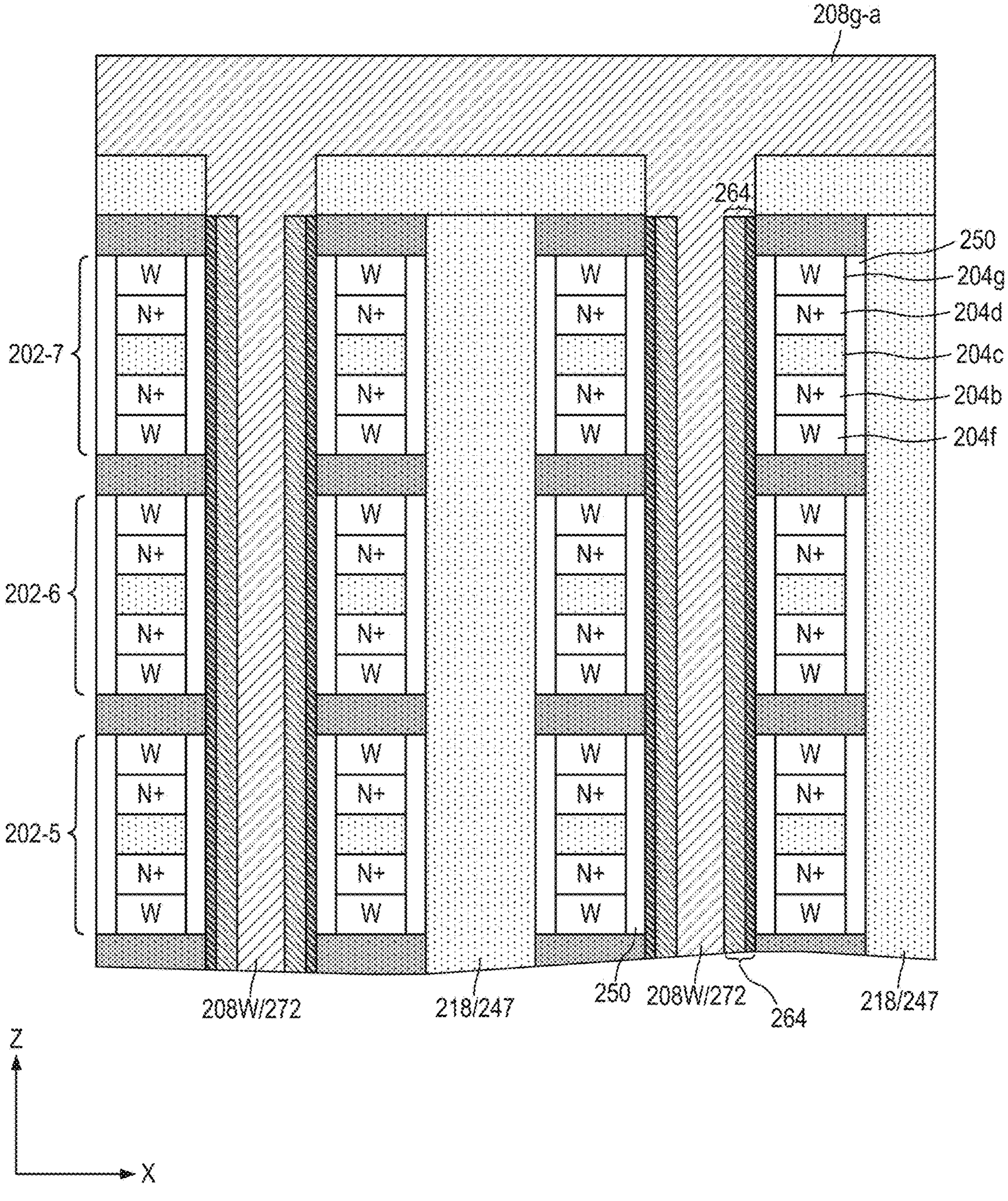

208g-a
264
250
204g
204d
204c
204b
204f
202-7
202-6
202-5
W
N+
208W/272
218/247
250
208W/272
264
218/247
Z
X

FIG. 2I

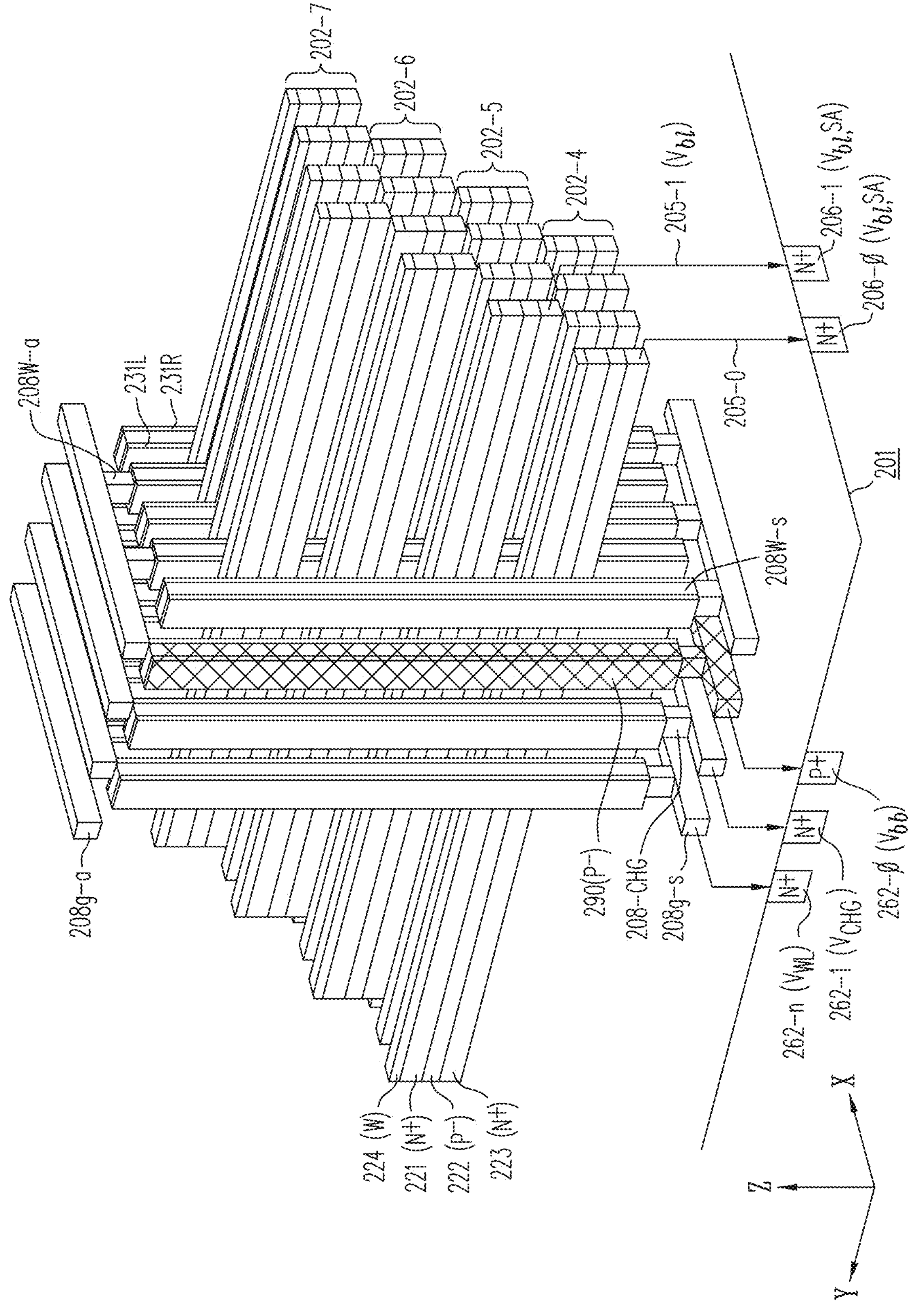
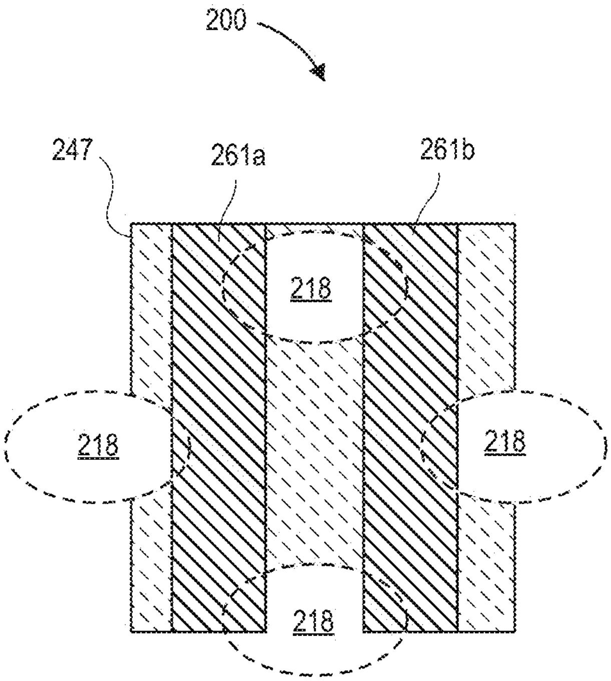
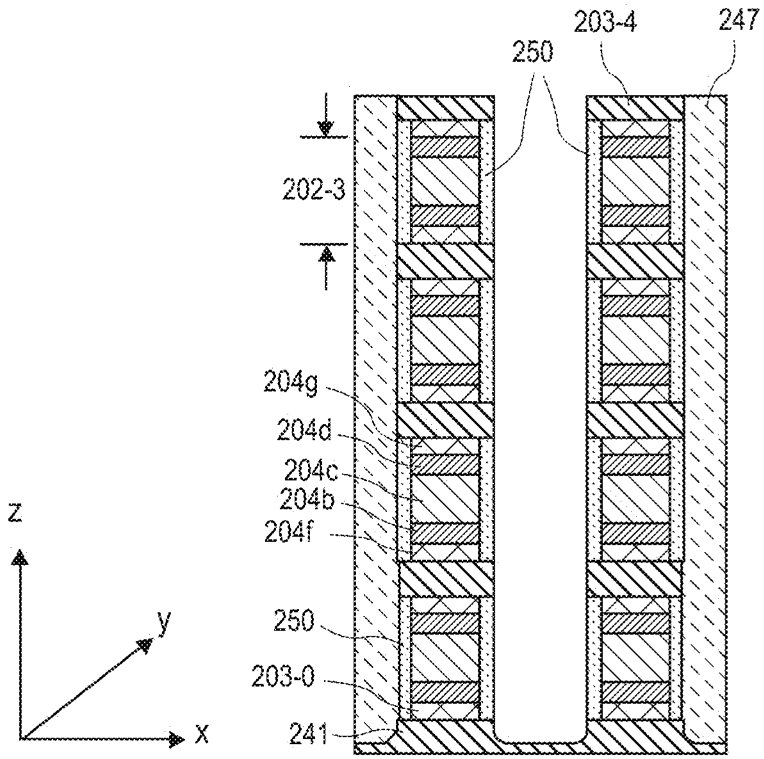

202-7
202-6
202-5
202-4
205-1 (V_bl)
206-1 (V_bl, SA)
206-0 (V_bl, SA)
205-0
N+
N+
201
208W-a
231L
231R
208W-s
P+
N+
N+
262-0 (V_bb)
262-1 (V_CHG)
262-n (V_WL)
290(P−)
208-CHG
208g-s
208g-a
224 (W)
221 (N+)
222 (P−)
223 (N+)
X
Y
Z

FIG. 2m


200
247
261a
261b
218
218
218
218

(i)


203-4
247
250
202-3
204g
204d
204c
204b
204f
250
203-0
241
z
y
x

(ii)

FIGURE 3a

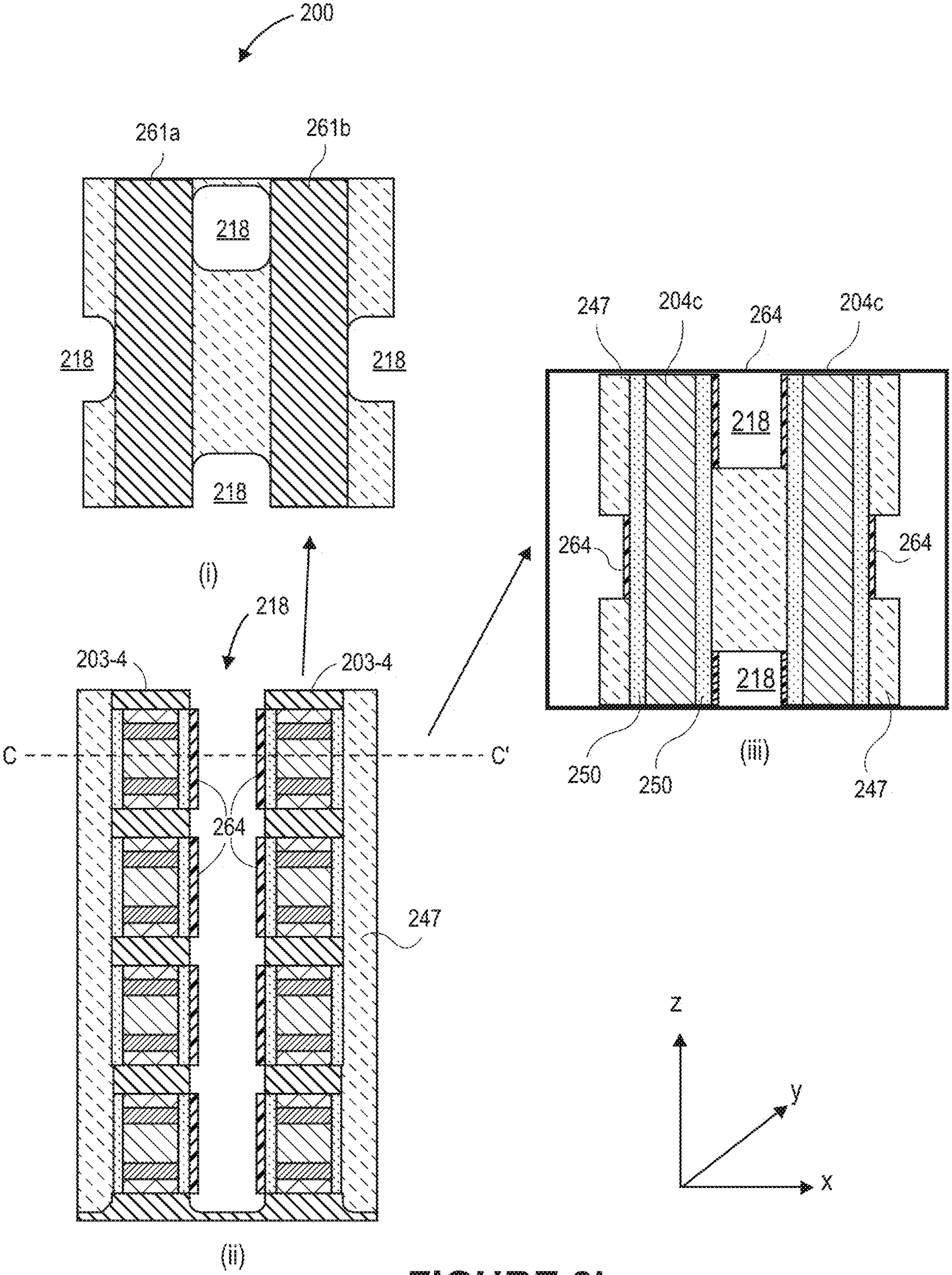

(i)
(iii)
203-4
203-4
C
C'
(ii)
z
y
x

FIGURE 3b

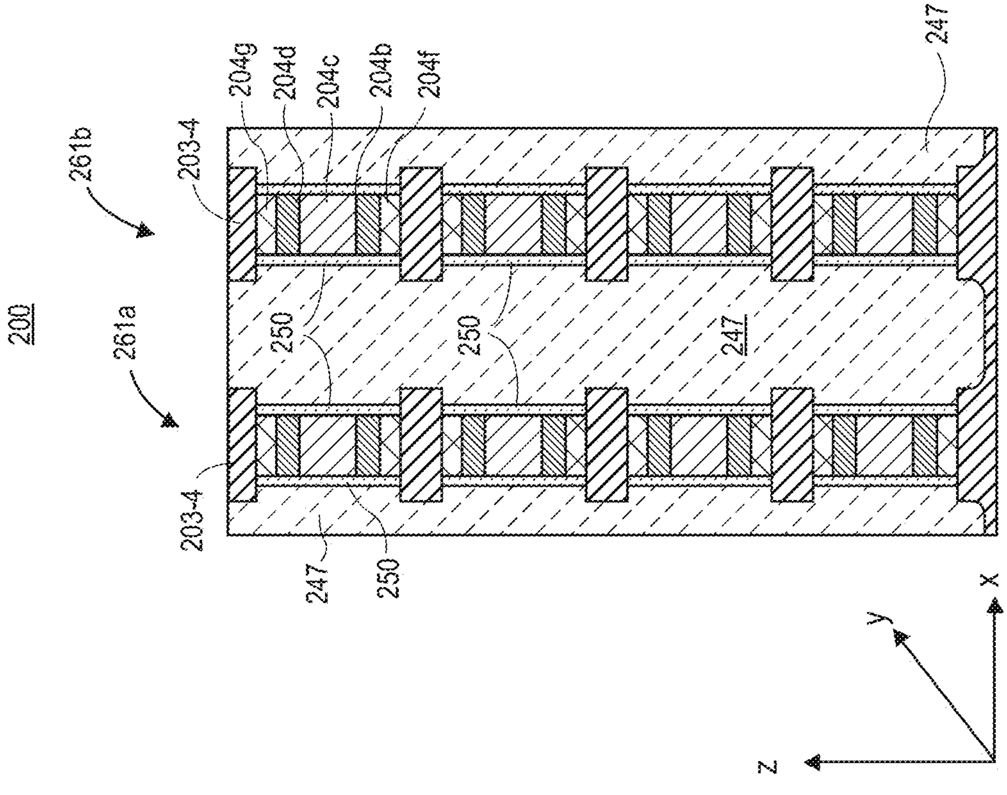

200
261a
261b
203-4
204g
204d
204c
204b
204f
247
250
203-4
247
250
250
250
247
x
y
z

FIGURE 4a

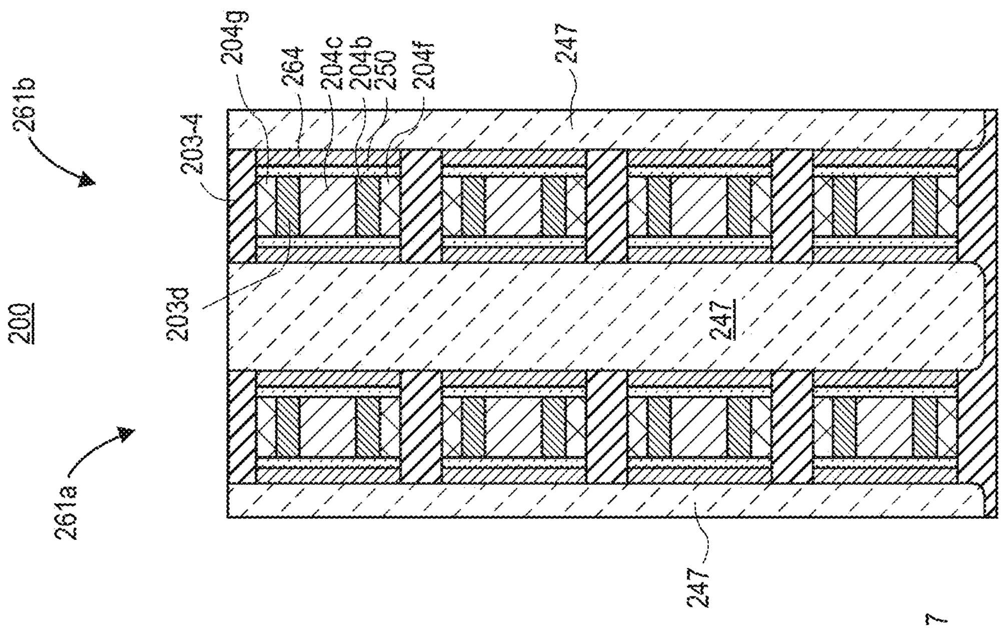

200
261b
261a
204g
264
204c
204b
250
204f
247
203-4
203d
247
247

FIGURE 4b

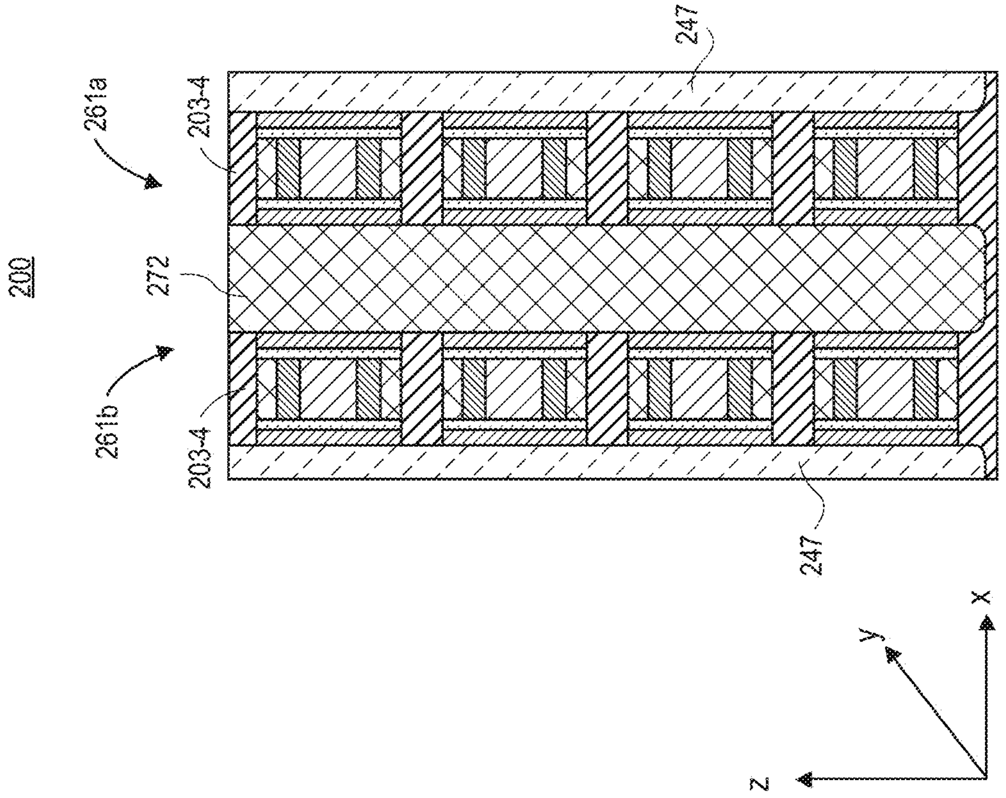

200
261a
203-4
247
272
261b
203-4
247
x
y
z

FIGURE 4C

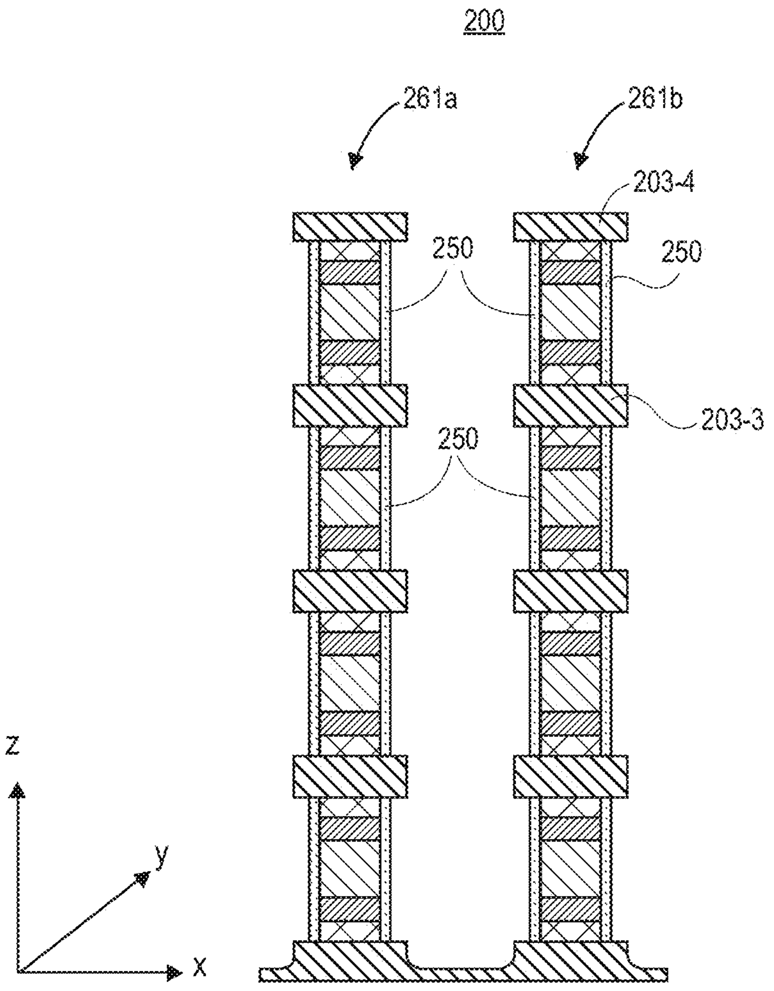

200
261a
261b
203-4
250
250
203-3
250
z
y
x

FIGURE 5a

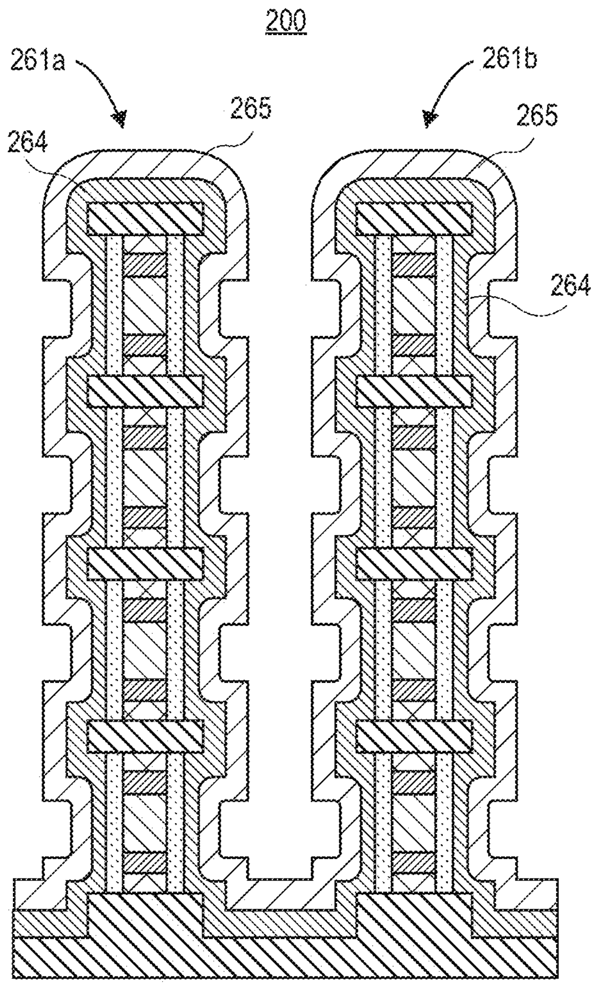

200
261a
261b
264
265
265
264

FIGURE 5b

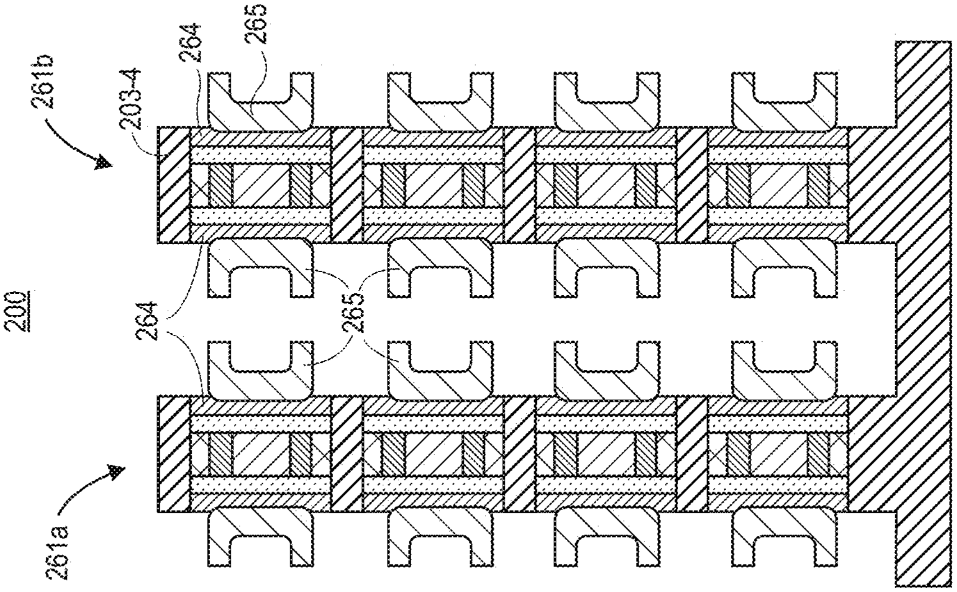

200
261a
261b
203-4
264
265
264
265

FIGURE 5d(i)

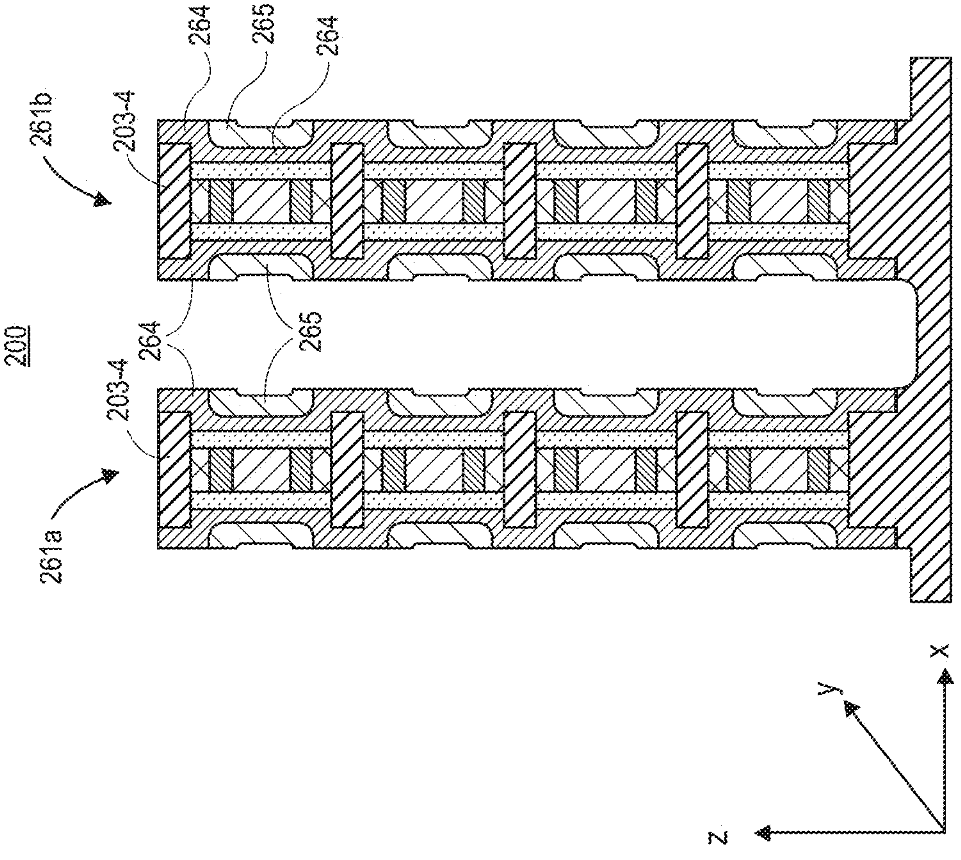

200
261a
261b
203-4
203-4
264
265
264
265
264
x
y
z

FIGURE 5c

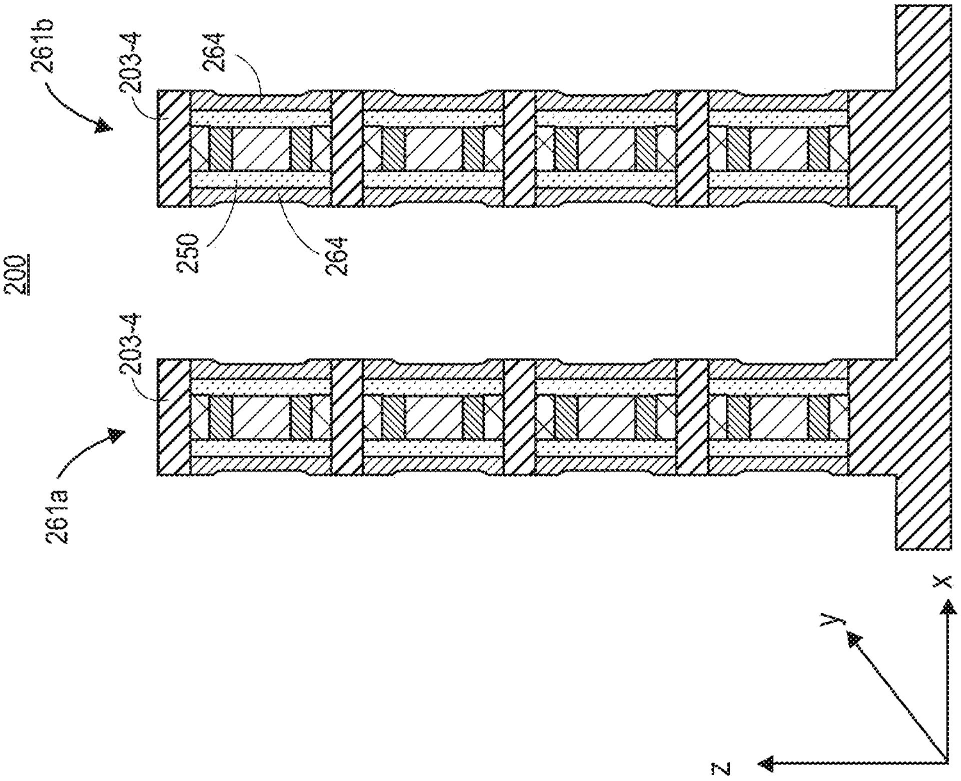

200
261a
261b
203-4
203-4
264
250
264
x
y
z

FIGURE 5e(ii)

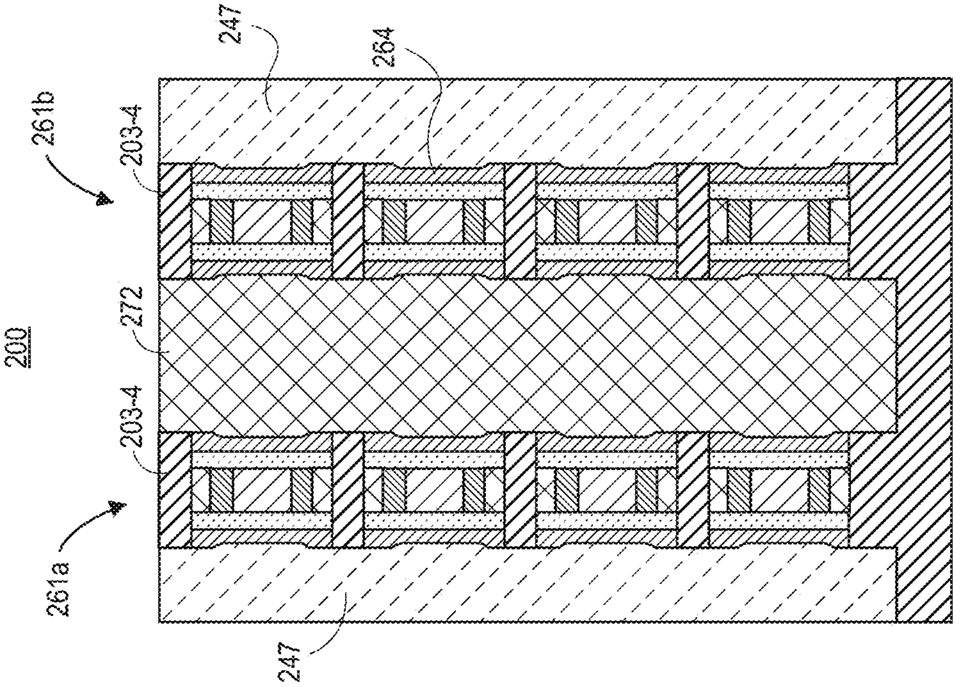

200
261a
261b
203-4
203-4
272
264
247
247

FIGURE 5f(ii)

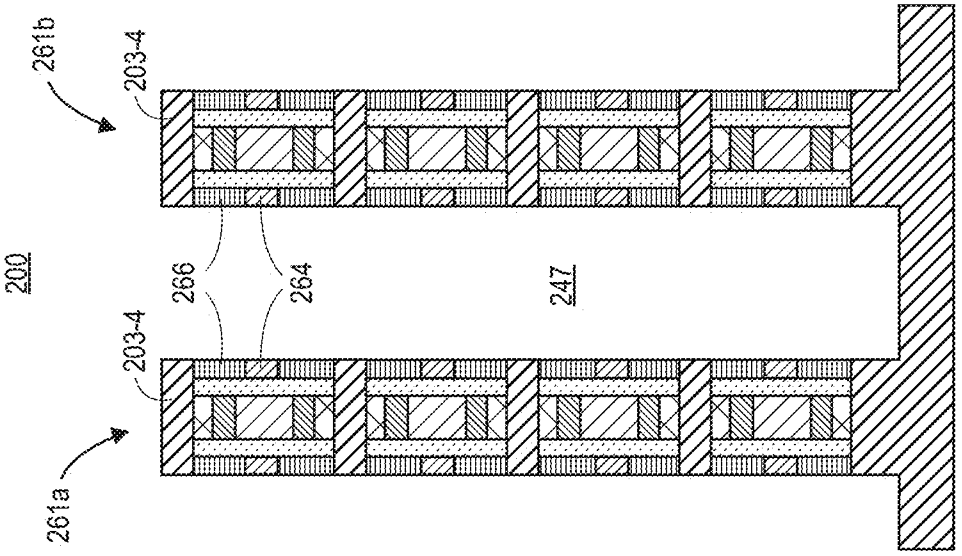

200
261b
203-4
266
264
247
203-4
261a

FIGURE 5e(iii)

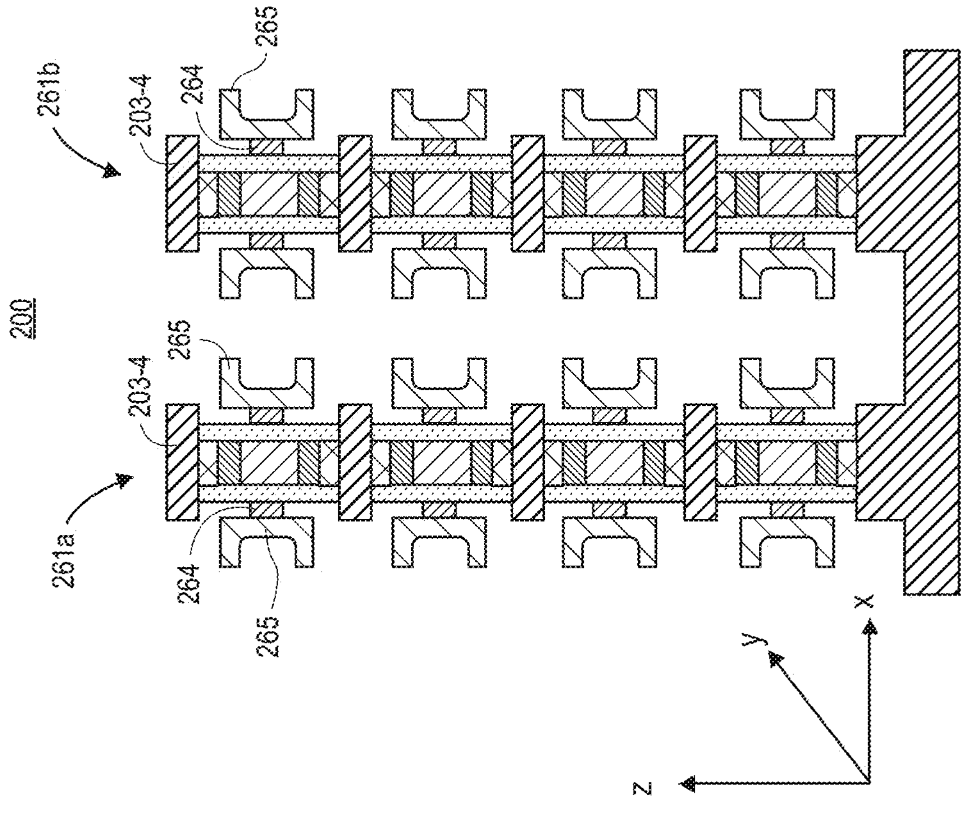

200
261b
203-4
264
265
203-4
265
261a
264
265
x
y
z

FIGURE 5d(iii)

