45nm 体硅工艺下使用双-栅氧化层厚度降低 SRAM 的泄漏功耗

杨 松[1,2,†] 王 宏[1] 杨志家[1]

(1 中国科学院沈阳自动化研究所，沈阳 110016)
(2 中国科学院研究生院，北京 100049)

摘要：提出了一种在 45nm 体硅工艺下使用双-栅氧化层厚度来降低整体泄漏功耗的方法.所提方法具有不增加面积和延时、改善静态噪声边界、对 SRAM 设计流程的改动很小等优点.提出了三种新型的 SRAM 单元结构，并且使用这些单元设计了一个 32kb 的 SRAM，仿真结果表明，整体泄漏功耗可以降低 50%以上.

关键词：栅极泄漏电流；SRAM；栅氧化层厚度；静态噪声边界
EEACC：2560F；2570D
中图分类号：TN402 **文献标识码**：A **文章编号**：0253-4177(2007)05-0745-05

1 引言

随着工艺尺寸的不断缩小，超大规模集成电路(very large-scale integration，VLSI)的性能和晶体管的密度得到了极大的提高，但同时芯片的整体泄漏功耗也会随之快速的增长.在近年的微处理器设计中，为了提高整体性能，芯片内部的存储器容量在飞速增长.根据 2006 年 ITRS[1] 的预测，到 2013 年存储器的面积将会占据整个存储器芯片的 90%以上.在一个存储器(通常为静态随机存储器，简记为 SRAM)占据如此大比例的芯片中，SRAM 的泄漏电流会对整个芯片的泄漏功耗起着决定性的作用.因此，降低 SRAM 的泄漏电流已经成为 VLSI 设计中必须要考虑的关键问题.国内外有大量的研究者都在尝试解决这个难题，Azizi[2] 等人提出了使用不对称的 SRAM 单元来减小泄漏电流的方法，这种方法主要是基于数据和指令存储器在待机状态下存储的状态多数为“0”这一特点而提出来的；Nii 等人[3] 则提出了使用与二极管并联的 PMOS 偏置晶体管来控制虚拟地的电压，从而减小 SRAM 单元泄漏电流的方法；Takeyama 等人[4] 设计了一个由复制晶体管所组成的偏置产生电路来给单元阵列提供待机状态时的偏置电压，这个技术可以有效地抑制制造过程中可能造成的阈值电压的波动，保证在各种工艺及制造条件下均可以最大化地减小泄漏电流.

然而，上面提到的低泄漏功耗 SRAM 的设计方法，大多数都会增加 SRAM 的面积，造成硬件的额外开支.本文提出了一种在 45nm 体硅工艺下通过对组成 SRAM 单元的晶体管使用不同栅氧化层厚度的技术来减小 SRAM 泄漏功耗的方法.此方法有不增加硬件开支、不增加 SRAM 整体延时、改善静态噪声边界以及对 SRAM 的设计流程改动很小等优点.

2 SRAM 结构

一个典型的 SRAM 结构如图 1 所示，它是由单元阵列、地址解码器、列选择器、敏感放大器、I/O 以及一个控制电路等模块所组成的.每一个模块的具体功能以及设计方法在相关的文献[5] 中都有详细的介绍.其中，单元阵列模块是用来存储数据的核心模块，在 SRAM 中占据的面积最大，消耗的泄漏功耗也最多.传统的由 6 个晶体管组成的 SRAM 单元(组成阵列的基本单元)电路如图 2 所示.

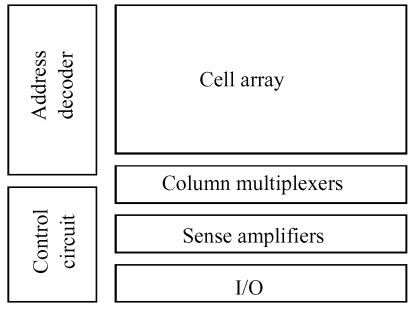

图 1 SRAM 模块框图
Fig.1 An SRAM block

† 通信作者.Email：yangsong@sia.cn
2006-11-13 收到，2007-01-16 定稿

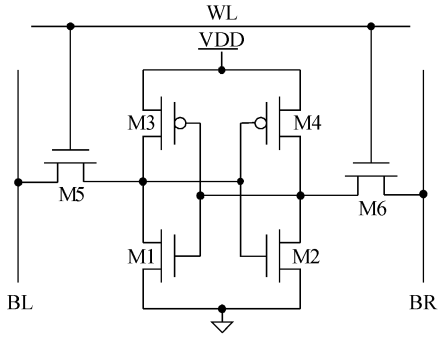

图 2 6 管 SRAM 单元
Fig. 2 A 6T SRAM cell

3 泄漏电流成分及其减小技术

深亚微米级 CMOS 晶体管的泄漏电流包括三个主要的成分:结隧道泄漏电流、亚阈值泄漏电流以及栅极泄漏电流.接下来分别对每一个成分进行简要的论述.

