文章编号: 1007 - 4252(2022)03 - 0003 - 200　　DOI: 10.20027/j.gncq.2022.0030

反熔丝: 器件及应用

兰　峻[1,2], 陈振杰[1,2], 郭昊玥[1,2], 李毅达[1,2], 沈　美[1,3], 毛　伟[1,2],
周生明[1,2], 靳　超[4], 冯圣中[4], 张国飙[1,2,*]

(1. 广东省三维集成工程研究中心, 深圳 518000; 2. 南方科技大学 深港微电子学院, 深圳 518000; 3. 南方科技大学 前沿与交叉科学研究院, 深圳 518000; 4. 国家超级计算深圳中心, 深圳 518000)

摘要: 本文综述了反熔丝在现场可编程门阵列(FPGA)和一次可编程存储器(OTP)中的应用, 以及未来作为三维通用人工智能(3D-AI)推理处理器的前景。本文还对反熔丝 ON 态的导通和老化机理进行了分析, 并建立了紧凑模型和失效模型。

关键词: 反熔丝; 导通机理; 老化机理; FPGA; OTP; 3D-AI 处理器
中图分类号: TN3　　**文献标志码:** A

Antifuses: Devices and Applications

LAN Jun[1,2], CHEN Zhen-jie[1,2], GUO Hao-yue[1,2], LI Yi-da[1,2], SHEN Mei[1,3], MAO Wei[1,2], ZHOU Sheng-ming[1,2], JIN Chao[4], FENG Sheng-zhong[4], ZHANG Guo-biao[1,2,*]

(1. Guangdong Provincial Engineering Research Center of 3-D Integration, Shenzhen 518000, China; 2. School of Microelectronics, Southern University of Science and Technology, Shenzhen 518000, China; 3. SUSTech academy for advanced interdisciplinary studies, Shenzhen 518000, China; 4. National Supercomputing Center in Shenzhen, Shenzhen 518000, China)

Abstract: This paper reviews the applications of antifuse in field programmable gate array (FPGA) and one time programmable (OTP), and its potential application in 3-D general-purpose AI (3D-AI) inference processor. The conduction and aging mechanisms of an antifuse in ON state are studied. Compact and failure models are also built.

Key words: Antifuse; Conduction; Aging; FPGA; OTP; 3D-AI processor

0　反熔丝简介

反熔丝(antifuse)面积小、工艺简单, 是一种重要微电子器件。顾名思义, 反熔丝的主要功能与熔丝(fuse)正好相反: 熔丝未编程时为“通(ON)”, 编程后为“断(OFF)”; 反熔丝未编程时为“断

收稿日期: 2022-04-12;　　修订日期: 2022-05-05
基金项目: 深圳市南山区区校共建微电子学院产学研合作平台“5G 前沿”项目(K20799105; K21799107).
作者简介: 兰　峻(1998-), 男, 硕士, 主要研究方向为存储器件(E-mail: 12032647@mail.sustech.edu.cn).
通信作者: 张国飙(1971-), 男, 博士, 教授, 主要研究方向为三维存储和三维计算(E-mail: zhanggb@sustech.edu.cn).

(OFF) ",编程后为"通(ON) "。反熔丝采用类似电容的结构,通过在两个电极之间夹一层薄薄的绝缘层(即反熔丝层) 来实现。对于刚出产的反熔丝,其绝缘层未被击穿,这时反熔丝处于未编程状态,即断开(高阻) 状态。编程时,在两个电极之间施加一个足够大的电压将绝缘层击穿,这时反熔丝处于编程状态,即导通(低阻) 状态。

1　反熔丝的应用: FPGA 与 OTP

目前,反熔丝已经在电子信息、军品以及航空航天等诸多领域得到了广泛的应用。在电子信息领域,其主要应用为现场可编程门阵列(FPGA) 以及一次可编程存储器(OTP) 。

1.1　FPGA

FPGA 属于半定制电路,最早由 Xilinx 公司于 1989 年提出[1] 。它与 ASIC 不同之处主要体现在其逻辑单元模块以及逻辑单元模块之间的互联线均可编程,这使得 FPGA 可以灵活地设置调用逻辑单元模块来实现各种特定功能。这一特点令 FPGA 在逻辑电路设计时具备极大的灵活性,在开发小型逻辑电路时能缩短开发周期、节省开发成本。

FPGA 主要由可编程逻辑单元模块和可编程互联线构成(图 1) 。可编程互联线在没有编程时,横线和竖线没有耦合; 可编程互联线被编程后,横线和竖线耦合。举一个例子,如果图 1 中可编程互联线 a 和 b 被编程,则逻辑单元模块 1 与 5 耦合; 另一方面,如果可编程互联线 a 和 c 被编程,则逻辑单元模块 1 与 6 耦合。这样实现了逻辑单元模块之间的可编程联接。

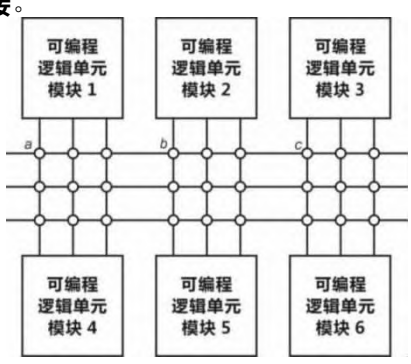

图 1　FPGA 的工作原理(圆圈代表可编程互联线)

Fig. 1 A typical FPGA and its programmable interconnects (circles)

Xilinx 的 FPGA 主要采用 SRAM 等实现可编程互联线,但是 SRAM 需占用较大的芯片面积。为了减少芯片面积,同时也为了提高抗辐射能力,Actel 和 QuickLogic 等公司提出了基于反熔丝的 FPGA。反熔丝在 FPGA 中主要作为可编程互联线。在 FPGA 执行编程操作时,被选中的反熔丝在编程后导通,将相应的逻辑单元模块互联,从而实现特定的逻辑功能。

目前,应用于 FPGA 中的反熔丝主要分为两种: 一种是 Actel 公司采用的 ONO 反熔丝,另一种是 QuickLogic 公司采用的非晶硅反熔丝。

1.1.1 ONO 反熔丝

ONO 反熔丝由 Actel 公司于 1988 年提出并应用于 FPGA [2] ,该反熔丝为硅-硅反熔丝。具体说来,其上下电极分别为 n+ 硅扩散区和多晶硅,反熔丝层(绝缘层) 为 ONO 电介质(图 2) 。ONO 反熔丝的制备工艺与 CMOS 工艺兼容: 在制备过程中首先在硅衬底上形成 n+ 扩散区作为下电极,而后先后沉积氧化硅层、氮化硅层和氧化硅层作为反熔丝层,最后沉积多晶硅作为上电极,得到 ONO 反熔丝。ONO 反熔丝具有可靠性高、稳定性强等优势[3] ,但由于其编程电压较高,ONO 反熔丝需要使用高压晶体管来进行编程,这会带来额外的芯片面积。此外,反熔丝一般位于 FPGA 中金属互联线(如图 1 中横线和竖线) 的交叉点处,n+扩散区和多晶硅还需要与金属互联线进行耦合,这也会带来额外的芯片面积。