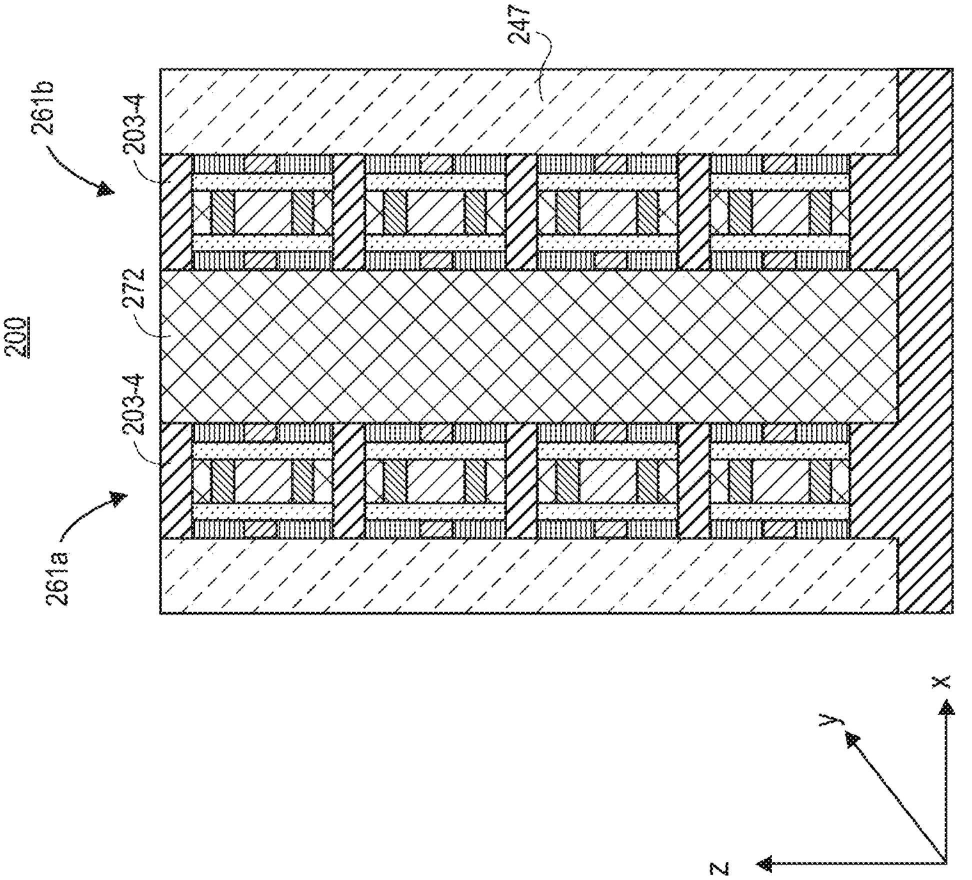

200
261b
203-4
247
272
203-4
261a
x
y
z

FIGURE 5f(iii)

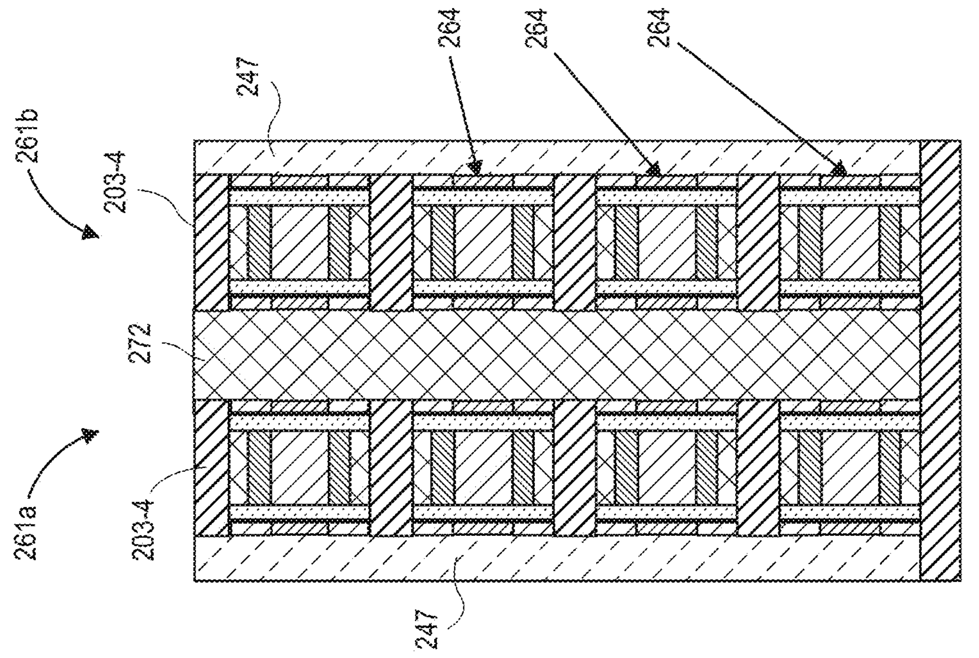

261b
247
264
264
264
203-4
272
261a
203-4
247

FIGURE 6b

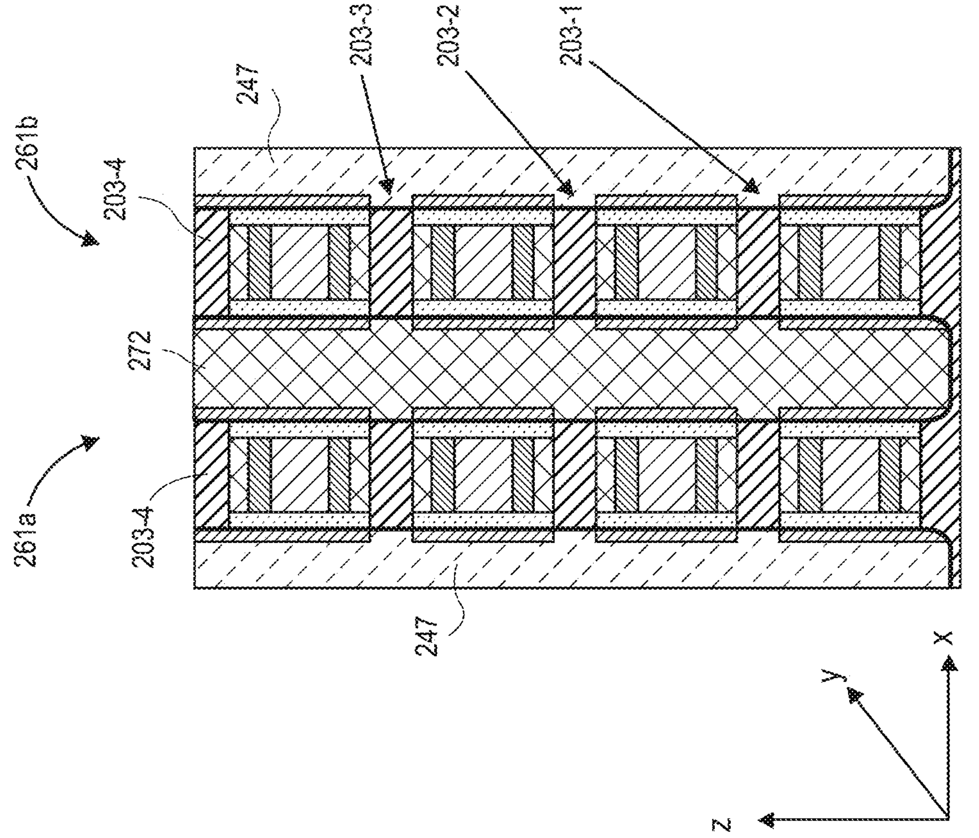

261b
247
203-3
203-2
203-1
203-4
272
261a
203-4
247
x
y
z

FIGURE 6a

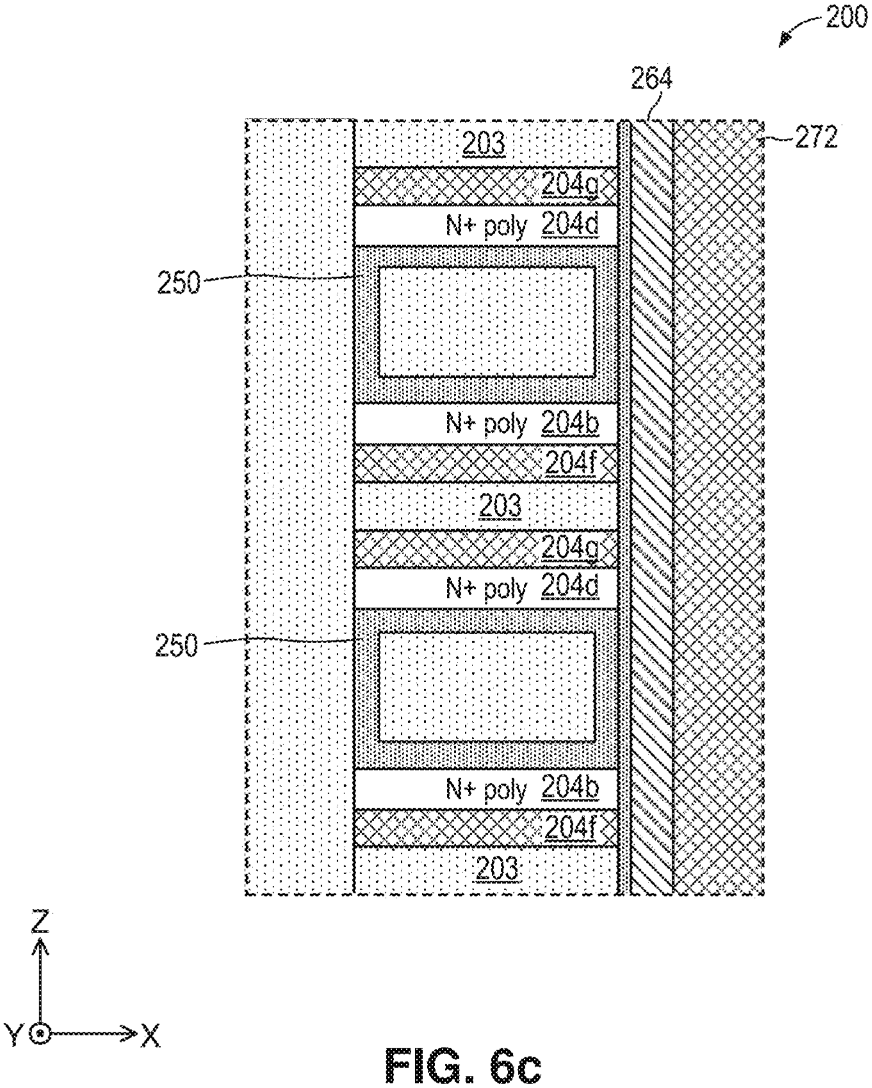

N+ poly 204d
N+ poly 204b
N+ poly 204d
N+ poly 204b
Z
Y
X

FIG. 6c

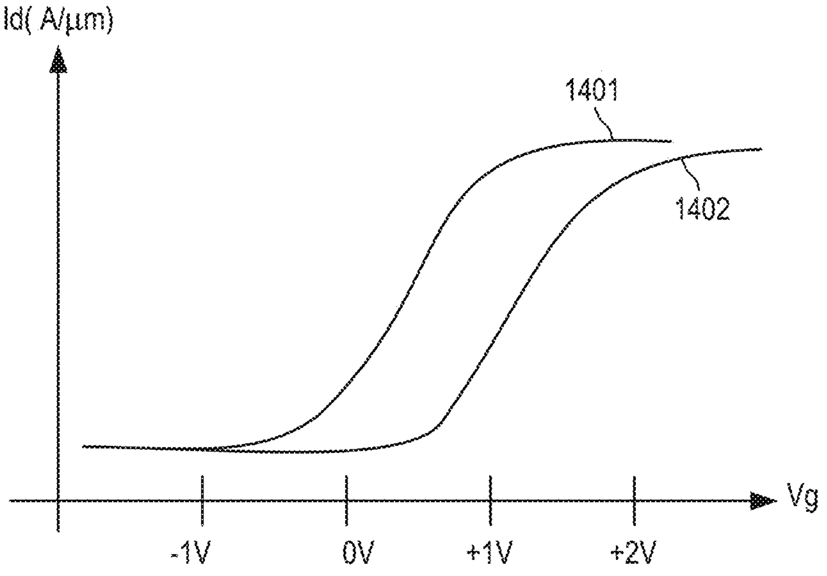


FIG. 7a

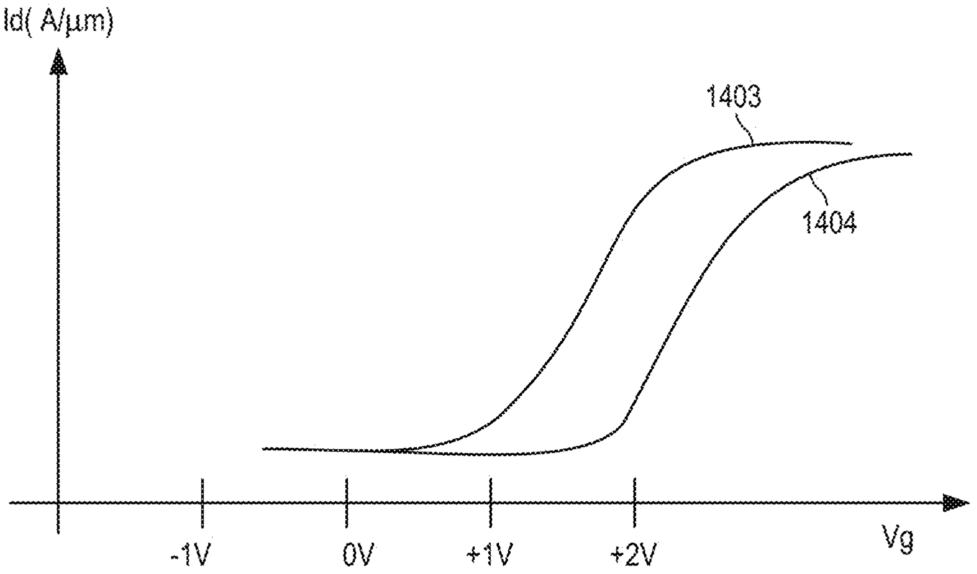


FIG. 7b

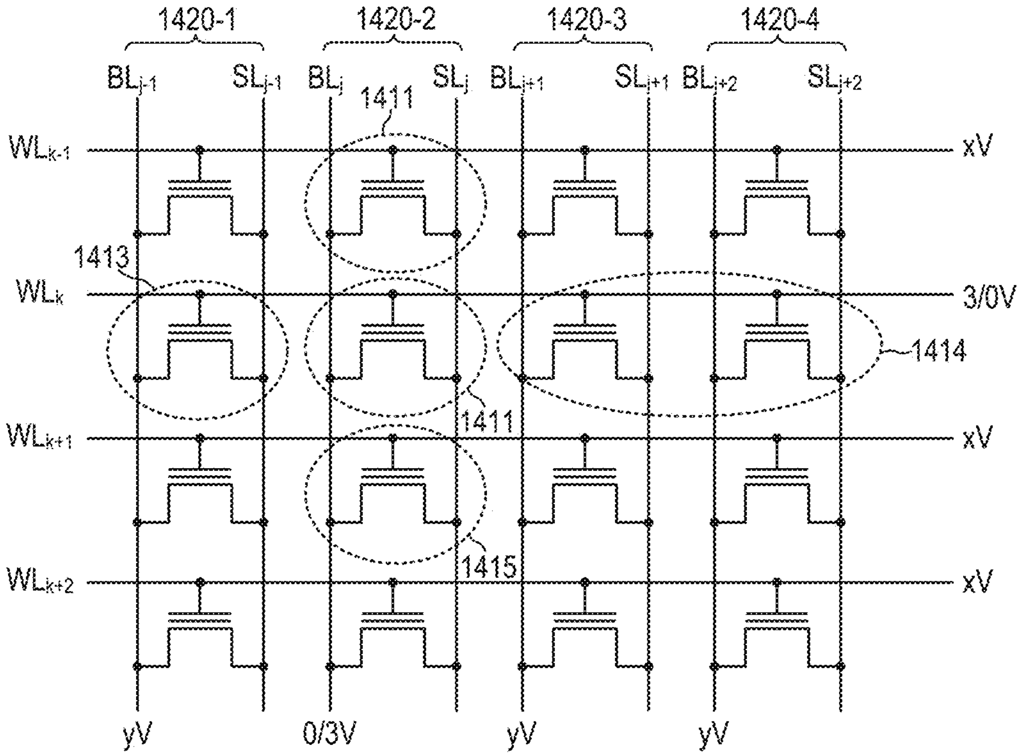

1420-1
1420-2
1420-3
1420-4
BL_{j-1}
SL_{j-1}
BL_j
SL_j
BL_{j+1}
SL_{j+1}
BL_{j+2}
SL_{j+2}
1411
WL_{k-1}
xV
1413
WL_k
3/0V
1414
1411
WL_{k+1}
xV
1415
WL_{k+2}
xV
yV
0/3V
yV
yV

FIG. 7c

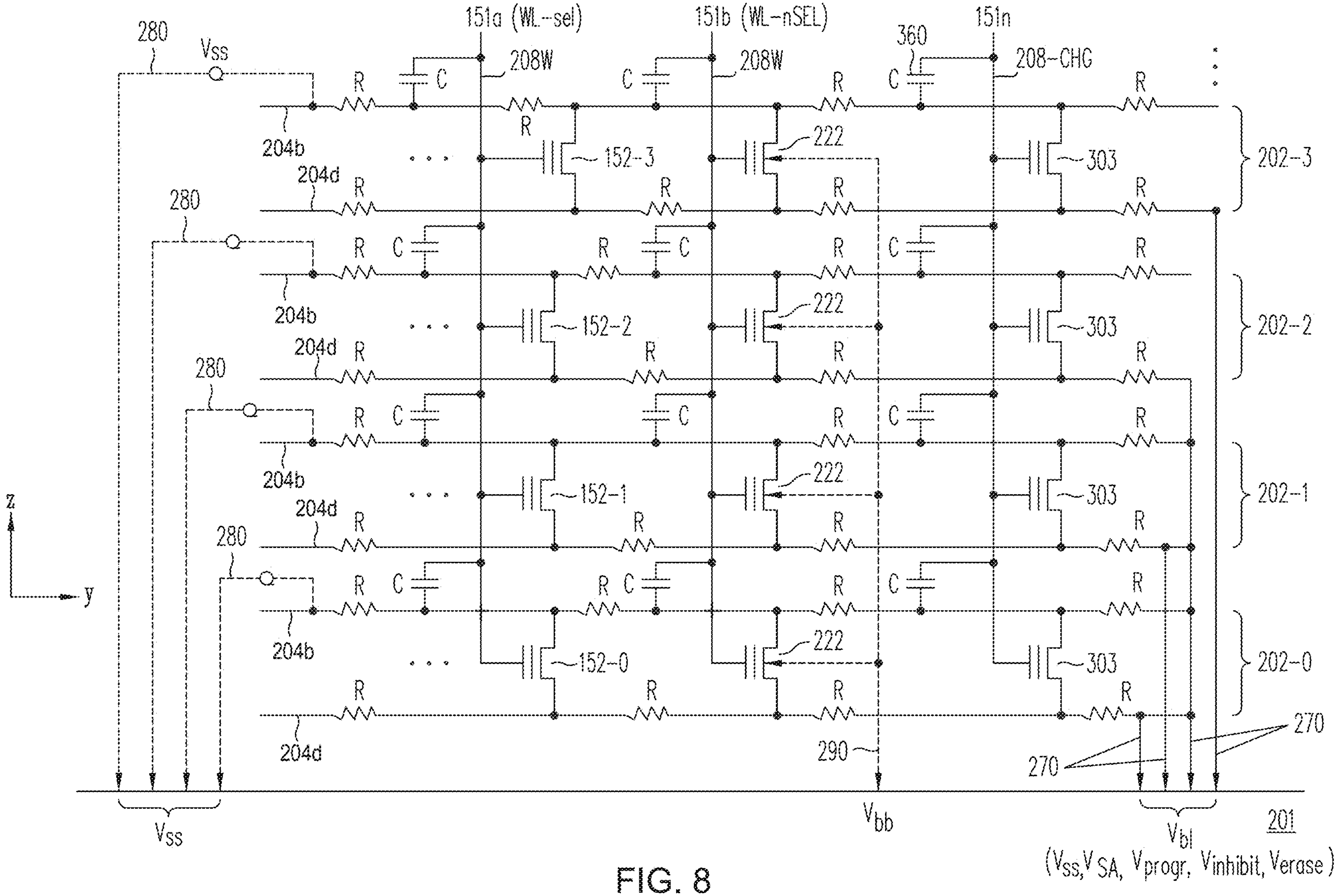

280
Vss
151a (WL-sel)
208W
151b (WL-nSEL)
208W
360
151n
208-CHG
C
R
204b
204d
152-3
152-2
152-1
152-0
222
303
202-3
202-2
202-1
202-0
280
290
270
270
z
y
Vss
Vbb
Vbl
201
(Vss,VSA, Vprogr, Vinhibit, Verase)

FIG. 8

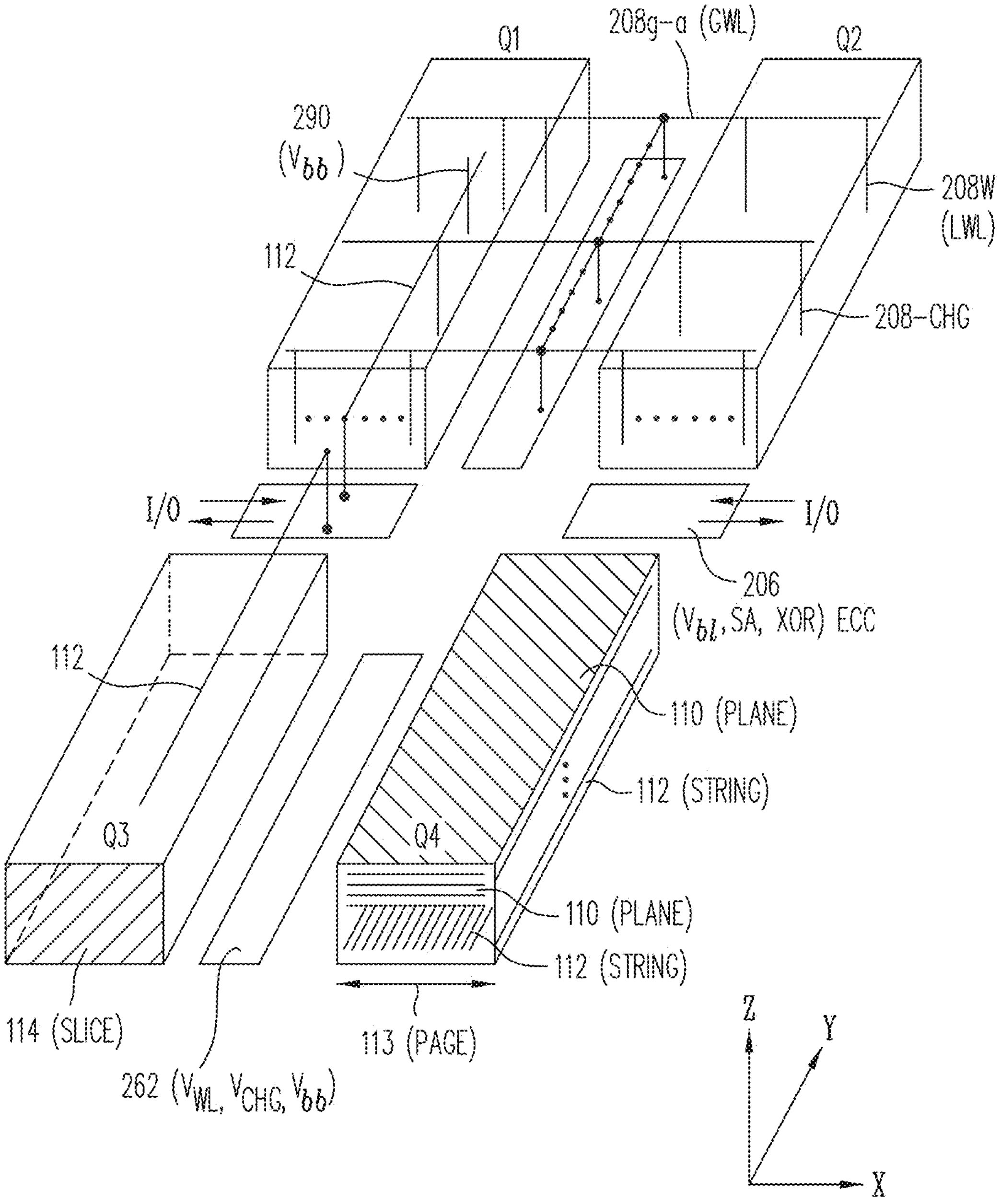

208g-a (GWL)
Q1
Q2
290
(V_{bb})
208W
(LWL)
112
208-CHG
I/O
I/O
206
(V_{bl}, SA, XOR) ECC
112
110 (PLANE)
112 (STRING)
Q3
Q4
110 (PLANE)
112 (STRING)
114 (SLICE)
113 (PAGE)
262 (V_{WL}, V_{CHG}, V_{bb})
Z
Y
X

FIG. 9A

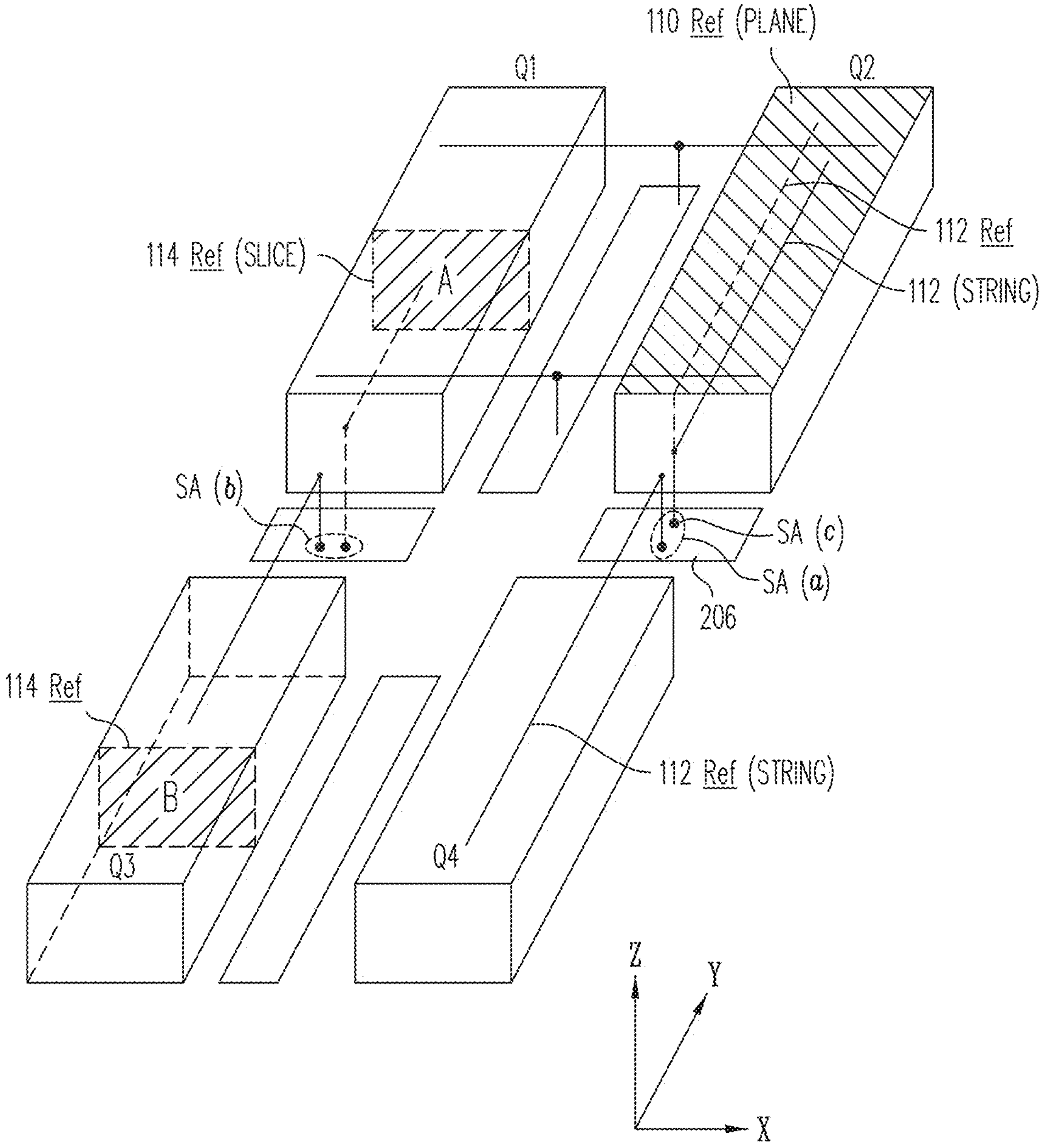

110 Ref (PLANE)
Q1
Q2
114 Ref (SLICE)
A
112 Ref
112 (STRING)
SA (b)
SA (c)
SA (a)
206
114 Ref
B
Q3
Q4
112 Ref (STRING)
Z
Y
X

FIG. 9b

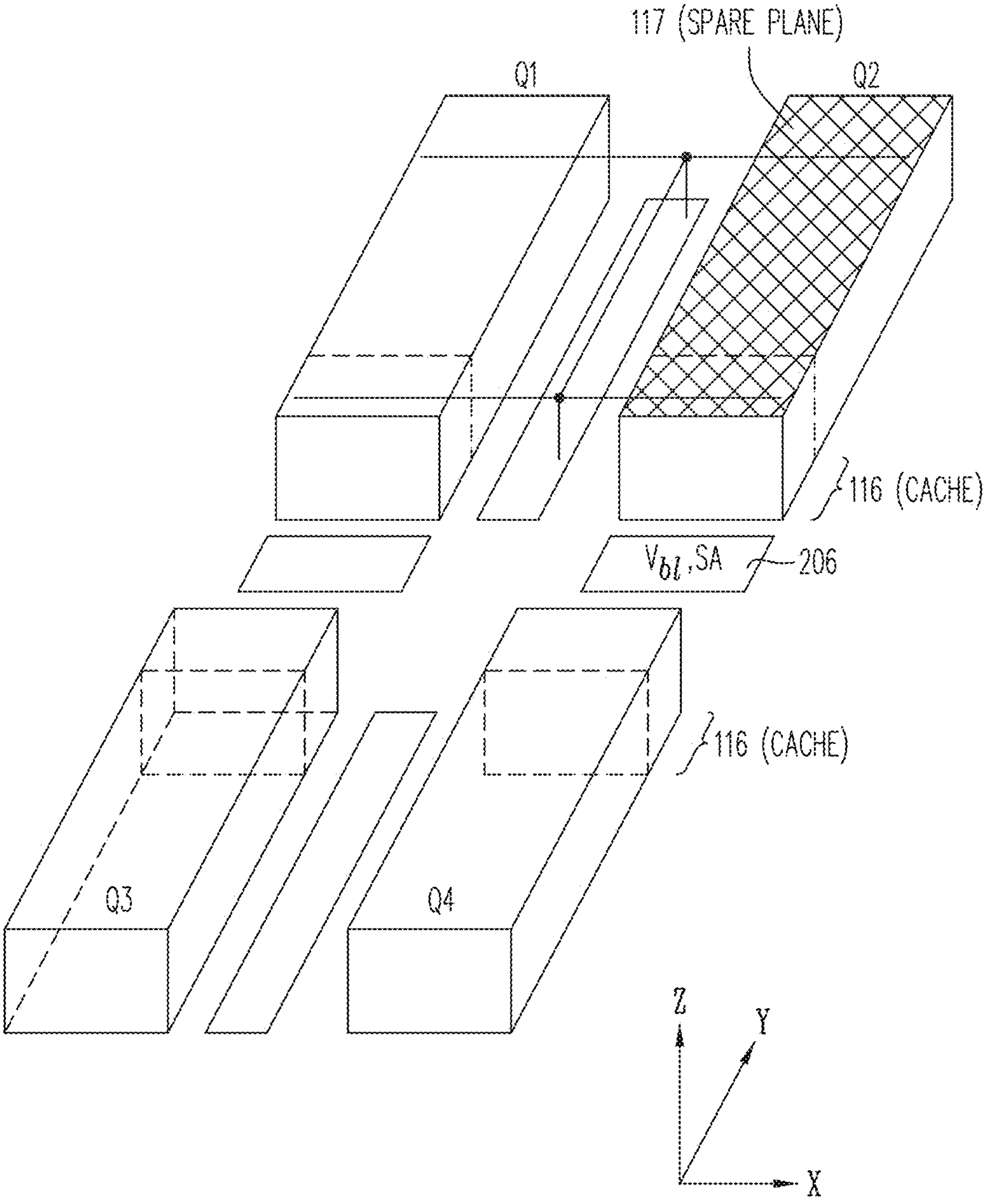

117 (SPARE PLANE)
Q1
Q2
116 (CACHE)
V_{bl},SA
206
116 (CACHE)
Q3
Q4
Z
Y
X

FIG. 9c

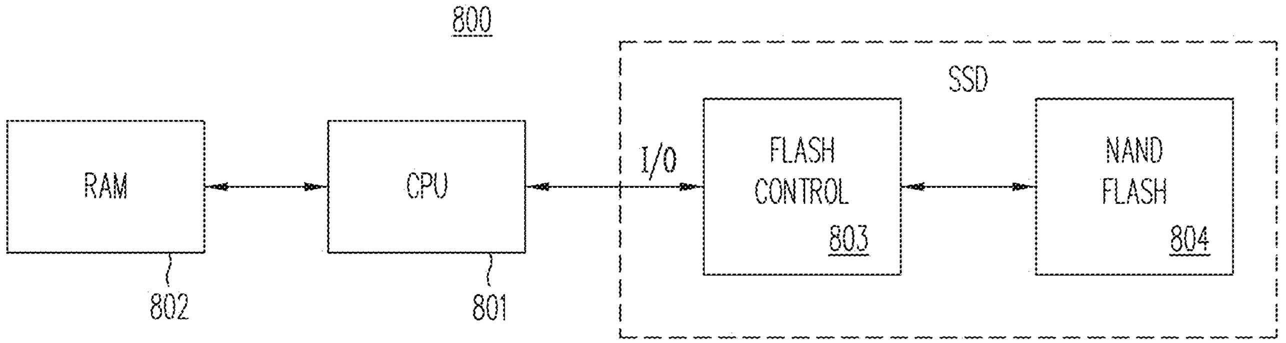

800
RAM
802
CPU
801
SSD
I/O
FLASH CONTROL
803
NAND FLASH
804

FIG. 10a

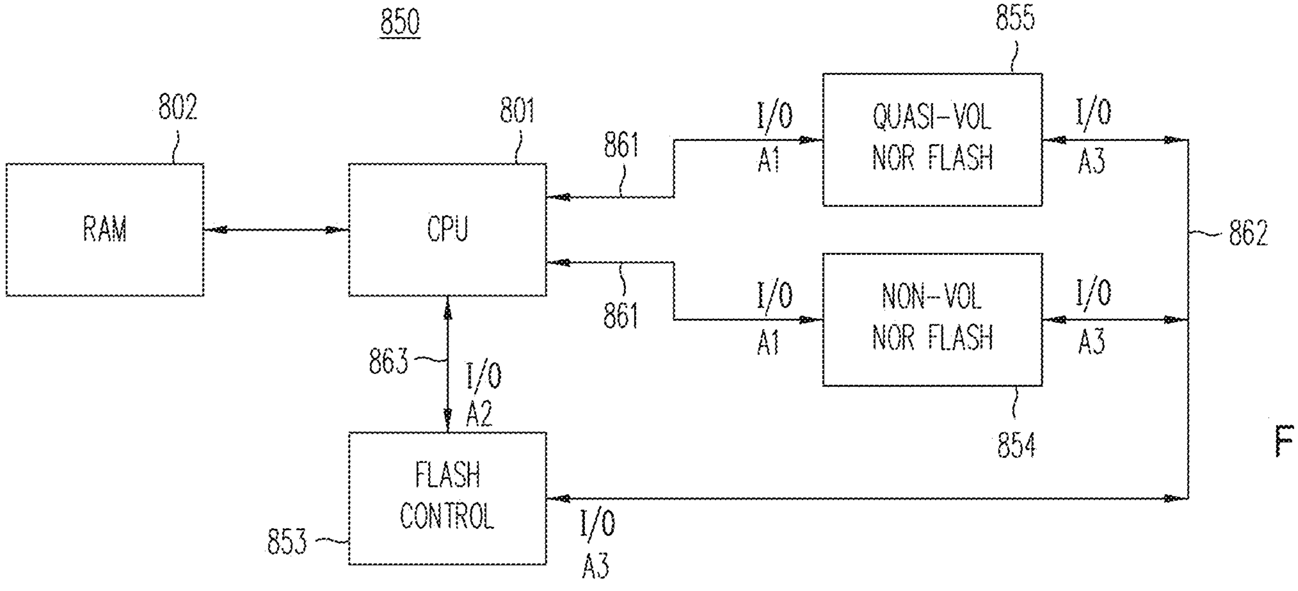

850
802
RAM
801
CPU
861
861
I/O
A1
I/O
A1
855
QUASI-VOL NOR FLASH
NON-VOL NOR FLASH
854
I/O
A3
I/O
A3
862
863
I/O
A2
FLASH CONTROL
853
I/O
A3

FIG. 10b

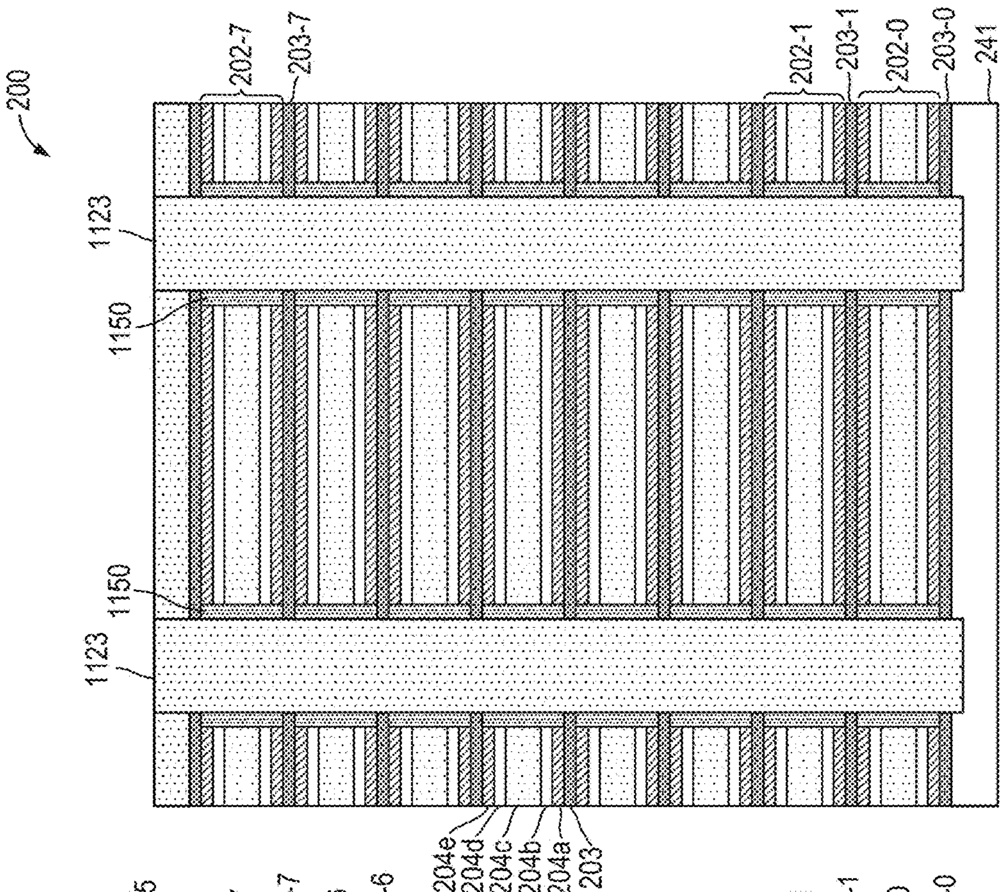

200
202-7
203-7
202-1
203-1
202-0
203-0
241
1123
1150
1150
1123
204e
204d
204c
204b
204a
203