3.1 结隧道泄漏电流

反向偏置 p-n 结隧道泄漏电流的产生主要有两个原因:一个是耗尽区边缘少数载流子的扩散,另一个则是由于反偏结耗尽区电子-空穴对的产生.结隧道泄漏电流是与结掺杂以及通过结的反向偏置电压呈指数关系,最常见的减小结隧道泄漏电流的方法就是在结上加正向衬底偏置电压.

3.2 亚阈值泄漏电流

亚阈值泄漏电流是出现在晶体管栅极到源极的电压低于阈值电压时,工作在弱反型状态下的漏极到源极的电流.基于 BSIM4 的模型[6],亚阈值泄漏电流的计算公式为:

$$I_{sub} = I_0 \exp\left(\frac{V_{gs} - V_{th}}{nkT/q}\right)\left(1 - \exp\left(\frac{-V_{ds}}{kT/q}\right)\right) \tag{1}$$

式中 V_{gs}和 V_{ds}分别为栅极与源极以及漏极与源极之间的电压;V_{th}为阈值电压.由(1)式可以看出,亚阈值泄漏电流会随着阈值电压的增长而呈指数增加,因此,减小亚阈值泄漏电流最有效和最常见的方法就是增加晶体管的阈值电压.

3.3 栅极泄漏电流

随着 CMOS 体硅工艺逐渐缩小到 45nm,栅氧化层厚度已经变得非常薄,达到 1.2～1.6nm[1].栅氧化层厚度的变薄会造成通过氧化层的电场增加,较高的电场加上很薄的栅氧化层就会导致在 nMOS 晶体管的沟道和栅极之间有电子隧道(对于 pMOS 就是空穴隧道)产生,这被称为栅氧化层隧道电流,也是栅极泄漏电流最主要的组成成分.栅氧化层隧道电流密度的简化计算公式为[7]:

$$J_T = A\left(\frac{V_{ox}}{T_{ox}}\right)^2 \exp\left(\frac{-B\left(1-\left(1-\frac{V_{ox}}{\phi_{ox}}\right)^{\frac{3}{2}}\right)}{\frac{V_{ox}}{T_{ox}}}\right) \tag{2}$$

其中 J_T 代表隧道电流密度;V_{ox}为栅氧化层电压;ϕ_{ox}是隧道粒子(电子或空穴)的势垒高度;T_{ox}代表栅氧化层的厚度,A 和 B 是与工艺相关的常数.减小栅极泄漏电流的常见技术包括增加栅氧化层厚度、减小栅电压以及使用具有较高相对介电常数(high-k)的材料来代替 SiO_2 栅氧化层等.然而,减小栅电压的技术往往要引入控制电路,会造成面积的增加;使用 high-k 材料的方法同样受到严重的限制,这是因为对于特征尺寸在 65nm 及以下的器件中,源极和漏极区域的结必须要非常的浅[1],这种非常浅的结是在 450～800℃ 的温度下通过对离子注入的掺杂进行后退火而实现的.不幸的是,最有可能替代 SiO_2 的 high-k 材料,如 ZrO_2 和 HfO_2 等在温度低于 500℃时会出现晶化的现象,而 Ta_2O_5 与硅衬底分界面的热稳定性又比较差.因此,尽管使用 high-k 材料是未来减小栅极泄漏电流的必然趋势,但就目前半导体制造工艺而言,实现起来仍然有相当大的难度.

4 低泄漏功耗 SRAM 单元结构

一个传统的 6 管 SRAM 单元在待机状态下的泄漏电流流向如图 3 所示.有研究表明[8],随着体硅工艺尺寸逐渐缩小到 45nm,栅极泄漏电流将占据整体泄漏电流的 80% 以上.结隧道泄漏电流(图 3