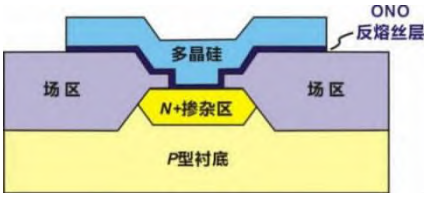

图 2　ONO 反熔丝结构图

Fig. 2 Cross-section of an ONOantifuse

1.1.2 非晶硅反熔丝

QuickLogic 公司于 1992 年提出了一种非晶硅反熔丝结构[4] ,该反熔丝为金属-金属反熔丝(图 3) 。具体说来,该反熔丝的上下电极均使用钨化钛(TiW) ,并采用非晶硅作为反熔丝层。由于上下电极为金属互联线(如图 1 中的横线和竖线) 的一部分,这将有效地减少器件面积。相比于 ONO 反熔

丝，非晶硅反熔丝层较厚，能保证良率。同时，它还具有较低的编程电压，且集成密度更高。这使得它比 ONO 反熔丝更适合应用于更加先进的工艺制程。

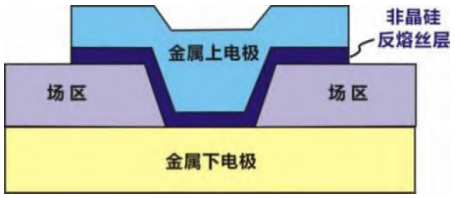

图 3 非晶硅反熔丝结构图

Fig. 3 Cross-section of an amorphous-siliconantifuse

上述两种反熔丝共同构成了如今基于反熔丝的 FPGA，它具有抗辐射、耐高低温、低功耗和速度快等优点，已经广泛应用于多种领域中。但因为反熔丝只能一次编程，这种 FPGA 不能改写，在开发初期设计改动频繁时比较麻烦，整体费用也比较昂贵。

1.2 OTP

反熔丝的另一个重要的应用是 OTP。反熔丝是理想的 OTP 存储元。OTP 只能编程一次，它是除掩膜编程只读存储器(mask-ROM)以外最简单的存储器。OTP 阵列含有多个规整排列的 OTP 存储元，每个存储元含有一个反熔丝和一个开关(图 4)。其中，反熔丝存储“0”和“1”；开关可以避免读写时存储元之间的串扰。

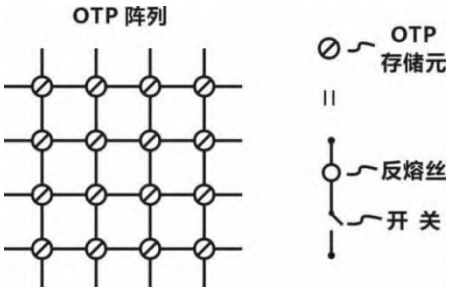

图 4 典型的 OTP 阵列及 OTP 存储元的构成

Fig. 4 A typical OTP array and OTP memory cells

相对于可擦除可编程只读存储器(EPROM)和带电可擦可编程只读存储器(EEPROM)，OTP 只能写一次，不能改写，这对于安全性要求高的应用来说是一个优势。更为重要的是，基于反熔丝的 OTP 可以做到完全与 CMOS 工艺兼容，不需要增加新的光罩。对于仅需要有限嵌入式 OTP 容量(用于存储芯片 ID、密钥、代码等信息)的处理器，这是一个巨大的成本优势。

1.2.1 2T-OTP

Kilopass 公司于 2004 年提出一种基于反熔丝的 2T-OTP[5]。其 OTP 存储元由两个晶体管(2T)构成：反熔丝晶体管和开关晶体管(图 5)。它在制备上完全兼容 CMOS 工艺。当把 2T-OTP 存储元置于一个 OTP 阵列进行写操作时，开关晶体管的栅极(连接字线)被施加一个开电压而被导通，同时反熔丝晶体管的栅极(连接位线)被施加一个写电压，反熔丝晶体管的源极接地，这样一来反熔丝晶体管的栅极绝缘层因为两边压差过大而被击穿，从而实现编程。在读操作时，将开关晶体管导通，同时给反熔丝晶体管施加一个读电压，通过读取流出源极的电流大小来判断该存储元的状态。2T-OTP 的优势在于其完美地兼容 CMOS 工艺。但是，由于其存储元包含两个晶体管，集成度稍显不高。

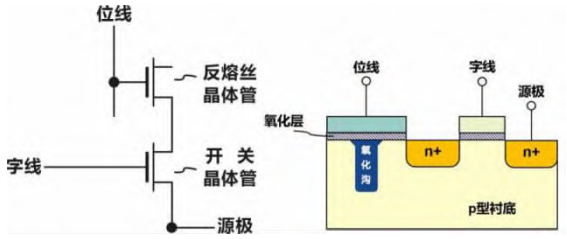

图 5 2T-OTP 存储元示意图及其横截面图

Fig. 5 A 2T-OTP memory cell: schematic and structure

1.2.2 1T-OTP

为了进一步提高集成度，Sidense 公司于 2008 年提出了一种 1T-OTP，它将反熔丝晶体管和开关晶体管的功能集成在一个晶体管上[6]。其特点在于它通过在同一个晶体管上沉积两个厚度不同的栅极氧化层来将两种不同的晶体管集成在一个晶体管上(图 6)。其制作工艺与普通晶体管大体相似，唯一不同之处在于，在沉积栅极氧化层时通过两张掩膜版来先后沉积出厚氧化层和薄氧化层，厚氧化层部分作为开关晶体管，薄氧化层部分作为反熔丝晶体管。通常来说，厚氧化层可采用 3.3 V 高压晶体管的栅极氧化层，薄氧化层可采用 1.8 V 低压晶体管的栅极氧化层。