FIG. 11b

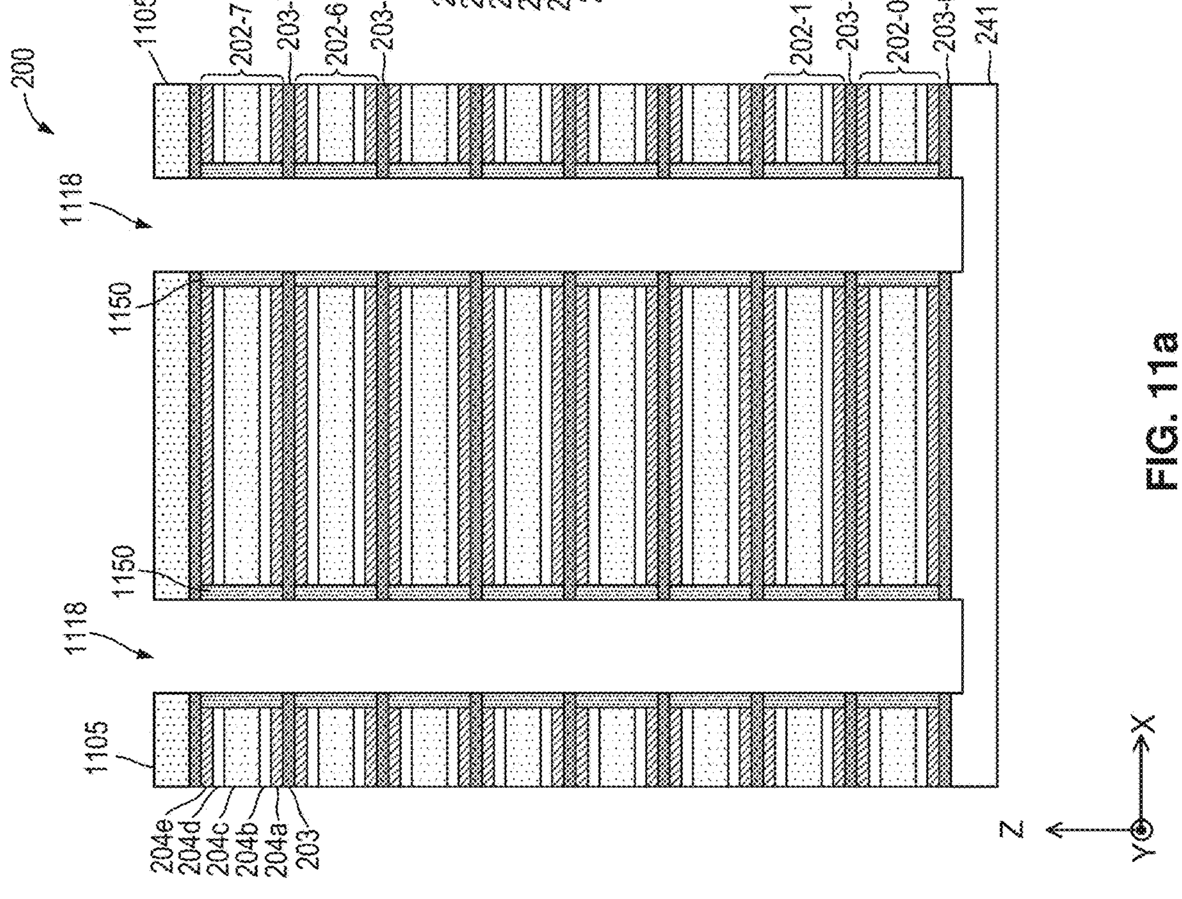

200
1105
202-7
203-7
202-6
203-6
202-1
203-1
202-0
203-0
241
1118
1150
1150
1118
1105
204e
204d
204c
204b
204a
203
Z
X
Y

FIG. 11a

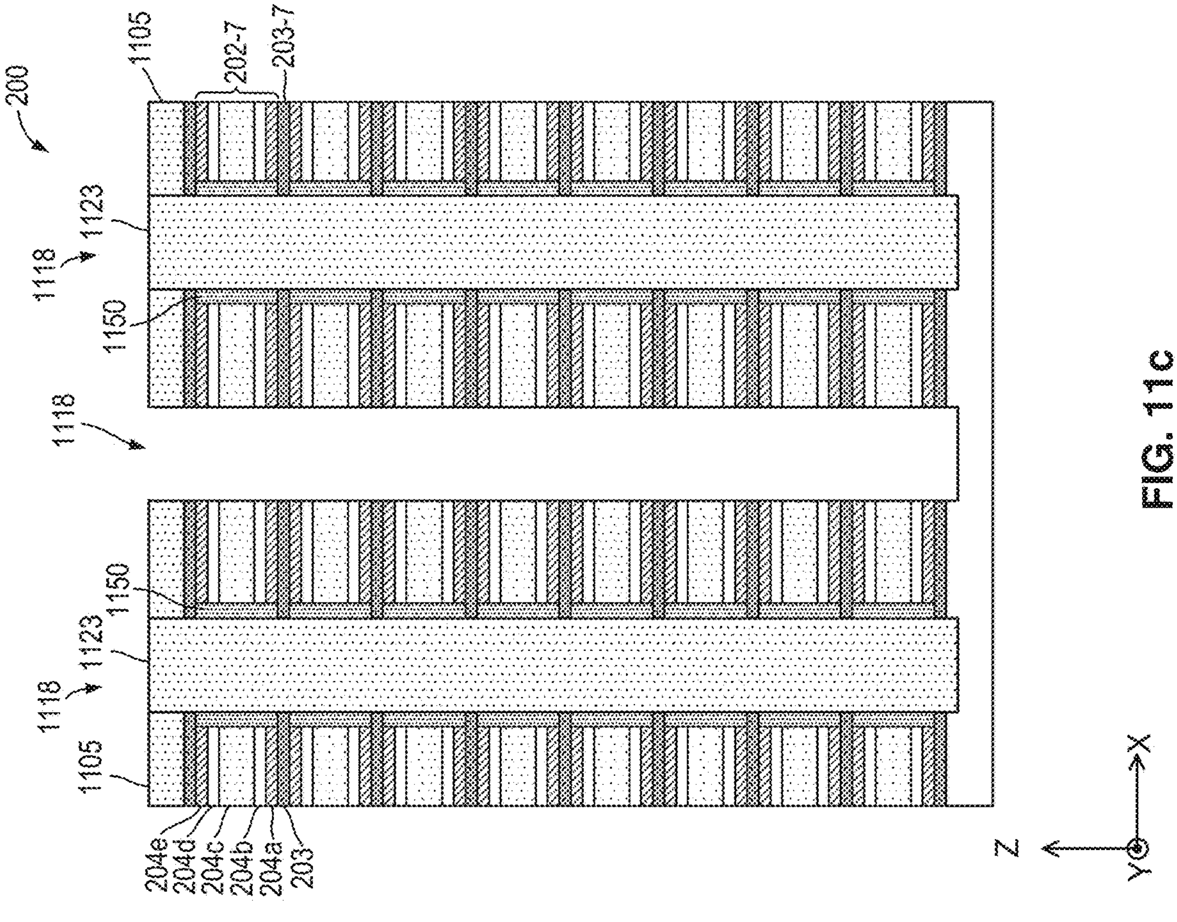

200
1105
202-7
203-7
1123
1118
1150
1118
1150
1123
1118
1105
204e
204d
204c
204b
204a
203
Z
X
Y

FIG. 11c

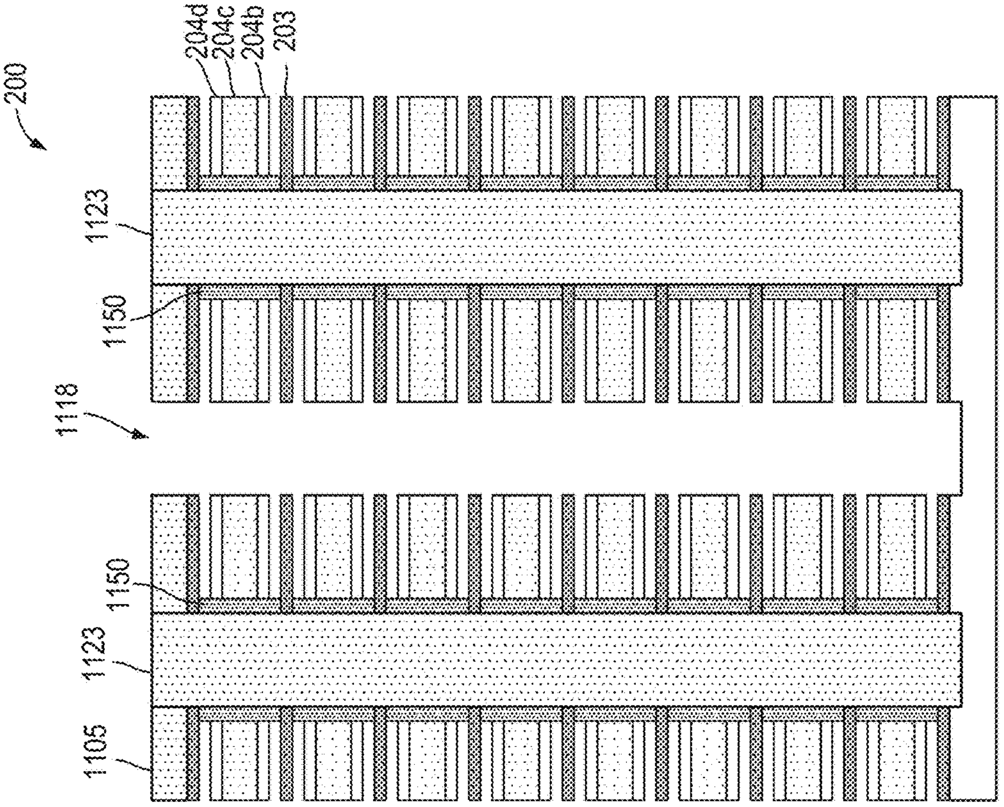

200
204d
204c
204b
203
1123
1150
1118
1150
1123
1105

FIG. 11d

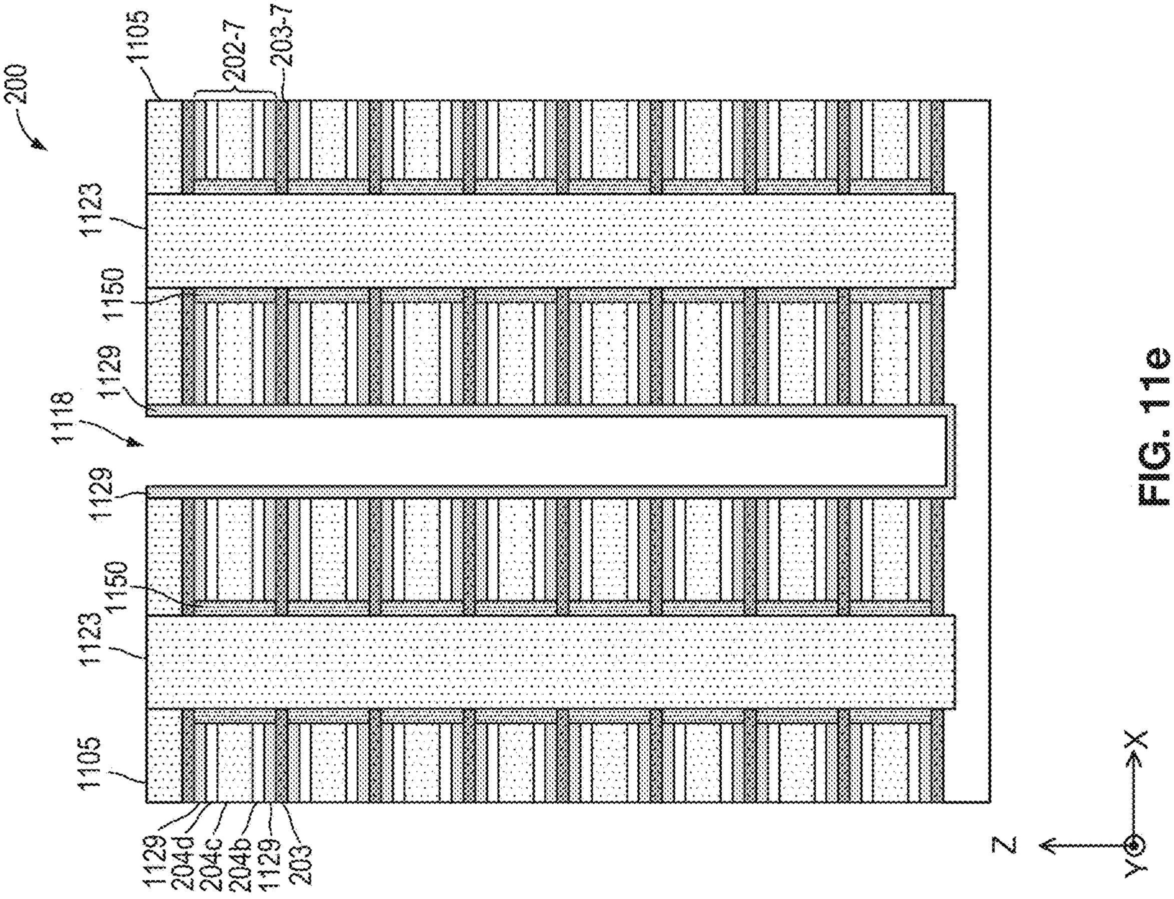

200
1105
202-7
203-7
1123
1150
1129
1118
1129
1150
1123
1105
1129
204d
204c
204b
1129
203
Z
X
Y

FIG. 11e

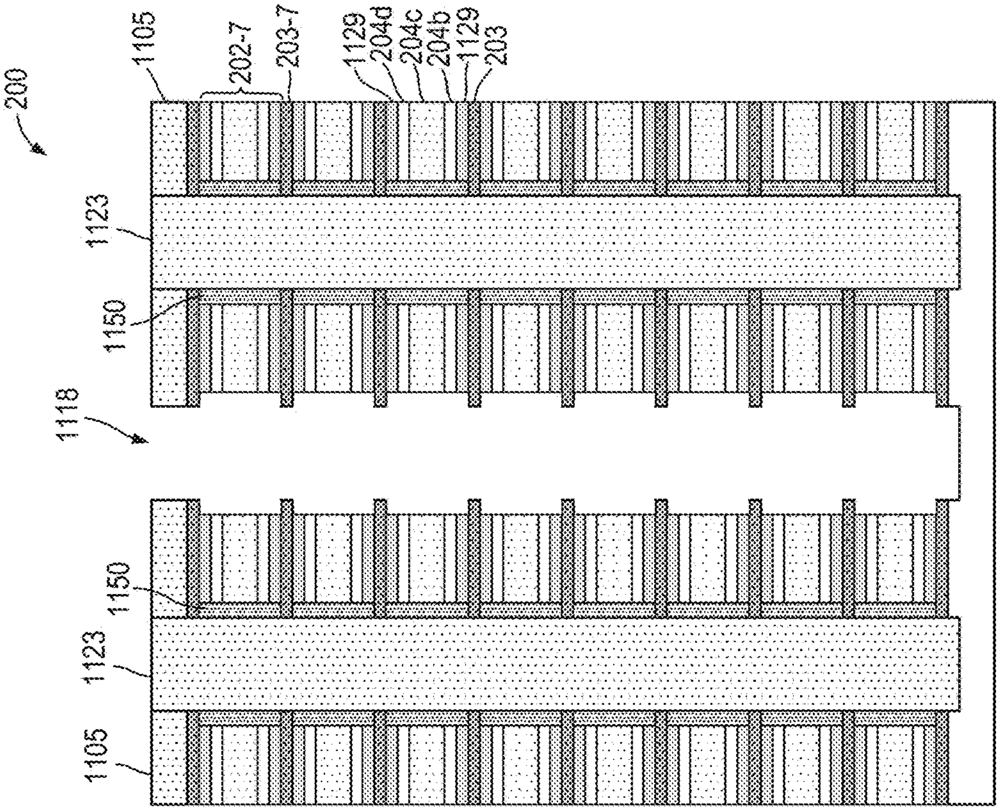

200
1105
202-7
203-7
1129
204d
204c
204b
1129
203
1123
1150
1118
1150
1123
1105

FIG. 11f

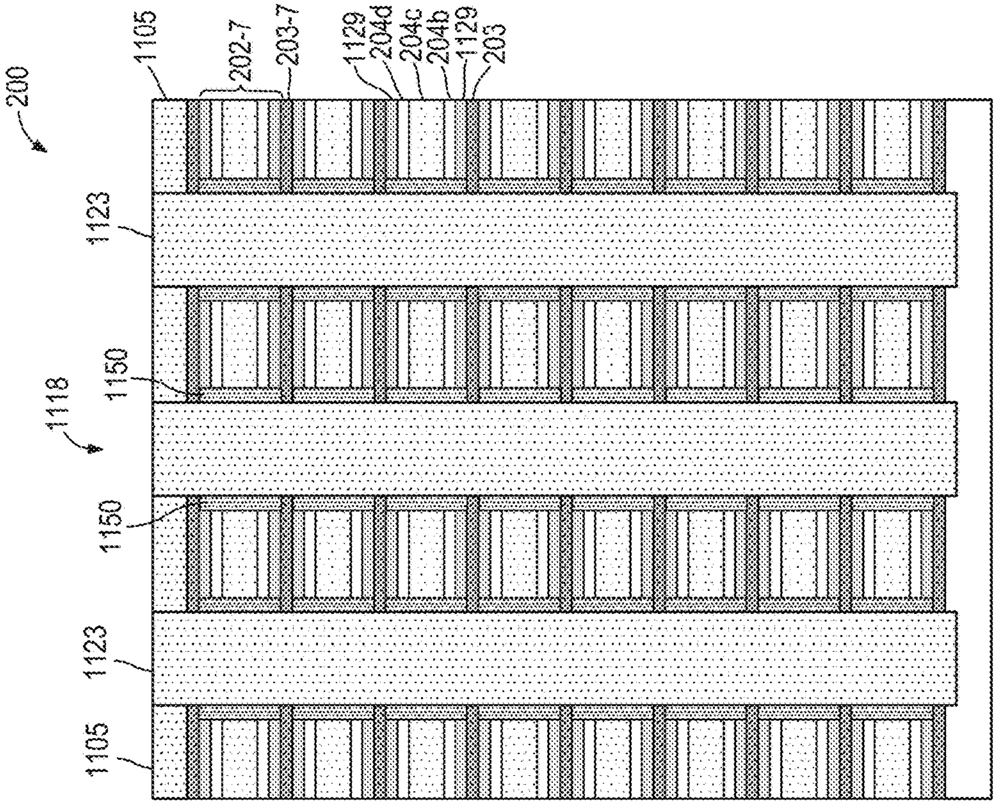

200
1105
202-7
203-7
1129
204d
204c
204b
1129
203
1123
1150
1118
1150
1123
1105

FIG. 11h

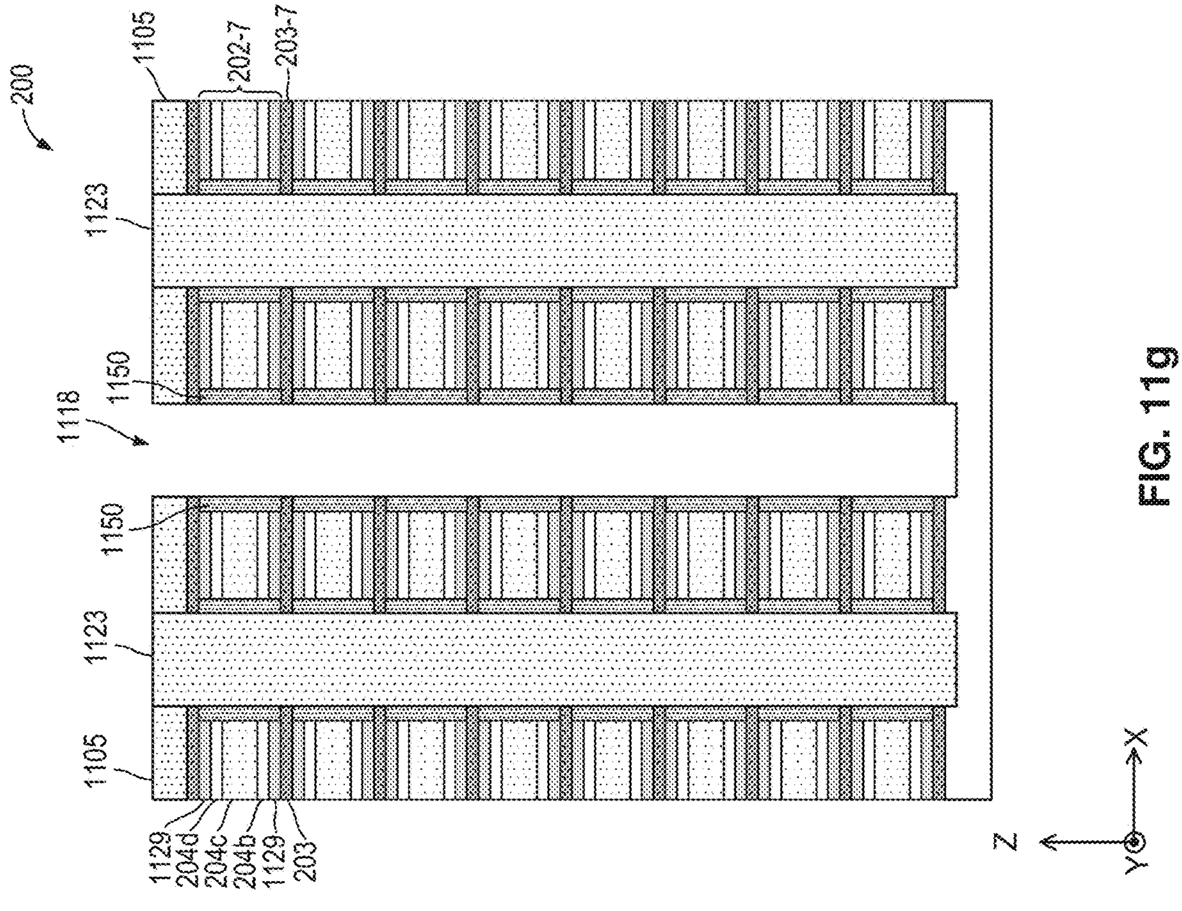

200
1105
202-7
203-7
1123
1150
1118
1150
1123
1105
1129
204d
204c
204b
1129
203
Z
X
Y

FIG. 11g

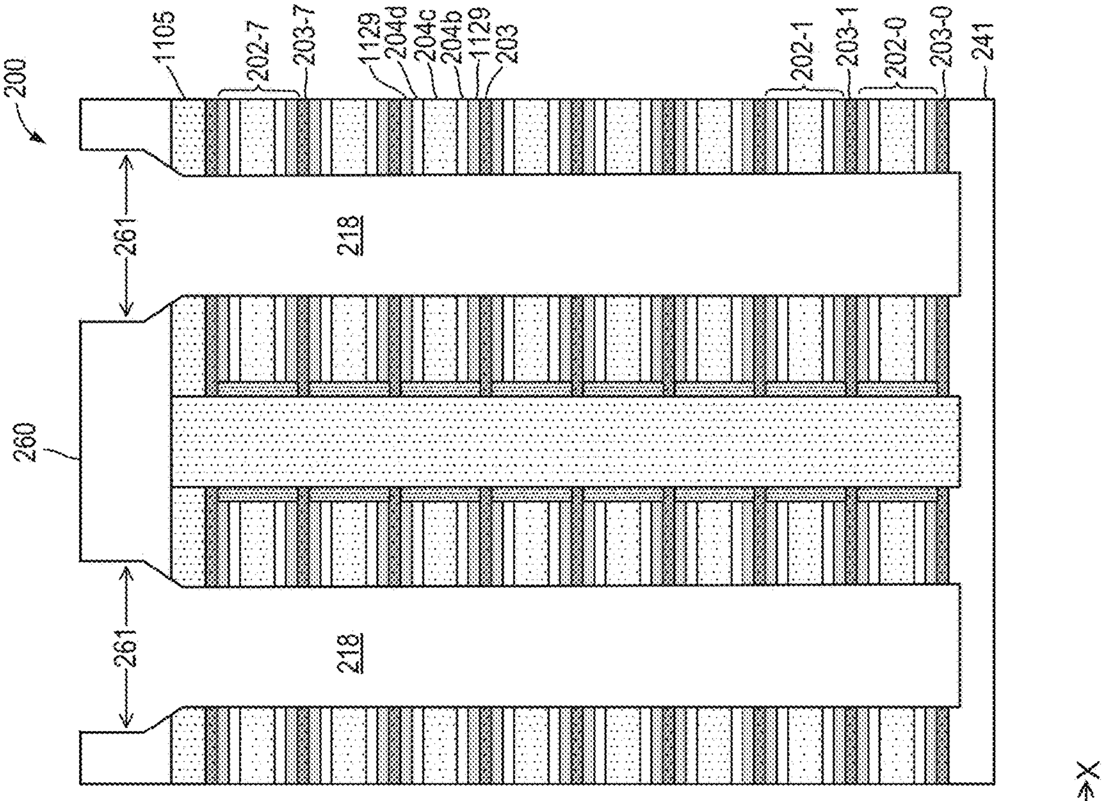

200
1105
202-7
203-7
1129
204d
204c
204b
1129
203
202-1
203-1
202-0
203-0
241
261
218
260
261
218
Z
X
Y

FIG. 11i(ii)

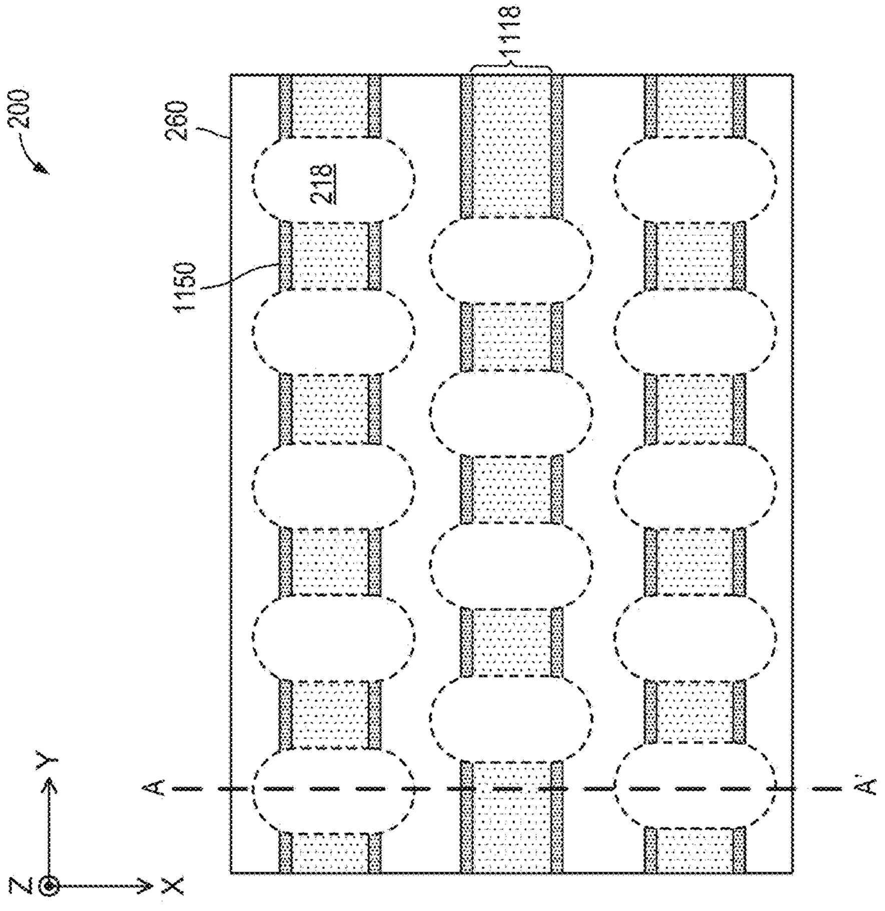

200
260
218
1150
1118
A
A'
Y
Z
X

FIG. 11i(i)

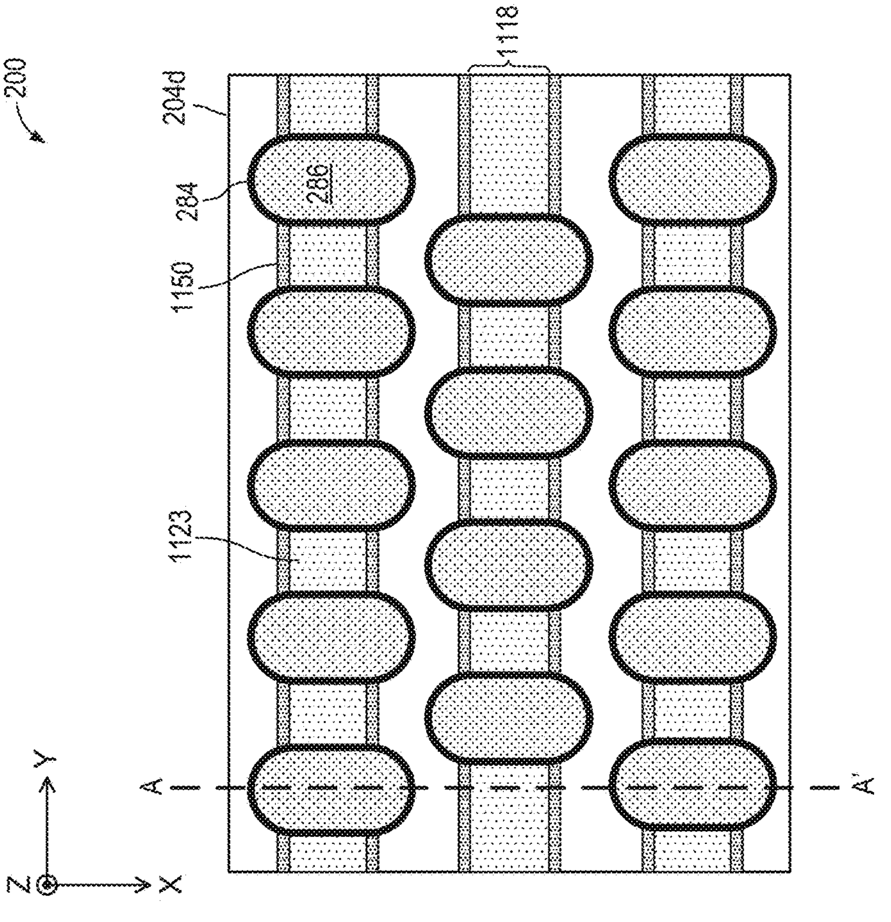

200
204d
284
286
1150
1123
1118
A
A'
Y
X
Z

FIG. 11J(i)

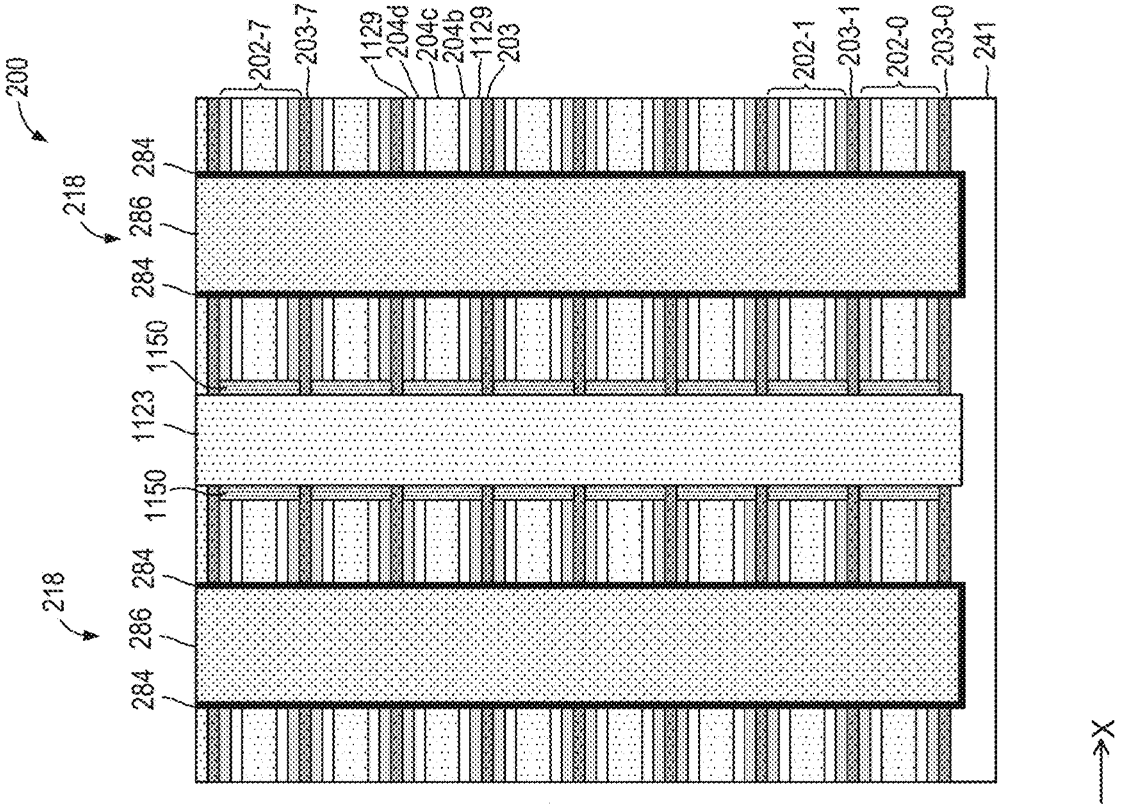

200
218
284
286
284
1150
1123
1150
218
284
286
284
202-7
203-7
1129
204d
204c
204b
1129
203
202-1
203-1
202-0
203-0
241
Z
X
Y

FIG. 11J(ii)

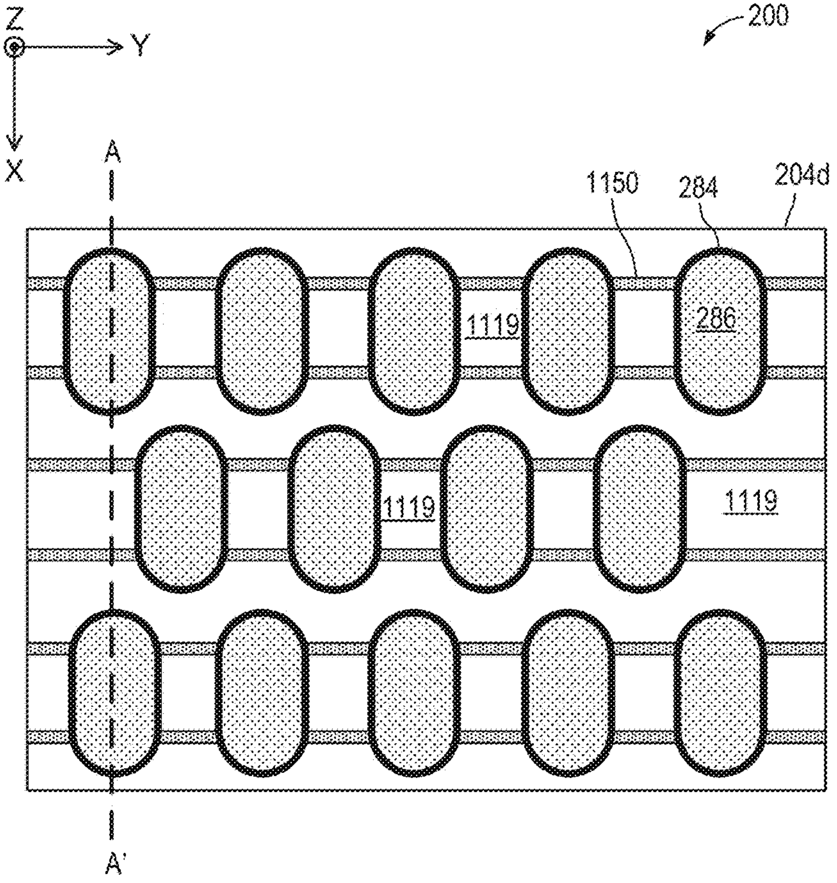

Z
Y
X
200
A
A'
1150
284
204d
286
1119
1119
1119

FIG. 11k(i)

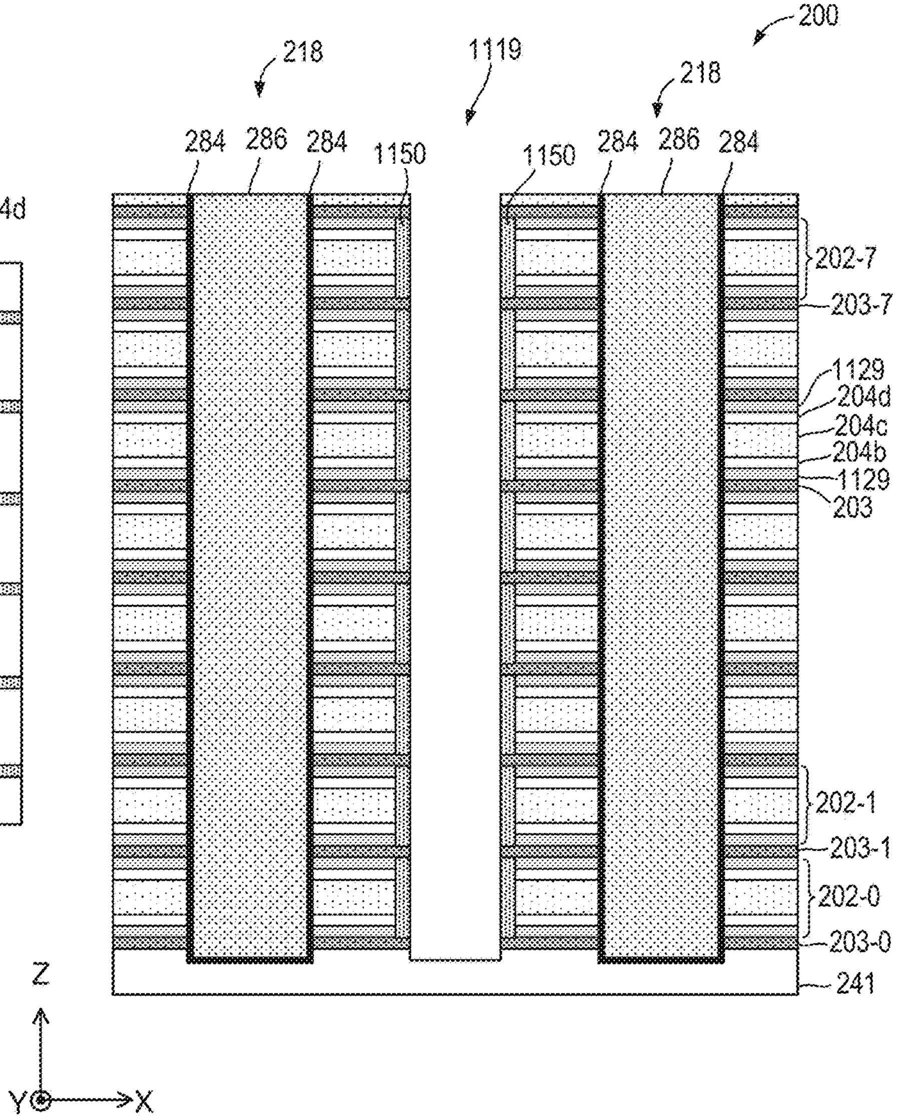

200
218
1119
218
284
286
284
1150
1150
284
286
284
202-7
203-7
1129
204d
204c
204b
1129
203
202-1
203-1
202-0
203-0
241
Z
Y
X

FIG. 11k(ii)

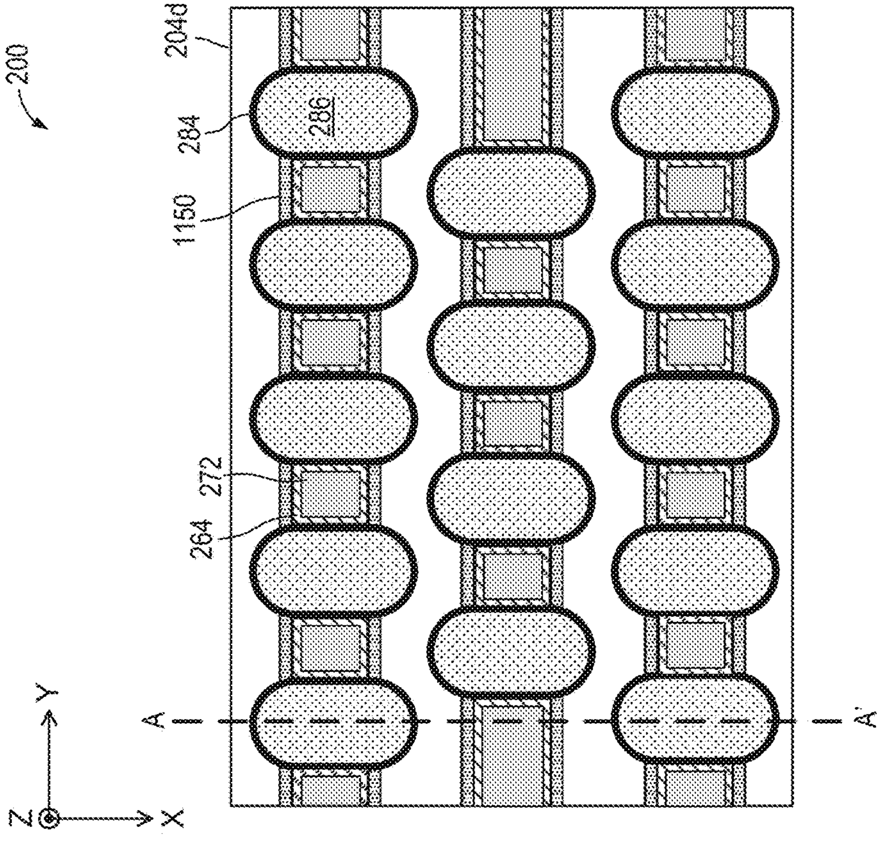

200
204d
284
286
1150
272
264
A
A'
Y
X
Z

FIG. 11I(i)