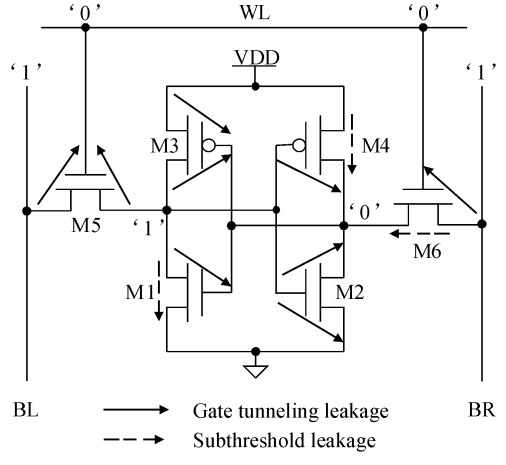

图 3 6 管 SRAM 单元待机状态泄漏电流
Fig. 3 Leakage current of a 6T SRAM cell in standby mode

并没有画出)和亚阈值泄漏电流在整体泄漏电流中的比例已经变得非常小,几乎可以被忽略,因此,本文没有作特殊的考虑.基于 3.3 节中提到的原因,我们选择了通过增加栅氧化层厚度来降低 SRAM 单元栅极泄漏电流的方法,此方法不会增加面积和整体延时,对 SRAM 设计流程的改动也很小.采用 SiO_2 作为栅氧化层,pMOS 晶体管与 nMOS 晶体管相比栅极泄漏电流会小一个数量级以上[8],通过增加 pMOS 晶体管的栅氧化层厚度而得到的泄漏功耗节省是很小的.因此,为了减小 SRAM 单元的栅极泄漏电流,通过增加传输和下拉 nMOS 晶体管栅氧化层的厚度就可以很好地实现.众所周知,每增加一个不同厚度的栅氧化层就必须在制造的过程中额外增加一层掩模版.为了节约制造成本,采用了双-栅氧化层厚度的技术来进行低泄漏功耗 SRAM 单元的设计.

为了便于 SRAM 单元的制造及应用,此次设计的单元全部采用对称结构,即在对称位置上的晶体管具有相同的栅氧化层厚度.这样,一个 SRAM 单元就可以有四种不同的电路结构,如图 4 所示.A 为原始的所有晶体管全部使用薄栅氧化层的结构,B 结构是将传输晶体管和下拉晶体管全部替换成厚栅氧化层的情况,C 结构是仅对两个传输晶体管使用厚栅氧化层进行替换的情况,D 则是仅将两个下拉晶体管替换成厚栅氧化层的结构.这些新型的 SRAM 单元结构,会导致泄漏功耗、读和写操作延时以及静态噪声边界等性能指标发生变化.

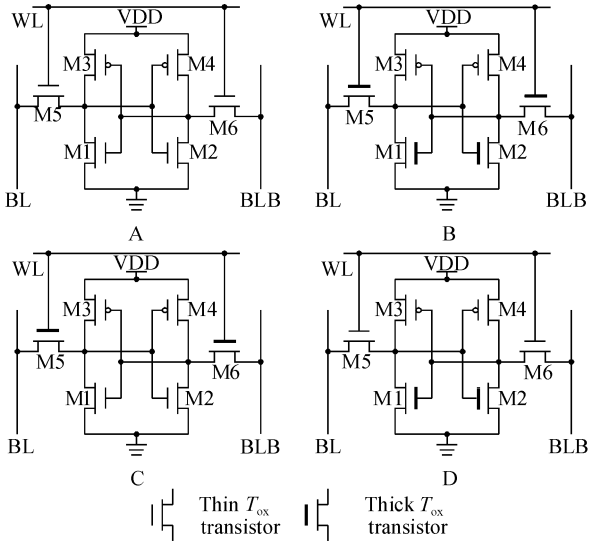

图 4　几种不同的 SRAM 单元结构

Fig.4　Several SRAM cell configurations

5　实验与分析

实验结果是使用 HSPICE[9] 采用最新版本的 45nm 体硅工艺 BSIM4 模型[6](加入了栅极泄漏电流的准确预估模型)仿真得到的.晶体管的阈值电压全部为 0.22V,薄的栅氧化层厚度为 1.4nm,厚的栅氧化层为 1.6nm,供电电压为 1.0V,仿真温度根据情况的不同分别为 27 和 110℃.

5.1　SRAM 单元的仿真

5.1.1　泄漏功耗的降低

表 1 和图 5 分别给出了在 27℃(室温)和 110℃(最坏情况)时,四种电路结构的泄漏电流值以及整体泄漏功耗的量化对比.通过表 1 可以知道当温度为 27℃时,B,C,D 结构的整体泄漏功耗与原始的 A 结构相比分别降低了 77.84%,21.28% 及 56.43%;而在 110℃的情况下,B,C,D 结构的泄漏功耗节省则分别为 75.95%,21.32%及 54.09%.从表中还可以知到,当温度从 27℃升高到 110℃之后,整体泄漏功耗平均只增加了 10.06%,这是因为亚阈值泄漏电流是由载流子的扩散而形成的,会随着温度的升高而呈指数的增长;而隧道电流是由电子穿过势垒电压形成的,由于通过栅氧化层的电场与温度并没有直接的关系,因此栅极泄漏电流对温度基本上是不敏感的.在 45nm 体硅工艺下,栅极泄漏电流已经在整体泄漏电流中占据了支配的地位,随着温度的升高而导致的整体泄漏功耗的轻微增加,主要是由处于次要地位的亚阈值泄漏电流的指数增长所造成的.