2 反熔丝的导通机理及紧凑模型

为建立反熔丝的紧凑模型，有必要对其导通机理进行分析。图 7 是一种典型的反熔丝的多次读电

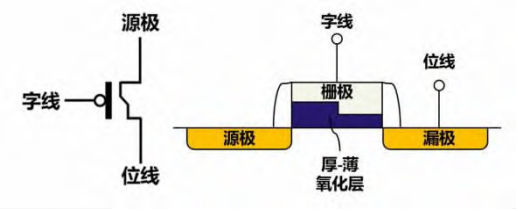

图 6　1T-OTP 存储元示意图及其横截面图

Fig. 6 A 1T-OTP memory cell: schematic and structure

阻-读电流(R-I) 扫描曲线[7]。该反熔丝的上下电极为钨,反熔丝层为 3 nm 厚的氧化硅。在第一次扫描之前,先使用 5 mA 的电流对器件进行编程(I_p = 5mA)。第一次扫描时最大读电流为 5 mA。第二次到第四次扫描的最大读电流分别为 10 mA、15 mA 和 20 mA。第一次和第二次扫描时,在读电流低于 5 mA 时 R-I 曲线基本一致,这说明了在低于编程电流时器件没有任何结构性变化。为了区分第一次和第二次扫描曲线,图 7 对第一次扫描曲线加了 5 Ω 的偏置。第二次扫描时,当读电流小于 5 mA(或小于前一次扫描的最大读电流) 时,读电阻阻值随读电流增加而上升; 当读电流大于 5 mA 时,读电阻阻值随读电流增加而下降。第三次和第四次扫描时,当读电流小于前一次扫描的最大读电流时,读电阻阻值随读电流增加而上升; 当读电流大于前一次扫描的最大读电流时,读电阻阻值随读电流增加而下降。从这些扫描曲线可以看出,前一次扫描的最大读电流可以被认为是反熔丝的编程电流。当读电流小于编程电流时,读电阻阻值随读电流增加而上升; 当读电流大于编程电流时,读电阻阻值随读电流增加而下降。

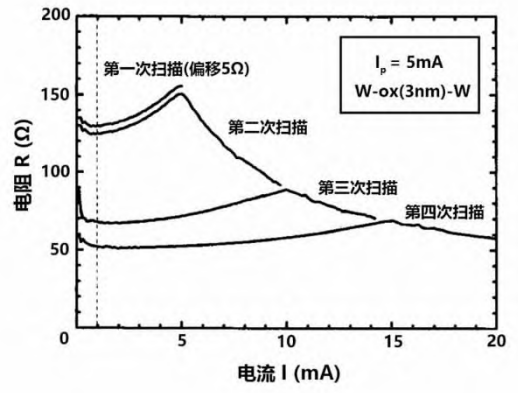

图 7　一种典型反熔丝的四次 R-I 扫描曲线

Fig. 7 R-I scans of a typical antifuse

为了解释上述现象,我们构筑了如下物理图像,并结合 QuickLogic 公司所提出的反熔丝结构加以验证[8]。在编程操作中,编程电压将反熔丝层在某个薄弱点击穿。当编程电流流过这个击穿点时,由于击穿点很小,流过击穿点的电流密度很大,其产生的焦耳热会使击穿点附近的温度高于反熔丝材料与电极材料的反应温度,导致反熔丝材料与电极材料发生化学反应,从而在两个电极和反熔丝层之间形成一个导电通道(图 8)。这个导电通道由反熔丝材料与电极材料的反应物构成,如金属硅化物和金属氧化物的混合物。

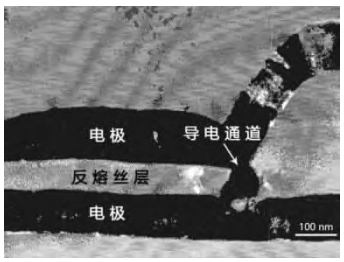

图 8　反熔丝被编程后的透射电镜图像[8]

Fig. 8 A TEM image of a programmedantifuse [8]

这里,我们用“核”来代表这个导电通道(图 9)。在编程时,核为热核,其边界温度 T_c 为反熔丝材料与电极材料的反应温度 T_{c0}。这时,核内部温度高于反应温度,核材料为反应物; 核外部温度低于反应温度,没有反应发生,其材料仍为电极。若不再施加编程电流,核温度降低,最后留下一个反应后的冷核。

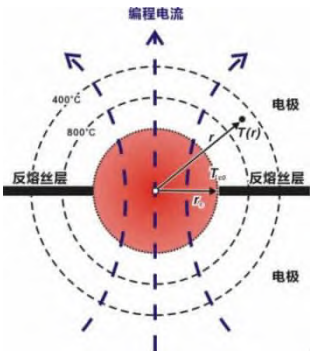

图 9　反熔丝编程时的电流密度和内部温度分布图

Fig. 9 Current density and internal temperature in anantifuse during programming

在对编程后的反熔丝重新进行读操作时(图 10),整个反熔丝会被读电流重新加热,温度升高导

致电阻率上升，因此读电阻阻值升高。当读电流等于编程电流时，核温重新达到反应温度。此时若读电流持续增加，则会产生更多的焦耳热，核温将继续上升并超过反应温度，这时核边界将扩大，导致导电通道的面积更大，从而降低读电阻阻值。基于上述物理图像，可以把编程电流定义为反熔丝经历的最大读电流，这与图 7 中观察到的 R–I 曲线相符。

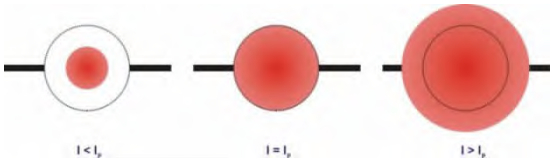

图 10 对反熔丝进行读操作时的温度变化（红色部分温度大于 T_{c0}）

Fig. 10 Temperature change during read (The temperature of the red portion is greater than T_{c0})

上述物理图像说明反熔丝 ON 态的导通机理为电–热效应：电流密度增加→核温升高→反熔丝材料与电极材料发生化学反应→核扩大→读电阻阻值下降。因此，我们可以对反熔丝进行电–热分析。基于热传导方程：

$$\nabla[\kappa(T)\ \nabla T(r)\ ] = -\rho(T)\cdot J(r) \tag{1}$$

为求得温度场 $T(r)$，我们需要知道电流密度 $J(r)$。利用器件的对称性，我们将三维系统近似为一维系统。在核外部，电流近似于从中心向外辐射流动，在点 r 的电流密度为：

$$J = \frac{I}{2\pi r^2} r > r_c \tag{2}$$

在完整求解热传导方程 Eq.（1）时，需要考虑电极材料电阻率 ρ 和热导率 κ 的温度效应。通过求解公式（1）和公式（2），我们可以得到核半径 r_c：

$$r_c = I_p \cdot \left[\int_{T_a}^{T_{c0}} \frac{\kappa(T)\ dT}{\sqrt{\left(\frac{V_f}{4\pi}\right)^2 - \int_{T_a}^{T} \frac{\kappa(T)\ \rho(T)}{2\pi^2} dT}} \right]^{-1} \tag{3}$$