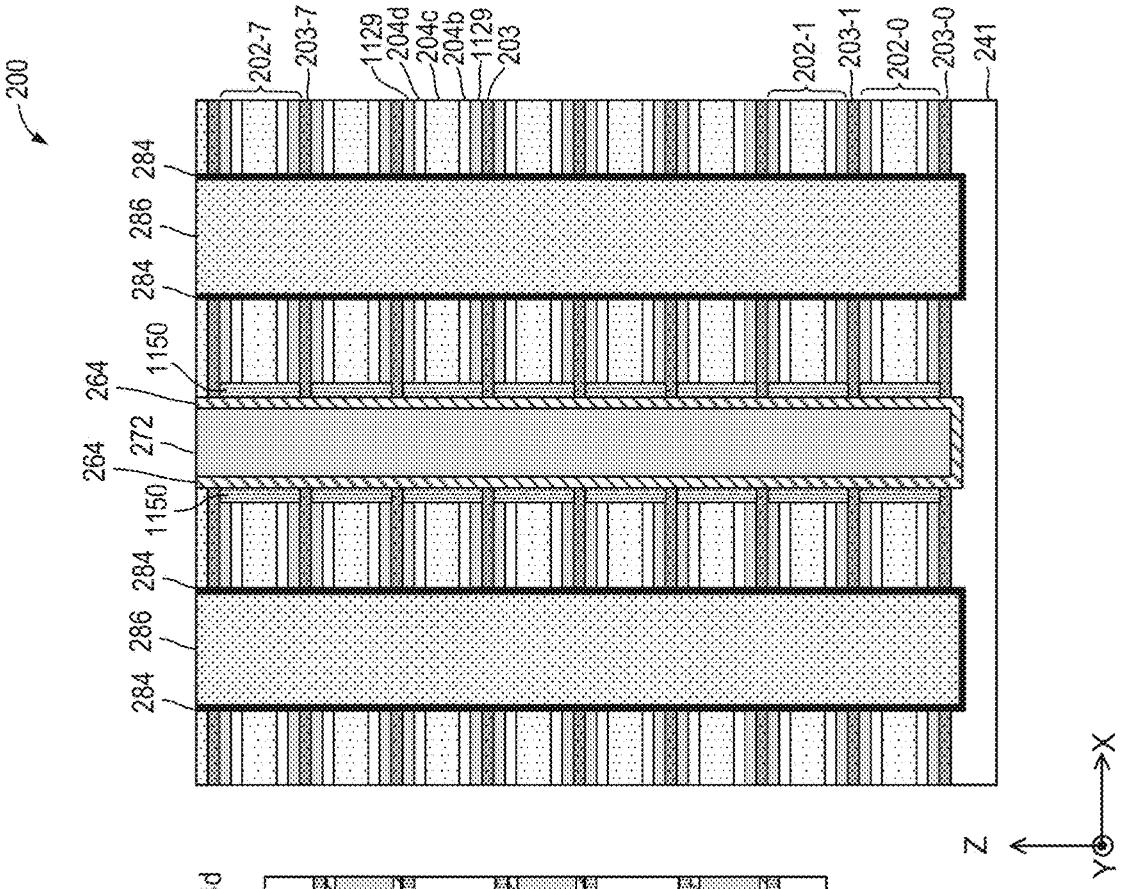

200
202-7
203-7
1129
204d
204c
204b
1129
203
202-1
203-1
202-0
203-0
241
284
286
284
1150
264
272
264
1150
284
286
284
Z
X
Y

FIG. 11I(ii)

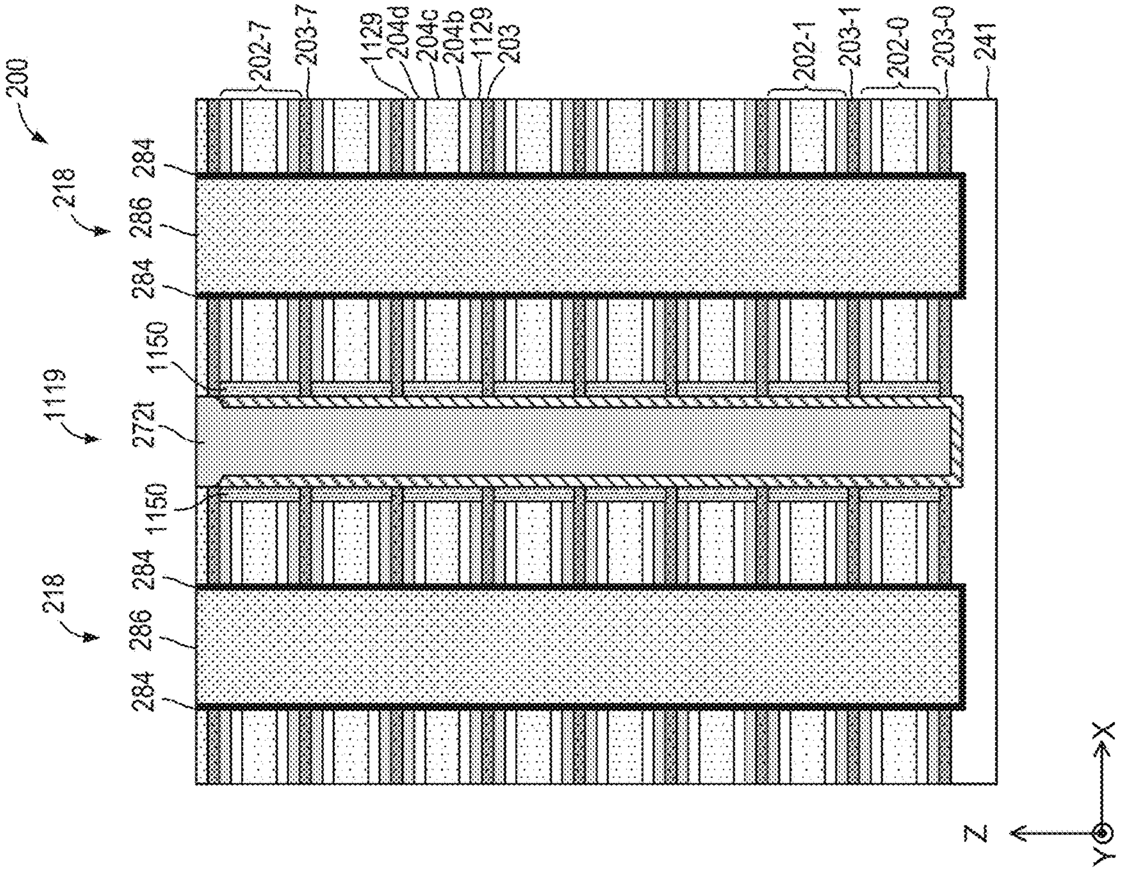

200
218
284
286
284
1150
1119
272t
1150
284
218
286
284
202-7
203-7
1129
204d
204c
204b
1129
203
202-1
203-1
202-0
203-0
241
Z
X
Y

FIG. 11m(ii)

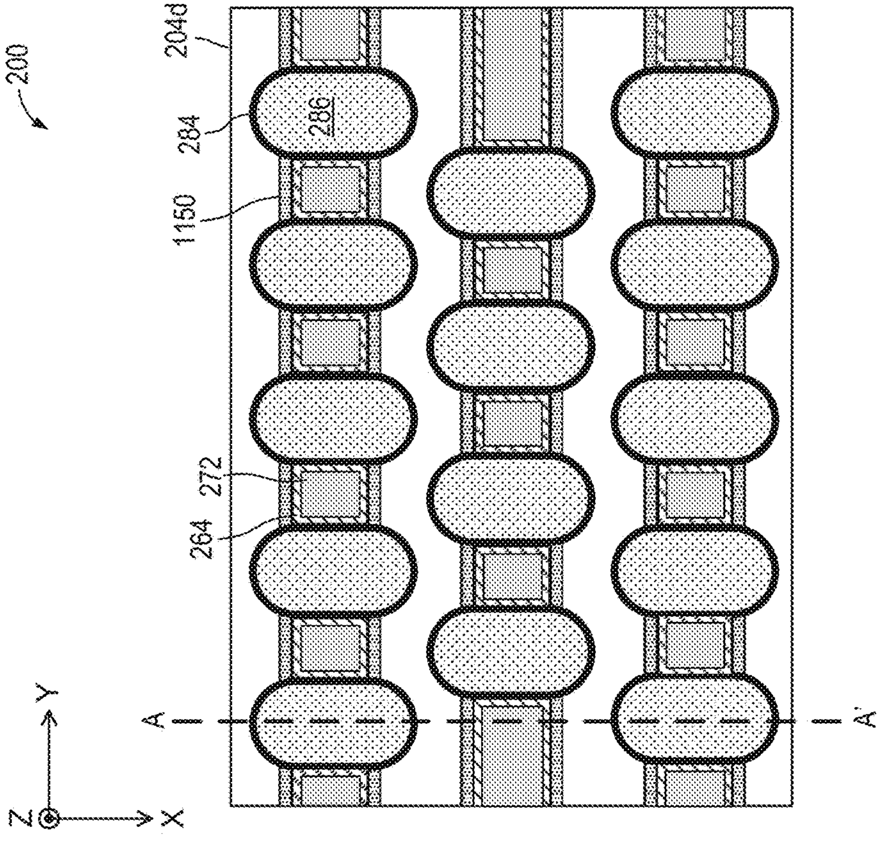

200
204d
284
286
1150
272
264
A
A'
Y
Z
X

FIG. 11m(i)

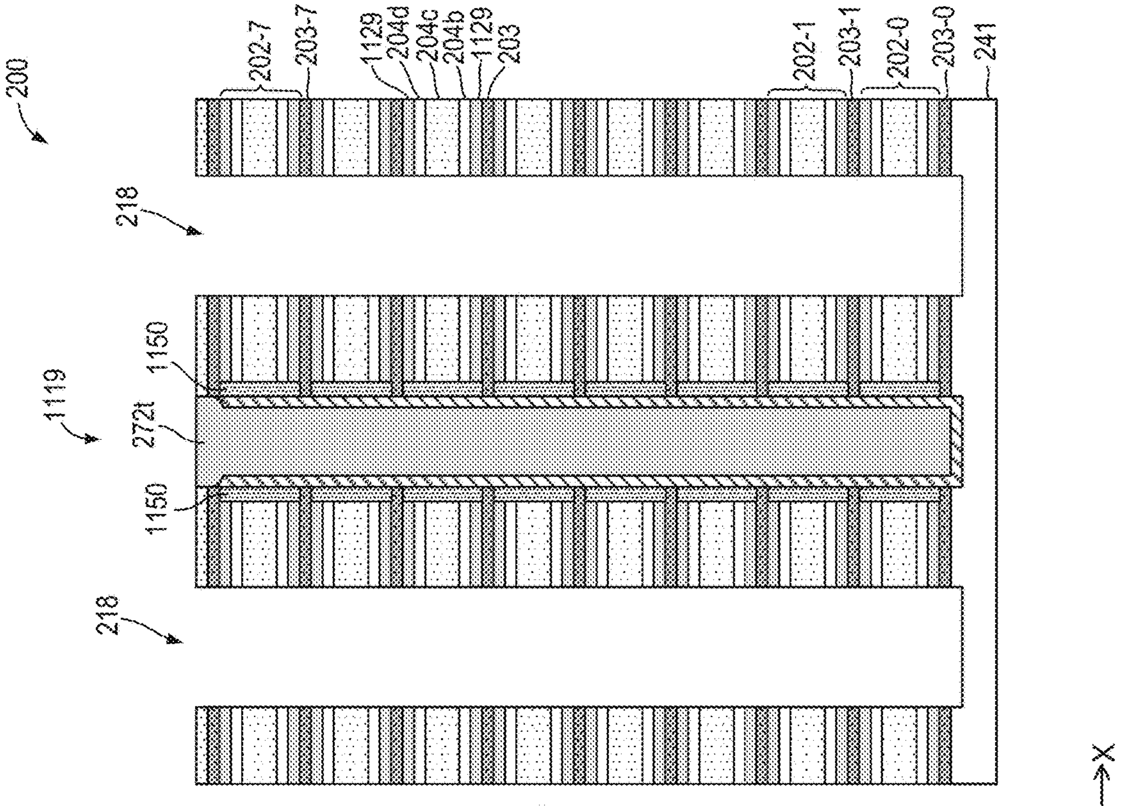

200
202-7
203-7
1129
204d
204c
204b
1129
203
202-1
203-1
202-0
203-0
241
218
1119
272t
1150
1150
218
Z
X
Y

FIG. 11n(ii)

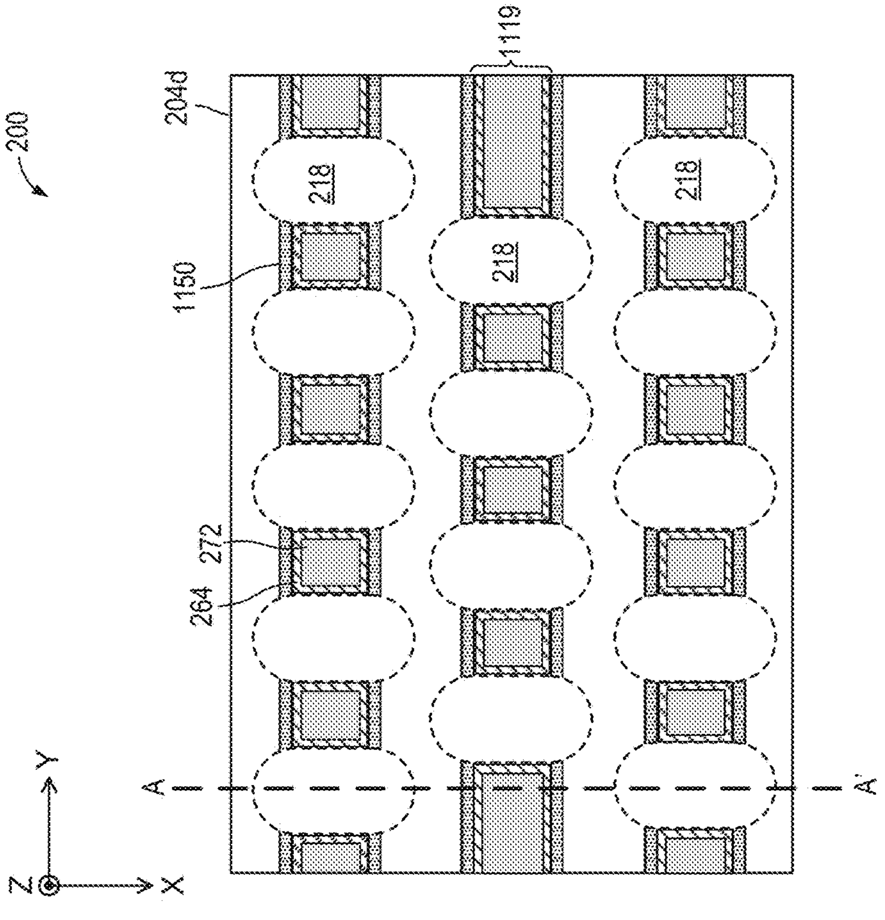

200
204d
1119
218
1150
218
218
272
264
A
A'
Y
Z
X

FIG. 11n(i)

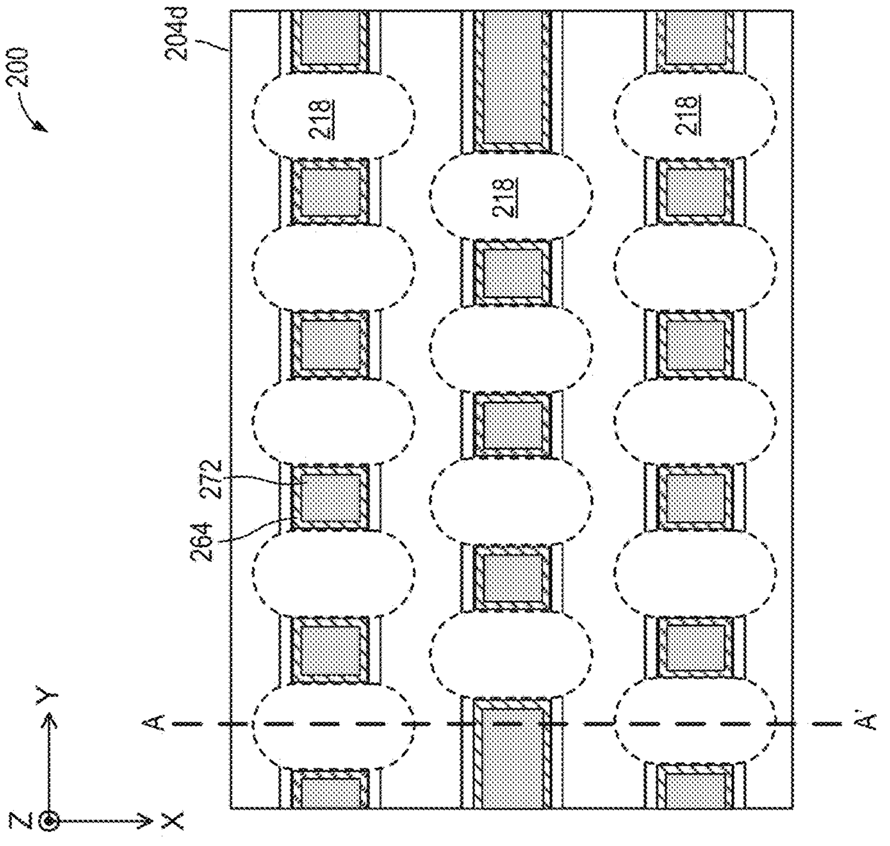

200
204d
218
218
218
272
264
A
A'
Y
Z
X

FIG. 11o(i)

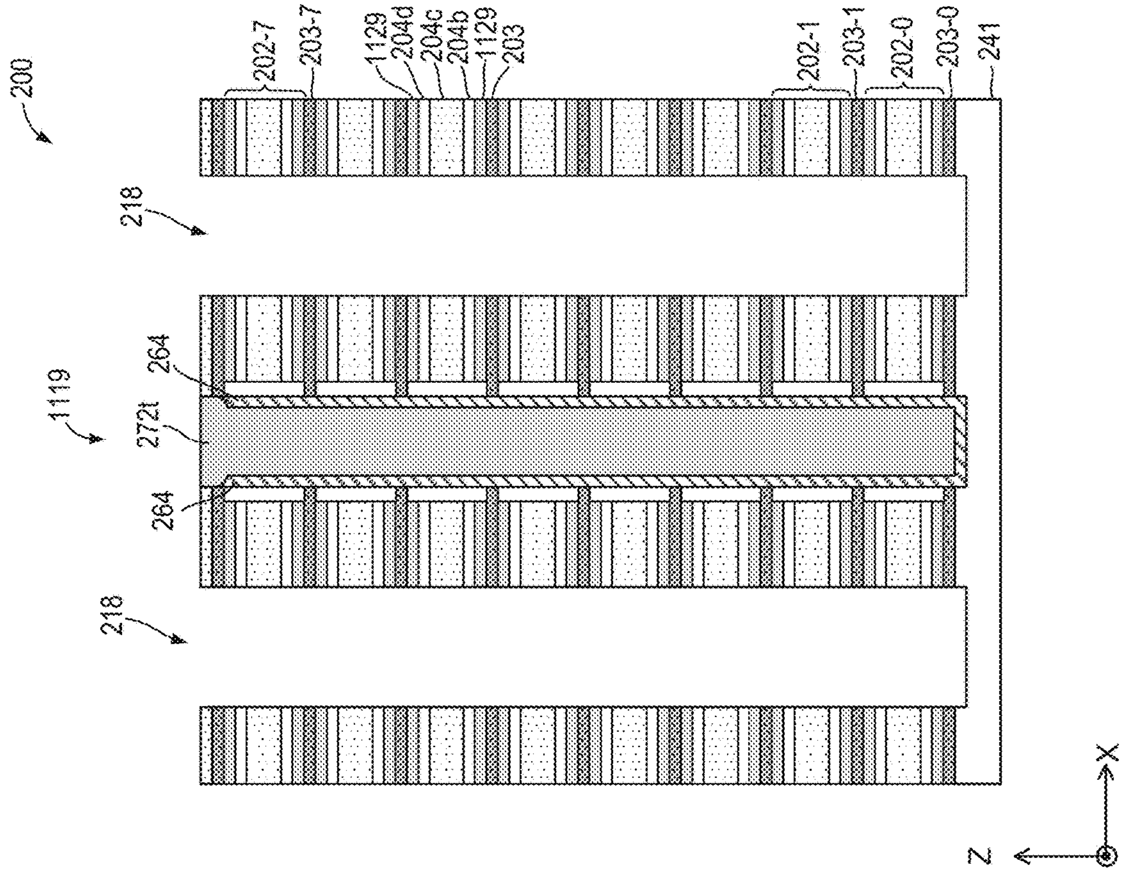

200
202-7
203-7
1129
204d
204c
204b
1129
203
202-1
203-1
202-0
203-0
241
218
218
1119
264
272t
264
Z
Y
X

FIG. 11o(ii)

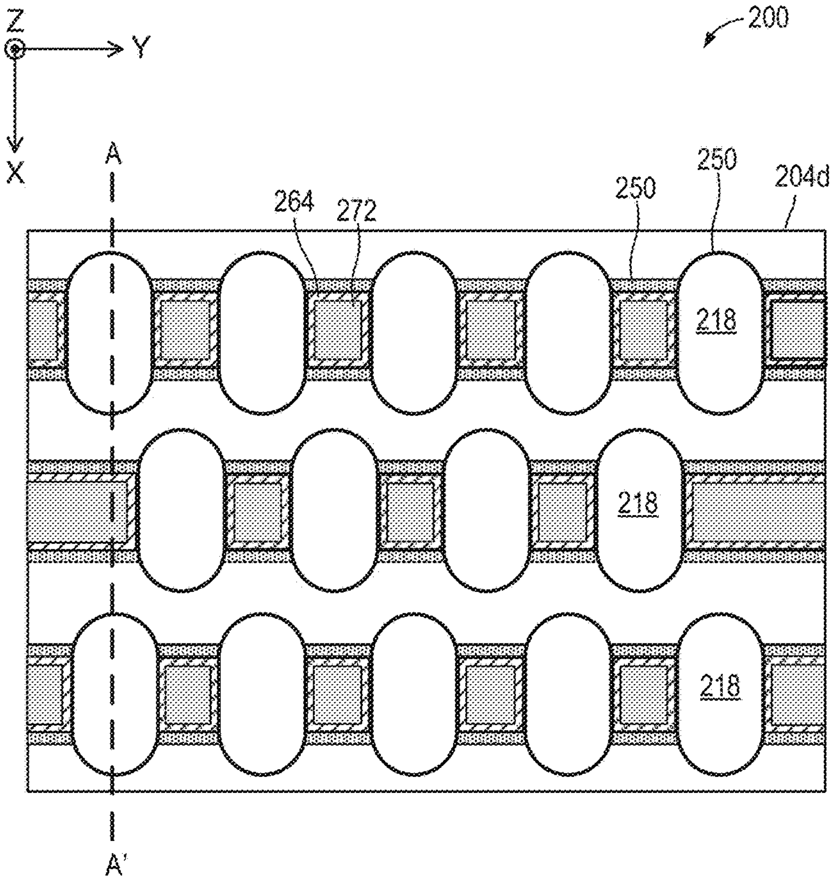

Z
Y
X
200
A
A'
264
272
250
250
204d
218
218
218

FIG. 11p(i)

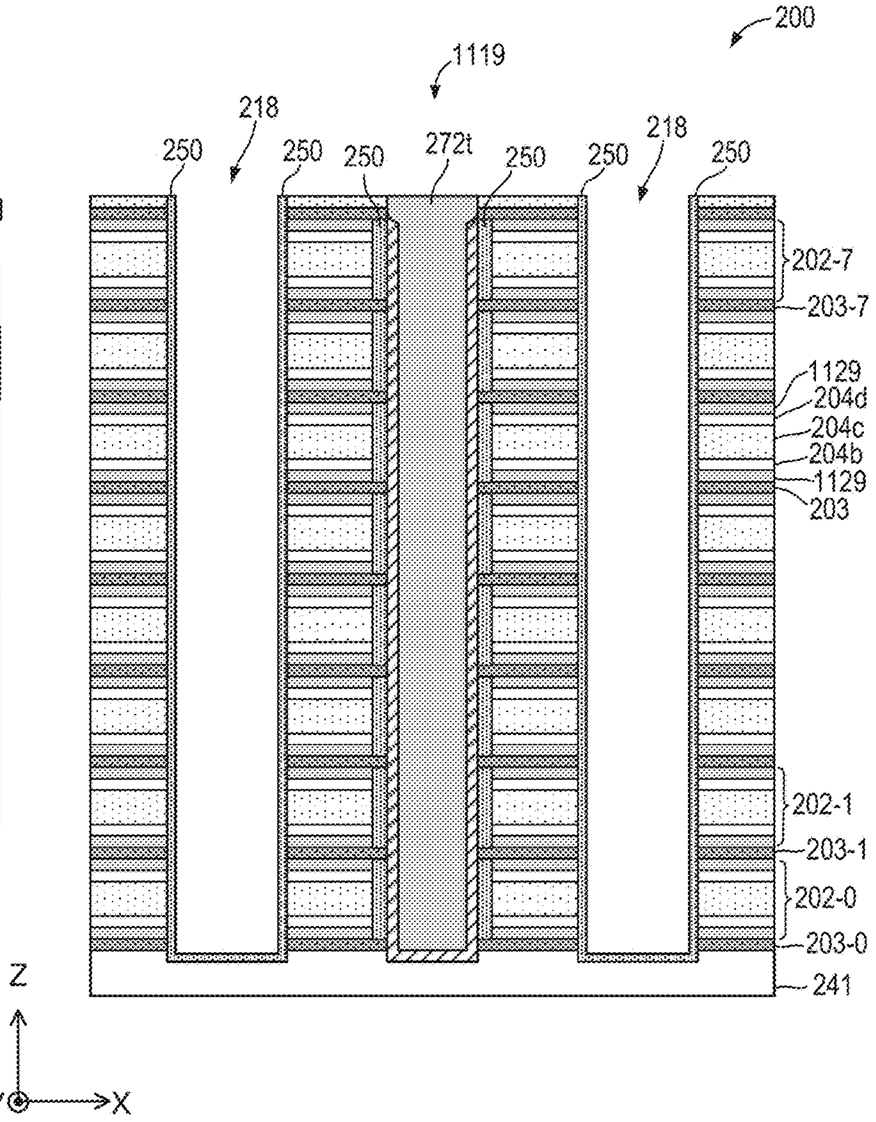

200
1119
218
218
250
250
250
272t
250
250
250
202-7
203-7
1129
204d
204c
204b
1129
203
202-1
203-1
202-0
203-0
241
Z
Y
X

FIG. 11p(ii)

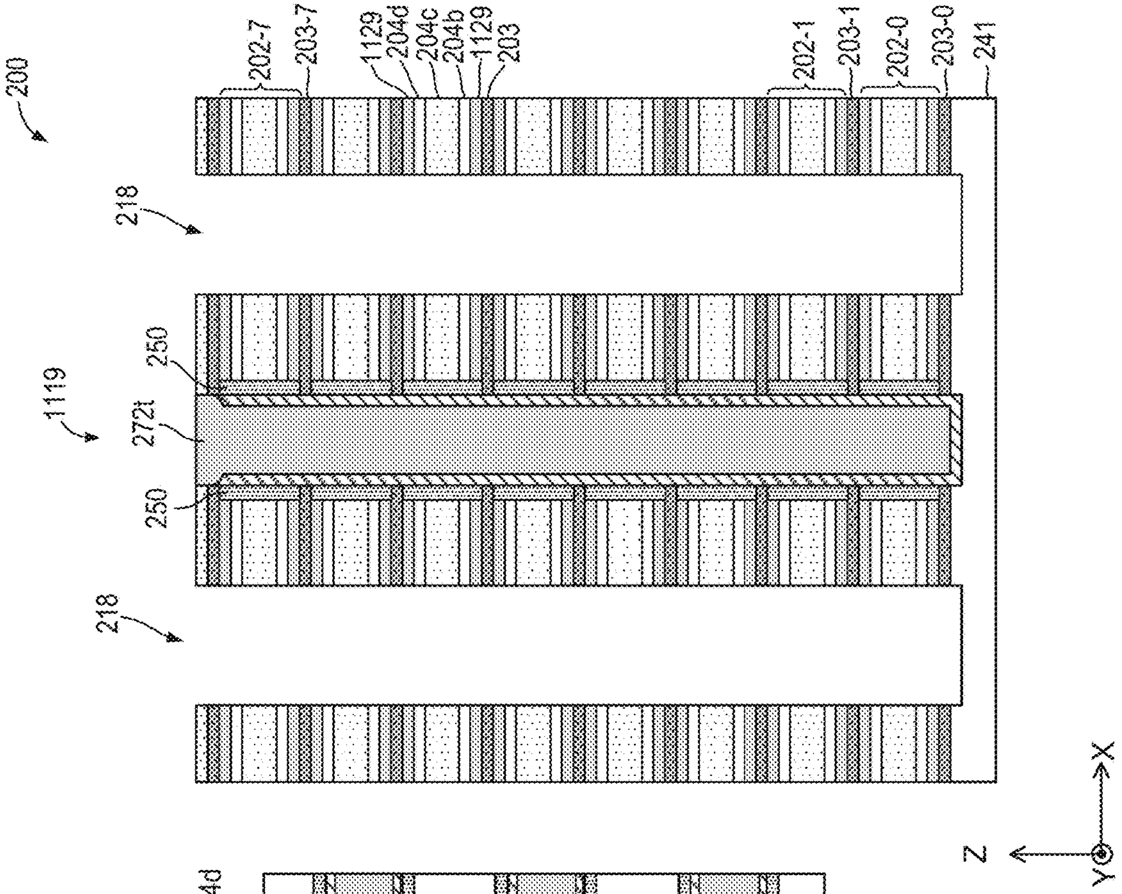

200
202-7
203-7
1129
204d
204c
204b
1129
203
202-1
203-1
202-0
203-0
241
218
1119
250
272t
250
218
Z
X
Y

FIG. 11q(iii)

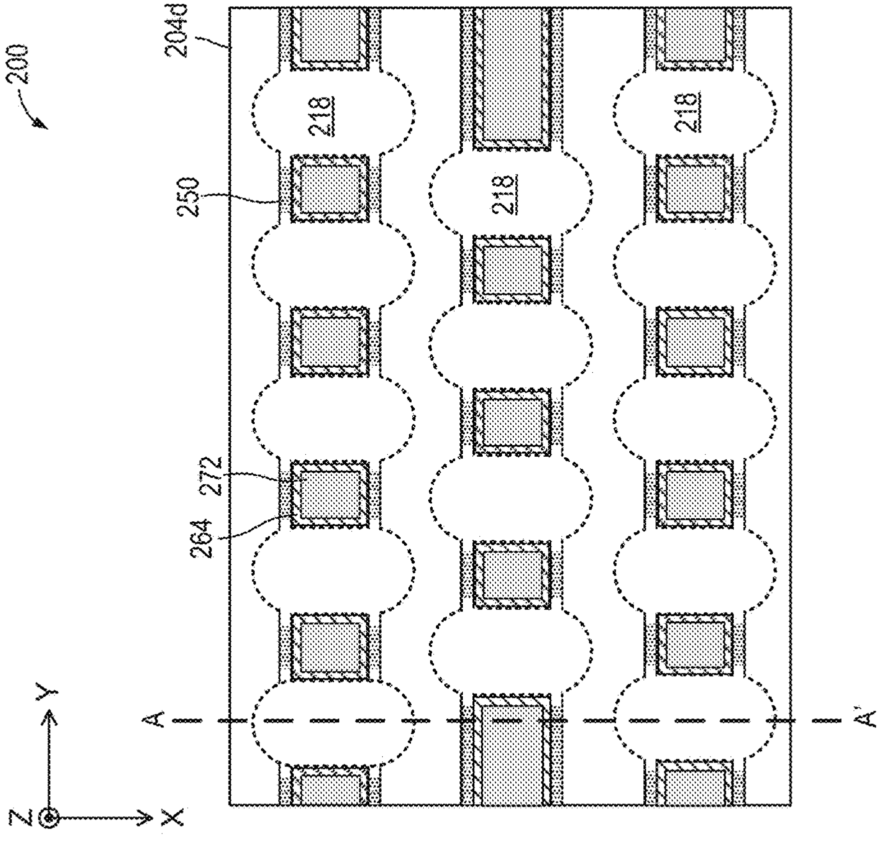

200
204d
250
218
218
218
272
264
Y
Z
X
A
A'

FIG. 11q(i)

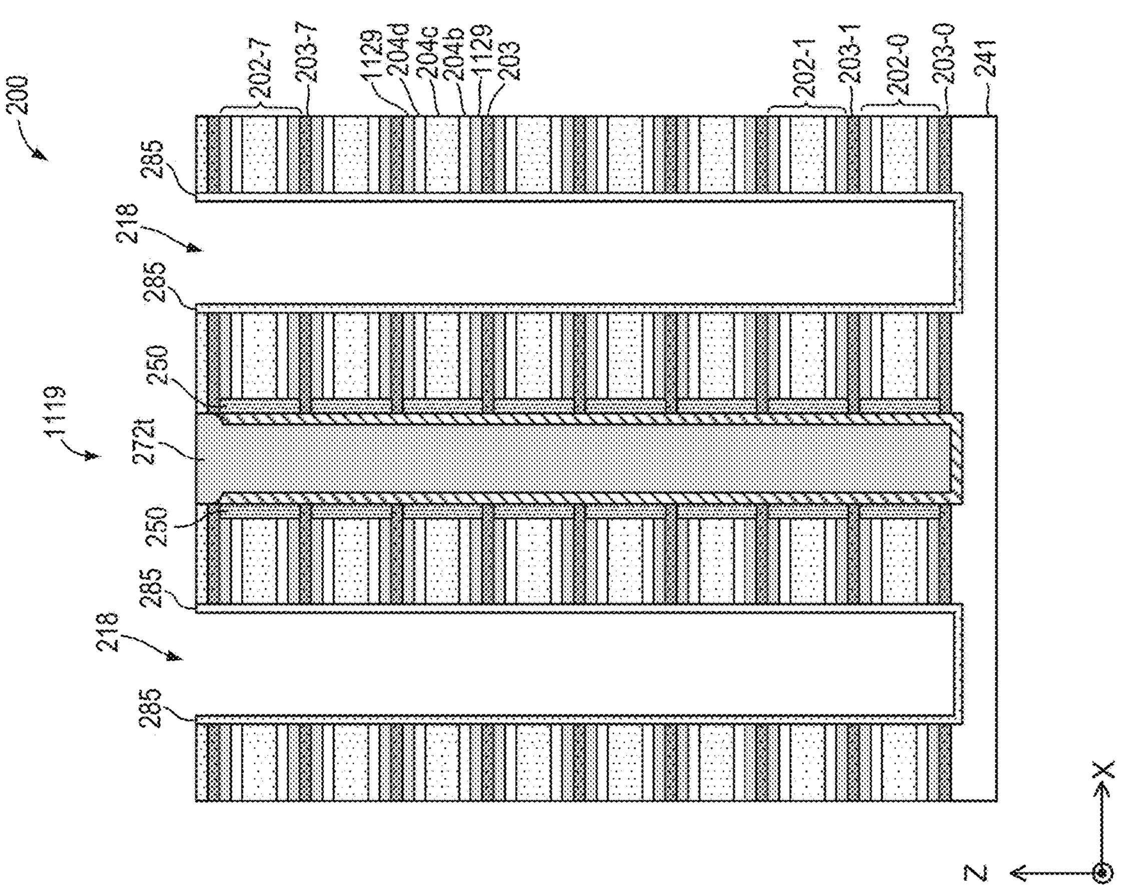

200
1119
202-7
203-7
1129
204d
204c
204b
1129
203
202-1
203-1
202-0
203-0
241
285
218
285
250
272t
250
285
218
285
Z
X
Y

FIG. 11r(ii)

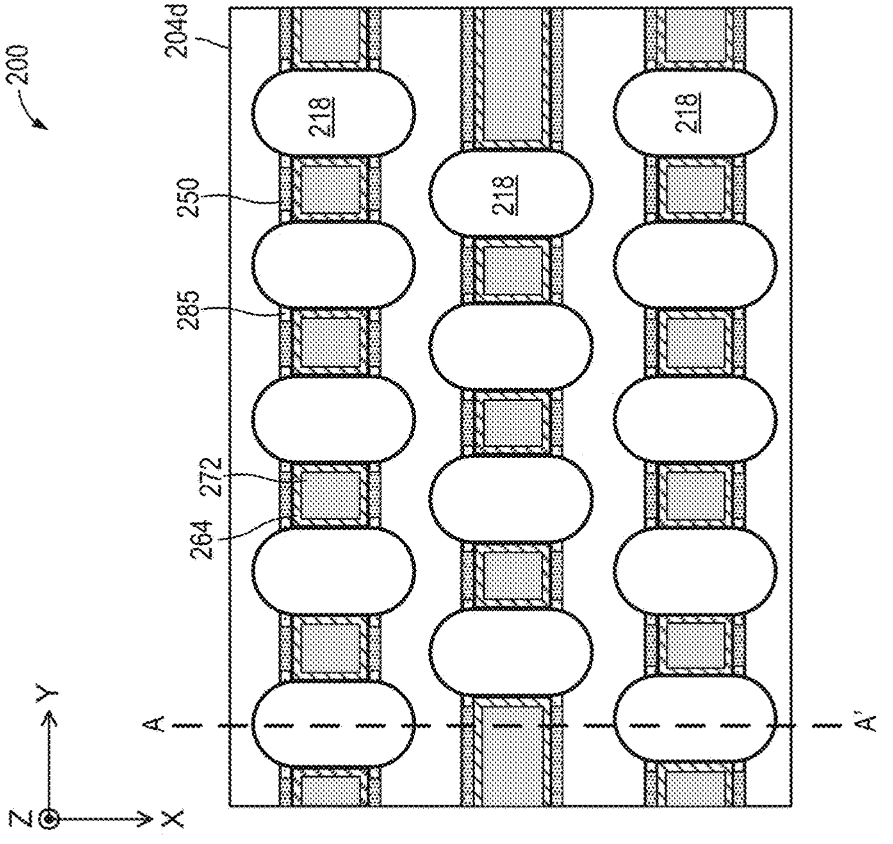

200
204d
250
285
272
264
218
218
218
A
A'
Y
Z
X

FIG. 11r(i)

3-DIMENSIONAL MEMORY STRING ARRAY OF THIN-FILM FERROELECTRIC TRANSISTORS

CROSS REFERENCE TO RELATED APPLICATIONS

The present application claims priority to U.S. provisional application ("Parent Application"), Ser. No. 63/222,926, entitled "3-Dimensional Memory String Array of Thin-Film Ferroelectric Transistors," filed on Jul. 16, 2021.

The present application is related to U.S. patent application ("Non-Provisional Application I"), Ser. No. 17/674,137, entitled "Thin-film Storage Transistor With Ferroelectric Storage Layer," filed on Feb. 17, 2022, which claims priority to (i) U.S. provisional application ("Provisional Application I"), Ser. No. 63/152,266, entitled "Thin-Film Storage Transistor With Ferroelectric Storage Layer," filed on Feb. 22, 2021, and (ii) U.S. provisional application ("Provisional Application II"), Ser. No. 63/153,273, entitled "Thin-Film Storage Transistor With Ferroelectric Storage Layer," filed on Feb. 24, 2021.

The present application is related to U.S. patent application ("Non-Provisional Application II"), Ser. No. 17/382,064, entitled "Methods For Fabricating A 3-Dimensional Memory Structure Of NOR Memory Strings," filed on Jul. 221, 2021, which claims priority to U.S. provisional application ("Provisional Application III"), Ser. No. 63/054,743, entitled "Methods For Fabricating A 3-Dimensional Memory Structure Of NOR Memory Strings," filed on Jul. 21, 2020. The present application is also related to U.S. provisional application ("Provisional Application IV"), Ser. No. 63/054,750, entitled "Methods for Fabricating A 3-Dimensional Memory Structure of NOR Memory Strings," filed on Jul. 21, 2020, now U.S. patent application Ser. No. 17/382,126, filed Jul. 21, 2021; U.S. provisional application ("Provisional Application V"), Ser. No. 63/139,435, entitled "Vertical NOR Thin-film Transistor Strings and Fabrication Thereof," filed on Jan. 20, 2021, now U.S. patent application Ser. No. 17/559,101, filed Dec. 22, 2021; and U.S. provisional application ("Provisional Application VI"), Ser. No. 63/114,958, entitled "Methods for Reducing Disturb Errors by Refreshing Data Alongside Programming or Erase Operations," filed on Nov. 17, 2020, now U.S. patent application Ser. No. 17/525,712, filed Nov. 12, 2021.

The present application is also related to U.S. non-provisional patent application ("Related Application"), Ser. No. 16/894,596, entitled "Capacitive-Coupled Non-Volatile Thin-Film Transistor Strings in Three Dimensional Arrays," filed on Jun. 5, 2020, which is a continuation of U.S. patent application Ser. No. 16/107,118, entitled "Capacitive-Coupled Non-Volatile Thin-Film Transistor Strings in Three Dimensional Arrays," filed on Aug. 21, 2018, which is a divisional application of U.S. non-provisional patent application Ser. No. 15/248,420, entitled "Capacitive-Coupled Non-Volatile Thin-Film Transistor Strings in Three Dimensional Arrays," filed on Aug. 26, 2016, which is related to and claims priority of (i) U.S. provisional application Ser. No. 62/235,322, entitled "Multi-gate NOR Flash Thin-film Transistor Strings Arranged in Stacked Horizontal Active Strips With Vertical Control Gates," filed on Sep. 30, 2015; (ii) U.S. provisional patent application Ser. No. 62/260,137, entitled "Three-dimensional Vertical NOR Flash Thin-film Transistor Strings," filed on Nov. 25, 2015; (iii) U.S. non-provisional patent application Ser. No. 15/220,375, "Multi-Gate NOR Flash Thin-film Transistor Strings Arranged in Stacked Horizontal Active Strips With Vertical Control Gates," filed on Jul. 26, 2016; and (vi) U.S. provisional patent application Ser. No. 62/363,189, entitled "Capacitive Coupled Non-Volatile Thin-film Transistor Strings," filed Jul. 15, 2016.

The disclosures of Parent Application, Related Application, Non-Provisional Applications I and II, and Provisional Applications I, II, III, IV, V, and VI are hereby incorporated by reference in their entireties.

BACKGROUND OF THE INVENTION

1. Field of the Invention

The present invention relates to high-density memory structures. In particular, the present invention relates to high-density, low read-latency memory structures formed by interconnected thin-film storage elements (e.g., 3-dimensional array of thin-film storage transistors), including those organized as NOR-type memory strings ("NOR memory strings").

2. Discussion of the Related Art

The memory circuit structures of this detailed description may be fabricated above the planar surfaces of one or more semiconductor substrates (e.g., silicon wafers) using conventional fabrication processes. In this description, "vertical" refers to a direction substantially normal to a planar surface of a semiconductor substrate, and "horizontal" refers to any direction that is orthogonal to "vertical." A Cartesian coordinate system may therefore be adopted, in which "horizontal" refers to any direction substantially parallel to the planar surface of semiconductor substrate, which is a surface spanning any two non-colinear horizontal directions X and Y (also referred to as the "X-Y plane"). Similarly, with respect to that Cartesian coordinate system, "vertical" refers to a normal direction (also referred to as the "Z-direction") to the planar surface.

A NOR-type memory string includes storage transistors that share a common source region and a common drain region, while allowing each storage transistor to be individually addressed and accessed. The Related Application discloses 3-dimensional arrays of NOR memory strings ("NOR memory strings") of charge storage transistors. The Related Application discloses, for example, a NOR memory string that includes (i) a common source region and a common drain region both running lengthwise along a horizontal direction and (i) the gate electrodes for the charge storage transistors each running along a vertical direction.

Advances in electrically polarizable materials ("ferroelectric materials"), especially those used in electronic circuits, suggest new potential applications in ferroelectric memory circuits. For example, the article "Ferroelectricity in Hafnium Oxide: CMOS compatible Ferroelectric Field Effect Transistors," by T. S. Böscke et al., published in 2011 *International Electron Devices Meeting* (*IEDM*), pp. 24.5.1-24.5.4 (doi: 10.1109/IEDM.2011.6131606), discloses a ferroelectric field effect transistor ("FeFET") that uses hafnium oxide as a gate dielectric material. By controlling polarization in a ferroelectric gate dielectric layer, the FeFET may be programmed to have one of two or more selectable threshold voltages. Each threshold voltage of the FeFET constitutes a state (e.g., a "programmed" state, an "erased" state, or any other state that represents a designated binary value). Such an FeFET has application in high-density memory circuits. For example, U.S. patent application Ser. No. 13/897,037, entitled "Apparatuses having a ferroelectric field-effect transistor memory array and related method," by D. V. Nirmal Ramaswamy et al., filed on May 17, 2013, now U.S. Pat. No. 9,281,044, discloses a 3-dimensional array of FeFETs.

The FeFETs of the prior art, however, suffer from low endurance. For example, the article "Vertical Ferroelectric HfO_2 FET based on 3-D NAND Architecture: Towards Dense Low-Power Memory," by K. Florent et. al., published in 2018 *IEEE International Electron Devices Meeting* (*IEDM*), 2018, pp. 2.5.1-2.5.4 (doi: 10.1109/IEDM.2018.8614710), discloses an endurance of merely 10^4 cycles. Such low endurance renders the memory circuits practically unsuitable for many memory applications. It is highly desirable to have a memory circuit that can serve as a building a block of a main memory system. Current building blocks of main memory systems (e.g., dynamic random-access memory (DRAM)) have low retention time and low memory density.

SUMMARY

According to one embodiment of the present invention, an array of memory cells includes FeFETs formed in stacks of horizontal active strips, with vertical control gate electrodes (also referred to as "local word lines") provided along one or both sidewalls of the active strips, the control gate electrodes being separated from their associated active strips by one or more polarizable or ferroelectric elements (also referred to as "ferroelectric or polarization layers"). Polarization of the polarizable element in an FeFET changes the threshold voltage the FeFET. Each active strip may include a channel layer formed adjacent both a common source region and a common drain region. In one embodiment, in a given stack of active strips (also referred to as an "active stack"), the polarizable or ferroelectric elements associated with the FeFETs of each active strip are separated from the polarizable or ferroelectric elements associated with the FeFETs of other active strips. The FeFETs of each active strip may be organized as one or more NOR memory strings. Further, independently addressable FeFETs may be formed along opposite sides of an active strip to achieve twice the memory density than if the FeFETs are formed as a single file along the length of the common bit line. Furthermore, by forming channel regions of the FeFETs along the vertical sidewalls of each active strip (i.e., between the common drain region and the common source region), a memory density that is substantially higher than is achieved in the prior art is possible because the volume of the channel region, even with a large extent along the Z-direction, has an insignificant footprint in the X-Y plane.

In one embodiment, only the common drain region (also referred to as a "common bit line") of the FeFETs in a NOR memory string is electrically and selectively connected by one or more conductors to one of several voltage sources. In one embodiment, the common source region may be pre-charged to a predetermined voltage prior to a read, program (sometimes also referred to as "write"), or erase operation, and is held to that predetermined voltage by the parasitic capacitance associated with the common source region during the read, program (sometimes also referred to as "write"), or erase operation. In one embodiment, one or more arrays of NOR memory strings may be organized and formed as an independently addressable modular memory circuit (also referred herein as a "tile"). A group of tiles may be formed above the same substrate but are configured to each be independently controlled and operated, thereby allowing massively parallel read, program or erase operations to be carried out concurrently on the tiles. The NOR memory strings within each tile may also be organized in groups that are configured to each be independently controlled and operated.

In one embodiment, in a 3-dimensional array of NOR memory strings, FeFETs are formed on both side edges along the lengths of each active strip, with vertical local word lines being provided along both side edges. High density is achieved by sharing the local word lines between their adjacent active stacks, each local word line serving as a common gate electrode to vertically aligned FeFETs in the active strips of both active stacks. The vertical local word lines may be contacted by interconnection conductors (also referred to as "global word lines") provided above or below the 3-dimensional array of NOR memory strings. The global word lines may run along a direction transverse to the lengths of the contacted active strips.

Organizing the FeFETs as 3-dimensional arrays of NOR memory strings—rather than as NAND memory strings, for example—results in (i) reduced read-latencies approaching those of dynamic random access memory (DRAM) arrays, (ii) reduced sensitivities to read-disturb and program-disturb conditions, (iii) reduced power dissipation and a lower cost-per-bit metric relative to planar NAND or DRAM memory arrays, and (iv) the ability to read, write or erase FeFETs on multiple active strips concurrently, thereby significantly increasing data throughput.

According to one embodiment of the present invention, undesirable effects on a memory operation due to variations among FeFETs within an array of NOR memory strings may be overcome by processing the memory operation in conjunction with FeFETs in a designated reference NOR memory string within the same array. For example, effects on a read operation of an FeFET of a NOR memory string due to a background leakage current present in NOR memory strings can be substantially mitigated by comparing the sensed result of a concurrently read FeFET within a reference NOR memory string. In some embodiments, the ferroelectric or polarizable elements may have a data retention time that requires refreshing to meet data stability requirements in some applications. Relative to storage transistors in conventional non-volatile memory circuits (e.g., non-volatile NAND memory strings), which do not require refreshing in those applications, the FeFETs may be considered "quasi-volatile" (QV) storage transistors. The FeFETs are considered quasi-volatile because, as compared to conventional volatile memory circuits (e.g., DRAM circuits), the FeFETs require refreshing significantly less frequently. For example, FeFETs in high density arrays need only be refreshed, restored, or replaced, once every few minutes, hours or every few days, and may be accomplished by accessing the FeFETs individually or accessed concurrently as one or more groups. Such refresh operations ensure long term stable and reliable operations. The refresh operations may also be performed in the background, without interfering regular user read, erase or write operations.

The present invention is better understood upon consideration of the detailed description below, in conjunction with the accompanying drawings.

BRIEF DESCRIPTION OF THE DRAWINGS

Various embodiments of the invention are disclosed in the following detailed description and the accompanying drawings. Although the drawings depict various examples of the invention, the invention is not limited by the depicted examples. It is to be understood that, in the drawings, like reference numerals designate like structural elements. Also, it is understood that the depictions in the FIGS. are not necessarily to scale.

FIG. **1**, which includes FIGS. **1***a*-**1** and **1***a*-**2**, show conceptualized memory structure **100**, which illustrates an organization of ferroelectric field-effect transistors (FeFETs) in an array of NOR memory strings, according to one embodiment of the present invention.

FIG. **1***b* shows a basic circuit representation of four NOR memory string-pairs, each NOR memory string-pair including two NOR memory strings provided on the same plane (i.e., one of planes **159-1**, **159-2**, **159-3** and **159-4**), according to one embodiment of the present invention.

FIG. **1***c* shows a basic circuit representation of four NOR memory strings, each NOR memory string being provided on a respective one of planes **159-1**, **159-2**, **159-3** and **159-4**, according to one embodiment of the present invention.

FIG. **2***a* shows a cross section in a Y-Z plane of semiconductor structure **200**, after active layers **202-0** to **202-7** (each separated from the next active layer by one of isolation layers **203-0** to **203-7**) have been formed on semiconductor substrate **201**, but prior to formation of individual active strips, in accordance with one embodiment of the present invention.

FIGS. **2***b*-**1**, **2***b*-**2**, **2***b*-**3**, **2***b*-**4** and **2***b*-**5** show various examples of active layer **220***a*, according to one embodiment of the present invention.

FIG. **2***c* shows cross section in a Y-Z plane of structure **200** of FIG. **2***a* through buried contacts **205-0** and **205-1**, which connect n$^+$ silicon layers **204***b* of active layers **202-0** and **202-1** to circuitry **206-0** and **206-1** in semiconductor substrate **201**.

FIGS. **2***d*(*i*), **2***d*(*ii*), **2***e*(*i*), **2***e*(*ii*), **2***f*(*i*), **2***f*(*ii*), **2***g*(*i*), **2***g*(*ii*), **2***h*(*i*), **2***h*(*ii*), **2***i*(*i*), **2***i*(*ii*), **2***j*(*i*), **2***j*(*ii*), **2***k*(*i*) and **2***k*(*ii*) illustrate steps of an exemplary process for forming a modular circuit or tile that includes an array of NOR memory strings, including an array portion (i.e., memory structure **200**), according to one embodiment of the present invention; in particular, FIG. **2***k* illustrates, preceding steps that form the gate stacks of the FeFETs, a step for providing a channel region for each FeFET; the channel region may be provided by any one group of the groups of process steps described in conjunction with FIGS. **3***a* to **3***b*, FIGS. **4***a* to **4***c* and FIGS. **5***a* to **5***f*, respectively.

FIG. **2***l* shows an X-Z plane cross section through a row of local word lines **208***g*-*a*, showing active strips formed out of active layers **202-7** and **202-6**, according to one embodiment of the present invention.

FIG. **2***m* shows a 3-dimensional view of horizontal active layers **202-4** to **202-7** in the embodiment of FIG. **2***k*, with local word lines or local pre-charge word lines connected to global word lines **208***g*-*s* and global word lines **208***g*-*a*, and showing each active strip as having its N$^+$ layer **204***d* (acting as a drain region) connected through select circuits to any of voltage supplies (e.g., V_{ss}, V_{bl}, V_{pgm}, $V_{inhibit}$, and V_{erase}), decoding, sensing and other circuits arranged either adjacent or directly underneath the memory arrays.

FIGS. **3***a* and **3***b* illustrate a first group of process steps for forming the gate stacks in the FeFETs of an array of NOR memory strings, according to one embodiment of the present invention.

FIGS. **4***a*, **4***b* and **4***c* illustrate a second group of process steps for forming the gate stacks in the FeFETs in an array of NOR memory strings, according to one embodiment of the present invention.

FIGS. **5***a*, **5***b*, **5***c*, **5***d*(*i*), **5***d*(*ii*), **5***e*(*i*), **5***e*(*ii*), **5***f*(*i*) and **5***f*(*ii*) illustrate a third group of process steps for forming the gate stacks in the FeFETs of an array of NOR memory strings, according to one embodiment of the present invention.

FIG. **6***a* shows an X-Z cross sectional view of memory structure **200**, which represents active stacks **261***a* and **261***b* fabricated under any group of process steps illustrated by FIG. **3***a*-**3***b*, FIGS. **4***a* to **4***c* and FIGS. **5***a* to **5***f*(*i*), in accordance with one embodiment of the present invention.

FIG. **6***b* shows an X-Z cross sectional view of memory structure **200**, which represents active stacks **261***a* and **261***b* fabricated under the group of process steps illustrated by FIGS. **5***a*-**5***f*(*ii*), in accordance with one embodiment of the present invention.

FIG. **6***c* shows an X-Z cross sectional view of memory structure **200**, fabricated under a channel-last fabrication process, in accordance with one embodiment of the present invention.

FIG. **7***a* shows hysteresis in the drain current (I_d) in response to an applied gate voltage (V_g) in a conventional FeFET.

FIG. **7***b* shows a desirable hysteresis in the drain current (I_d) response to an applied gate voltage (V_g) in a thin-film FeFET in a NOR memory array, according to one embodiment of the present invention.

FIG. **7***c* shows a circuit schematic diagram of FeFETs in NOR memory strings **1420-1** to **1420-4**, according to one embodiment of the present invention.

FIG. **8** illustrates the methods and circuit elements used for setting source voltage V_{ss} in n$^+$ silicon layers **204***d*; specifically, source voltage V_{ss} may be set through hard-wire decoded source line connections **280** (dashed line) or alternatively, by activating pre-charge FeFETs **303** and decoded bit line connections **270**, to any one of voltage sources for bit line voltages V_{ss}, V_{bl}, V_{pgm}, $V_{inhibit}$ and V_{erase}.

FIG. **9***a* shows semiconductor structure **600**, which is a three-dimensional representation of an array of NOR memory strings organized into quadrants **Q1**-**Q4**; in each quadrant, (i) numerous NOR memory strings are each formed in an active strip extended along the Y-direction (e.g., NOR memory string **112**), (ii) pages extending along the X-direction (e.g., page **113**), each bit in the page corresponds to an FeFET from each NOR memory string at a corresponding Y-position, the NOR memory strings in the page being of the same corresponding Z-position (i.e., of the same active layer); (iii) slices extending in both the X- and Z-directions (e.g., slice **114**), with the bits in each slice provided by the pages of the same corresponding Y-position, one page from each of the planes, and (iv) planes extending along both the X- and Y-directions (e.g., plane **110**), the bits in each plane being provided by all pages at a given Z-position (i.e., formed out of the same active layer).

FIG. **9***b* shows structure **600** of FIG. **9***a*, showing FeFETs in programmable reference string **112**-Ref in quadrant **Q4** and FeFETs in NOR memory string **112** in quadrant **Q2** coupled to sense amplifiers SA(a), **Q2** and **Q4** being "mirror image quadrants"; FIG. **9***b* also shows (i) programmable reference slice **114**-Ref (indicated by area A) in quadrant **Q3** similarly providing corresponding reference FeFETs for slice **114** in mirror image quadrant **Q1**, sharing sense amplifiers SA(b), and (ii) programmable reference plane **110**-Ref in quadrant **Q2** providing corresponding reference FeFETs to plane **110** in mirror image quadrant **Q1**, sharing sense amplifiers SA(c), and also providing corresponding reference FeFETs for NOR memory strings in the same quadrant (e.g., NOR memory string **112**).

FIG. **9***c* shows structure **600** of FIG. **9***a*, showing slices **116** being used as a high speed cache because of their close proximity to their sense amplifiers and voltage sources **206**;

FIG. **9***c* also show spare planes **117**, which may be used to provide replacement or substitution NOR memory strings or pages in quadrant Q2.

FIG. **10***a* shows in simplified form prior art storage system **800** in which microprocessor (CPU) **801** communicates with system controller **803** in a flash solid state drive (SSD) that employs NAND flash chips **804**; the SSD emulates a hard disk drive and NAND flash chips **804** do not communicate directly with CPU **801** and have relatively long read latency.

FIG. **10***b* shows in simplified form system architecture **850** using the memory devices of the present invention, in which FeFETs of NOR memory string arrays **854** and **855** communicate directly with CPU **801** through one or more input and output (I/O) ports **861**, and indirectly through controller **863**.

FIGS. **11***a*, **11***b*, **11***c*, **11***d*, **11***e*, **11***f*, **11***g*, **11***h*, **11***i*(*i*), **11***i*(*ii*), **11***j*(*i*), **11***j*(*ii*), **11***k*(*i*), **11***k*(*ii*), **11***l*(*i*), **11***l*(*ii*), **11***m*(*i*), **11***m*(*ii*), **11***n*(*i*), **11***n*(*ii*), **11***o*(*i*), **11***o*(*ii*), **11***p*(*i*), **11***p*(*ii*), **11***q*(*i*), **11***q*(*ii*), **11***r*(*i*), and **11***r*(*ii*) illustrate an additional process in which, subsequent to forming ferroelectric layers, the channel regions for the FeFETs of an array of NOR memory strings are formed by replacing an earlier-formed sacrificial layer, according to one embodiment of the present invention.

DETAILED DESCRIPTION OF THE PREFERRED EMBODIMENTS

In this detailed description, process steps described for one embodiment may be used in a different embodiment, even if the process steps are not expressly described in the different embodiment. When reference is made herein to a method including two or more defined steps, the defined steps can be carried out in any order or simultaneously, except where the context dictates or specific instruction otherwise are provided herein. Further, unless the context dictates or express instructions otherwise are provided, the method can also include one or more other steps carried out before any of the defined steps, between two of the defined steps, or after all the defined steps

The present inventors realize that certain polarizable materials (also referred to as "ferroelectric materials") may be used to form storage elements in memory applications. Such ferroelectric materials include, for example, hafnium oxide, which have been used in forming charge-trapping layers in non-volatile or quasi-volatile memory circuits. According to one embodiment of the present invention, high-density data storage may be achieved using ferroelectric field-effect transistors ("FeFETs") as the storage transistors of a 3-dimensional memory array (e.g., an array of NOR memory strings). The present invention, as illustrated herein, provides high endurance, long data retention, and relatively low voltage operations for both erase (under 7.0 volts) and programming (e.g., under −7.0 volts) operations. By combining their ferroelectric or polarization characteristics with the 3-dimensional organization (e.g., as thin-film NOR memory strings described herein), memory arrays formed out of the FeFETs of the present invention achieve the additional benefits of high-density, low-cost, high-speed randomly accessed memory circuits (e.g., low read latency).

FIGS. **1***a*-**1** and **1***a*-**2** show conceptualized memory structure **100**, which illustrates an organization of ferroelectric field-effect transistors (FeFETs) in an array of NOR memory strings, according to one embodiment of the present invention. This detailed description is facilitated by adopting as reference a Cartesian coordinate system, as indicated in FIGS. **1***a*-**1** and **1***a*-**2**. Conceptualized memory structure **100**, provided herein as an abstraction of certain salient characteristics of a memory structure of the present invention, represents an array of NOR memory strings fabricated over a surface of substrate **101**. Conceptualized memory structure **100**, being an abstraction, is not drawn to scale along any of the X-, Y-, and Z-directions. Substrate **101** may be, for example, a conventional silicon wafer having a top layer with a planar surface for fabricating integrated circuits, as is familiar to those of ordinary skill in the art. Thus, under this adopted Cartesian coordinate system, the term "horizontal" refers to any direction parallel to the planar surface of substrate **101**, which is a surface of an X-Y plane. Similarly, the term "vertical" refers to the Z-direction or a normal to the planar surface.

As shown in both FIGS. **1***a*-**1** and **1***a*-**2**, four planes (e.g., plane **110**) of conceptualized memory structure **100** are isolated from each other and stacked one on top of another in the vertical direction. FIGS. **1***a*-**1** and **1***a*-**2** each show four horizontal active strips of NOR memory strings in each plane (e.g., active strip **112**). As discussed below in further detail in this detailed description, each active strip—which includes several material layers extending lengthwise along the Y-direction—may have formed therein one or more NOR memory strings. Each NOR memory string includes numerous FeFETs (e.g., FeFET **111**) sharing a common drain region and a common source region, the FeFETs being formed side-by-side along one or both sides of the active strip, with the thin-film FeFET current flowing through a channel region along the vertical direction. In an FeFET, the polarization in the ferroelectric element affects its threshold voltage.

In data storage or memory applications, two different threshold voltage values are selected to represent the values of a binary data bit. In such applications, the polarization states of the FeFET may be referred to as being in the "erased state" or the "programmed" state. In a NOR memory string, programming (also referred to as "writing"), reading or erasing any one of the FeFETs does not require activating any other FeFETs in the same NOR memory string. Accordingly, each FeFET of a NOR memory string is randomly addressable and accessible. FeFETs in close proximity that are not intended for the read, program or erase operation, however, are protected from unintended programming or erasing by having their transistor terminals suitably biased, often referred to as "program-inhibit" or "erase-inhibit" operations. Furthermore, an FeFET that has been accessed repeatedly may experience a gradual shift in its threshold voltage due to a phenomenon commonly referred to as "read-disturb." Indeed, even programmed or erased FeFETs need to be protected from shifts in their threshold voltages ("disturbed") during read, write or erase operations on nearby FeFETs. One effective way to mitigate these "disturb" conditions is a "refresh" operation on the FeFET. The refresh operation is designed to read and restore the programmed or erased state of the FeFET before its threshold voltage shifted too much. In some embodiments, an error detection and correction (ECC) operation may accompany the refresh operation, thereby allowing the correct data to be restored, even if the disturb condition has shifted the FeFET to an incorrect polarization state.

FIG. **1***a*-**2** introduces additional randomly addressable units of memory cells that may be organized in an array of NOR memory strings: "columns," "pages" and "slices". For example, each column (e.g., column **115**) represents the vertically-aligned FeFETs of multiple NOR memory strings in the active strips of an active stack; these vertically-aligned FeFETs share a common control gate or local word line.

Each "page" refers to a fixed number FeFETs (e.g., a fixed number of FeFETs arranged along the X-direction, indicated in FIG. **1***a*-**2** as page **113**), and each "slice" refers to an organization of contiguous FeFETs that extend along both the X- and Z-directions and one FeFET deep along the Y-direction, indicated in FIG. **1***a*-**2** by reference numeral **114**.

Although illustrated by conceptualized memory structure **100**, which is shown in the figures as an array having a 4 planes by 4 active strips organization, a memory structure of the present invention may have any number of FeFETs along any of the X-, Y- and Z-directions. For example, there may be 1, 2, 4, 8, 16, 32, 64 . . . planes in the Z direction, 2, 4, 8, 16, 32, 64, . . . active strips along the X-direction, and each NOR memory string in each active strip may have 2, 4, 8, 16, . . . 8192 or more FeFETs formed along the Y-direction. The use of numbers that are integer powers of 2 (i.e., 2^n, where n is an integer) herein follows a customary practice in conventional memory design. It is customary to access each addressable unit of memory by decoding a binary address. Thus, for example, a memory structure of the present invention may have M NOR memory strings along each of the X and Z directions, with M being a number that is not necessarily 2^n, for any integer n. FeFETs of the present invention may be organized such that reading, programming or erasing may be performed simultaneously on an individual page or individual slice basis. In some embodiments, an erase operation can also be performed in simultaneously for FeFETs in an even larger addressable unit of the array of NOR memory strings.

FIG. **1***b* shows a basic circuit representation of four NOR memory string-pairs, with each NOR memory string-pair including two NOR memory strings located on the same plane (i.e., one of planes **159-1**, **159-2**, **159-3** and **159-4**), according to one embodiment of the present invention. Within each NOR memory string-pair, corresponding FeFETs of the NOR memory strings share a common local word line (e.g., local word line **151***n*). The detailed structure of this configuration is discussed and illustrated below. FIG. **1***b* shows a configuration which includes four NOR memory string-pairs on four separate planes (e.g., NOR memory strings **150**L and **150**R on plane **159-4**) that are provided in adjacent columns of conceptualized memory structure **100** and share common local word line **151***a*.

As shown in FIG. **1***b*, NOR memory strings **150**L and **150**R may be NOR memory strings formed along two active strips located on opposite sides of shared local word line **151***a*. FeFETs **152**R-**1** to **152**R-**4** and **152**L-**1** to **152**L-**4** may be FeFETs located in the four pairs of active strips on opposite sides of local word line **151***a*. In this embodiment, local word line **151***a* controls FeFETs **152**R-**1**, **152**R-**2**, **152**R-**3** and **152**R-**4** from four NOR memory strings located on planes **159-1**, **159-2**, **159-3** and **159-4**, respectively. Local word line **151***a* also controls FeFETs **152**L-**1**, **152**L-**2**, **152**L-**3** and **152**L-**4** from four NOR memory strings on corresponding planes **159-1**, **159-2**, **159-3** and **159-4**, respectively. As discussed in greater detail below, in some embodiments, the parasitic capacitance C intrinsic to each NOR memory string (e.g., the distributed capacitance between the common n^+ source region of a NOR memory string and its multiple associated local word lines) may be used as a virtual voltage source, under some operating conditions, to provide common source voltage V_{ss}. For that matter, the common source voltage can be set through a pre-charge operation that provides it the voltage of the common n+ drain, for example. This flexibility in readily setting the source voltage and the drain voltage to a voltage that is more positive than the voltage on the local word line (e.g., during an erase operation), or to a voltage that is more negative than the voltage on the local word line (e.g., during a programming operation) is particularly advantageous because the polarization states (e.g., the "erased" and "programmed" states) of the FeFET are primarily symmetrical. That is, a first polarization state of the FeFET can be set by imposing a first voltage on both the source and the drain regions—as a first electrode—and a second voltage on the local word line, serving as a second electrode. This first polarization state can be readily reversed (i.e., set to the other polarization state by exchanging the voltages on the first and the second electrodes (i.e., imposing (i) the second voltage on both the source and the drain regions, and (ii) the first voltage on the local word line). According to one embodiment of the present invention, the NOR memory string does not require a hard-wired connection to a ground potential (Vss) for its common source region, common drain region, or any of its channel regions. Consequently, both polarization states can be achieved on any of the FeFETs of the NOR memory string using only positive voltages (or only negative voltages) applied to the corresponding local word line and the common source and the common drain regions. This single voltage polarity capability is advantageous because the typical FeFET has polarization states characterized by a voltage hysteresis loop that between a negative voltage in the "programmed" state to a positive voltage in the "erased" state.

FIG. **1***c* shows a basic circuit representation of four NOR memory strings, with each NOR memory string being located on a respective one of planes **159-1**, **159-2**, **159-3** and **159-4**, according to one embodiment of the present invention. In FIG. **1***c*, corresponding FeFETs (e.g., FeFETs **152-1**, **152-2**, **152-3** and **152-4**) from the NOR memory strings share a common local word line (e.g., **151***a*). Each NOR memory string may run horizontally along the Y-direction, with storage elements (i.e., FeFETs) of the NOR memory string connected between common source line **153**-*m* and common bit line **154**-*m*, where in takes a value between 1 and 4. Corresponding FeFETs in the four NOR memory strings share corresponding one of local word lines **151**-*n*, where n is the index of a local word line. The FeFETs in the NOR memory strings of the present invention are variable threshold voltage thin-film ferroelectric or polarization transistors wherein current to voltage hysteresis is created when the FeFET is electrically programmed, erased, or read, using programming, inhibition, erasure and read voltages that may be readily generated in the "circuitry under array" ("CUA") formed in the semiconductor substrate.

Fabrication Processes

FIG. **2***a* shows a cross section in a Y-Z plane of semiconductor structure **200**, after active layers **202-0** to **202-7** (each separated from the next active layer respectively by isolation layers **203-0** to **203-7**) have been formed on semiconductor substrate **201**, but prior to forming individual active strips, in accordance with one embodiment of the present invention. Semiconductor substrate **201** represents, for example, a p^- doped bulk silicon wafer on which support circuits for memory structure **200** may be formed as CUA prior to forming the active layers. Such support circuits, which may be formed alongside contacts **206-0** and **206-1** in FIG. **2***c*, may include both analog and digital circuits. Some examples of such support circuits include shift registers, latches, sense amplifiers, reference cells, power supply lines, bias and reference voltage generators, inverters, NAND, NOR, Exclusive-Or and other logic gates, input/output drivers, address decoders (e.g., bit line and word line decoders), other memory elements, sequencers and state machines. These support circuits may be formed out of the building blocks for conventional devices (e.g., n-wells, p-wells, triple wells, n^+, p^+ diffusions, isolation regions, low and high voltage transistors, capacitors, resistors, vias, interconnects and conductors), as is known to those of ordinary skill in the art.

After the support circuits have been formed in and on semiconductor substrate **201**, an isolation layer is provided, which may be a deposited or grown thick silicon oxide, for example.

Next, in some embodiments, one or more layers of interconnect conductors (e.g., metallic interconnect lines) may be formed, including "global word lines" for interconnecting the support circuits to the "local word lines" within the array of NOR memory strings. The layers of interconnect conductors are further discussed below. In other embodiments, global word lines are not provided under the memory array. Such metallic interconnect lines (e.g., global word line landing pads **263** of FIG. **2***c*, discussed below) may be provided as horizontal long narrow conductive strips running along a predetermined direction that may be perpendicular to the NOR memory strings to be formed at a later step. To facilitate discussion in this detailed description, the global word lines are presumed to run along the X-direction. The metallic interconnect lines may be formed by applying photo-lithographical patterning and etching steps on one or more deposited metal layers. (Alternatively, these metallic interconnect lines can be formed using a conventional damascene process, such as a copper or tungsten damascene process). A thick oxide is deposited to form isolation layer, followed by a planarization step using conventional chemical mechanical polishing (CMP) techniques.

Over isolation layer **203-0**, active layers **202-0** to **202-7** are then successively formed, with each active layer being electrically insulated from the previous active layer underneath by a corresponding one of isolation layers **203-1** to **203-7**. In this detailed description, the term "active layer" refers to material layers that form circuit elements (e.g., thin-film ferroelectric or polarization transistors) in the array of NOR memory strings. In FIG. **2***a*, although eight active layers are shown, any number of active layers may be provided. In practice, the number of active layers to be formed may depend on the process technology, such as the availability of a well-controlled anisotropic etching process that allows cutting through a tall stack of active layers to reach semiconductor substrate **201**.

FIGS. **2***b***-1** to **2***b***-5** show examples of active layer **220***a*, according to one embodiment of the present invention. In the example of FIG. **2***b***-1**, active layer **220***a* includes the material layers: (i) n^+ polysilicon or silicon layers **204***d* and **204***b* and (ii) sacrificial dielectric layer **204***c* (which is to be subsequently replaced at least in part by a p^- doped channel polysilicon layer). In one implementation, n^+ silicon layers **204***d*-**204***b* may be deposited successively in the same process chamber without removal in between. N^+ silicon layer **204***b* may be formed by depositing 10-100 nm of in-situ doped n^+ polysilicon. Dielectric layer **204***c* may be formed by depositing a sacrificial material (e.g., silicon oxide or SiO_2) in the thickness range of 10-100 nm. N^+ silicon layer **204***d* (i.e., the top portion of the deposited polysilicon) may be formed by depositing 10-100 nm of in-situ doped n^+ polysilicon. Both silicon layers **204***b* and **204***d* may be in situ n^+-doped. N^+ dopant concentrations in silicon layers **204***d* and **204***b* may be, for example between $1\times10^{20}/cm^3$ and $1\times10^{21}/cm^3$. One consideration for the dopant concentration is the sheet resistivity in n^+ silicon layers **204***d* and **204***b*. The n^+ doping may be achieved by either (i) a low-energy shallow high-dose ion implantation of phosphorus, arsenic or antimony, or (ii) in-situ phosphorus or arsenic doping of the deposited polysilicon, forming a 10-100 nm thick n^+ silicon layer **204***d* on top. The boron or p^--dopant concentration of the channel polysilicon layer—replacing at least in part sacrificial dielectric layer **204***c*—can be in the range of $1\times10^{16}/cm^3$ to $1\times10^{18}/cm^3$. The actual boron concentration in layer **204***c* determines the native transistor threshold voltage, channel mobility, $n^+p^-n^+$ punch-through voltage, n^+p^- junction leakage and reverse diode conduction characteristics, and channel depletion depth under the various operating conditions for the $n^+p^-n^+$ FeFETs formed along active strips formed out of active layers **202-0** to **202-7**.

Thermal activation of the n^+ and p^- implanted species and recrystallization of silicon layers **204***d*, **204***c* and **204***b* should preferably take place all at once after all active layers **202-0** to **202-7** have been formed, using a conventional rapid thermal annealing technique (e.g., at 700° C. or higher) or a conventional rapid laser annealing technique, thereby ensuring that all active layers experience elevated temperature processing in roughly the same amount.

Alternatively, n^+ and p^- dopants of each of active layers **202-0** to **202-7** can be activated individually by shallow rapid thermal annealing using, for example, an excimer laser annealing (ELA) technique at an ultraviolet wavelength (e.g., 308 nanometer). The annealing energy which is absorbed by the polysilicon or amorphous silicon to partially melt layer **204***d* and part or all of layer **204***c*, optionally penetrating layer **204***b* to affect volume **205** (see FIG. **2***b***-4**) without unduly heating other active layers lying below portion **223** of n^+ silicon layer **204***b* of the annealed active layer **220***a*.

Although the use of successive layer-by-layer excimer laser shallow rapid thermal anneal is more costly than a single deep rapid thermal anneal step, ELA has the advantage that the localized partial melting of polysilicon (or amorphous silicon) can result in recrystallization of annealed volume **205** to form larger silicon polycrystalline grains having substantially improved mobility and uniformity, and reduced FeFET leakage due to reduced segregation of n^+ dopants at the grain boundaries of the affected volume. Such shallow excimer laser low-temperature anneal technique is well-known to those of ordinary skill in the art. For example, such technique is used to form polysilicon or amorphous silicon films in solar cell and flat panel display applications. See, for example, the article "Comprehensive Study of Lateral Grain Growth in Poly-Si Films by Excimer Laser Annealing (ELA) and its applications to Thin Film Transistors" by H. Kuriyama et al., published in Japanese Journal of Applied Physics, Vol. 33, Part 1, Number 10, 20 Aug. 1994, or the article "Annealing of Silicon Backplanes with 540W Excimer Lasers", published in technical publication by Coherent Inc. on their website.

The thickness of layer **204***c* roughly corresponds to the channel length of the FeFETs to be formed, which may be 10 nm thick or less over long active strips. In one embodiment, for example, diffusion in active layer **220***a* of FIG. **2***b***-5** may be controlled so that the channel length of the FeFETs remains near the target of 20.0-100.0 nm, even after several thermal process cycles, by depositing an ultra-thin (from one or a few atomic layers thick) film of silicon nitride (e.g., SiN or Si_3N_4), or another suitable diffusion-blocking film following the formation of n^+ silicon layer **204***b* (e.g., layer **204***b*-*d* in FIG. **2***b***-5**). A second ultra-thin film of silicon nitride, or another suitable diffusion-blocking film (e.g., layer **204***d-d* in FIG. **2***b*-**5**), may optionally be deposited following deposition of p^- doped layer **204***c*, before depositing n^+ silicon layer **204***d*. The ultra-thin dopant diffusion-blocking layers **204***d-d* and **204***b-d* can be deposited by chemical vapor deposition, atomic layer deposition or any other suitable means (e.g., high pressure nitridation at low temperature). Each ultra-thin dopant diffusion-blocking layer acts as a barrier that prevents the n^+ dopants in n^+ silicon layers **204***d* and **204***b* from diffusing into p^- layer **204***c*, yet are sufficiently thin to only marginally impede the MOS transistor action in the channel region between n^+ silicon layer **204***d* (acting as a source) and n^+ silicon layer **204***b* (acting as a drain). These additional ultra-thin dopant diffusion-blocking layers increase the manufacturing cost, but such ultra-thin dopant diffusion-blocking layers may serve to significantly reduce the cumulative leakage current from the multiple FeFETs along the active strips that are in the "off" state. However, if that leakage current is tolerable then these ultra-thin layers can be omitted. Indeed, relative to conventional non-ferroelectric field-effect transistors, the "off" state leakage currents in FeFETs are significantly smaller because of the large bandgap in the ferroelectric storage material (e.g., HfO_2).

Sheet resistance in n^+ silicon layers **204***b* and n^+ silicon layers **204***d* may be reduced by adding optional conductive or metallic layers **204***g* or **204***f* to each active layer, adjacent to n^+ silicon layers **204***d* or **204***b* (e.g., metal layer **204***a*, labeled as W (for tungsten) in FIGS. **2***b*-**2** and **2***b*-**3**). Metal layer **204***g* or **204***f* may be provided by one or more deposited metal layers. For example, metal layer **204***f* or **204***g* may be provided by depositing 1-2 nm thick liner layer of TiN followed by depositing a 1-40 nm thick layer of tungsten, a similar refractory metal, or a polycide or silicide (e.g., nickel silicide). Metal layer **204***f* or **204***g* is more preferably in the 10.0-40.0 nm thickness range. Preferably, a sacrificial material (e.g., silicon nitride or SiN) is deposited to stand-in for metal layers **204***f* and **204***g*, which is subsequently replaced by a conductive material to form low-resistivity layers **204***f* and **204***g*. Even very thin layer **204***f* and **204***g* (e.g., 20.0-50.0 nm) can significantly reduce the line resistance of a long active strip, while allowing the use of less heavily doped n^+ silicon layers **204***d* and **204***b*.

As shown in FIG. **2***c*, the conductor inside contact opening **205-1** can become quite long for a tall stack, which may adversely increase the line resistance. In that case, metal layer **204***a* (e.g., a tungsten layer) may preferably be included below n^+ silicon layer **204***b*, so as to substantially fill contact opening **205-1**, rather than placing it above n^+ silicon layer **204***d*, as is shown in FIG. **2***c*.

In the embodiments where metal layers **204***f* and **204***g* are not present, there are several tradeoffs that can be made: for example, longer active strips are possible if the resulting increased read latency is acceptable. In general, the shorter the active strip, the lower the line resistance and therefore the shorter the latency. (The trade-off is in array efficiency). In the absence of metallic layers **204***f* and **204***g*, the thickness of n^+ silicon layers **204***d* and **204***b* can be increased (for example, to 100.0 nanometers) to reduce the intrinsic line resistance, at the expense of a taller stack to etch through. The line resistance can be further reduced by increasing the n^+ doping concentration in n^+ silicon layers **204***d* and **204***b* and by applying higher anneal temperatures in excess of 1,000° C. (e.g., by rapid thermal anneal, deep laser annealing or shallow excimer laser annealing techniques) to enhance recrystallization and dopant activation and to reduce dopant segregation at the grain-boundaries.

Shorter active strips also have superior immunity to leakage between n^+ silicon layer **204***b* and n^+ silicon layer **204***d*. A thicker n^+ layer provides reduced strip line resistance and increased strip parasitic capacitance, which is not desirable for dynamic sensing. The integrated circuit designer may opt for a shorter active strip (with or without metal layer **224**) when low read latency is most valued. Alternatively, the strip line resistance may be reduced by contacting both ends of each active strip, rather than just at one end.

Block-formation patterning and etching steps define separate blocks in each of the active layers formed. Each block occupies an area in which a large number (e.g., thousands) of active strips running in parallel may be formed, as discussed below, with each active strip running along the Y-direction, eventually forming one or more NOR memory strings that provide a large number (e.g., thousands) of individually randomly accessible FeFETs.

Each of active layers **202-0** to **202-7** may be successively formed by repeating the steps described above. In addition, in the block-formation patterning and etching steps discussed above, each next higher active layer may be formed with an extension slightly beyond the previous active layer (see, e.g., as illustrated in FIG. **2***c*, discussed below, layer **202-1** extends beyond layer **202-0**) to allow the upper active layer to access its specific decoders and other circuitry in semiconductor substrate **201** through designated buried contacts.