其中，T_a 是环境温度；V_f 为反熔丝的特征电压。注意到，V_f 为编程时反熔丝上的电压降，为一常数，参见公式（6）。

公式（3）虽然给出了 r_c 的解析解，但很难获得直观的结果。我们需要对其进行近似。因为电流密度在核内最大，大部分焦耳热在核内产生，可以假设焦耳热仅在核内部产生。忽略各种温度效应（包括电阻率 ρ 和热导率 κ 的温度效应），可解出温度场 $T(r)$：

$$T(r) = \frac{I_p^2 \cdot R}{4\pi\kappa r} + T_a$$

相应地，核半径 r_c 为：

$$r_c = \frac{I_p^2 \cdot R}{4\pi\kappa(T_{c0} - T_a)} \tag{4}$$

反熔丝的 ON 态电阻由核电阻和电极扩散电阻组成，它们均反比于 r_c（图 9）：

$$R = R_c + R_s = \frac{(\rho_c + \rho_s)}{\pi r_c} \tag{5}$$

结合公式（4）和公式（5），得到反熔丝 ON 态电阻：

$$R = \frac{\sqrt{4\kappa(\rho_c + \rho_s)(T_{c0} - T_a)}}{I_p} = \frac{V_f}{I_p} \propto \frac{1}{I_p} \tag{6}$$

从公式（6）可以看出，反熔丝 ON 态电阻 R 与编程电流 I_p 呈反比关系，这与实验结果相符（图 11）。公式（6）也是反熔丝 ON 态的紧凑模型。

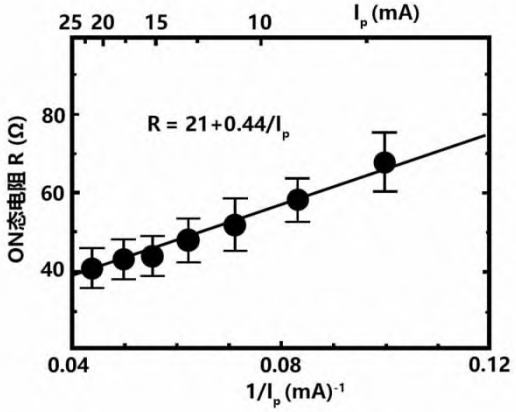

图 11 反熔丝 ON 态电阻 R 与编程电流 I_p 具有反比关系

Fig. 11 The ON-state resistance R of the antifuse has an inverse relationship with the programming current I_p

从公式（6）可以看到，V_f 与 I_p 无关，仅与电极和反熔丝的材料特性相关。有意思的是，基于 Wiedemann–Franz 定律[9]，导电的金属也导热，即金属的电导率与热导率成正比（或，热导率与电阻率之积为常数），V_f 在一级近似下与电极材料仅弱相关。例如，虽然金的电阻率不到钨的一半，但金电极与钨电极反熔丝的 V_f 均为 ~0.5 V[10]。这是因为虽然金的电阻率较低（即电导率较高），但其热导率也高，导致金电极较散热，核较小，电阻与钨电极反熔丝接

近。另一方面,V_f随反熔丝层的厚度增加而增加,这是因为导电通道的厚度会增加反熔丝电阻。实验条件下观察到的 V_f一般在 0.4 ~ 0.9 V 之间。

3　反熔丝的老化机理和失效模型

反熔丝 ON 态存在稳定性问题[11]。当施加在反熔丝上的施压电流 I_s越大,其失效时间越短(图 12)。为了掌握使用反熔丝时的安全操作窗口,有必要对反熔丝进行失效分析,了解其老化机理,建立其失效模型,以便在实际应用时使用加速老化测试。

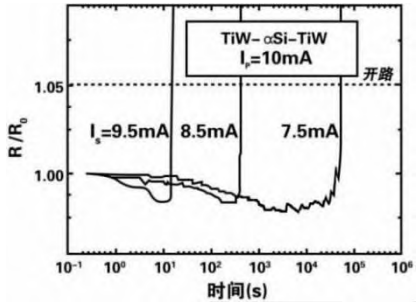

图 12　不同施压电流 I_s下反熔丝的失效时间

Fig. 12 Time – to – failure (TTF) ofantifuse under different stress current I_s

施压电流对反熔丝有两个影响: 1、导电通道处的电流密度高达 100 MA/cm^{2}或甚至更高,这会导致电迁移劳损; 2、导电通道处的温度将上升几百摄氏度,这会导致热应力劳损(图 13)。在分析反熔丝的老化机理时需要了解这两种劳损(电迁移劳损和热应力劳损)各自带来的老化效果。

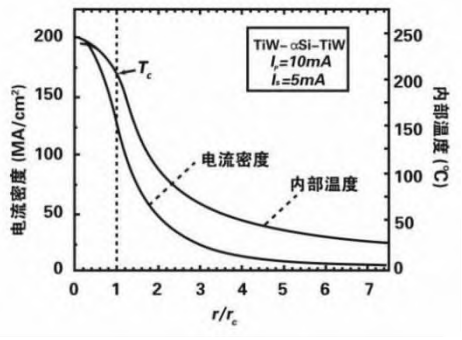

图 13　电流施压时,反熔丝内部的电流密度和温度

Fig. 13 Current density and temperature inside theantifuse under a stress current

为了分析电迁移劳损带来的老化效果,我们设计了一个双向电流施压实验(图 14)。反熔丝已被 I_p = 10 mA 编程,然后施加 | I_s | = 8.5 mA 的双向 I_s: 在施加一段时间的正向 I_s后,I_s极性反转,再施加一段时间的负向 I_s。可以观察到: 施加正向 I_s时,开始反熔丝电阻降低,然后升高。如果没有极性反转,反熔丝将于 700 s 处失效。随着极性反转,反熔丝的电阻降低,这说明相反 I_s极性可以治愈前一段时间造成的劳损。由于热应力劳损没有这种对 I_s极性的依赖,表明电迁移劳损是导致反熔丝老化的一个重要机理。

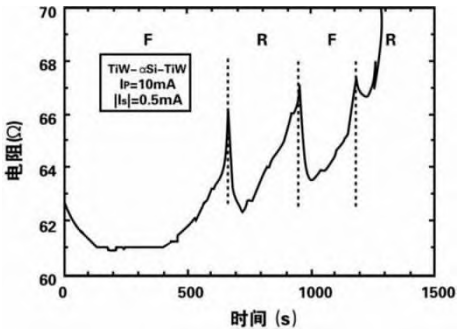

图 14　反熔丝施加双向施压电流实验

Fig. 14 A bidirectional current stress test

另一方面,为了分析热应力劳损带来的老化效果,我们设计了一种高频交流施压实验(图 15)。可以观察到,在高频交流施压时,反熔丝电阻一开始便出现轻微下降,在稳定循环一定周期之后,反熔丝电阻阻值迅速增加,进而失效。如果施压电流的幅度越大,其失效周期次数越少,由于在高频交流施压条件下,电迁移劳损不明显。在高频交流施压下能观察到反熔丝失效,表明热应力也是导致反熔丝失效的另一个重要机理。