As shown in FIG. **2***c*, buried contacts **205-0** and **205-1** connect contacts **206-0** and **206-1** in semiconductor substrate **201**, for example, to the local bit lines or source lines formed out of n^+ silicon layer **204***b* in each of active layers **202-0** and **202-1**. Buried contacts for active layers **202-2** to **202-7** (not shown) may be similarly provided to connect active layers **202-2** to **202-7** to contacts **206-2** to **206-7** in semiconductor substrate **201** in an inverted staircase-like structure in which the active layer closest to the substrate has the shortest buried contact, while the active layer furthest from the substrate has the longest buried contact. Alternatively, in lieu of buried contacts, conductor-filled vias extending from the top of the active layers may be etched through isolation layers **203-0** and **203-1**. These vias establish electrical contact from substrate circuitry **206-0**, for example, to top n^+ silicon layers **204***d*-**0** (or metal layer **224**, if provided). The vias may be laid out in a "staircase" pattern with the active layer closest the substrate connected by the longest via, and the active layer closest to the top connected by the shortest via. The vias (not shown) have the advantage that more than one plane can be contacted in one masking-and-etch step, as is well-known to a person of ordinary skill in the art.

In an integrated circuit incorporating multiple arrays of NOR memory strings, the arrays of NOR memory strings may each be provided as a modular circuit or tile, as discussed above. Each array of NOR memory strings may include an "array" portion, in which the FeFETs of the NOR memory stings are provided, and one or more "staircase" portions for routing signals into or out of the NOR memory strings, using either the buried contact inverted staircase-like structure, or by conductor-filled vias in a staircase structure, as known to those of ordinary skill in the art. Each tile may have, for example, a regular-shape footprint (e.g., a rectangular footprint) on the planar surface of the semiconductor substrate, thus allowing many tiles to be formed and arranged on a substrate in a regular manner (e.g., a m×n matrix, where m and n are integers.)

Through a switch circuit, each of contacts **206-0** to **206-7** may apply a pre-charge voltage V_{bl} to the respective bit line or source line of the corresponding NOR memory strings. During a read operation, each selected bit line may be connected by the switch circuit to an input terminal of a sense amplifier or a latch. The switch circuit may selectively connect each of contacts **206-0** to **206-7** to any of a number of specific voltage sources, such as a programming voltage (V_{pgm}), inhibit voltage ($V_{inhibit}$), erase voltage (V_{erase}), or any other suitable predetermined or pre-charge reference voltage V_{bl} or V_{ss}. In some embodiments, discussed below, taking advantage of the relatively large parasitic distributed capacitance along a bit line or source line in an active strip, a virtual voltage reference (e.g., a virtual ground, providing ground voltage V_{ss}) may be created in the source line (i.e., n^+ silicon layer **204***d*, or a metal line) of each active strip by pre-charging the source line, as discussed below. The virtual ground approach eliminates the need for hard-wiring n^+ silicon layer **204***d* to a voltage source in the substrate, making it possible to more efficiently use the staircase via structure to connect each active strip from the top to the substrate.

FIG. **2***c* also shows buried contacts **261-0** to **261-***n* for connecting global word lines **208***g-s*—which are to be formed running along the X-direction—to contacts **262-0** to **262-***n* in semiconductor substrate **201**. Global word lines **208***g-s* are provided to connect corresponding local word lines **208**W-s yet to be formed to circuits **262-***n* in substrate **201**. Landing pads **263** are provided on the global word lines to allow connection to local word lines **208**W-s, which are yet to be formed vertically on top of horizontally running global word lines **208***g-s*. Through a switch circuit and a global word line decoder, each of global word line contacts **262-0** to **262-***n* may be selectively connected, either individually, or shared among several global word lines, to any one of a number of reference voltage sources, such as stepped programming voltages ($V_{program}$), program-inhibit voltage ($V_{inhibit}$), read voltages (V_{read}) and erasure voltages (V_{erase}).

The buried contacts, the global word lines and the landing pads may be formed using conventional photo-lithographical patterning and etching steps, followed by deposition of one or more suitable conductors or by alloying (e.g., tungsten metal, alloy or tungsten silicide).

FIGS. **2***d* to **2***k* illustrates a portion of an exemplary process for forming a modular circuit or tile that includes an array of NOR memory strings, including the array portion (i.e., memory structure **200**), according to one embodiment of the present invention; FIG. **2***k* illustrates a step the exemplary process that provides a channel layer for the FeFETs. The immediate steps following the step providing the channel layer in the exemplary process are illustrated by three groups of process steps, each group representing a different approach to "gate stack" formation. The groups of process steps are illustrated, respectively, in FIGS. **3***a* to **3***b*, **4***a* to **4***c* and **5***a* to **5***f*.

FIG. **2***d*(*i*) shows a cross section in the Z-X plane of memory structure **200** after depositions of active layers **202-0** to **202-7**, in accordance with one embodiment of the present invention. Initially, in this embodiment, pad oxide **240** (e.g., silicon oxide) is provided over planar surface **201** of the semiconductor substrate. Etch-stop layer **241** (e.g., tungsten (W), tungsten nitride (WN), aluminum oxide (AlO), aluminum nitride (AlN), or any suitable etch-stop material) is then provided. Silicon oxycarbide (SiOC) layer **203-0** is then provided to isolate etch-stop layer **241** from the next active layer (i.e., active layer **202-0**) to be deposited. Thereafter, active layers **202-0** to **202-7** (eight in total in this example; generically, each may be referred to as "active layer **202**") are successively deposited, each active layer being isolated from the next active layer by a corresponding one of isolation layers **203-1** to **203-7** (active layers **203-0** to **203-7** are also each generically referred to as "isolation layer **203**" and, collectively, "isolation layers **203**"). Active layers **202** each include, in order of deposition, (i) silicon nitride (SiN) layer **204***a*, (ii) n^+-doped amorphous silicon (or polysilicon) layer **204***b*, (iii) sacrificial oxide layer **204***c*, (iv) n^+-doped amorphous silicon (or polysilicon) layer **204***d*, and (v) SiN layer **204***e*. Isolation layer **245** (e.g., an additional SiOC layer) is then deposited on top of active layer **202-7**. The resulting structure is shown as memory structure **200** of FIG. **2***d*(*i*).

FIG. **2***d*(*ii*) illustrates successive recessing and etching steps to create staircase portion **200-***sc*, in accordance with one embodiment of the present invention. As shown in FIG. **2***d*(*ii*), on the top surface of memory structure **200** (i.e., isolation layer **245**) is formed mask layer **210**, which exposes first portion **211** of memory structure **200**, as shown in FIG. **2***d*(*ii*)(**1**). The exposed portion of isolation SiOC layer **245** is then removed to expose a portion of active layer **202-7** underneath. That exposed portion of active layer **202-7** is then removed, exposing a portion of SiOC layer **203-6** underneath. The resulting structure is shown in FIG. **2***d*(*ii*)(**2**). Mask layer **210** is then recessed to expose a further portion of isolation SiOC layer **245**. Removal of the exposed SiOC layers (e.g., SiOC layers **245** and **203-7**), removal of exposed active layer **202** and recessing mask layer **210** are then repeated 7 more times, thereby forming a staircase structure. Thereafter, an oxide is deposited to fill the portions of active layers **204-1** to **204-7** removed. A chemical-mechanical polishing (CMP) step is carried out to remove mask layer **210** and to planarize the top surface of memory structure **200**.

The steps illustrated in FIG. **2***d* form the staircase structures (e.g., on opposite sides of the array portion) prior to processing the array portion. However, the array portion may also be processed prior to formation of the staircase structures.

The surface of SiOC layer **245** is patterned and etched to create openings **218**. FIG. **2***e*(*i*) is a top view of memory structure **200**, showing shafts **218**, after patterning and etching. In FIG. **2***e*(*i*), shafts **218** are shown to have a circular X-Y plane cross section. However, any suitable X-Y plane cross section shape (e.g., oval) may also be used. A series of etching steps remove the exposed portions of isolation layers **203-0** to **203-7**, and active layers **202-0** to **202-7** until etch-stop layer **241** is reached, thereby extending shafts **218** along the Z-direction. FIG. **2***e*(*ii*) shows an X-Z plane cross-section of memory structure **200** along line A-A' in FIG. **2***e*(*i*), after the etching steps through active layers **202-0** to **202-7** and isolation layers **203-0** to **203-7**, in accordance with one embodiment the present invention. (In this detailed description, a "column" of objects denotes objects aligned along the Y-direction, whereas a "row" of objects denotes objects aligned along the X-direction.) In FIG. **2***e*(*i*), shafts of each column is staggered relative to the shafts of each of its immediate neighbor column, such that the closest shafts in adjacent columns have a greater separation between them than if such shafts are aligned along the X-direction. In one embodiment adjacent columns of shafts are 160 nm apart along the X-direction, and each of shafts **218** may have a diameter of 100.0 nm. Shafts **218** may each be up to 2.0 microns deep, so that the etching steps has an aspect ratio that is less than 20.

Shafts **218** may then be lined using an oxide liner **242** and filled using sacrificial material **243** (e.g., carbon or tungsten (W) with an adhesive layer of titanium nitride (TiN)). The surface of memory structure **200** may then be planarized using CMP or, when carbon is used as the sacrificial material, etched back. After shafts **218** are filled with sacrificial material **243**, hard mask **244** (e.g., a carbon hard mask layer) is then deposited and photo-lithographically patterned over memory structure **200**. Hard mask **244** may be patterned to provide rectilinear openings **246**, each extending lengthwise along the Y-direction, exposing portions of memory structure **200** (e.g., each 60.0 nm wide, in one embodiment), and overlapping but not removing the columns of filled shaft **218**. A top view and a cross-section view of resulting memory structure **200**, after hard mask **244** is patterned, are shown in FIGS. **2***f*(*i*) and **2***f*(*ii*), respectively, according to one embodiment of the present invention. The cross-section view of FIG. **2***f*(*ii*) cuts through the X-Z plane along line A-A' of FIG. **2***f*(*i*).

From underneath rectilinear openings **246**, successive etching steps remove portions of SiOC layer **245**, isolation layers **203-0** to **203-7**, active layers **202-0** to **202-7**, and etch-stop layer **241**, thereby dividing memory structure **200** into numerous active stacks separated by trenches **267**. The etching steps are designed to be selective to oxide liner **242** and sacrificial material **243** in filled shafts **218**, so that filled shafts **218** become pillars **248** of sacrificial material **243**, surrounded on its cylindrical surface by oxide liner layer **242**. An additional wet etch may be applied to remove any stringers of the various material. Hard mask **244** may then be removed (e.g., by CMP or by ashing, as appropriate). A top view and an X-Z plane cross section view along line A-A' of resulting memory structure **200**, after hard mask **244** has been removed, are shown in FIGS. **2***g*(*i*) and **2***g*(*ii*), respectively, according to one embodiment of the present invention.

Trenches **267** may then be filled by dielectric material **247** (e.g., silicon oxide) using, for example, atomic layer deposition (ALD) or a spin-on technique. Excess oxide on the top surface of memory structure **200** may then be removed using, for example, CMP. Thereafter, pillars **248** are removed using, for example, a suitable wet etch to restore shafts **218**. FIGS. **2***h*(*i*) and **2***h*(*ii*) show a top view and an X-Z plane cross-section view along line A-A' of resulting memory structure **200**, respectively, after removal of pillars **248**, according to one embodiment of the present invention.

Removal of pillars **248** restores shafts **218** between filled portions of trenches **267** and exposes active layers **202-0** to **202-7** as sidewalls of shafts **218**. Through shafts **218**, SiN layers **204***a* and **204***e* of each of active layers **202-0** to **202-7** may be removed using, for example, a wet etch of silicon nitride. FIGS. **2***i*(*i*) and **2***i*(*ii*) show a top view and an X-Z plane cross-sectional view along line A-A' of resulting memory structure **200**, respectively, after removal of SiN layers **204***a* and **204***e* from each of active multi-layers **202-0** to **202-7**, according to one embodiment of the present invention.

An ALD step then deposits a conductive material into the cavities resulting from removal of SiN layers **204***a* and **204***e* in each of active layers **202-0** to **202-7**. The conductive material may include, for example, a metal liner (e.g., titanium (Ti), titanium nitride (TiN), tantalum (Ta) or tantalum nitride (TaN)), followed by a refractory metal (e.g., tungsten (W), tungsten nitride (WN) or molybdenum (Mo)). Thereafter, an anisotropic separation etch removes the conductor material from shafts **218**. Conductor layers **204***f* and **204***g* that respectively replace SiN layers **204***a* and **204***e* in each active layer may be recessed, for example, by 10.0 nm, to allow a subsequent filler material (e.g., silicon nitride) to prevent corrosion in the conductive material. FIGS. **2***j*(*i*) and **2***j*(*ii*) show a top view and an X-Z plane cross-sectional view along line A-A' of resulting memory structure **200**, respectively, after replacement of SiN layers **204***a* and **204***e* in each active layer **202** by conductor layers **204***f* and **204***g*, according to one embodiment of the present invention. Conductor layers **204***f* and **204***g* are optional conductor layers that are formed adjacent and in contact with n^+-doped silicon layers **204***b* and **204***d*. As n^+ silicon layers **204***b* and **204***d* of each active layer **202** would become the common bit line and the common source line of a NOR memory string to be formed, conductor layers **204***f* and **204***g* desirably reduce the resistivities in the common bit line and the common source line.

Then, sacrificial oxide layer **204***c*, n^+ silicon layer **204***b*, and n^+ silicon layer **204***d* of each active layer **202** may be recessed (e.g., by 20.0 nm) to create a cavity into which a channel semiconductor material may be deposited. Polysilicon layer **250**, intended for providing the channel regions of the thin-film storage transistors, is then deposited, etched back for separation, and recessed to remove stringers. Channel polysilicon layer **250** may be, for example, p^--doped poly silicon (e.g., 7.0-14.0 nm thick, preferably about 10.0 nm thick). In some embodiments, the final thickness is preferably the depth of the cavity created in active layer **204** (e.g., 20.0 nm). FIGS. **2***k*(*i*) and **2***k*(*ii*) show a top view and an X-Z plane cross-sectional view along line A-A' of resulting memory structure **200**, respectively, after channel polysilicon layer **250** is recessed, according to one embodiment of the present invention.

The next steps in the exemplary process form the "gate stacks" for FeFETs in the NOR memory strings. The gate stack for an FeFET includes (i) optionally, an interface layer, (ii) a ferroelectric or polarization layer, and (iii) a gate electrode. Optionally, a thin metal oxide (e.g., 1.0-2.0 nm thick aluminum oxide) may also be provided between the gate electrode and the polarization layer to reduce tunneling of charged species into the polarization layer. The inventors discover that superior performance in 3-dimensional arrays of NOR memory strings may be achieved when the polarization layers in the FeFETs are separated from each other. FIGS. **3***a* to **3***b*, **4***a*-**4***c* and **5***a*-**5***f* illustrate three exemplary, alternative groups of process steps suitable for fabricating the gate stacks of FeFETs in a NOR memory string, in accordance with some embodiments of the present invention. To provide greater clarity, each of FIGS. **3***a* to **3***b*, **4***a* to **4***c* and **5***a* to **5***f* shows only active stacks **261***a* and **261***b* and only active strips formed out of active layers **202-0** to **202-3**. However, it is understood that the teachings provided therein apply to all active stacks and active strips in any array of NOR memory strings in that embodiment of the present invention.

FIGS. **3***a* to **3***b* illustrate a first group of process steps for forming the gate stacks in FeFETs of an array of NOR memory strings, according to one embodiment of the present invention. FIGS. **3***a* to **3***b* each provide (a) an X-Y plane view across isolation layer **203-4** between the active strips that are formed out of active layers **202-3** and **202-4** in FIG. **2***k*(*ii*), and (b) an X-Z plane cross-sectional view, each showing active stacks **261***a* and **261***b* of memory structure **200** after completing the respective steps in this first group. FIG. **3***b* also includes a third view (iii), which is an X-Y plane cross-sectional view along line C-C' of FIG. **3***b*(*ii*).

After channel polysilicon layer **250** is conformally deposited, an etch-back step is performed, as discussed above in conjunction with FIGS. **2***k*(*i*) and **2***k*(*ii*). The etch-back step also exposes isolation layers **203-0** to **203-7** on the sidewalls of shafts **218**, as illustrated on in FIG. **3***a*(*ii*). In some embodiments, the etch-back step may also be designed to form a recess in active layers **202** between exposed portions of adjacent isolation layers **203**. Thereafter, a thin silicon nitride layer (e.g., 1 nm thick) may be formed on the exposed surfaces of channel polysilicon layer **250** in shafts **218**. The silicon nitride layer may be formed using an ammonia nitridation step on a native oxide on the surface of polysilicon layer **250**. (Alternatively, a thin oxide layer may be formed on polysilicon layer **250** by an oxidation step at a temperature of 400-650° C., and then converted to silicon nitride by any suitable nitridation process.) This nitride layer is used as an interface dielectric layer between channel polysilicon layer **250** and a polarization or polarization layer to be formed. Other suitable materials for the interface dielectric layer includes zirconium oxide (ZrO_2), silicon oxynitride (SiON), silicon nitride (Si_3N_4), or silicon oxide (SiO_2).

The native oxide layer—from which the interface dielectric layer is formed—may be formed by a chemical clean of the surface of polysilicon channel material **250**. (The silicon nitride in interface dielectric layer may also be any combination of silicon oxynitride (SiON) or silicon nitride (Si_3N_4), and silicon oxide (SiO_2), 0.0 to 2.0 nm thick, with a refractive index between 1.5 to 2.0.) A densification treatment may be performed on the silicon nitride layer using, for example, pulsed ozone or by thermal annealing in a hydrogen or deuterium ambient, or any other techniques known to a person of ordinary skill in the art. The densification treatment reduces electronic leakage through the interface dielectric layer, reduces the surface states at the interface between the third semiconductor layer and the polarization layer, or both. The inventors believe that the interface dielectric layer should have a refractive index between 1.4 and 1.8, preferably 1.6, which may be achieved, for example, by annealing at 630° C. in ozone or deuterium gas.

Selective deposition techniques may be used to deposit polarization layer **264** on the exposed surfaces, other than on isolation layers **203-0** to **203-7**. Suitable techniques include "surface poisoning" and self-assembled monolayers ("SAMs") techniques. In the SAM technique, for example, the exposed surfaces are treated with reactive species having active hydroxyl (—OH) bonds. Resulting memory structure **200** is shown in FIGS. **3***a*(*i*). In this embodiment, as the interface dielectric layer is deposited in a recess in each active strip between adjacent isolation layers, the interface dielectric layer is not seen in the X-Y plane cross-sectional view of FIG. **3***a*(*ii*), which cuts through isolation layer **203-4**.

Polarization layer **264** is then selectively deposited on the surfaces of the silicon nitride layer and channel material **250**. (e.g., the treatment by SAMs prevents deposition of polarization layer **250** on the sidewalls of isolation dielectric layers **203**.) Resulting memory structure **200** is shown in FIG. **3***b*(*i*). In this embodiment, the X-Y plane cross-sectional view of FIG. **3***b*(*iii*) shows polarization layer **264** deposited in the recessed portion of active layer **203-3**, between adjacent exposed portions of isolation layers **203-3** and **203-4**. Polarization layer **264** may be formed by the selective deposition using ALD techniques in an ozone ambient at temperature between 200° C. to 330° C. (e.g., around 300° C.), with a post-deposition annealing step at a temperature between 400° C. and 1000° C., based on the desired crystallization phase requirement of the ferroelectric material.

Deposition of the ferroelectric material may be achieved, for example, using any suitable deposition technique (e.g., chemical or physical vapor deposition, ALD, or evaporation) repeated cycles of hafnium oxide depositions and zirconium oxide depositions (e.g., in a HfO_2:ZrO_2 ratio of 4:1). Polarization layer **264** may be, for example, 2.0 to 8.0 nm thick, preferably, 2.5-5.0 nm thick). For the thicker polarization layers (e.g., greater than 4.0 nm), additional SAM treatments between deposition cycles may be advisable. Post deposition annealing may be performed, for example, in a temperature of 400-1000° C. to achieve a desired crystallization phase. Other suitable ferroelectric materials include aluminum zirconium-doped Hafnium oxide (HfZrAlO), aluminum-doped hafnium oxide (HfO_2:Al), silicon-doped hafnium oxide (HfO_2:Si) or lanthanum-doped hafnium oxide (HfO_2:La). The term HZO may encompass hafnium zirconium oxide (HfZrO), hafnium zirconium oxynitride (HfZrON), hafnium zirconium aluminum oxide (HfZrAlO) or any hafnium oxide that includes zirconium impurities.

The "surface poisoning" technique may be used to provide both the interface dielectric layer and polarization layer **264**. In one embodiment, HZO is deposited directly onto polysilicon channel layer **250** after its formation. After a few atomic layers of HZO have been deposited, a layer of silicon oxide is formed as native oxide at the interface between channel polysilicon layer **250** and the atomic layers of HZO. Nitridation of the silicon oxide layer can proceed in the manner discussed above to form the interface dielectric layer, which is silicon nitride in this case. The remainder of polarization layer **264** may then proceed to completion.

The interface dielectric layer significantly reduces tunneling of charged species (i.e., electrons or holes) between channel polysilicon layer **250** and polarization layer **264**. The charged species that tunnel into polarization layer **264** may adversely affect its polarization. Over time, the interface dielectric layer may break down as a result of a high number of program/erase cycles, thereby reducing the endurance of the FeFET. The interfacial layer may be formed out of a material with a dielectric constant greater than that of silicon oxide ("high-k" material; preferably, greater than 3.9). For example, aluminum oxide is a high-k material, having a dielectric constant as high as 9.0. A high-k material reduces both the electric field during program or erase operations and tunneling from the channel region. For a 0.0 nm thick interface dielectric layer, the polarization layer may be directly deposited by ALD onto the channel region (e.g., polysilicon). A native oxide of a self-limiting thickness (e.g., 1.0 to 10.0 angstrom) would be inherently formed at the interface between the channel region and the polarization layer. While this native oxide may be of low quality or prone to create a leakage current, it can be substantially improved by a self-limiting nitridation step. This approach is particularly advantageous when the channel region is formed after high temperature steps, such that contamination by dopant diffusion is a lesser concern. In some embodiments, a bandgap-engineered tunneling layer (e.g., a silicon oxide (SiO_2) and zirconium oxide (ZrO_2) multi-layer) may serve as the interface dielectric layer, providing the advantage of reduced tunneling into the polarization layer. In some embodiments, aluminum oxide (e.g., Al_2O_3) or silicon oxide nitride may be suitable. The high-k dielectric properties of the zirconium oxide and other high-k material reduce the electric field in the interface dielectric layer.

In other embodiments, polysilicon channel **250** may be replaced by a junction-less oxide semiconductor (e.g., indium gallium zinc oxide ("IGZO"), and indium aluminum zinc oxide ("IAlZnO")). With such a channel material, the interfacial layer between it and polarization layer **264** (e.g., HZO) may be subdued or even non-existent. The inventors believe that the oxide semiconductor channel region improves the endurance of the FeFET.

Thereafter, conductive material **272** (i.e., gate electrode) is deposited to fill shafts **218**. Conductive material **272** may include tungsten (W), molybdenum (Mo), aluminum (Al), ruthenium (Ru), tantalum (Ta), titanium (Ti), or any combination or alloy of these metals.

FIGS. **4***a* to **4***c* illustrate a second process flow for forming the gate stacks of the FeFETs in an array of NOR memory strings, according to one embodiment of the present invention. FIG. **4***a* shows an X-Z plane cross-sectional view of active stacks **261***a* and **261***b* of memory structure **200**, after the interface layer (e.g., a 1.0 nm thick silicon nitride layer) has been formed, for example, in substantially the same manner as described above in conjunction with FIG. **3***a*. Relative to memory structure **200** of FIG. **3***a*, however, the recesses in the active strips that are formed out of active layers **202-0** to **202-3** in FIG. **4***a* are deeper. For example, for channel polysilicon layer **250** to have an ultimate thickness of 10.0 nm, the recesses in the active strips that are formed out of active layers **202-0** to **202-3** are made to be 20.0 nm deep. An etch-back step applied to polysilicon channel layer **250** (e.g., a wet etch) may reduce the thickness of channel polysilicon layer **250** to 10.0 nm.

Thereafter, polarization layer **264** may be formed by depositing using, for example, ALD techniques. Specifically, polarization layer **264** may be deposited on the interface layer in the recesses of active strips formed out of active layers **202-0** to **202-3**. As discussed above, the polarization layer may be formed, for example, using repeated cycles of hafnium oxide depositions and zirconium oxide depositions (e.g., in a HfO_2:ZrO_2 ratio of 4:1). Resulting memory structure **200** is shown in FIG. **4***b*. Other suitable ferroelectric materials that may be used include aluminum-doped hafnium oxide (HfAlO) and aluminum zirconium-doped hafnium oxide (HfZrAlO). As shown in FIG. **4***b*, because of the recesses, polarization layer **264** in the active strips are separated from each other by the adjacent isolation dielectric layers **203**. Thereafter, conductive material **272** is deposited to fill shafts **218**, forming thereby the gate electrodes or local word lines.

FIGS. **5***a* to **5***f* illustrate a third group of process steps for forming the gate stacks in FeFETs of an array of NOR memory strings, according to one embodiment of the present invention. FIG. **5***a* shows an X-Z plane cross-sectional view of active stacks **261***a* and **261***b* of memory structure **200**, after deposition of channel polysilicon layer **250** into deep recesses of the active strips that are formed out of active layers **202-0** to **202-3** (e.g., 20 nm deep), an etch-back step and the formation of an interface dielectric layer in preparation for the ferroelectric material deposition. The etch-back step thins channel polysilicon layer **250** (e.g., to 10.0 nm thick) to expose the sidewalls of isolation dielectric layers **203-0** to **203-4**. Memory structures **200** of FIG. **5***a* may be formed using substantially the same process steps as those discussed above that are used to form memory structure **200** of FIG. **4***a*.

Thereafter, polarization layer **264** is formed on memory structure **200**, using the ALD techniques discussed, for example, in conjunction with FIG. **4***b* above, followed by a conformal deposition of amorphous silicon liner **265**. Resulting memory structure **200** is shown in FIG. **5***b*. Then, an anisotropic dry etch step removes portions of amorphous silicon liner **265**, which exposes portions of polarization layer **264** on the sidewalls of isolation dielectric layers **203** in shaft **218**, while allowing the remainder of amorphous silicon liner **265** protect the portions of polarization layer **264** in the recesses of the active strips. The anisotropic dry etch also sputters away amorphous silicon liner **265** and polarization layer **264** from top of memory structure **200**. Resulting memory structure **200** is shown in FIG. **5***c*.

Thereafter, in a first variation, a wet etch step removes the ferroelectric material (e.g., hafnium zirconium oxide (HZO)) of polarization layer **264** from the sidewalls of isolation dielectric layers **203** in shafts **218**. In a second variation, the wet etch step is allowed to remove a portion of the ferroelectric material of polarization layer **264** underneath amorphous liner layer **265**. The inventors believe that reducing the extent of polarization layer **264** along the Z-direction allow greater control of the resulting polarization properties. In addition, the interface dielectric layer for the diminished polarization layer **264** results in a reduced electric field between the source region and the drain region of an FeFET. Resulting memory structure **200** for the first variation and resulting memory structure **200** for the second variation are respectively shown in FIGS. **5***d*(*i*) and **5***d*(*ii*). Thereafter, the remainder of amorphous silicon liner **265** may be removed by a wet etch. In the second variation, oxide **266** (e.g., silicon oxide) is deposited using, for example, ALD techniques to fill the recesses in active layers **202**. An anisotropic etch-back step exposes diminished polarization layer **264** on the sidewalls of shafts **218**. Resulting memory structure **200** for the first variation and resulting memory structure **200** for the second variation are shown in FIGS. **5***e*(*i*) and **5***e*(*ii*), respectively.

Conductive material **272** is then deposited into and fills shafts **218** (e.g., a titanium nitride (TiN) adhesion layer, followed by tungsten). Conductive material **272** provides a common gate electrode—i.e., the local word line—for vertically aligned FeFETs in the NOR memory strings of an active stack. Resulting memory structure **200** for the first and the second variations are shown in FIGS. **5***f*(*i*) and **5***f*(*ii*), respectively. In summary, memory structure **200** of FIG. **6***a* represents active stacks **261***a* and **261***b* fabricated under the first group of process steps, the second group of process steps, or the first variation of the third group of process steps, in accordance with one embodiment of the present invention. Memory structure **200** of FIG. **6***b* represents active stacks **261***a* and **261***b* fabricated under the second variation of the third group of process steps, in accordance with one embodiment of the present invention.

The device structures for FeFETs shown in FIGS. **6***a* and **6***b* suggest that the 3-dimensional NOR memory strings disclosed herein have a significant advantage in that they offer a relatively large surface area of polarizable material in polarization layer **264**. A large surface area in polarization layer **264** provides a tight distribution of voltages in the erased and programmed states, which are difficult to achieve in planar FeFETs. Moreover, that relatively large surface area of polarizable material is provided entirely on the sidewalls of the active strips, and thus provides a high storage density without the penalty of substrate silicon real estate. Specifically, as the normal to polarization layer **264** is substantially horizontal, the polarization layer incurs little silicon real estate. In one embodiment, for example, the footprint of polarization layer **264** of each FeFET, as projected onto the planar surface of the semiconductor substrate, is merely 10.0 nm by 45.0 nm. (However, the Z-Y plane surface area of ferroelectric is 100.0 nm×45.0 nm, with a 100.0 nm channel length.) In some embodiments, the projected footprint of the polarization layer even lies within the projected footprint of the common bit line, as the ferroelectric or polarization layer is provided in a recess between the common bit line and the common source line. The advantage is amplified in a 3-dimensional array of NOR memory strings, as the footprint of the polarization material is substantially independent of the number of active strips in an active stack. Furthermore, the substantial advantage over planar storage transistors is achieved even if only one polarization layer is provided along the vertical or Z-direction. In that regard, the FeFETs of the present invention is advantageous even in embedded single-layer memory application.

In the process illustrated above, the channel region and the gate stack of each thin-film FeFETs are formed inside one of shafts **218**, which may have a curved contour in an X-Y plane cross section. In that embodiment, each rectilinear portion of trenches **267** between shafts **218** provide an isolation between adjacent FeFETs in a NOR memory string. Alternatively, replacement of SiN layers **204**_a_ and **204**_e_ may begin from the rectilinear portions of trenches **267** between adjacent shafts **218**, rather than through shafts **218**. Subsequent to metal replacement, channel polysilicon **250** and the gate stacks may be formed as described above, except that the subsequent process steps take place in the rectilinear portions of trenches **267** between shafts **218**, rather than in shafts **218**.

After the gate stacks are formed using any group of the process steps illustrated herein by FIGS. **3**_a_-**3**_b_, FIGS. **4**_a_-**4**_c_, FIGS. **5**_a_-**5**_c_, **5**_d_(_i_), **5**_e_(_i_) and **5**_f_(_i_), and FIGS. **5**_a_-**5**_c_, **5**_d_(_ii_), **5**_e_(_ii_) and **5**_f_(_ii_)), an isolation layer is provided over memory structure **200**. Thereafter, global interconnect conductors may be formed using any suitable conventional technique to provide connections to the local word lines in the array portion, and to the common bit lines in the staircase portions. FIG. **2**_l_ shows a X-Z plane cross section through a row of local word lines **208**_g-a_, showing active strips formed out of active layers **202-7** and **202-6**. Each local word line is formed by conductive material **272** in a rectilinear portion of one of trenches **267**, between adjacent shafts **218**. As shown in FIG. **2**_l_, each active strip includes n^+ polysilicon layers **204**_b_ and **204**_d_, and low-resistivity metal layer **204**_f_. (In FIG. **2**_k_, low-resistivity metal layer **204**_g_ is not shown.) Embedded in dielectric layers **231**L and **231**R are gate stack elements formed out of channel polysilicon layer **250** and polarization layer **264**, each element being located between n^+ silicon layers **204**_b_ and **204**_d_ (along the Z-direction). The gate stack of each FeFET is formed by one such gate stack element and its adjacent local word line **208**_g-a_ (connected to a global word line conductor running above the array portion) or **208**_g-s_ (connected to a global word line conductor running between the array portion and the silicon substrate). This configuration represents a high packing density configuration for FeFETs.

N^+ silicon layer **204**_b_ (i.e., the common source line) can be pre-charged to a suitable voltage required for an operation of the FeFETs at hand (e.g., program voltage V_{prog}, inhibition voltage $V_{inhibit}$, erase voltage V_{erase}, or the read reference voltage V_{bl}). During a read operation, any of FeFETs that are in the "on" state conduct current in the vertical or Z-direction between n^+ silicon layers **204**_d_ and **204**_b_.

Optional metal layers **204**_f_ and **204**_g_ reduce the resistance of n^+ silicon layers **204**_b_ and **204**_d_, respectively, thus facilitating fast memory device operations. N^+ silicon layer **204**_b_ in any active strip **202** may be left floating. One or more local word lines (referred to as a "pre-charge word line") may be used to bias n^+ silicon layer **204**_b_. When a suitable voltage is applied to selected pre-charge word lines (i.e., rendering the associated pre-charge FeFET conducting), each pre-charge word line momentarily inverts its channel region (i.e., channel polysilicon layer **250** in its associated gate stack), so that n^+ silicon layer **204**_b_ (i.e., the common source line) may be pre-charged to the pre-charge voltage V_{ss} in the associated n^+ silicon layer **204**_d_ from voltage source V_{bl} in the substrate. When the voltage on the pre-charge word line is withdrawn, (i.e., when the pre-charge FeFET is returned to its non-conducing state) and all the other word lines on both sides of the active strip are also "turned off", device operation may proceed with n^+ silicon layer **204**_d_ left electrically charged to provide a virtual voltage reference at the pre-charged voltage V_{ss} (typically, ~0V) because the distributed parasitic capacitor formed between n^+ silicon layer **204**_b_ and its multiple local word lines is sufficiently large to hold the virtual voltage reference long enough to support the program, program-inhibit or read operation (see below). Although the FeFETs in a NOR memory string may also serve as pre-charge FeFETs along each NOR memory string, to speed up the pre-charge phase for a read operation (read pre-charge requires a lower word line voltages of typically ~1.0 volts), some of the FeFETs (e.g., one in every 32 or 64 FeFETs along the NOR memory string) may also be activated. It is preferable that FeFETs dedicated entirely to serve as pre-charge FeFETs are provided, as they are more tolerant of program-disturb conditions than the FeFETs used in memory operations.

FIG. **2**_m_ shows a three-dimensional view of horizontal active layers **202-4** to **202-7** in the embodiment of FIG. **2**_k_, with local word lines or local pre-charge word lines connected to global word lines **208**_g-s_ and global word lines **208**_g-a_, and showing each active strip as having its N^+ layer **204**_d_ (acting as a drain region) connected through select circuits to any of voltage supplies (e.g., V_{ss}, V_{bl}, V_{pgm}, $V_{inhibit}$, and V_{erase}), decoding, sensing and other circuits arranged either adjacent or directly underneath the memory arrays. The substrate circuitry is represented schematically by **206-0** and **206-1** in substrate **201**.

In the exemplary fabrication processes described in detail above, channel polysilicon **250** is formed prior to forming polarization layer **264**. Higher quality channel regions for the FeFETs may be achieved in a process in which the channel region is formed after forming polarization layer **264**. In such a "channel-last" approach, the polarization layer and the channel layer may be optimized independently, which improves both endurance and disturb immunity. FIGS. **11**_a_ to **11**_r_ illustrate a process in which the channel region for an FeFET is formed after forming the polarization layer. The channel region in such a process is achieved by replacing at least in part a previously formed sacrificial layer with a channel material after the polarization layer is formed, according to one embodiment of the present invention.

FIG. **11**_a_ illustrates, for example, at the beginning of processing the array portion of the 3-dimensional array of NOR memory strings (e.g., as described above in conjunction with FIG. **2**_d_ above), isolation layer **1105** (e.g., a SiOC layer) and a hard mask layer (e.g., a carbon hard mask) are deposited over the active layers. The hard mask layer is then photo-lithographically patterned, so that its pattern may be transferred to allow etching a first group of trenches (i.e., trenches **1118**) in memory structure **200**. Trenches **1118** each extend through isolation layers **1105** and **203-7** to **203-0**, active layers **202-7** to **202-0** and etch-stop layer **241**. Trenches **1118** are each substantially rectilinear in an X-Y plane cross section, each trench being 70.0 nm wide along the X-direction, for example, with corresponding edges of adjacent trenches separated 190.0 nm from each other. In that embodiment, trenches **1118** are etched at an aspect ratio that is less than 50 (and, in some embodiments, even less than 30).

A series of etching steps then recess silicon nitride layers **204***a* and **204***e*, n^+-doped silicon layers **204***b* and **204***d* and sacrificial oxide layer **204***c* of each active layer **202** by, for example, 10.0 nm. Thereafter, sacrificial layer **1150** (e.g., amorphous silicon or polysilicon layer, or any other suitable sacrificial material) is conformally deposited to provide a "stand-in" layer for the channel regions to be subsequently formed. Sacrificial layer **1150** is then etched back (i.e., a separation etching step). As described above, n^+ silicon layers **204***b* and **204***d* in each active layer **202** would provide the common bit line and the common source line for the FeFETs in a NOR memory string to be formed. The hard mask and any excess sacrificial material of sacrificial layer **1105** on top of memory structure **200** are then removed (e.g., by CMP). Resulting memory structure **200** is shown in the X-Z plane cross section of FIG. **11***a*.

Thereafter, dielectric material **1123** is deposited to fill trenches **1118**. Any excess dielectric material is removed from the top surface of semiconductor structure **200**, which is then planarization (e.g., by CMP). FIG. **11***b* shows resulting memory structure **200** in an X-Z plane cross section.

Subsequently, one or more additional groups of trenches may be etched in the same manner as described above for trenches **1118**. Forming trenches in multiple groups allow each group of trenches to be etched at a favorable aspect ratio (e.g., less than 50). FIG. **11***c* shows resulting memory structure **200** in an X-Z plane cross section, after the second group of trenches are formed, according to one embodiment of the present invention. Both groups of trenches are shown in FIG. **11***c* and each indicated by reference numeral **1118**.

In this embodiment, through the unfilled ones of trenches **1118**, SiN layers **204***a* and **204***e* of each active layer **202** may be removed using a wet etch, for example. FIG. **11***d* shows resulting memory structure **200** in an X-Z plane cross section, after removal of SiN layers **204***a* and **204***e* from each of active layers **202-0** to **202-7**, according to one embodiment of the present invention.

Conductive material **1129** may then be deposited using, for example, an ALD technique. Conductor material **1129** may be, for example, a liner (e.g., titanium, titanium nitride (TiN), tantalum (Ta), tantalum nitride (TaN), or any combination thereof), followed by a refractory metal (e.g., tungsten (W), tungsten nitride, or Molybdenum (Mo)). Conductive material **1129** may be deposited into the cavities resulting from removing SiN layers **204***a* and **204***e* from each of active layers **202-0** to **202-7**. Thereafter, an etch-back step or an anisotropic etching step removes conductive material **1129** from the bottom of trenches **1118**, leaving a substantially conformal layer on their sidewalls. FIG. **11***e* shows resulting memory structure **200** in an X-Z plane cross section, after replacement of SiN layers **204***a* and **204***e* from each of active layers **202-0** to **202-7** by conductive material **1129**, according to one embodiment of the present invention.

Conductive material **1129** form optional conductor layers in contact with N^+ polysilicon layers **204***b* and **204***d* to reduce the resistivity in the common bit line and the common source line of a NOR memory string to be formed. Conductive material **1129** may be removed from the sidewalls of trenches **218** and to be further recessed by additional etching. A series of etching steps then recess n^+-doped silicon layers **204***b* and **204***d*, and sacrificial oxide layer **204***c* of each active layer **202** by, for example, 10 nm. FIG. **11***f* shows resulting memory structure **200** in an X-Z cross section, after recessing conductive material **1129**, n^+-doped silicon layers **204***b* and **204***d* and sacrificial oxide layer **204***c* of each active layer **202**, according to one embodiment of the present invention.

Thereafter, channel stand-in sacrificial layer **1150** may be conformally deposited on the sidewalls of trenches **1118**, in substantially the same manner as described above in conjunction with FIG. **11***a*. FIG. **11***g* shows resulting memory structure **200** in an X-Z cross section, after deposition of sacrificial material **1150** into trenches **1118**, according to one embodiment of the present invention.

Trenches **1118** may then be filled by dielectric material **1123** and planarized in the same manner as described above in conjunction with the steps for providing dielectric material **1123** of FIG. **11***b*. FIG. **11***h* shows resulting memory structure **200** in an X-Z cross section, after deposition of dielectric material **1123** into trenches **1118**, according to one embodiment of the present invention.

After silicon oxide **1123** is deposited into trenches **1118** and planarized, the next steps form the gate stacks (i.e., polarization layer **264** and the gate electrode (i.e., the "word line" or the "local word line") for each FeFET of the NOR memory strings. After planarization of dielectric material **1123**, hard mask **260** is provided over memory structure **200**, photolithographically patterned and developed. Hard mask **260** includes columns of oval openings **261**. (In this description, a "column" of objects denotes objects aligned along the Y-direction, whereas a "row" of objects denotes objects aligned along the X-direction.) FIG. **11***i*(*i*), oval openings **261** in each column are provide staggered relative to oval openings **261** in the immediately adjacent columns along the X-direction, such that the closest openings in adjacent columns have a greater separation between them than if such openings are aligned along the X-direction. In one embodiment, adjacent openings within each column are located at a 110.0 nm pitch along the Y-direction, while adjacent columns are also provided at 110.0 nm pitch along the X-direction. In this embodiment, the major and minor axes of each opening may be, for example, 100.0 nm and 60.0 nm along the X-direction and the Y-direction, respectively. A series of etchings through openings **261** excavates corresponding shafts **218**, removing dielectric material **1123** from trenches **1118** and portions of isolation layers **203-0** to **203-7**, and active layers **202-0** to **202-7**, reaching down to etch step layer **241**. FIGS. **11***i*(*i*) and **11***i*(*ii*) show, respectively, a top view and an X-Z plane cross section view (along line A-A' of FIG. **11***i*(*i*)) of resulting memory structure **200**, after shafts **218** are formed, according to one embodiment of the present invention. The aspect ratio of the etch steps creating shafts **218** have an aspect ratio that is still substantially within the desirable range (e.g., less than 50).

Next, silicon nitride liner **284** (e.g., 5 nm thick) is deposited conformally on the sidewalls of shafts **218**, which are then filled by sacrificial material **286**. Sacrificial material **286** may be, for example, a deposited amorphous silicon. Hard mask **260** is then removed and the surface of memory structure **200** is planarized (e.g., by CMP). FIGS. **11***j*(*i*) and **11***j*(*ii*) show a top view and an X-Z plane cross section view (along line A-A' of FIG. **11***j*(*i*)) of resulting memory structure **200**, after shafts **218** are filled by sacrificial material **286**, respectively, according to one embodiment of the present invention.

A wet etch then removes dielectric material **1123** from portions **1119** of trenches **1118** between filled shafts **218**. Portions **1119** are also referred to as shafts **1119**. FIGS. **11***k*(*i*) and **11***k*(*ii*) show, respectively, a top view and an X-Z plane cross section view (along line A-A' of FIG. **11***k*(*i*)) of resulting memory structure **200**, after dielectric material **1123** is removed from shafts **1119**, according to one embodiment of the present invention.

Thereafter, within each of shafts **1119**, gate stacks for FeFETs in the adjacent active strips are provided. As discussed above, each gate stack may be formed using any of the techniques illustrated above in conjunction with FIGS. **3***a*-**3***b*, FIGS. **4***a*-**4***c*, FIGS. **5***a*-**5***c*, **5***d*(*i*), **5***e*(*i*) and **5***f*(*i*), or FIGS. **5***a*-**5***c*, **5***d*(*ii*), **5***e*(*ii*) and **5***f*(*ii*). In some embodiments, as discussed above, each gate stack may include (a) polarization layer **264** having, for example, both an interface dielectric layer and the ferroelectric material (e.g., zirconium-doped hafnium oxide) and (b) a gate electrode formed out of conductive material **272** common to FeFET in the adjacent active strips. Conductive material **272** may include a metal liner (e.g., titanium (Ti), titanium nitride (TiN), tantalum (Ta) or tantalum nitride (TaN)) and a refractory metal (e.g., tungsten (W), tungsten nitride (WN) or molybdenum (Mo)). Formed in this manner, the gate electrode may be 60.0 nm×60.0 nm or less. A planarization step (e.g., CMP) may remove excess conductive material **272** from the top surface of memory structure **200**. FIGS. **11***l*(*i*) and **11***l*(*ii*) show a top view and an X-Z plane cross section view (along line A-A' of FIG. **11***l*(*i*)) of resulting memory structure **200**, respectively, after the gate stacks to FeFETs have been formed in each of shafts **1119**, according to one embodiment of the present invention.

Cap **272***t* may be provided to facilitate contact to the underlying local word line and to protect polarization layer **264** in subsequent processing steps. Cap **272***t* may be formed using additional masking, patterning, deposition (e.g., conductive material **272**), and planarization steps. FIGS. **11***m*(*i*) and **11***m*(*ii*) show a top view and an X-Z plane cross section view (along line A-A' of FIG. **11***m*(*i*)) of resulting memory structure **200**, respectively, after cap **272***t* has been provided in each of shafts **1119**, according to one embodiment of the present invention.

Sacrificial material **286** and silicon nitride liner layer are then removed from shafts **218** to provide access to sacrificial layer **1150**. FIGS. **11***n*(*i*) and **11***n*(*ii*) show a top view and an X-Z plane cross section view (along line A-A' of FIG. **11***n*(*i*)) of resulting memory structure **200**, respectively, after sacrificial material **286** and silicon nitride liner layer **284** have been removed from each of shafts **1119**, according to one embodiment of the present invention.

Thereafter, sacrificial material **1150** is removed by a wet etch, for example. FIGS. **11***o*(*i*) and **11***o*(*ii*) show a top view and an X-Z plane cross section view (along line A-A' of FIG. **11***o*(*i*)) of resulting memory structure **200**, respectively, after sacrificial material **1150** is removed from each of shafts **218**, according to one embodiment of the present invention.

P^--doped polysilicon is then deposited as channel polysilicon layer **250** into the cavities resulting from removing sacrificial material **1150** from underneath polarization layer **264**. At the same time, p^--doped polysilicon is also deposited up to 10.0 nm on the sidewalls of shafts **218**. FIGS. **11***p*(*i*) and **11***p*(*ii*) show a top view and an X-Z plane cross section view (along line A-A' of FIG. **11***p*(*i*)) of resulting memory structure **200**, respectively, after a wet etch recesses channel polysilicon layer **250** underneath polarization layer **264** in shafts **1119**, according to one embodiment of the present invention.

Channel polysilicon layer **250** may be further recessed to provide greater isolation. This etching step also likely removes p^--doped channel polysilicon layer **250** from the sidewalls of shafts **218**. FIGS. **11***q*(*i*) and **11***q*(*ii*) show a top view and an X-Z plane cross section view (along line A-A' of FIG. **11***q*(*i*)) of resulting memory structure **200**, respectively, after channel polysilicon layer **250** is deposited into the cavities resulting from removing sacrificial material **1150** from shafts **1119**, according to one embodiment of the present invention.

Thereafter, an ALD silicon oxide liner **285** (e.g., 10 nm) is deposited into the space under polarization layer **264** resulting from recessing channel polysilicon layer **250**. At the same time, silicon oxide liner **285** is also deposited on the sidewalls of shafts **218** to seal access to channel polysilicon **250** from shafts **218**. FIGS. **11***r*(*i*) and **11***r*(*ii*) show a top view and an X-Z plane cross section view (along line A-A' of FIG. **11***r*(*i*)) of resulting memory structure **200**, respectively, after channel polysilicon material **250** is deposited into the cavities resulting from removing sacrificial material **1150** from shafts **1119**, according to one embodiment of the present invention.

FIG. **6***c* shows memory structure **200**, fabricated under a channel-last fabrication process, in accordance with one embodiment of the present invention. As shown in FIG. **6***c*, channel polysilicon layer **250** is deposited conformally to the walls of the cavity resulting from removal of sacrificial layer **1150**. In the example of FIG. **6***c*, channel polysilicon layer **250** does not completely fill the cavity.

A silicon oxide may then be deposited into shafts **218** to provide further isolation between the FeFETs. Alternatively, shafts **218** may also be left unfilled, allowing the air gaps to serve as the isolation between thin-film storage transistors.

According to yet another embodiment of the present invention, the FeFETs of the present invention need not have a silicon channel region. Instead, the channel region may be formed using an 8.0 to 15.0 nm thick oxide semiconductor material. Suitable oxide semiconductor materials include (i) indium zinc oxide (InZnO, or "IZO") and (ii) indium gallium zinc oxide (InGaZnO, or "IGZO"). An IZO or IGZO channel region has the advantage of a high mobility for greater switching performance and without concern for electron or hole tunneling. For example, a 10.0 nm thick IZO film has an electron mobility of 40.6 cm^2/V, relative to an electron mobility of 5.6 cm^2/V in an aluminum zirconium oxide (AZO) film of comparable thickness, and an electron mobility of 1.0-4.0 cm^2/V in a polysilicon film with the same comparable thickness. Similarly, an IGZO film has an electron mobility of 10.0-100.0 cm^2/V, depending on the relative compositions of indium, gallium and zinc.

IZO and IGZO each can provide a junction-less transistor channel, and each may be deposited at a temperature lower than 400° C., even as low as 250° C. Thus, they may be used to provide the channel regions of storage transistors in a 3-dimensional memory array, regardless of whether such storage transistors are FeFETs (i.e., transistors with a polarizable layer) or a quasi-volatile (QV) storage transistors, such as those described in the Related Application. Modifying conventional storage transistors to use IZO or IGZO as channel material imposes only a relatively modest burden on the thermal budget. This is particularly advantageous in applications where planar support logic and analog circuitry is provided at the surface of the substrate beneath the 3-dimensional array of memory strings.

Furthermore, in a junction-less transistor, the source region and the drain region (i.e., bit line) are redundant, and may be replaced by a metal layer (e.g., Mo), thereby dispensing with the need for the n^+ polysilicon layers to be strapped to metal strapping layers (e.g., metal layers **204***f* and **204***g*) in each active strip. These metal layers are referred to as the "metal common bit line" and the "metal source line." The gate stack may be substantially the same as that described above with respect to the FeFETs with the silicon channel (e.g., an silicon nitride interface dielectric layer, a polarization layer, and the gate electrode). As this alternative FeFET does not have a p/n junction, any leakage current from an FeFET in the programmed state is relatively small. Consequently, such an FeFET is particularly advantageous for high-temperature applications. Such an FeFET may also have a relatively shorter channel length, as there is no need for a margin in the channel length to allow for undesirable dopant diffusion during thermal cycles from the heavily-doped semiconductor common bit line and the common source line. The metal common bit line and metal common source line may each also have a reduced thickness. For example, in one embodiment, a 30.0-40.0 nm thick metal common bit line and a 30.0-40.0 nm thick metal common source line, a 30.0-40.0 nm thick IZO channel layer, and a 30.0 nm thick isolation layer (e.g., SiOC), for a relatively thinner total of 120.0-150.0 nm. The common source line and the common bit line may be formed by replacing a sacrificial material (e.g., carbon) by metal in a metal-replacement step carried out in a later portion of the fabrication process. These thinner layers reduce the aspect ratio for anisotropic etching of trenches through multiple layers of memory strings. Furthermore, the thinner layer of each string results in reduced parasitic capacitance C between local word-lines and the side walls of each string, which reduces the RC delays as well the power dissipation for the 3-dimensional memory array.

Programming, Erasing and Read

An FeFET may be polarized to either (i) a conducting or "erased" state, or (ii) a non-conducting or "programmed" state. In an FeFET, its threshold voltage in the erased state is lower than its threshold voltage in the programmed state. FIG. **7***a* shows hysteresis in the drain current (I_d) in response to an applied gate voltage (V_g) in a conventional FeFET. (The conventional FeFET is formed at the planar surface of a monocrystalline semiconductor substrate and is not formed as a thin-film field effect transistor.) In FIG. **7***a*, waveform **1401** traces the drain current of the FeFET in its erased state, as the gate voltage increases from less than −1.0 volts to greater than 1.0 volts, and waveform **1402** traces the drain current of the FeFET in its programmed state, as the gate voltage decreases from greater than 1.0 volts to less than −1.0 volts. As seen in FIG. **7***a*, the conventional FeFET has a negative threshold voltage (V_t).

It is desirable in some applications, however, for an FeFET (e.g., a thin-film FeFET in a NOR memory string) to have a positive threshold voltage (V_t), such as around 0.5 volts, to prevent undesirable leakage current when it is subject to a disturb condition (e.g., as experienced by a neighboring non-selected FeFET in the NOR memory string) during a read operation.

FIG. **7***b* shows a desirable hysteresis in the drain current (I_d) response to an applied gate voltage (V_g) in a thin-film FeFET in a NOR memory array, according to one embodiment of the present invention. In FIG. **7***b*, waveform **1403** traces drain current I_d of the FeFET in its low-threshold voltage polarization state ("erased state", in this embodiment), as the gate voltage increases from less than −1.0 volts to greater than 1.0 volts, and waveform **1404** traces the drain current of the FeFET in its high-threshold voltage polarization state ("programmed" state, in this embodiment), as the gate voltage decreases from greater than 1.0 volts to less than −1.0 volts. As seen in FIG. **7***b*, the FeFET has a positive threshold voltage (V_t) of about 0.5 volts, and 1.0-2.5 volts threshold voltage difference ("window") between the polarization states. With a p^- polysilicon channel region (e.g., boron-doped), this window may be achieved by (i) increasing the boron dopant concentration in the channel region, (ii) providing a gate electrode that is formed out of a conductive material that has a high work function (e.g., tungsten (W), molybdenum (Mo), aluminum (Al), ruthenium (Ru), tantalum (Ta), or titanium (Ti), (iii) appropriate biasing in the common source region (see below), or (iv) the combination of (i), (ii) and (iii). In an embodiment in which the FeFET channel is an oxide semiconductor material (e.g IGZO), the FeFET's low-threshold voltage state can be permanently shifted to a positive voltage by reducing the IGZO thickness or by a thermal anneal at a specific temperature.

In an array of NOR memory strings, its FeFETs share a common bit line and a common source line. FeFETs in different NOR memory strings may share a common word line. For example, FeFETs at corresponding positions in NOR memory strings in the same active stack, or in NOR memory strings of adjacent active strips, may share a common word line. FIG. **7***c* shows a circuit schematic diagram of FeFETs in NOR memory strings **1420-1** to **1420-4**, according to one embodiment of the present invention. In FIG. **7***c*, reference numerals **1411**, **1412** and **1413** indicate the FeFETs that share common bit line BL_j and common source line SL_j. Likewise, reference numerals **1412**, **1413** and **1414** indicate FeFETs that share common word line WL_k. Even though not selected for a program, erase or read operation, an FeFET may experience a "disturb" condition affecting the stability of the threshold voltage state it is set. The disturb condition arises because the voltages on the FeFET's word line, source line or bit line may be subject to interference by signal changes in its vicinity as to affect a change in its polarization state during the program, erase or read operation. The signal changes are typically associated with the voltages imposed on the word line or bit line the FeFET shares with the selected FeFET.

To facilitate reference, in this detailed description, the term "affected FeFET" refers to a non-selected FeFET that shares a bit line or a word line with a selected FeFET of an intended program, erase or read operation. Any of the FeFETs indicated by reference numerals **1411**, **1413**, **1414** and **1415** is an affected FeFET. Likewise, the term "once-removed FeFET" refers to an FeFET that shares a word line or a bit line with an affected FeFET. All the unlabeled FeFETs shown in FIG. **7***b* are once-removed FeFETs.

Without mitigation, a disturb condition may cause an affected FeFET to move away from its polarization state and thereby loses the data value it is holding. To mitigate a disturb condition, the non-selected word line or the non-selected bit line of an affected FeFET may be biased to a predetermined voltage ("inhibit voltage") during the program, erase or read operation to counteract the effect of the voltage imposed on the word line or bit line it shares with the selected FeFET.

Table 1 provides exemplary bias voltages for the gate electrodes (or word lines), the source regions (or the common source lines), the drain regions (or the common bit lines) for the selected and affected FeFETs in a 3-dimensional array of NOR memory strings during erase, program and read operations, under a first disturb mitigation scheme ("V/2 scheme"; in which an affected FeFET is subject to one-half the "stress voltage" of a selected FeFET, during a program or erase operation). The stress voltage refers to the difference in voltage between the gate electrode and the bit line in an FeFET, during a program, erase or read operation. The bias voltages of Table 1 are based on the case when (i) an FeFET holding logic value '1' (i.e., in the low-threshold voltage polarization state, with a threshold voltage of approximately −0.25 volts), and (ii) an FeFET holding logic value '0' (i.e., in the high-threshold voltage polarization state, with a threshold voltage of approximately 0.6 volts).

TABLE 1

Operation	Selected FeFET			Non-selected FeFETs that Share a Bit Line or a Gate Electrode with the Selected FeFET		
	Gate/Word Line	Source	Drain/Bit Line	Word Line	Source Line	Drain/Bit Line
Create Low-threshold Voltage Polarization State (‘1’)	0.0 volts	3.0 volts	3.0 volts	1.5 volts	1.5 volts	1.5 volts
Create High-threshold Voltage Polarization State (‘0’)	3.0 volts	0.0 volts	0.0 volts	1.5 volts	1.5 volts	1.5 volts
Read	1.0 volts	0.75 volts	1.25 volts	0.0 volts	0.0	0.0

Under the V/2 scheme, during a program or erase operation, an affected FeFET (e.g., in FIG. 7*c*, any of the FeFETs indicated by reference numerals **1411**, **1413**, **1414** and **1415**) experiences a stress voltage of −1.5 volts (write ‘1’) or 1.5 volts (write ‘0’). The once-removed FeFETs each experience no stress voltage (i.e., a stress voltage of 0.0 volts).

In this example, as the threshold voltage of an FeFET in the ‘0’ state is negative, during a read operation, the source voltage is set to 0.75 volts, or any suitable positive voltage sufficient to prevent conduction prior to the gate electrode achieving the read voltage. The higher bias voltage (the “boost”) avoids both having to generate a negative voltage and having to provide corresponding additional isolation structures to protect the devices in the silicon substrate. During a read operation, to reduce read disturb under the V/2 scheme, the voltage difference between the word line and the common bit line or the common source line in the selected FeFET is maintained at less than the voltage that may change the polarization state of the selected FeFET.

Table 2 provides, during erase, program and read operations under a second disturb mitigation scheme (“V/3 scheme”), exemplary bias voltages for the gate electrodes (or word lines), the source regions (or the common source lines), the drain regions (or the common bit lines) of selected and affected FeFETs in a 3-dimensional array of NOR memory strings. Under the V/3 scheme, an affected FeFET is subject to one-third the stress voltage of a selected FeFET during a program or erase operation. The bias voltages of Table 2 are also based on the case when (i) an FeFET holding logic value ‘0’ (i.e., in the “programmed” state) has a threshold voltage of approximately 0.6 volts, and (ii) an FeFET holding logic value ‘1’ (i.e., in the “erased” state) has a threshold voltage of approximately −0.25 volts.

TABLE 2

Operation	Selected FeFET			Non-selected FeFETs that Share a Bit Line or a Gate Electrode with the Selected FeFET		
	Gate/Word Line	Source	Drain/Bit Line	Word Line	Source Line	Drain/Bit Line
Create Low-threshold Voltage Polarization State (‘1’)	0.0 volts	3.0 volts	3.0 volts	2.0 volts	1.0 volts	1.0 volts
Create High-threshold Voltage Polarization State (‘0’)	3.0 volts	0.0 volts	0.0 volts	1.0 volts	2.0 volts	2.0 volts
Read	1.0 volts	0.75 volts	1.25 volts	0.0 volts	0.0 volts	0.0 volts

Under the V/3 scheme, during a program or erase operation, an affected FeFET experiences a stress voltage of 1.0 volts (write ‘0’) or −1.0 volts (write ‘1’). A once-removed FeFETs experiences a stress voltage of −1.0 volts (write ‘0’) or 1.0 volts (write ‘1’). Thus, the V/3 scheme provides a lesser stress voltage in the affected FeFETs. However, because a voltage difference of 1.0 volt (write ‘1’) or −1.0 volt (write ‘0’) exists between the non-selected bit lines and the non-selected word lines, which accounts for the charge separation in the parasitic capacitors between the non-selected word lines and the non-selected bit lines, the V/3 scheme has a greater power dissipation than the V/2 scheme during a program, erase or read operation.

Under either the V/3 scheme or the V/2 scheme, the stress voltage for an affected FeFET sharing a word line with the selected FeFET is 1.0 volts during a read operation, which is equal to the lower stress voltage under the V/3 scheme for a program or erase operation. As a program or erase operation is expected to have a significantly longer duration (e.g., 100.0 ns) than a read operation (e.g., 10.0 ns), limiting the overall effects of the disturb conditions may be achieved by adequately addressing the disturb conditions during the program and erase operations.

Even though the data retention time of an FeFET of the present invention is expected to be significantly longer (e.g., tens of days to tens of years) than the data retention time of a quasi-volatile (QV) storage transistor (e.g., any of those transistors described in the Related Application), the methods for maintaining data integrity beyond the data retention time for QV storage transistors are also applicable to maintaining data retention in the FeFETs of the NOR memory strings of the present invention. For example, Provisional Application VII discloses one or more methods for reducing the effects of disturb conditions by refreshing a subset of the storage transistors that share a word line with a selected storage transistor during a program or erase operation (“refresh in-place”). Refresh in-place takes advantage of the activated selected word line in the program or erase operation to write back the data values in the storage transistors in the affected subset.

In some mobile applications, the device's battery life is shorter than the data retention time, periodic refresh may be triggered when the battery is detected to be connected to a charger. In this manner, one can be assured that the data in the memory arrays in a fully charged device will persist until the next charge cycle. In addition, refresh operation is initiated when the time since the last refresh is within a threshold (e.g., 10% of the data retention time).

When the body region of an FeFET in a NOR memory string is floating, its programming speed may be slower than its erase speed. In such a condition, the gate-induced drain leakage (GIDL) effect may be harnessed to improve the programming speed. The GIDL effect may be activated by creating a voltage difference (e.g., 0.5 to 2.0 volts) between the common bit line and the common source line (V_{ds}) during programming, for example. This voltage difference may be achieved by first momentarily pre-charging the common source line to a predetermined source line voltage through the common bit line, and then setting the common bit line to its target voltage, as disclosed in the Related Application.

FIG. **8** illustrates the methods and circuit elements used for setting source voltage V_{ss} in n^+ silicon layer **204***b*. Specifically, source voltage V_{ss} may be set through hard-wired decoded source line connections **280** (shown in dashed line) or alternatively, by activating pre-charge thin film transistors **300** (which, preferably, is not implemented by a FeFET) and decoded bit line connections **270** to any one of bit line voltages V_{ss}, V_{bl}, V_{pgm}, $V_{inhibit}$ and V_{erase}. Alternatively, source reference voltage V_{ss} may be accessed through a metal or n^+-doped polysilicon conductor connecting from the top of the memory array through staircase vias. Each of the conductors in hard-wired connections **280** may be independently connected, so that the source voltages for different planes or within planes need not be the same. In some embodiments, it is advantageous to dispense with the hard-wired source voltage V_{ss} connections, by taking advantage of a virtual voltage source in the intrinsic parasitic capacitance of the NOR memory string.

The present invention takes advantage of the cumulative intrinsic parasitic capacitance that is distributed along each NOR memory string to increase the number of FeFETs that can be programmed, read or erased in parallel in a single operation (though, preferably, on separate strings), while also significantly reducing the operating power dissipation. As shown in FIG. **8**, local parasitic capacitor **360** (contributing to a cumulative capacitance C) exists at each overlap between a local word line (as one plate) and the $n^+/p^-/n^+$ active layer (as the other plate). Each local parasitic capacitor may be approximately 0.005 femtofarads (each femtofarad is 1×10^{-15} farad), too small to be of much use for temporary storage of charge. However, since there may be a thousand or more FeFETs contributing capacitance from one or both sides of an active strip, the total distributed capacitance C of n^+ silicon layer **204***b* (the common source line) and n^+ silicon layer **204***d* (the common bit line) in a long NOR memory string can be in the range of ~10.0 to 100.0 femtofarads. This is also roughly the capacitance at sensing circuitry connected through connections **270** (e.g., voltage source V_{bl}.

During a program or erase operation in a selected FeFET, and their respective inhibit operations on a non-selected FeFETs, both n^+ silicon layers **204***d* and **204***b* are preferably held at the same voltage, therefore the leakage current between n^+ silicon layers **204***d* and **204***b* is insignificant. During a program or erase operation, charge leakage from cumulative capacitance C flows primarily to the substrate through the substrate selection circuitry, which has very little transistor leakage, as it is formed in single crystal or epitaxial silicon. Nevertheless, even a 100-microsecond polarization state retention time is sufficient to complete the sub-100 nanosecond read operation or the sub-100 microsecond program operation (see below) of the selected FeFET on the NOR memory string.

An FeFET in a NOR memory string, unlike a DRAM cell, may operate as a non-volatile memory transistor, so that, even if parasitic capacitor C of the NOR memory string is completely discharged, the information stored in the selected FeFET remains intact in the ferroelectric storage material (i.e., polarization layer **264**). In a DRAM cell, however, the information would be lost without frequent refreshes. Accordingly, distributed capacitance C of a NOR memory string of the present invention is used to temporarily hold the pre-charge voltage on n^+ silicon layers **204***d* and **204***b* at one of voltages V_{ss}, V_{bl}, V_{progr}, $V_{inhibit}$, or V_{erase}, and is not used to store actual data for any of the FeFETs in the NOR memory string. Pre-charge transistor **303** of FIG. **3***a*, controlled by word line **151***n* (i.e., word line **208**-CHG), is activated momentarily immediately preceding each read, program, or erase operation to transfer voltage V_{bl} (e.g., through connections **270**) from the substrate circuitry (not shown) to n^+ silicon layer **204***b*. For example, under the V/3 scheme described above, voltage V_{bl} can be set at 0.75 V to pre-charge n^+ silicon layer **204***b* to a virtual ground voltage ~0.75 V during a read operation, or to pre-charge both n^+ silicon layers **204***d* and **204***b* to either 1.0 Volts and 2.0 volts in non-selected bit lines involved in refresh in-place during a program or erase operation.

The value of cumulative capacitance C may be increased by lengthening the NOR string to accommodate the thousands more FeFETs along each side of the active strip, correspondingly increasing the retention time of pre-charge voltage V_{ss} on n^+ silicon layer **204***b*. However, a longer NOR memory string suffers from an increased line resistance as well as higher leakage currents between n^+ silicon layer **204***d* and n^+ silicon layer **204***b*. Such leakage currents may interfere with the sensed current when reading the one FeFET being addressed with all other FeFETs of the NOR memory string in their "off" (and somewhat leaky) states. Also, the potentially longer time it takes to pre-charge a larger capacitor during a read operation can conflict with the desirability for a low read latency (i.e., a fast read access time). To speed up the pre-charging of the cumulative capacitance C of a long NOR memory string, pre-charge transistors—which may be constructed separately from the memory FeFETs—may be provided spaced apart along either side of the active strip (e.g., once every 128, 256 or more FeFETs).

Because the variable-threshold FeFETs in a long NOR memory string are connected in parallel, the read operating condition for the NOR memory string should preferably ensure that all FeFETs along both edges of an active strip operate in enhancement mode (i.e., they each have a positive threshold voltage, as applied between control gate **151***n* and voltage V_{ss} at source **204***d*). With all FeFETs being in enhancement mode, the leakage current between n^+ silicon layer **204***d* and n^+ silicon layer **204***b* of the active strip is suppressed when all control gates on both sides of the active strip are held at, or below V_{ss}~0V. This enhancement threshold voltage can be achieved by providing polysilicon channel layer **250** with a suitable dopant concentration (e.g., a boron concentration between 1×10^{16} and 1×10^{18} per cm^3 or higher, which results in an intrinsic FeFET threshold voltage of between ~0.5 V and ~1 V). Alternatively, to substantially shut off a source-drain leakage current in any of the FeFETs along a NOR memory string that is programmed into a depletion threshold voltage state, the common source region and the common drain region may be boosted into a "virtual enhancement" threshold voltage state by pre-charging the common source line and the common bit line of the NOR memory string to voltages V_{ss} and V_{bl}, respectively. This biasing condition substantially shuts off source-drain conduction currents in the FeFETs of the same NOR memory string. The "virtual enhancement" mode of operation is advantageous because it allows the same voltage polarity to be applied to source, drain and gate terminals for all program, erase and read operations of the FeFETs in the NOR memory string. Furthermore, implementation of the "virtual enhancement" mode of operation provides the flexibility of not having to rely on a factory-based operation to apply a fixed, potentially sub-optimal dopant concentration to the polysilicon channel of the FeFET.

Quasi-Volatile NOR Memory Strings

Endurance is a measure of a storage transistor's performance degradation after some number of write-erase cycles. Endurance of less than around 10^{12} cycles is considered too low for some applications. Greater endurance in an FeFET of the present invention may be achieved by, for example, increasing the dielectric constant k in the interface dielectric layer, typically a silicon oxide-nitride or silicon nitride film of thickness 0.1 to 3.0 nm. The interface dielectric layer may also be kept thin (i.e., even less than 0.1 nm), or eliminated altogether or replaced with another dielectric film of a higher dielectric constant k significantly greater than 4 (e.g., aluminum oxide).

A FeFET having a low data retention time (i.e., less than, for example, one year), may be referred to as "quasi-volatile FeFET". Such an FeFET may require refreshes to restore the weakened polarization. Because the quasi-volatile FeFET of the present invention provides a DRAM-like fast read access time with a low latency, the resulting quasi-volatile NOR memory strings may be suitable for use in some applications that currently require DRAMs. The advantages of quasi-volatile FeFETs in NOR memory strings, arranged in a planar structure or in 3-dimensional memory arrays, over DRAMs include: (i) a much lower cost-per-bit figure of merit because DRAMs cannot readily be built in 3-dimensional blocks, and (ii) a much lower power dissipation, as the refresh cycles need only be run approximately once every tens of minutes or once every few days, as compared to every ~64 milliseconds, as required in state-of-the-art DRAM technology. 3-dimensional memory arrays of FeFETs may also have advantages over the 3-dimensional NOR memory arrays of the Related Application ("3-D QV NOR memory strings"). For example, (i) FeFETs can be fabricated at significantly lower thermal cycles (e.g., at 400.0-500.0° C.) than the thermal cycle for fabricating direct tunneling, charge-trapping thin-film transistors (e.g. at 600.0-860.0° C.), and (ii) the erase and programming polarization states of some FeFETs can be switched at significantly lower voltages (e.g., less than 4.0 volts), as compared to the 6.0-8.0 volts required to inject electrons (i.e., to program) or eject trapped electrons (i.e., to erase) in 3-dimensional QV NOR memory strings.

The NOR memory string of FeFETs in the present invention appropriately adopt the program/read/erase conditions to incorporate the periodic data refreshes. For example, because each NOR memory string of quasi-volatile FeFETs is frequently read-refreshed or program-refreshed, it is not necessary to "hard-program" such FeFETs to open a large threshold voltage window between the '0' and '1' states, as compared to non-volatile FeFETs where a minimum 10-year data retention time may be required. Quasi-non-volatile threshold voltage window may be as narrow as 0.2 volts to 1.5 volts. The reduced threshold voltage window allows such FeFETs to be programmed and erased at lower programming/erase voltages and by shorter-duration programming pulses, which reduce the stress voltage, thereby extending endurance. Periodic refresh operations can be performed in a background mode so as to minimize collision with erase/program/read operations. During such a refresh operation, it is possible for the system to check the health of each FeFET, to restore its full window of operation or altogether retire the FeFET by replacing it with a redundant fresh FeFET.

Source-Drain Leakage in Long NOR Memory Strings

In long NOR memory strings, the current of the one accessed FeFET in a read operation competes with the cumulative subthreshold leakage currents from the thousand or more parallel non-selected FeFETs. Similarly, precharged strip capacitor C has to contend with charge leakage not just of one transistor (as in a DRAM circuit) but the charge leakage through the thousand or more transistors in the NOR memory string. That charge leakage reduces substantially the charge retention time on C to perhaps a few hundred microseconds, potentially requiring counter measures to reduce or neutralize such leakage. However, the leakage from a thousand or so transistors is only significant during read operations. During program, program-inhibit or erase operations, source layer **204***b* and bit line layer **204***d* are preferably held at the same voltage, therefore transistor leakage between the two layers is insignificant (the leakage of charge from capacitor C during program, program-inhibit or erase operations is primarily leaked to the substrate through the substrate selection circuitry, which is formed in single-crystal or epitaxial silicon where transistor leakage is very small). For a read operation, even a relatively short 100-microsecond retention time of charge on the source and bit line capacitors may be sufficient time to complete the sub-100 nanosecond read operation (see below) of the FeFETs of the present invention. One difference between a FeFET in a NOR memory string of the present invention and a DRAM cell is that, relative to the latter, the former is a non-volatile memory transistor for DRAM suitable applications, so that even if parasitic capacitor C is completely discharged, the information stored in the selected FeFET is not lost from the ferroelectric or polarizable element, unlike a DRAM cell where the stored information is lost unless refreshed. Capacitor C is used solely to temporarily hold the pre-charge voltage on n^+ silicon layers **204***d* and **204***b* at one of voltages V_{ss}, V_{bl}, V_{progr}, $V_{inhibit}$, or V_{erase}; capacitor C is not used to store actual data for any of the non-volatile FeFETs in the string. Pre-charge transistor **303**, controlled by word line **151***n* (**208**-CHG) (FIG. **8**) is activated momentarily immediately preceding read, program, program-inhibit or erase operations to transfer through connections **270** the voltage V_{bl} from the substrate circuitry (not shown) to capacitor C of n^+ silicon layer **204***d*. For example, voltage V_{bl} can be set at ~0V to pre-charge n^+ silicon layer **204***b* to a virtual ground voltage during read, or to pre-charge both n^+ silicon layers **204***d* and **204***b* to their programming or erase bias voltages during program inhibit. The value of cumulative capacitors C may be increased by lengthening the active string to accommodate thousands more FeFETs along each side of the string, correspondingly increasing the retention time of pre-charge voltage V_{ss} on n^+ silicon layer **204***d*. To speed up the pre-charging of the capacitance C of a long active strip, several pre-charge thin-film transistor

303 may be provided spaced apart along either side of the active strip (e.g., once every 128, 256 or more FeFETs).

Revolving Reference NOR Memory String Address Locations to Extend Cycle Endurance.

In applications requiring a large number of write/erase operations, the threshold-voltage window of operation for the FeFETs in the NOR memory strings may drift with cycling, away from the threshold-voltage window that is programmed into the FeFETs of the reference NOR memory strings at the device's beginning of life. The growing discrepancy between FeFETs on the reference NOR memory strings and FeFETs on the addressed memory NOR memory strings over time, if left unattended, can be detrimental to data integrity. To overcome this drift, reference NOR memory strings in a block need not always be mapped to the same physical address and need not be permanently programmed for the entire life of the device. Since the programmable reference NOR memory strings are practically identical to the memory NOR memory strings sharing the same plane in a block, reference NOR memory strings need not be dedicated for that purpose in any memory array block. In fact, any one of the NOR memory strings can be set aside as a programmable reference NOR memory string. In fact, the physical address locations of the programmable reference NOR memory strings can be rotated periodically (e.g., changed once every 100 times the block is erased) among the sea of memory NOR memory strings, so as to level out the performance degradation of memory NOR memory strings and reference NOR memory strings as a result of extensive program/erase cycles. For example, the reference NOR memory strings may be rotated among the 8 planes (e.g., the reference NOR memory strings may be mapped in a round-robin fashion to one of the 8 planes of an active stack, kept track by a 3-bit counter).

According to the current invention, any NOR memory string can be rotated periodically to be designated as a programmable reference NOR memory string, and its address location may be stored inside or outside the addressed block. The stored address may be retrieved by the system controller when reading the NOR memory string. Under this scheme, rotation of reference NOR memory strings can be done either randomly (e.g., using a random number generator to designate new addresses), or systematically among any of the active memory NOR memory strings. Programming of newly designated reference NOR memory strings can be done as part of the erase sequence when all FeFETs on a slice or a block may be erased together, to be followed by setting anew the reference voltages on the newly designated set of reference NOR memory strings. In this manner, all active memory NOR memory strings and all reference NOR memory strings in a block drift statistically more or less in tandem through extensive cycling.

Programmable Reference Slices.

In some embodiments of the present invention, a block may be partitioned into four equal-size quadrants, as illustrated in FIG. **9***a*. FIG. **9***a* show semiconductor structure **600**, which is a three-dimensional representation of a NOR memory string array of FeFETs organized into quadrants **Q1**-**Q4**. In each quadrant, (i) numerous NOR memory strings are each formed in active strips extending along the Y-direction (e.g., NOR memory string **112**), (ii) pages extending along the X-direction (e.g., page **113**), each page consisting of one FeFET from each NOR memory string at a corresponding Y-position, the NOR memory strings in the page being of the same corresponding Z-position (i.e., of the same active layer); (iii) slices extending in both the X- and Z-directions (e.g., slice **114**), with each slice consisting of the pages of the same corresponding Y-position, one page from each of the planes, and (iv) planes extending along both the X- and Y-directions (e.g., plane **110**), each plane consisting of all pages at a given Z-position (i.e., of the same active layer).

FIG. **9***b* shows structure **600** of FIG. **9***a*, showing FeFETs in programmable reference NOR memory string **112**-Ref in quadrant Q**4** and FeFETs in NOR memory string **112** in quadrant Q**2** coupled to sense amplifiers SA(a), Q**2** and Q**4** being "mirror image quadrants." FIG. **9***b* also shows (i) programmable reference slice **114**-Ref (indicated by area B) in quadrant Q**3** similarly providing corresponding reference FeFETs for slice **114** in mirror image quadrant Q**1**, sharing sense amplifiers SA(b), and (ii) programmable reference plane **110**-Ref in quadrant Q**2** providing corresponding reference FeFETs to plane **110** in mirror image quadrant Q**1**, sharing sense amplifiers SA(c), and also providing corresponding reference FeFETs for NOR memory strings in the same quadrant (e.g., NOR memory string **112**).

As shown in FIG. **9***b*, programmable reference NOR memory strings **112**Ref of FeFETs may be provided in each quadrant to provide reference voltages for the NOR memory strings of FeFETs on the same plane in the same quadrant, in the manner already discussed above. Alternatively, programmable reference slices (e.g., reference slice **114**Ref) are provided on mirror-image quadrants for corresponding memory slices. For example, when reading a memory slice in quadrant Q**1**, programmed reference slice **114**Ref (area B) in quadrant Q**3** is simultaneously presented to sense amplifiers **206** that are shared between quadrants Q**1** and Q**3**. Similarly, when reading a memory slice in quadrant Q**3**, reference slice **114**Ref (area A) of quadrant Q**1** is presented to the shared sense amplifiers **206**. There can be more than one reference slice distributed along the length of NOR memory strings **112** to partially accommodate mismatched in RC delay between the slice being read and its reference slice. Alternatively, the system controller can calculate and apply a time delay between the global word line of the addressed slice and that of the reference slice, based on their respective physical locations along their respective NOR memory strings. Where the number of planes is large (e.g., 8 or more planes), one or more planes can be added at the top of the block to serve either as a redundant plane (i.e., to substitute for any defective plane) in the quadrant, or as programmable reference pages, providing reference threshold voltages for the addressed pages sharing the same global word line conductor **208***g*-*a*. The sense amplifier at the end of each NOR memory string receives the read signal from the addressed page at the same time as it receives the signal from the reference page at the top of the block, since both pages are activated by the same global word line.

In one embodiment, each memory block consists of two halves, e.g., quadrants Q**1** and Q**2** constitute an "upper half" and quadrants Q**3** and Q**4** constitute a "lower half." In this example, each quadrant has 16 planes, 4096 (4K) NOR memory strings in each plane, and 1024 (1K) FeFETs in each NOR memory string. It is customary to use the unit "K" which is 1024. Adjacent quadrants Q**1** and Q**2** share 1K global word lines (e.g., global word line **208***g*-*a*) driving 2048 (2K) local word lines **208**W per quadrant (i.e., one local word line for each pair of FeFETs from two adjacent NOR memory strings). 4K FeFETs from quadrant Q**1** and 4K FeFETs from quadrant Q**2** form an 8K-bit page of FeFETs. 16 pages form a 128K-bit slice, and 1K slices are provided in a half-block, thus providing 256 Mbits of total storage per block. (Here, 1 Mbits is 1K×1 Kbits.) The 4K strings in each plane of quadrants Q**2** and Q**4** share substrate circuitry **206**, including voltage sources for voltage V_{bl} and sense amplifiers (SA). Also included in each quadrant are redundant NOR memory strings that are used as spares to replace faulty NOR memory strings, as well to store quadrant parameters such as program/erase cycle count, quadrant defect map and quadrant ECC. Such system data are accessible to a system controller. For blocks with high plane counts, it may be desirable to add one or more planes to each block as spares for replacing a defective plane.