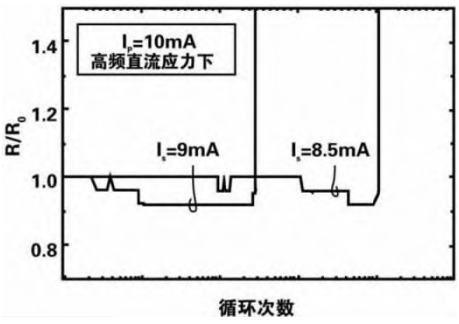

图 15　反熔丝的高频交流施压实验

Fig. 15 A high frequency AC current stress test

根据广义 Black 方程,反熔丝 ON 态的失效时间

(TTF)与核电流密度 J_c 和核温 T_c 具有如下关系:

$$TTF = \frac{A}{J_c^n} \times e^{E_a/k_B T_c} \quad (7)$$

其中,A 是材料常数,E_a 是活化能。公式(7)中的第一项因子考虑了电迁移的作用,第二项因子考虑了热应力的作用。

根据电-热模型,核温 T_c 可表示为:

$$T_c \approx (T_{c0} - T_a)\frac{I_s^2}{I_p^2} + T_a \quad (8)$$

将公式(8)带入公式(7)中,反熔丝的失效模型为:

$$TTF \approx TTF_0 \cdot e^{\left(\frac{E_a}{k_B T_{c0}}\right)\cdot\left(\frac{I_s}{I_p}\right)^2} \quad (9)$$

注意到,反熔丝的失效时间对施压电流 I_s 的依赖性远强于传统的电迁移($TTF \propto 1/J^n$),同时失效时间对环境温度的敏感性远低于传统热应力($TTF \propto e^{E_a/k_B T_a}$)。这是因为施压电流 I_s 主要通过提升核温 T_c 来影响失效时间。上述失效模型可以用于找到反熔丝的工作电流窗口(图 16)。为确保反熔丝 10 年工作寿命,工作电流应小于编程电流的 60%。

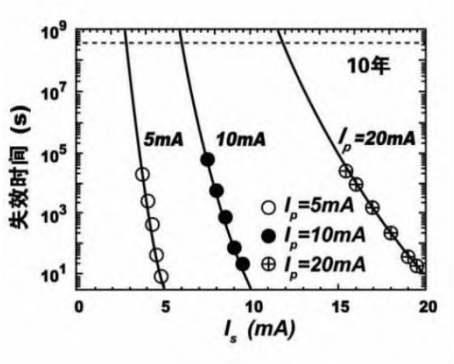

图 16 反熔丝的失效时间与施压电流的关系

Fig. 16 The relationship between TTF and stress current

4 展望:三维集成

随着"微缩"的潜力被日益挖尽,"增维"(2D→3D)集成,成为集成电路的必然趋势。"增维"对我国芯片产业有重要意义。由于 EUV 光刻机等关键半导体设备被禁运,我国不宜单纯追求"7 nm",而应将"7 nm 等效"作为追求目标。三维集成采用多层堆叠,可用较易国产化的设备(如 80 nm)达到7 nm 等效(例如,3D-NAND 采用 80 nm 技术,通过 128 层的堆叠,实现 7 nm 等效,128/80 nm^2 = 1/7 nm^2)。"增维"可为我国发展集成电路另辟蹊径。

反熔丝及 OTP 由于结构和功能简单,对材料和工艺的要求较低,研发周期较短,是突破三维集成关键技术的理想载体。国际上已实现 3-D OTP 的量产,它采用 P/i/N 二极管作为开关元件。团队最近发现将 OTS 器件作为开关元件,其各项指标均比 P/i/N 二极管有较大优势。基于 OTS 器件的 3D-OTP 可用于实现三维通用 AI(3D-AI)推理处理器。

4.1 3D-OTP

1.2 节中的 OTP 均为二维集成。在 90 年代末,业界先行者意识到随着线宽的不断缩小,总有一天"微缩"难以持续,有必要开拓"增维"路径。

4.1.1 3-D OTP

在对三维集成在存储领域的探索中,张国飙于 1996 年提出三维存储器,并于 1998 年获得美国专利 US 5,835,396 [12]。Matrix 半导体公司获得张国飙的专利许可并于 2003 年量产了业界第一款三维存储产品——3-D OTP [13]。3-D OTP 由四个垂直堆叠的存储层构成,其存储元采用多晶硅 P/i/N 二极管及反熔丝结构(图 17) [14]。Matrix 的 1 Gb 芯片面积仅为 31 mm^2,为当时世界上面积最小的 1 Gb 存储芯片。三维堆叠技术比平面二维结构能获得更大密度,可以最大化利用空间,被认为是延续摩尔定律的一种关键技术。

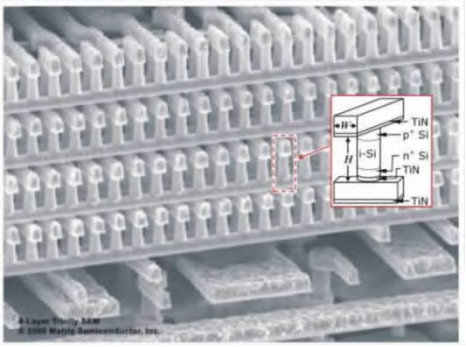

图 17 Matrix 公司 3-D OTP 截面图及其存储元示意图

Fig. 17 Matrix's 3-D OTP and its memory cell

4.1.2 OTS

虽然 P/i/N 二极管在 3-D OTP 中有效地起到开关单元的作用,但是其 ON 态电流密度不能满足高速 OTP 存储阵列的需求。相比之下,Ovonic 阈值开关器件(OTS)因其出色的 ON 态电流密度而被认

为是 OTP 存储元的理想开关元件。

OTS 是一个阈值开关,它采用硫系化合物(如 GeS、GeSe 等)作为其介质材料,其 ON 态电流密度远高于 P/i/N 二极管(图 18)[15-22]。对于 P/i/N 二极管来说,虽然选择性超过 10^6,但 ON 态电流密度仅能达到 800 A/cm^2[14]。而对于 OTS 来说,选择性往往可以超过 10^4,ON 态电流密度虽然根据材料的不同而有优劣,但整体远高于 P/i/N 二极管。尤其值得注意的是,GeS 的 ON 态电流密度达到 34 MA/cm^2,而且选择性也高达 10^6[15],它是 OTP 存储元的潜在材料。

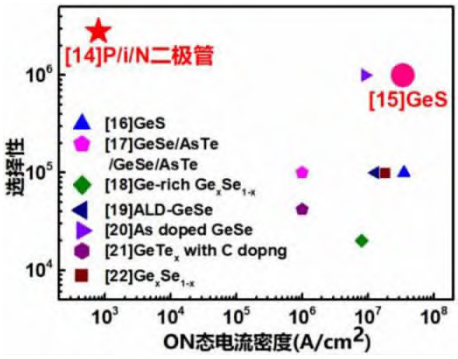

图 18　多种 OTS 与 P/i/N 二极管选择性和 ON 态电流密度的比较

Fig. 18 S (selectivity) and J_{ON} (ON-state current density) comparison between various OTS's and P/i/N diode

4.2　3D-AI 推理处理器

人工神经网络是 AI 研究的热点。一个神经网络由大量神经元组成。在每个神经元中,输入矢量 x 与突触权重 w 通过线性组合 $\Sigma\times$ 后,由一个非线性激活函数 $f(\cdot)$ 处理,产生输出矢量 y。AI 处理器的主要研究内容包括矩阵乘法($\Sigma\times$)和激活函数 $f(\cdot)$ 的芯片实现。

$$y_j = \sum_i w_{ji} \times x_i;\ Z_j = f(y_j) \tag{10}$$

4.2.1 面临的挑战

矩阵乘法($\Sigma\times$)芯片实现的研究热点为存内计算架构,它们多基于模拟乘法器[23-27],即采用模拟交叉存储阵列来实现矢量矩阵乘法。其中,存储阵列以模拟形式存储突触权重。具体芯片实现时多采用高值存储(高值为多值存储的一种,每个存储元存储的态数≥8,如 32[26])。虽然高值存储在存储产业早已成熟,如 3D-NAND 已量产 QLC 闪存(16 值),正在研发可编程逻辑控制器 PLC(32 值),这归功于存储器可用 ECC(Error Correcting Code)电路纠错,每态有较大的操作窗口(容错度)。但是,乘法器却不具有 ECC 机制,每态的操作窗口(容错度)较小,这将限制存储阵列大小(如 64×64),并对器件失配(性能一致性)提出较高要求。此外,可多次编程(可擦除重复写)的要求对存储元的材料和工艺带来更大挑战。

另一方面,激活函数 $f(\cdot)$ 的芯片实现却鲜有人问津,这是因为绝大多数激活函数很难通过传统的嵌入式处理器实现。在研的 AI 处理器均采用 ReLU[28]。ReLU 为线性整流函数,芯片实现成本极低(可以用一个晶体管实现),这是其巨大的优势。但是,ReLU 应用场景单一,主要用于图像分类,并不具备通用性。在很多应用场景需要其他激活函数(表 1)。例如,语言翻译需要使用 Tanh,图像识别和语音识别需要使用 GeLU[29]。而且,AI 算法研发人员正针对不同应用场景不断提出新的激活函数。在研的激活函数包括: Sigmoid/Tanh 类激活函数及其变种(sTanh、ReSech、sSigmoid、pTahnh、Hexpo、ISigmoid、SiLU、LiSHT、Elliot、SRS 等)[30-33]、线性整流类激活函数及其变种(ReLU、LReLU、PReLU、RReLU、CReLU、BReLU、PTELU、FReLU、EReLU、RTReLU、DualReLU、PairedReLU、ABReLU、NLReLU、MTLU、L-ReLU 等)[34-38]、指数类激活函数及其变种(ELU、SELU、CELU、PELU、PELU、FELU、PREU、EELU、PDELU 等)[39-43]、自适应类激活函数及其变种(APL、SAF、BDAA、AAF、Swish、ESwish、TAF、SLAF、MeLU)[44-46]、混杂激活函数(Softplus、Mish、GeLU、RePU、VAF、DY-ReLU、WHE、KAF)[47-48] 等。上述激活函数包含多种超越函数,芯片实现成本极高。

表 1　各种激活函数(AF)的函数形式及其应用场景

Table 1 The functional forms of various activation functions (AF) and their application scenarios

激活函数(AF)	$Z=f(y)$	应用场景
Sigmoid	$1/(1+e^{-y})$	-
Tanh	$(e^{y}-e^{-y})/(e^{y}+e^{-y})$	语言翻译
ReLU	$y\ (y\geq 0);\ 0\ (y<0)$	图像分类
GeLU	$0.5y(1+\tanh(\sqrt{2}/(y+0.044715y^3)))$	语音识别 图像识别
针对不同应用场景,新的激活函数不断出现		

4.2.2 3D-AI 推理处理器

在研 AI 处理器仅支持 ReLU 激活函数，应用场景单一（图像分类），不具备通用性（如自然语言处理需要 GeLU 激活函数）。为了解决上述问题，我们提出一种三维通用 AI（3D-AI）推理处理器。它采用一种新的设计方法——“存储换计算（或，查表计算）”：利用三维集成带来的海量存储，通过读取查找表来代替复杂计算，减轻处理器的负担。3D-AI 处理器将 CMOS 层（模拟电路）、第一 3D-OTP 层（存储突触权重）、第二 3D-OTP 层（存储激活函数查找表）三维集成（图 19）。它根据应用场景将相应的激活函数查找表（AF-LUT）录入到 3D-OTP 中，从而支持多种激活函数和应用场景，保障通用性。

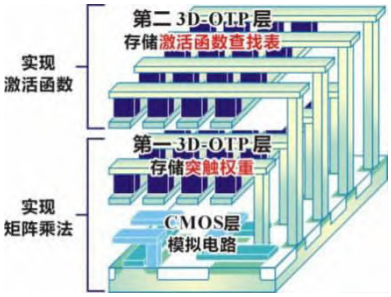

图 19 一种典型 3D-AI 推理处理器的三维集成结构

Fig. 19 A typical 3D-AI inference processor

3D-AI 处理器通过 3D-OTP 来分别实现激活函数 $f(\cdot)$ 和矩阵乘法（$\Sigma\times$）。激活函数 $f(\cdot)$ 是“存储换计算”的一个典型应用，它利用查找表存储阵列来实现激活函数 $f(\cdot)$（图 20）：其输入为 y，输出为 $Z=f(y)$，存储阵列以数字形式存储激活函数查找表（AF-LUT）。不同的激活函数 $f(\cdot)$（如 ReLU、ELU、GeLU）可以通过改变 AF-LUT 来实现。归纳起来，3D-AI 处理器根据应用场景将相应的 AF-LUT 录入到 3D-OTP 中，从而支持多种激活函数和应用场景，以保障通用性。

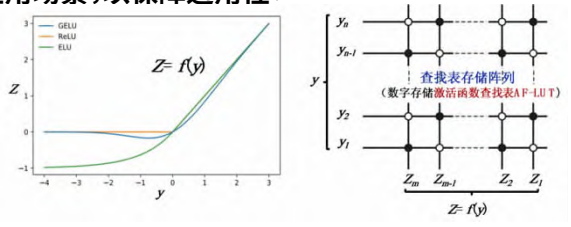

图 20 激活函数 $Z=f(y)$（左）及其查找表存储阵列（右）

Fig. 20 Activation function $Z=f(y)$ (left) and its lookup table (AF-LUT) storage array (right)