Programmable Reference Planes, Spare Planes

High-capacity storage systems based on 3-dimensional arrays of FeFETs organized as NOR memory strings of the present invention require a high-speed system controller for error-free massively parallel erase and erase-inhibit, program and program-inhibit, and read operations that may span many "chips" including many memory blocks. In fact, in some applications, a hierarchy of controllers (e.g., organized as a tree structure) may be provided, with the controllers at the leaf level each managing the operations of multiple chips, each containing a large number of tiles. Controllers at a higher level in the hierarchy (i.e., closer to the root or roots) control controllers closer to the leaves. At the chip level, each tile is provided logic functions implemented in the memory circuits dedicated to assist the controllers in managing operations carried out by the tiles of the chip. Also, each memory circuit stores system parameters and information related to the files stored in the memory circuit. Such system information is typically accessible to the system controller, but not accessible by the user. It is advantageous for the system controller to quickly read the memory circuit-related information. A block under this organization (i.e., 256 Megabits) provides 2K slices. A terabit memory circuit may be provided by including 4K blocks.

As shown in FIGS. **9***a* and **9***b*, the FeFETs in quadrants Q**2** and Q**4** share voltage source V_{bl}, sense amplifiers SA, data registers, XOR gates and input/output (I/O) terminals to and from substrate circuitry **206**. According to one organization, FIG. **9***a* shows NOR memory strings **112**, quarter-planes **110**, half-slices **114**, and half-pages **113**. Also shown are optional pillars **290** supplying back-bias voltage V_{bb} from the substrate. FIG. **9***b* shows examples of locations of reference strings **112**(Ref), reference slices **114**(Ref) and reference planes **110** (Ref). In the case of reference strings, reference string **112** (Ref) of quadrant Q**4** can serve as a reference string to NOR memory string **112** on the same plane in quadrant Q**2**, the two NOR memory strings being presented to a shared differential sense amplifier SA in circuitry **206**. Similarly, reference slice **114** Ref (area A) in quadrant Q**1** can serve as reference for a slice in quadrant Q**3**, while a reference slice B in quadrant Q**1** can serve as reference for slices in quadrant Q**3**, again sharing differential sense amplifiers SA provided between quadrants Q**1** and Q**3**. Global word lines **208***g-a* are connected to local word lines **208**W and local pre-charge word lines **208**-CHG. Substrate circuitry and input/output channels **206** are shared between FeFETs in quadrants Q**2** and Q**4**. Under this arrangement, their physical locations allow cutting by half the resistance and capacitance of NOR memory strings **112**. Similarly, global word line drivers **262** are shared between quadrants Q**1** and Q**2** to cut by half the resistance and capacitance of the global word lines, and pillars **290** (optional) connect p^- layers of NOR memory strings **112** to the substrate voltage.

Since silicon real estate on an integrated circuit is a valuable resource, rather than adding reference strings or reference pages to each plane, it may be advantageous to have some or all reference strings or reference pages provided in one or more additional planes. The additional plane or planes consume minimal additional silicon real-estate and the reference plane has the advantage that the addressed global word line **208***g-a* accesses a reference page at the same time it accesses an addressed page on any of the planes at the same address location along the active strings in the same quadrant. For example, in FIG. **9***b*, reference string **112**Ref, which is shown as dashed line in quadrant Q**2**, resides in reference plane **110**Ref in this example. NOR memory string **112**Ref tracks memory NOR memory string **112** being selected for read in the same quadrant and the read signals from the two NOR memory strings reach the differential sense amplifiers SA for that quadrant practically at the same time. Although reference plane **110**Ref is shown in FIG. **9***b* as being provided in the top plane, any plane in the quadrant can be designated a reference plane. In fact, it is not necessary for every NOR memory string on the reference plane to be a reference string: e.g., every one in eight NOR memory strings can be designated as a reference NOR memory string that is shared by eight NOR memory strings in other planes. The remainder of NOR memory strings in the reference plane may serve as spare strings to substitute for defective strings on the other planes in the block.

Alternatively, one or more additional planes (e.g., plane **117** in FIG. **9***c*) can be set aside to serve as spare memory resources to substitute for defective NOR memory strings, defective pages or defective planes in the same quadrant.

As related to electrically programmable reference strings, slices, pages or planes, once set in their designated threshold voltage states, care must be exercised at all times to inhibit their inadvertent programming or erasing during programming, erasing or reading the non-reference strings.

A very large storage system of 1 petabyte (8×10^{15} bits) requires 8,000 1-terabit memory circuits ("chips"), involving 32M blocks or 64 G slices. (1 Gbits is 1K×1 Mbits). This is a large amount of data to be written (i.e., programmed) or read. Therefore, it is advantageous to be able to program or erase, and read in parallel a great many blocks, slices or pages on numerous chips at once, and to do so with minimum power dissipation at the system level. It is also advantageous for a terabit capacity memory chip to have many input/output channels such that requested data can be streamed in and out in parallel concurrently from and to a large number of blocks. The time required to track down the physical location of the most current version of any given stored file or data set would require a significant amount of time for the system controller to maintain, such as the translation the logical address into the most current physical addresses. The translation between logical to physical addresses would require, for example, a large look-up file allocation table (FAT) to correctly access the slice in the right block on the right chip. Such a search could add considerable read latency (e.g., in the range of 50-100 microseconds) which would impede a fast-read access goal (e.g., under 100 nanoseconds). Accordingly, one aspect of the present invention significantly reduces the search time by introducing a system-wide parallel on-chip rapid file searches, so as to significantly reduce the latency associated with a large FAT, as described below.

Fast Reads: Pipelined Streaming and Random Access

At system initiation of a multi-chip storage system of the present invention, all chips may be erased and reference strings, reference slices or reference planes may be uploaded to their reference states. In one embodiment, one such multi-chip storage system includes a memory module that includes multiple memory chips connected to a module controller which interfaces with the host system controller and manages the multiple memory chips without participation by the host system controller. The memory circuits in the memory module may communicate with each other and with the module controller through an interposer substrate, through-silicon vias (when the memory circuits are stacked one on top of another), hybrid bonding or any other suitable multi-chip assembly technique.

The module controller (sometime also referred to as a "chiplet") designates as cache storage the memory slices (e.g., slice **116** in FIG. **9***c*) that are physically closest to the sense amplifiers and voltage sources **206**. Because of the RC delays along the length of each NOR memory string, the FeFETs in each NOR memory string that are physically closest to substrate circuitry **206** may have their voltages V_{bl} established a few nanoseconds sooner than the FeFETs furthest from substrate circuitry **206**. For example, the first ~50 slices or so (shown as slice **116** in FIG. **9***c*) out of the 1K slices in each quadrant have the shortest latency and can be designated as a cache memory or storage, to be used for storing quadrant operational parameters, as well as information regarding the files or data set stored in the quadrant. For example, each memory page (2×4 Kbits) or slice (2×4 Kbits×16=128 Kbits) written into the upper half-block (i.e., quadrants Q**1** and Q**2**) can have a unique identifier number assigned to it by the module controller, together with an index number that identifies the type of file that is stored.

The cache storage may be used to store on-chip resource management data, such as file management data. A file can be identified, for example, as "hot file" (i.e., associated with a large number of accesses, or a "high cycle count"), "cold file" (i.e., has not been altered for a long time, and is ready to be moved to slower storage or archival memory at a future time), "delete file" (i.e., ready for future erase in background mode), "defective file" (i.e., to be skipped over), or "substitute file" (i.e., replacing a defective file). Also included in the identifier may be a time stamp representing the last time and date the file associated with the identifier was written into the quadrant. Such unique identifier, typically between 32-bit and 128-bit long can be written into one or more of the cache slices as part of the writing of the file itself into the other memory slices in the same half-tile. Files are written sequentially into available erased space, and the identifiers can be assigned by incrementing the previous unique identifier by one for each new file written into memory. If desired, new files can be written into partial slices and the unwritten part of the slice can be used for writing part or whole of the next file, to avoid wasting storage space. Writing sequentially until the entire memory space of the system is used helps level out the wear-out of FeFETs throughout the system. Other on-chip resource management data may include chip, tile, plane, slice, page and string parameters, address locations of faulty strings and their replacement strings, defective pages, defective planes, defective slices and defective tiles and their substitute replacements, file identifiers for all files resident in the tile, look up tables and link lists for skipping over unusable memory, tile-erase cycle counts, optimum voltages and pulse shape and durations for erase, program, program-inhibit, program scrub, read, margin read, read refresh, read scrub operations, error correcting codes, and data recovery modes, and other system parameters.

Because of the modularity of each chip at the tile level and their low-power operations, it is possible to execute simultaneously erase of some tiles, programming at some other tiles, and reading one or more of remaining tiles. The module controller can use that parallelism of operations at the tile level to work in background mode; for example, the module controller may delete (i.e., erase, so as to free up space) some tiles or entire chips, de-fragment fragmented files into consolidated files, move files, tiles or chips that have been inactive for longer than a predetermined time to slower or archival storage, or to chips that group together files with close dates and time stamps, while rewriting the original file identifier with the latest time stamp into cache storage **116** of the next available physical tile.

To facilitate high-speed searches for the location of the most current version of any one file out of the many millions such files in a petabyte storage system, it is important that the unique identifier for each file, wherever it has been physically relocated to, be accessed quickly by the module controller. According to one embodiment of the present invention, a module controller broadcasts the unique identifier (i.e., the 32-128 bits word) for the file being searched simultaneously to some or all the chips in the system. Each chip is provided with a buffer memory to temporarily store that identifier and, using on-chip Exclusive-Or (XOR) circuits, compare the identifier in the buffer memory with all the identifiers stored on cache **116** of each tile and report to the module controller when a match has been found, together with the location where the corresponding file is located. If more than one match is found, the module controller picks the identifier with the most recent time-stamp. The search can be narrowed to just a few chips if the file being searched has been written within a known time period. For a 1-terabit chip, just one 128-Kbit slice or 16×8 Kb pages would be sufficient to store all the 64-bit identifiers for all 2K slices of each tile.

FeFET Pairs for Fast Read Cache Memory.

To reduce read latency for cache storage **116**, FeFETs in NOR memory strings that are physically nearest to sense amplifiers **206** can be arranged in pairs. For example, in adjacent NOR memory strings, two FeFETs sharing a common local word line may be used to store a single data bit between them. In a typical programming operation, FeFETs on both NOR memory strings are initialized to the erased state, then one of the FeFETs is programmed to the programmed state, while the other FeFET is program-inhibited, so as to remain in the erased state. Both FeFETs on the two adjacent active strips are read simultaneously by a differential sense amplifier in substrate circuitry when their shared local word line is raised to the read voltage, the first FeFET that start to conduct tips the sense amplifier into state '0' or state '1', depending on which FeFET is in the programmed FeFET.

This FeFET-pair scheme has the advantage of high-speed sensing and higher endurance because FeFETs of two adjacent NOR memory strings are almost perfectly matched, so that at the sense amplifier even a small programmed voltage differential between the two FeFETs being read will suffice to correctly trip the sense amplifier. In other words, among other roles, the FeFET-pair acts as its own reference, neutralizing any leakage current or noise along the FeFET-pair's respective NOR memory strings. In addition, as the threshold voltage of a programmable reference FeFET may drift over many write/erase cycle over the life of the device, under this scheme the reference FeFET and the read FeFET are both reset with each new cycle. In fact, either one of the two FeFETs in the pair can serve as the reference FeFET. If the two FeFETs making the pair are randomly scrambled to invert or not invert the data written in each cycle, to ensure that statistically each FeFET in each pair serves as the reference FeFET for approximately the same number of cycles as the other FeFET. (The invert/not invert code can be stored in the same page as the page being programmed, to assist in the descrambling during a read operation). Because the paired FeFETs are in close proximity to each other, i.e., on two adjacent active strips on the same plane, the FeFETs can best track each other for local variations in the manufacturing process or to best neutralize (i.e., cancel out) the strip leakage during a read operation. In one embodiment the first and second FeFETs of a pair share a common local word line and separate pre-charge transistors. The respective common bit lines of the first and second FeFETs of the pair function as the input or output terminals of a quasi-volatile data latch. Compared to an SRAM circuit with its low memory density and high leakage, or a DRAM circuit with its very short data retention time, a FeFET-pair in a 3-dimensional NOR memory strings of the present invention have a longer data retention time, lower power dissipation, and very high memory density as data latches.

Alternatively, the FeFET pairing scheme may be applied to FeFETs on different planes where the pair shares a common vertical local word line. The cost of this scheme is in silicon efficiency, as the two FeFETs are required to store one bit between them. For this reason, each tile can be organized such that only a small percentage (e.g., 1% to 10%) of the tile is used as high-speed dual FeFET pairs, while the rest of the tile is operated as regular NOR memory strings and programmable reference FeFET strings. The actual percentage set aside for the FeFET-pair scheme can be altered dynamically by the system controller, depending on the specific usage application. For example, the FeFET pairs can be used to store critical system parameters values and may be periodically refreshed, since they are quasi-volatile. Thus, these FeFET-pairs can maintain their polarization states for extended periods of time (e.g., hours to days) without consuming much power. In fact, such FeFET-pair implemented in 3-dimensional NOR memory arrays of the current invention (which are highly silicon efficient) can be employed as high-density data caches or buffer memory, to back up or altogether replace SRAM or DRAM cache memories that are manufactured on planar (i.e., two-dimensional) silicon substrate and are therefore costly and power consuming, even when they are in their standby modes. Such FeFET-pairs could be particularly useful in artificial intelligence and machine learning applications. The high level of flexibility for operating the NOR memory strings of FeFETs of the present invention result in part from the fact that the FeFETs in a NOR memory string are randomly addressable and operate independently of each other, or of FeFETs in other NOR memory strings, unlike conventional NAND strings.

Numerous applications of data storage, such as video or high-resolution imaging require data files that occupy many pages or even many slices. Such files can be accessed rapidly in a pipelined fashion, i.e., the system controller stores the first page or first slice of the file in the cache memory while storing the remaining pages or slices of the file in a low-cost memory and streaming out the data in a pipeline sequence. The pages or slices may thus be linked into a continuous stream, such that the first page of the file is read quickly into the sense amplifiers and transferred to a data buffer shift register to clock the first page out of the tile while pre-charging and reading the next, slower page in a pipeline sequence, thereby hiding the read access time of each page following the first page. For example, if the first page of 8 Kbits stored in the cache memory is read in 10.0 nanoseconds and then clocked out at 1 Gbit per second, the entire 8K bits would take approximately 1 microsecond to complete clocking out, which is more than sufficient time for the second page to be read from the slower, lower-cost pages. The flexibility afforded by pre-charging randomly selected FeFET strings makes it possible for one or more data files from one or more tiles to be written or read concurrently, with their data streams routed on-chip to one or more data input/output ports.

Random Access Reads

The pre-charging scheme of the current invention allows data to be programmed to be serially clocked into, or randomly accessed, and likewise read out serially in a stream or randomly accessed by words. For example, an addressed page in one plane can be read in one or more operations into the sense amplifiers, registers or latches of the addressed plane, after which it can be randomly accessed in 32-bit, 64-bit, 128-bit or 512-bit or more words, one word at a time, for routing to the input/output pads of the chip. In this manner, the delay attendant to streaming the entire page sequentially is avoided.

Reading FeFETs of the present invention is much faster than reading conventional NAND flash memory cells because, in a NOR memory string, only the FeFET to be read is required to be "on", as compared to a NAND string, in which all the FeFETs in series with the one FeFET being read must also be "on". To reduce read latency, some or all the planes in selected memory tiles may be kept pre-charged to their read voltages V_{ss} (source line) and V_{bl} (bit line), thereby rendering them ready to immediately sense the addressed FeFET (i.e., eliminating the time required for pre-charge immediately before the read operation). Such ready-standby requires very little standby power because the current required to periodically re-charge capacitor **360** to compensate for charge leakage is small. Within each tile, all NOR memory strings on all eight or more planes can be pre-charged to be ready for fast read; for example, after reading FeFETs in NOR memory strings of plane **207-0** (FIG. **2***a*), FeFETs in NOR memory strings of plane **207-1** can be read in short order because its source and bit line voltages V_{ss} and V_{bl} are already previously set for a read operation. The pre-charge operation of a selected FeFET in a NOR string can be performed as the final step following completion of a read operation so that it does not adversely impact the read latency.

In memory tile **100**, only one FeFET per NOR memory string can be read in a single operation. In a plane with eight thousand side by side NOR memory strings, the eight thousand FeFETs that share a common global word line may all be read concurrently, provided that each NOR memory string is connected to its own sense amplifier. If each sense amplifier is shared among, for example, four NOR memory strings in the same plane using a string decode circuit, then four read operations are required to take place in four successive steps, with each read operation involving two thousand FeFETs. Each plane can be provided its own set of dedicated sense amplifiers or, alternatively one set of sense amplifiers can be shared among NOR memory strings in the eight or more planes through a plane-decoding selector. Additionally, one or more sets of sense amplifiers can be shared between NOR memory strings in quadrants and their mirror image quadrants (see, e.g., sense amplifiers (SA) **206** in FIGS. **6***a*, **6***b*, and **6***c*). Providing separate sense amplifiers for each plane allows concurrent read operations of NOR memory strings of all planes, which correspondingly improves the read operation throughput. However, such higher data throughput comes at the expense of greater power dissipation and the extra chip area needed for the additional sense amplifiers (unless they can be laid out in substrate **201** underneath tile **100**). In practice, just one set of sense amplifiers per stack of NOR memory strings may suffice because of the pipeline clocking or data in and out of the memory tile, so that while a first page in one plane is being transferred out of its sense amplifiers to a high speed shift register or buffer memory, the first page of the second plane is being read into the second set of sense amplifiers, with the two sets sharing one set of input/output shift registers.

Parallel operations may also create excessive electrical noise through ground voltage bounces when too many FeFETs are read all at once. This ground bounce is substantially suppressed in all embodiments that rely on pre-charging capacitor **360** to set and temporarily hold the virtual V_{ss} voltage for each active strip. In this case, source voltage V_{ss} of all NOR memory strings is not connected to the chip's V_{ss} ground line, allowing any number of active strips to be sensed simultaneously without drawing charge from the chip ground supply.

Read, Program, Margin Read, Refresh and Erase Operations for Quasi-Volatile 3-Dimensional NOR Memory Strings of FeFETs.

The quasi-volatile NOR memory strings or slices of the current invention may be used as alternatives to some or all DRAMs in many memory applications, e.g., the memory devices for supporting central processing unit (CPU) or microprocessor operations on the main board ("motherboard") of a computer. The memory devices in those applications are typically required to be capable of fast random read access and to have very high cycle-endurance. In that capacity, the quasi-volatile NOR memory strings of FeFETs of the present invention employ similar read/program/inhibit/erase sequences as the non-volatile NOR implementation. In addition, since the polarization stored on programmed FeFETs slowly depolarizes, the weakened polarization state may need to be restored by reprogramming the FeFETs in advance of a read error. To avoid the read error, one may employ "margin read" conditions to determine if a program-refresh operation is required, as are well known to a person skilled in the art. Margin read is an early-detection mechanism for identifying which FeFET will soon fail, to trigger a restoration to the correct programmed state. Quasi-volatile FeFETs typically may be programmed, program-inhibited or erased at reduced programming voltage (V_{pgm}), program inhibit voltage ($V_{inhibit}$) or erase voltage (V_{erase}) or are programmed using shorter pulse durations. The reduced voltages or shorter pulse durations result in a reduced stress on the ferroelectric material and, hence, improvement in endurance. All slices in a tile may require periodic reads under margin conditions to early-detect excessive threshold voltage shifts of the programmed FeFETs due to damage in the ferroelectric storage material. For example, the erase threshold voltage may be 0.5V±0.2 V and the programmed threshold voltage may be 1.5V±0.2V, so that a normal read voltage may be set at ~1V while the margin-read may be set at ~1.2V. Any slice that requires a program-refresh needs to be read and then correctly reprogrammed into the same slice or into an erased slice in the same tile or in another previously erased tile. Multiple reads of quasi-volatile FeFETs can result in disturbing the erase or program threshold voltages, and may require rewriting the slice into another, erased slice. Read disturbs are suppressed by lowering the voltages applied to the control gate, and the source and drain regions during reads. However, repetitive reads may cumulatively cause read errors. Such errors can be recovered by requiring the data to be encoded with error correcting codes ("ECC"). Typically, each read operation involves decoding the retrieved encoded data. The decoding process computes a syndrome, which indicates if the retrieved data has been corrupted. Based on the syndrome, the decoding process determines where the corrupted bits ("error bits") are and makes the necessary correction, if possible. In one embodiment, for a 512-bit data word, one suitable ECC scheme detects up to 6 error bits and corrects up to 5 error bits. An FeFET that has been subject to a billion cycles or more of erase or programming operations may have one or both of its polarization states drifted outside of its designated operational threshold voltage window. This condition may result in an error bit. The system controller may replace the data word associated with the FeFET with a spare word and may adjust the erase/program bias conditions to achieve an operational window for the FeFET. This same repair/replace sequence may be included within the refresh operation (typically, operating in a background mode), as every FeFET in the entire memory circuit is read by the refresh process according to its schedule. The ECC encoded word has 10% or more bits than the data word, and thus imposes an overhead cost in word size and may increase the read latency (e.g., by a few nanoseconds). However, a suitable ECC scheme greatly enhances memory reliability over the life of the memory circuit.

One requirement for the proper operation of the quasi-volatile memory of the present invention is the ability to read and program-refresh a large number of FeFETs, NOR memory strings, pages or slices. For example, a quasi-volatile 1-terabit chip has ~8,000,000 slices of 128K bits each. Assuming that 8 slices (~1 million) of FeFETs can be program-refreshed in parallel (e.g., one slice in each of 8 tiles), and assuming a program-refresh time of 100.0 microseconds, then program-refreshing the entire chip may be achieved in ~100 seconds. This massive parallelism is made possible in memory devices of the present invention primarily because of two key factors; 1) the programming or erase mechanism requires very low current per FeFET, allowing 1 million or more FeFETs to be programmed together without expanding excessive power; and 2) the parasitic capacitor intrinsic to a long NOR memory string enables pre-charging and temporarily holding the pre-charged voltage on multiple NOR memory strings. These characteristics allow a multitude of pages or slices on different tiles to be first read in margin-read mode (or in ECC-check mode) to determine if a refresh is required, and if so, the pages or slices are individually pre-charged for program or program-inhibit and then program-refreshed in a single parallel operation. A quasi-volatile memory with average retention time of ~10 minutes or longer will allow the system controller to have adequate time for properly program-refresh, and to maintain a low error rate that is well within the ECC recovery capability.

FIG. **10***a* shows in simplified form prior art storage system **800** in which microprocessor (CPU) **801** communicates with system controller **803** in a flash solid state drive (SSD) that employs NAND flash chips **804**. The SSD emulates a hard disk drive and NAND flash chips **804** do not communicate directly with CPU **801** and have relatively long read latency. FIG. **10***b* shows in simplified form system architecture **850** using the memory devices of the current invention, in which 3-dimensional non-volatile FeFET NOR memory string arrays **854**, or quasi-volatile NOR memory string arrays **855** (or both) are accessed directly by CPU **801** through one or more of input and output (I/O) ports **861**. I/O ports **861** may be one or more high speed serial ports for data streaming in or out of NOR memory string arrays **854** and **855**, or they may be 8-bit, 16-bit, 32-bit, 64-bit, 128-bit, or any suitably sized wide words that are randomly accessed, one word at a time. Such access may be provided, for example, using a DRAM-compatible protocol (e.g., DDR4), and future higher speed industry standard memory interface protocols, or other protocols for DRAM, SRAM or NOR flash memories. I/O ports **862** handle storage system management commands, with flash memory controller **853** translating CPU commands for memory chip management operations and for data input to be programmed into the memory chips. In addition, CPU **801** may use I/O ports **862** to write and read stored files using one of several standard formats (e.g., CXL, PCIe, HBM, NVMe, eMMC, SD, USB, SAS, or multi-Gbit high data-rate ports). I/O ports **862** communicate between system controller **853** and FeFET NOR memory string arrays in the memory chips.

It is advantageous to keep the system controller or chiplet (e.g., system controller **853** of FIG. **8***b*) off the memory chips, as each system controller typically manages a large number of memory chips, so that it is disengaged as much as possible from the continuous ongoing margin-read/program-refresh operations, which can be more efficiently controlled by simple on-chip state machines, sequencers or dedicated microcontrollers. For example, parity-check bit (1-bit) or more powerful ECC words (typically, between a few bits to 70 bits or more) can be generated for the incoming data by the off-chip controller or on-chip by dedicated logic or state machines and stored with the page or slice being programmed. During a margin-read operation the parity bit generated on-chip for the addressed page is compared with the stored parity bit. If the two bits do not match, the controller reads again the addressed page under a standard read (i.e., non-margin). If that gives a parity bit match, the controller reprograms the correct data into the page, even though it is not yet fully corrupted. If the parity bits do not match, then on-chip dedicated ECC logic or the off-chip controller intervenes to detect and corrects the bad bits and rewrites the correct data preferably into another available page or slice, and permanently retiring the errant page or slice. To speed up the on-chip ECC operations, it is advantageous to have on-chip Exclusive-Or, or other logic circuitry to find ECC matches quickly without having to go off-chip. Alternatively, a memory chip can have one or more high-speed I/O ports dedicated for communication with the controller for ECC and other system management chores (e.g., dynamic defect management), so as not to interfere with the low latency data I/O ports. As the frequency of read or program-refresh operations may vary over the life of the memory chip due to FeFET wear-out after excessive program/erase cycling, the controller may store in each tile (preferably in the high-speed cache slices) a value indicating the time interval between refresh operations. Such a time interval tracks the cycle count of the tile. Additionally, the chip or the system may have a temperature monitoring circuit whose output data is used to modulate the frequency of refreshes according to the measured chip temperature.

In the example of a 1-terabit chip having only 8 tiles out of 4,000 tiles, or 0.2% or less of all tiles are being refreshed at any one time, program-refresh operations can be performed in a background mode, while all other tiles can proceed in parallel with their pre-charge, read, program and erase operations. In the event of an address collision between the 0.2% and the 99.8% of tiles, the system controller arbitrates one of the accesses is more urgent. For example, the system controller can pre-empt a program-refresh, storing away temporarily the state of the interrupted operation, to yield priority to a fast read, then return to complete the program-refresh.

In summary, in the integrated circuit memory chip of the present invention, each active strip and its multiple associated conductive word lines are architected as a single-port isolated capacitor that can be charged to pre-determined voltages which are held semi-floating (i.e., subject to charge leaking out through the string-select transistor in the substrate circuitry) during read, program, program-inhibit or erase operations. That isolated semi-floating capacitor of each active strip, coupled with the low current required to program or erase the FeFETs in a 3-dimensional NOR memory string associated with the active strip, makes it possible to program, erase or read a massive number of randomly selected tiles, sequentially or concurrently. Within the integrated circuit memory chip, the NOR memory strings of one or more of a first group of tiles are first pre-charged and then erased together, while the NOR memory strings of one or more other groups of tiles are first pre-charged and then programmed or read together. Furthermore, erasing of the first group of tiles and programming or reading of a second group of tiles can take place sequentially or concurrently. Tiles that are dormant (e.g., tiles that store archival data) are preferably held at a semi-floating state, preferably isolated from the substrate circuits after having their NOR memory strings and conductors set at ground potential. To take advantage of the massively parallel read and program bandwidths of these quasi-floating NOR memory strings, it is advantageous for the integrated circuit memory chip to incorporate therein multiple high-speed I/O ports such as high-bandwidth memory (HBM) employing dense through-silicon vias (TSVs)) or the much denser heterogeneous integration connectors (e.g., in die-to-die or die-to-wafer format). Data can be routed on-chip to and from these I/O ports, for example, to provide multiple channels for word-wide random access, or for serial data streams out of the chip (reading) or into the chip (programming or writing).

FeFETs of the present invention has the added advantage of being selectively configurable to a depletion threshold voltage state ("depletion mode"). When a depletion mode FeFET has negligible or insignificant source-drain leakage current. A NOR memory string that is not selected for read, write or erase operation may have its common drain region and its common source region biased to a predetermined voltage difference to prevent source-drain leakage currents in the FeFETs that are not in the depletion mode. While held in this bias condition, the NOR memory string has insignificant or negligible leakage current.

In some memory applications, FeFETs may be organized as differential pairs, in which one FeFET in each differential pair holds a first binary value and the other FeFET holds in the differential pair holds the opposite binary value. The differential pair may be used as input or output terminal of a latch. The differential pair may be formed by storage transistors sharing a common gate electrode.

In this detailed description, various embodiments or examples of the present invention may be implemented in numerous ways, including as a process; an apparatus; a system; and a composition of matter. A detailed description of one or more embodiments of the invention is provided above along with accompanying figures that illustrate the principles of the invention. The invention is described in connection with such embodiments, but the invention is not limited to any embodiment. Numerous modifications and variations within the scope of the present invention are possible. The scope of the invention is limited only by the claims and the invention encompasses numerous alternatives, modifications, and equivalents. Numerous specific details are set forth in the description in order to provide a thorough understanding of the invention. These details are provided for the purpose of example and the invention may be practiced according to the claims without some or all of these specific details. For the purpose of clarity, technical material that is known in the technical fields related to the invention has not been described in detail so that the invention is not unnecessarily obscured. The present invention is defined by the appended claims.

We claim:

1. A memory structure formed above a planar surface of a semiconductor substrate, organized as a 3-dimensional array of NOR memory strings, each NOR memory string comprising a plurality of thin-film ferroelectric transistors (FeFETs) sharing a common drain region and a common source region, the common source region and the common drain region each extending lengthwise along a first direction substantially parallel to the planar surface of the semiconductor substrate, the common source region and the common drain region being spaced apart in a second direction substantially perpendicular to the planar surface of the semiconductor substrate, wherein each FeFET further comprises (i) a channel region in contact along the second direction with both the common source region and the common drain region at a first side of the channel region and, (ii) an electrically polarizable layer in contact with the channel region at a second side, opposite the first side, of the channel region so that the channel region is provided between the electrically polarizable layer and the common source region and between the electrically polarizable layer and the common drain region, and (iii) a conductive layer serving as a gate electrode provided adjacent the electrically polarizable layer.

2. The memory structure of claim **1**, wherein the common source region, the common drain region and the channel region of each FeFET comprise crystalline silicon or polycrystalline silicon.

3. The memory structure of claim **1**, wherein each FeFET comprises a junction-less field-effect transistor without a p/n junction between the channel region and the common source region and without a p/n junction between the channel region and the common drain region.

4. The memory structure of claim **3**, wherein the channel region of the FeFET comprises an oxide semiconductor.

5. The memory structure of claim **4**, wherein the oxide semiconductor comprises one or more of: indium zinc oxide and indium gallium zinc oxide.

6. The memory structure of claim **4**, wherein the common drain region and the common source region each comprise a metal.

7. The memory structure of claim **6**, wherein the metal comprises one or more of: molybdenum, titanium, tungsten, lanthanum, tantalum, ruthenium, any silicide thereof, any nitride thereof and any combination thereof.

8. The memory structure of claim **1**, wherein the polarizable layer comprises one or more of: laminated layers of hafnium oxide (HfO2) and zirconium oxide (ZrO2), zirconium-doped hafnium oxide (HfO:Zr), aluminum zirconium-doped hafnium oxide (HfZrAlO), aluminum-doped hafnium oxide (HfO2:Al), silicon-doped hafnium oxide (HfO2:Si), lanthanum-doped hafnium oxide (HfO2:La), hafnium zirconium oxide (HfZrO), hafnium zirconium oxynitride (HfZrON), hafnium zirconium aluminum oxide (HfZrAlO), any hafnium oxide that includes zirconium impurities, and combinations thereof.

9. The memory structure of claim **4**, wherein the oxide semiconductor has a charge carrier mobility exceeding 4.0 cm2/V.

10. The memory structure of claim **4**, wherein the oxide semiconductor in the channel region is formed by deposition at a temperature between 200.0-450.0° C.

11. The memory structure of claim **1**, wherein the polarizable layers of the FeFETs are electrically isolated from each other.

12. The memory structure of claim **1**, wherein each FeFET further comprises an interface dielectric layer between the channel region and the polarizable layer, the interface dielectric layer having a dielectric constant greater than stoichiometric silicon oxide's dielectric constant.

13. The memory structure of claim **12**, wherein the interface dielectric layer comprises one or more of a silicon nitride layer and an aluminum oxide layer.

14. The memory structure of claim **12**, wherein the interface dielectric layer is formed by nitridation of a silicon oxide or silicon oxynitride.

15. The memory structure of claim **12**, wherein the interface dielectric layer is converted by nitridation from an oxide formed on the surface of the channel region.

16. The memory structure of claim **15**, wherein the nitridation is carried out using ammonia anneal.

17. The memory structure of claim **12**, wherein the interface dielectric layer has a dielectric constant greater than 3.9.

18. The memory structure of claim **12**, wherein the interface dielectric layer is treated using one or more of: pulsed ozone and thermal annealing in a hydrogen or deuterium ambient.

19. The memory structure of claim **12**, wherein the interface dielectric layer comprises one or more of: aluminum oxide, zirconium oxide (ZrO2), silicon oxynitride (SiON), silicon nitride (Si3N4), and silicon oxide (SiO2).

20. The memory structure of claim **1**, wherein the 3-dimensional array comprises stacks of NOR memory strings wherein, within each stack, the NOR memory strings are provided at successively greater predetermined distances along the second direction from the planar surface of the semiconductor substrate.

21. The memory structure of claim **20**, wherein each stack of NOR memory strings has a predetermined number of NOR memory strings.

22. The memory structure of claim **20**, wherein the stacks are arranged in rows substantially along a third direction parallel to the planar surface of the semiconductor substrate and orthogonal to both the first and second directions.

23. The memory structure of claim **22**, wherein the conductive layers of a first group of FeFETs in the stacks of NOR memory strings are connected to a first conductor extending along the third direction, each FeFET belonging to a different NOR memory string within the stacks.

24. The memory structure of claim **23**, wherein the conductive layers of a second group of FeFETs provided on an opposite side of the same stacks are connected to a second conductor extending along the third direction, each FeFET belonging to a different NOR memory string within the stacks.

25. The memory structure of claim **24**, wherein the first conductor and the second conductor are offset from each other along the first direction.

26. The memory structure of claim **1**, wherein current flows in the channel region of each FeFET substantially along the second direction.

27. A method for operating a selected FeFET within the memory structure of claim **24**, comprising, during a program or erase operation:

(i) applying a first voltage on the gate electrode associated with the selected FeFET, and applying a second voltage on both the common source region and the common drain region of the selected FeFET, wherein the first and second voltages each have a magnitude difference sufficient to carrying out programming or erasing on the selected FeFET;

(ii) for each FeFET sharing the first conductor with the selected FeFET, applying on the common drain region of the FeFET a third voltage such that the first voltage and the third voltage is different in magnitude by either one-half the magnitude difference of the selected FeFET, or one-third the magnitude difference of the selected FeFET; and

(ii) for each FeFET sharing a common drain region with the selected FeFET, applying on the common gate electrode associated with the FeFET a fourth voltage such that the second voltage and the fourth voltage is different in magnitude by either one-half the magnitude difference of the selected FeFET or one-third the magnitude difference of the selected FeFET.

28. The method of claim **27**, wherein (i) the selected FeFET has a threshold voltage that represents either a programmed state or an erased state, and (ii) when the selected FeFET is in a programmed state, the threshold voltage is a negative value, the method further comprising, during a read operation, applying a boost voltage to the common source region of the selected FeFET that is greater than the negative value.

29. The method of claim **28**, further comprising, during a program or erase operation on the selected FeFET, refreshing one or more FeFETs that share a common gate electrode of the selected FeFET.

30. The memory structure of claim **20**, wherein the NOR memory strings within each stack are isolated one from another by an isolation layer.

31. The memory structure of claim **30**, wherein the polarizable layer is deposited using a selective deposition technique that prevents the polarizable layer from being deposited on the isolation layer.

32. The memory structure of claim **31**, wherein the selective deposition is carried out in an ozone ambient at temperature between 200.0° C. to 350.0° C.

33. The memory structure of claim **31**, wherein the selective deposition includes a post-deposition annealing step carried out at a temperature between 400° C. and 1000° C.

34. The memory structure of claim **31**, wherein the isolation layer comprises silicon oxycarbide (SiOC).

35. The memory structure of claim **1**, wherein each NOR memory string further comprises one or more pre-charge transistors, each configured, when activated, for pre-charging the common source region to a voltage provided on the common drain region.

36. The memory structure of claim **1**, wherein circuitry for supporting memory operation is formed at the planar surface of the semiconductor substrate underneath the FeFETs.

37. The memory structure of claim **36**, wherein the circuitry for supporting memory operations includes both analog and digital circuits.

38. The memory structure of claim **36**, further comprising a layer of interconnect conductors formed above, and in electrically connection with, the array of NOR memory strings, the layer of interconnect conductors being provided for routing control and data signals among the NOR memory strings and the circuitry for supporting memory operations.

39. The memory structure of claim **38**, wherein the circuitry for supporting memory operations comprises two or more of: shift registers, latches, sense amplifiers, reference cells, power supply lines, bias and reference voltage generators, inverters, NAND, NOR, Exclusive-Or and other logic gates, input/output drivers, address decoders, other memory elements, sequencers, state machines, and data processing circuitry that receives control and data signal from an external control circuit.

40. The memory structure of claim **39**, wherein the external control circuit is formed on a separate semiconductor substrate.

41. The memory structure of claim **39**, wherein the semiconductor substrates are provided in an integrated circuit package.

42. The memory structure of claim **40**, wherein the semiconductor substrates are wafer-bonded.

43. The memory structure of claim **36**, further comprising interconnect conductors formed between the array of NOR memory strings and the planar surface, the interconnect conductors electrically connect among the support circuitry, the common drain regions of the FeFETs transistors, and the conductive layers of the FeFETs.

44. The memory structure of claim **1**, wherein each FeFET, upon application of a first voltage difference between the gate electrode and the common source region or the common drain region, is electrically placed in a first of two polarization states and, upon application of a second voltage difference between the gate electrode and the common source region or the common drain region is electrically placed in a second of the two polarization states.

45. The memory structure of claim **1**, wherein the polarizable layer contacts the channel region directly.

46. The memory structure of claim **1**, wherein the polarizable layer is deposited using chemical or physical vapor deposition, ALD, or evaporation.

47. The memory structure of claim **1**, wherein a portion of the polarizable layer is provided in a recess between the common source region and the common drain region.

48. The memory structure of claim **47**, wherein the polarizable layer is formed using a sacrificial material that protects a first portion of the polarizable layer inside the recess but allows a second portion of the polarizable layer to be removed from the isolation layers.

49. The memory structure of claim **47**, wherein, within the recess, the ferroelectric material is embedded in a dielectric material.

50. The memory structure of claim **49**, wherein the dielectric material comprises silicon oxide.

51. The memory structure of claim **47**, wherein the polarizable layer is embedded with a dielectric material within the recess.

52. The memory structure of claim **51**, wherein the dielectric material comprises silicon oxide.

53. The memory structure of claim **1**, wherein the channel region is formed subsequent to forming the polarizable layer.

54. The memory structure of claim **1**, further comprising a metal layer adjacent and in electrical contact with the common drain region.

55. The memory structure of claim **1**, wherein the FeFET polarization for erase and program states are achieved with single voltage polarity.

56. The memory structure of claim **1**, wherein programming of each FeFET takes advantage of a gate-induced drain leakage (GIDL) effect.

57. The memory structure of claim **1**, wherein programming of each FeFET comprises applying a voltage difference between the common drain region and the common source region, thereby generating charge carriers in the channel region.

58. The memory structure of claim **57**, wherein the voltage difference is between 0.5 volts and 2.0 volts.

59. The memory structure of claim **1**, wherein each FeFET is selectively configurable to a depletion threshold voltage state.

60. The memory structure of claim **59**, wherein one of the FeFET in one of the NOR memory strings is configured to the depletion threshold voltage state, and wherein, during operation, the common drain region and the common source region of that FeFET are pre-charged respectively to a first predetermined voltage and a second predetermined voltage.

61. The memory structure of claim **60**, wherein the FeFETs in the NOR memory string, other than the FeFET in the depletion threshold voltage state, each have a substantially zero conduction current between their respective common source regions and common drain regions during the operation.

62. The memory structure of claim **61**, wherein the FeFET in the depletion threshold voltage state has a substantially zero source-drain leakage current.

63. The memory structure of claim **1**, wherein a first FeFET of a first NOR memory string and a second FeFET of a second NOR memory string operate as a differential pair, in which the first FeFET holds a first binary value and the second FeFET holds a binary value different from the first binary value.

64. The memory structure of claim **63**, wherein the gate electrode of the first FeFET and the gate electrode of the second FeFET are electrically connected.

65. The memory structure of claim **63**, wherein the common drain region of the first FeFET and the common drain region of the second FeFET function as input or output terminals of a latch.

66. A memory structure formed above a planar surface of a semiconductor substrate, organized as a 3-dimensional array of NOR memory strings, each NOR memory string comprising a plurality of thin-film ferroelectric transistors (FeFETs) sharing a common drain region and a common source region, the common source region and the common drain region each extending lengthwise along a first direction substantially parallel the planar surface of the semiconductor substrate, wherein each FeFET further comprises (i) a channel region in contact with both the common source region and the common drain region, (ii) a conductive layer serving as a gate electrode, and (iii) an electrically polarizable layer in contact with the channel region, and wherein each NOR memory string further comprises one or more pre-charge transistors, each configured, when activated, for pre-charging the common source region to a voltage provided on the common drain region.

* * * * *