另一方面，数模乘法器用来实现矢量矩阵乘法（$\Sigma\times$），它含有权重存储阵列和乘法读出电路（图 21）。权重存储阵列以数字形式（一般为低值，即≤4 值）存储突触权重。相对于模拟乘法器以模拟形式（一般为高值，即≥8 值）存储突触权重，数模乘法器给芯片制造带来较小压力。

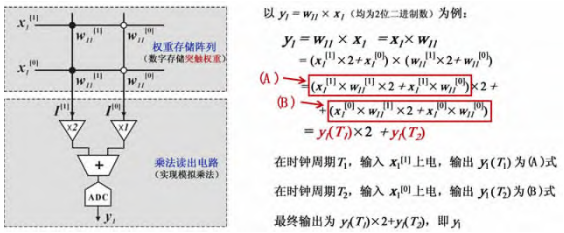

图 21 数模乘法器用于实现矢量矩阵乘法（$\Sigma\times$）及其工作原理

Fig. 21 A hybrid multiplier to implement vector-matrix multiplication ($\Sigma\times$)

之所以将 3D-OTP 作为首选的三维存储器并进而基于此设计 3D-AI 处理器，原因在于 3D-OTP 技术较简单成熟。OTP 只需一次编程，在存储材料上有更多选择，工艺开发周期更短。国外存储原厂已实现三维存储（3D-NAND、3D-XPoint）的量产，正将注意力转向三维存算。我国在三维存储领域还处于起步阶段。如继续按照常规的研发节奏（材料→器件→电路→阵列→中试→量产），当我国实现三维存储量产时，国外已占据三维存算的知识产权高地，我国又将陷于被动。因此，有必要找到一种具有良好可制造性且研发周期短的三维存储技术，并以此为平台抢占三维存算的知识产权高地。我们认为 3D-OTP 正是这样一个平台。

5 结论

反熔丝在 FPGA 和 OTP 中已获得广泛应用。FPGA 应用以 Actel 公司的 ONO 反熔丝和 QuickLogic 公司的非晶硅反熔丝为主要代表。嵌入式 OTP 应用以 Kilopass 公司的 2T-OTP 和 Sidense 公司的 1T-OTP 为代表。由于反熔丝结构简单，它成为业界首款实现“增维”（2D→3D）的集成电路产品——Matrix 半导体公司的 3-D OTP。

反熔丝 ON 态导通机理属于电-热效应：电流密度增加→核温升高→电极材料与反熔丝材料发生化学反应→核扩大→读电阻阻值下降。读电阻阻值 R

与编程电流 I_p(反熔丝经历的最大电流)的大小成反比,即 $R=V_f/I_p$。其中,V_f是反熔丝的特征电压,它为编程时反熔丝上的电压降。V_f为一常数,与 I_p 无关,仅与电极材料的特性弱相关,并随反熔丝层的厚度增加而增加。对于不同反熔丝,V_f介于 0.4 ~ 0.9 V 之间。

反熔丝 ON 态老化机理为电迁移劳损和热应力劳损的共同作用。由于电迁移劳损和热应力劳损均与温度强相关,而核温 T_c 由施压电流 I_S 控制,因此反熔丝的失效时间 TTF 与施压电流 I_S 具有如下关系: $TTF \propto e^{(I_p/I_S)^2}$。对于非晶硅反熔丝,为了保证其 10 年工作寿命,反熔丝的工作电流应小于编程电流的 60%。

在研的 AI 处理器仅支持 ReLU 激活函数,应用场景单一(图像分类),不具备通用性(如自然语言处理需要 GeLU 激活函数)。3D-AI 推理处理器采用"存储换计算(或,查表计算)",可根据应用场景将相应的激活函数查找表(AF-LUT)录入到 3D-OTP 中,从而支持多种激活函数和应用场景,以保障通用性。3D-AI 推理处理器通过三维集成,将神经网络计算所需的矩阵乘法和激活函数优美地结合在一起,开拓了一种能突破冯·诺伊曼能效瓶颈的新型计算和存储架构。

参考文献:

[1] Freeman R H. Configurable electrical circuit having configurable logic elements and configurable interconnects: U. S. Patent 4,870,302 [P]. 1989-9-26.

[2] Hamdy E, McCollum J, Chen S O, et al. Dielectric based antifuse for logic and memory ICs [C]. //Technical Digest., International Electron Devices Meeting. IEEE, 1988: 786-789.

[3] Chiang S, Wang R, Chen J, et al. Oxide-nitride-oxideantifuse reliability [C]. //28th Annual Proceedings on Reliability Physics Symposium. IEEE, 1990: 186-192.

[4] Gordon K E, Wong R J. Conducting filament of the programmed metal electrode amorphous siliconantifuse [C]. // Proceedings of IEEE International Electron Devices Meeting. IEEE, 1993: 27-30.

[5] Peng J Z. Semiconductor memory cell and memory array usinga breakdown phenomena in an ultra-thin dielectric: U. S. Patent US 6,822,888 B2 [P]. 2004-11-23.

[6] Kurjanowicz W. split-channel antifuse array architecture: U. S. Patent US 7.402,855 B2 [P]. 2008-06-22.

[7] Zhang G, Hu C, et al. An electro-thermal model for metal-oxide-metalantifuses [J]. IEEE Transactions on Electron Devices, 1995, 42(8): 1548-1558.

[8] Wong R J, Gordon K E, Chan A K. Time dependent reliability of the programmed metal electrodeantifuse [C]. // Proceedings of International Reliability Physics Symposium. IEEE, 1996: 22-26.

[9] Kittel C. Introduction to Solid State Physics, 6thed [M]. New York: John Wiley, 1986.

[10] Zhang G, Hu C, et al. Characteristic voltage of programmed metal-to-metalantifuses [J]. IEEE Electron Device Letters, 1994, 15(5): 166-168.

[11] Zhang G, King Y, Elfoukhy S, et al. On-state reliability of amorphous silicon antifuses [C]. //Proceedings of International Electron Devices Meeting. IEEE, 1995: 551-554.

[12] Zhang G. Three-dimensional read-only memory: US, US5835396 A [P]. 1998-11-10.

[13] Christopher M, et al. Matrix® 3-D memory roadmap advancements-white paper [Z]. CA: Matrix Semiconductor, Inc, 2005.

[14] Herner S B, Bandyopadhyay A, Dunton S V, et al. Vertical pin polysilicon diode with antifuse for stackable field-programmable ROM [J]. IEEE Electron Device Letters, 2004, 25(5): 271-273.

[15] Jia S, Li H, Gotoh T, et al. Ultrahigh drive current and large selectivity in GeS selector [J]. Nature communications, 2020, 11(1): 1-9.

[16] Jia S, Li H, Liu Q, et al. Scalability of Sulfur - Based Ovonic Threshold Selectors for 3D Stackable Memory Applications [J]. physica status solidi (RRL) - Rapid Research Letters, 2021, 15(6): 2100084.

[17] Adinolfi V, Laudato M, Clarke R, et al. ALD Heterojunction Ovonic Threshold Switches [J]. ACS Applied Electronic Materials, 2020, 2(12): 3818-3824.

[18] Hatem F, Chai Z, Zhang W, et al. Endurance improvement of more than five orders in Ge_xSe_{1-x} OTS selectors by using a novel refreshing program scheme [C]. //2019 IEEE International Electron Devices Meeting (IEDM). IEEE, 2019: 35.2.1-35.2.4.

[19] Yoo C, Kim W, Jeon J W, et al. Atomic Layer Deposition of Ge_xSe_{1-x} Thin Films for Endurable Ovonic Threshold Selectors with a Low Threshold Voltage [J]. ACS Applied Materials & Interfaces, 2020, 12(20): 23110-23118.

[20] Liu G, Li T, Wu L, et al. Increasing Trapped Carrier Density in NanoscaleGeSeAs Films by As Ion Implantation

for Selector Devices in 3D-Stacking Memory [J]. ACS Applied Nano Materials, 2019, 2(9): 5373-5380.

[21] Wang L, Cai W, He D, et al. Performance Improvement of $GeTe_x$-Based Ovonic Threshold Switching Selector by C Doping [J]. IEEE Electron Device Letters, 2021, 42(5): 688-691.

[22] Avasarala N S, Donadio G L, Witters T, et al. Half-threshold bias I off reduction down to nA range of thermally and electrically stable high-performance integrated OTS selector, obtained by Se enrichment and N-doping of thin GeSe layers [C]. //2018 IEEE Symposium on VLSI Technology. IEEE, 2018: 209-210.

[23] Sabry Aly M M, Gao M, Hills G, et al. Energy-Efficient Abundant-Data Computing: The N3XT 1,000x [J]. Computer, 2015, 48(12): 24-33.

[24] Cai F, Correll J M, Lee S H, et al. A fully integrated reprogrammable memristor-CMOS system for efficient multiply-accumulate operations [J]. Nature Electronics, 2019, 2(7): 290-299.

[25] Luo Q, Chen B, Cao R, et al. Complementary memory cell based on field-programmable ferroelectric diode for ultra-low power current-SA free BNN applications [C]. //2019 IEEE International Electron Devices Meeting (IEDM). IEEE, 2019: 38.5.1-38.5.4.

[26] Li Y, Tang J, Gao B, et al. Monolithic 3D Integration of Logic, Memory and Computing-In-Memory for One-Shot Learning [C]. //2021 IEEE International Electron Devices Meeting (IEDM). IEEE, 2021: 21.5.1-21.5.4.

[27] Luo J, Xu W, Du Y, et al. Energy-and Area-efficient Fe-Fin FET-based Time-Domain Mixed-Signal Computing In Memory for Edge Machine Learning [C]. //2021 IEEE International Electron Devices Meeting (IEDM). IEEE, 2021: 19.5.1-19.5.4.

[28] Nair V, Hinton G E. Rectified linear units improve restrictedboltzmann machines [C]. Icml, 2010.

[29] Hendrycks D, Gimpel K. Gaussian error linear units (gelus) [J]. arXiv preprint arXiv: 1606.08415, 2016.

[30] LeCun Y, Bottou L, Bengio Y, et al. Gradient-based learning applied to document recognition [J]. Proceedings of the IEEE, 1998, 86(11): 2278-2324.

[31] Samatin Njikam A N, Zhao H. A novel activation function for multilayer feed-forward neural networks [J]. Applied Intelligence, 2016, 45(1): 75-82.

[32] Xu B, Huang R, Li M. Revise saturated activation functions [J]. arXiv preprint arXiv: 1602.05980, 2016.

[33] Kong S, Takatsuka M. Hexpo: A vanishing-proof activation function [C]. //2017 International Joint Conference on Neural Networks (IJCNN). IEEE, 2017: 2562-2567.

[34] Maas A L, Hannun A Y, Ng A Y. Rectifier nonlinearities improve neural network acoustic models [C]. //Proc. icml. 2013, 30(1): 3.

[35] He K, Zhang X, Ren S, et al. Delving deep into rectifiers: Surpassing human-level performance onimagenet classification [C]. //Proceedings of the IEEE international conference on computer vision. 2015: 1026-1034.

[36] Shang W, Sohn K, Almeida D, et al. Understanding and improving convolutional neural networks via concatenated rectified linear units [C]. //international conference on machine learning. PMLR, 2016: 2217-2225.

[37] Liew S S, Khalil-Hani M, Bakhteri R. Bounded activation functions for enhanced training stability of deep neural networks on visual pattern recognition problems [J]. Neurocomputing, 2016, 216: 718-734.

[38] Duggal R, Gupta A. P-TELU: parametric tan hyperbolic linear unit activation for deep neural networks [C]. // Proceedings of the IEEE International Conference on Computer Vision Workshops. 2017: 974-978.

[39] Clevert D A, Unterthiner T, Hochreiter S. Fast and accurate deep network learning by exponential linear units (elus) [J]. arXiv preprint arXiv: 1511.07289, 2015.

[40] Klambauer G, Unterthiner T, Mayr A, et al. Self-normalizing neural networks [J]. Advances in neural information processing systems, 2017, 30.

[41] Barron J T. Continuously differentiable exponential linear units [J]. arXiv preprint arXiv: 1704.07483, 2017.

[42] Trottier L, Giguere P, Chaib-Draa B. Parametric exponential linear unit for deep convolutional neural networks [C]. //2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA). IEEE, 2017: 207-214.

[43] Li Y, Fan C, Li Y, et al. Improving deep neural network-with multiple parametric exponential linear units [J]. Neurocomputing, 2018, 301: 11-24.

[44] Agostinelli F, Hoffman M, Sadowski P, et al. Learning activation functions to improve deep neural networks [J]. arXiv preprint arXiv: 1412.6830, 2014.

[45] Mishra A, Chandra P, Ghose U, et al. Bi-modal derivative adaptive activation function sigmoidal feedforward artificial neural networks [J]. Applied Soft Computing, 2017, 61: 983-994.

[46] Qian S, Liu H, Liu C, et al. Adaptive activation functions

in convolutional neural networks [J]. Neurocomputing, 2018, 272: 204-212.

[47] Dugas C, Bengio Y, Bélisle F, et al. Incorporating second-order functional knowledge for better option pricing [J]. Advances in neural information processing systems, 2000, 13.

[48] Hendrycks D, Gimpel K. Gaussian error linear units (gelus) [J]. arXiv preprint arXiv: 1606.08415, 2016.