表 1　四种 SRAM 单元结构的泄漏电流　　nA

Table 1　Leakage current of four configurations　nA

Type	27℃			110℃		
	Sub	Gate	Total	Sub	Gate	Total
A	0.53	312.30	312.83	20.36	312.47	332.83
B	0.27	69.06	69.33	10.12	69.91	80.03
C	0.35	245.92	246.27	15.34	246.53	261.87
D	0.46	135.85	136.31	16.86	135.93	152.79

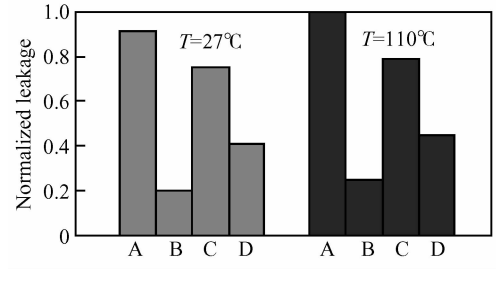

图 5　四种结构整体泄漏功耗的量化对比

Fig.5　Contrast of normalized leakage for four configurations

5.1.2　读操作延时的增加

由 BSIM4 模型中的公式可知,增加栅氧化层的

厚度会使晶体管的阈值电压升高，这会抑制部分亚阈值泄漏电流，但同时阈值电压的升高也会造成电路读和写操作延时的增加.由于读操作的时间往往会比写操作的时间长很多，因此，电路结构对读操作延时的影响就显得尤为重要.图 6 给出了在 110℃的温度下，B，C，D 三种结构与原始的 A 结构相比读操作延时增加的百分比.从图中可以看出，对全部 NMOS 晶体管采用厚栅氧化层(B)和只对 nMOS 传输管采用厚氧化层(C)的结构都有比较大的延时增加，分别增加了 10.13%和 9.89%；而采用只对下拉 nMOS 晶体管使用厚氧化层(D)的结构几乎没有延时的增加，只增加了 0.65%.需要注意的是，这里的仿真结果是假设在相同的字线和位线电容的条件下获得的，事实上，对于 B 和 C 结构，由于对传输晶体管采用了比较厚的栅氧化层，将会减小部分字线以及位线上的电容，也就意味着会在一定程度上减小字线和位线的延时.换句话讲，在实际的 SRAM 设计中，采用 B 或 C 结构所增加的延时会比图 6 所示的延时增加要小一些.

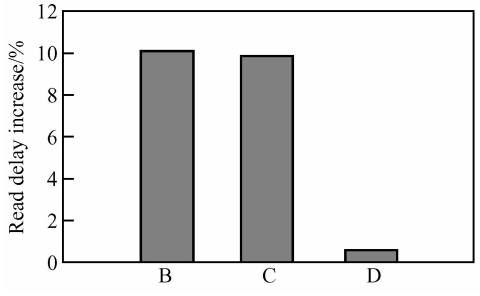

图 6 几种结构读操作延时的增加

Fig.6 Read delay increase for each configuration

5.1.3 静态噪声边界的改善

一个 SRAM 单元的静态噪声边界(static noise margin，SNM)被定义为可以使单元内的状态发生翻转所需的最小直流噪声电压.SRAM 单元在读操作时对噪声尤其敏感，这是因为此时单元内部存储“0”的节点会由于下拉 nMOS 晶体管和传输晶体管各自电阻的分压而升高到一个高于地节点电压的值，如果这个电压值足够高，单元内所存储的值就会发生改变.这里分别在待机状态下和读操作时对所有的结构进行了仿真，仿真的温度为 110℃.通过仿真计算得到 SNM 的结果见表 2.

由表 2 可以看出，无论是在待机模式还是在读操作时，文中提出的各种单元结构的 SNM 都有了一定程度的改善.在待机模式下，B，C，D 结构相对于 A 分别增加了 28.3%，3.4%以及 27%；而在读操作时，B，C，D 结构的 SNM 改善的程度和待机模式稍有不同，与 A 结构相比分别增加了 55.9%，33.3%以及 7.8%.

表 2 待机状态和读操作的 SNM 对比

Table 2 Contrast of SNM in standby and read mode

Type	A	B	C	D
Standby mode SNM/mV	237	304	245	301
Read mode SNM/mV	102	159	136	110

5.2 低泄漏功耗 SRAM 的设计及仿真

为了检验提出的电路结构能否有效降低泄漏功耗，我们按照图 1 的结构框图设计了一个 1GHz、32kb 的 SRAM，原始的 SRAM 单元全部使用阈值电压为 0.22V，薄栅氧化层厚度为 1.4nm 的晶体管.

5.2.1 SRAM 单元的替换

从前面的仿真结果中可以看出，C 结构有着与 B 结构相当的读操作延时的增加，而泄漏功耗的节省能力与 B 和 D 结构相比却差很多，因此，在此次低功耗 SRAM 的设计中没有使用 C 结构，也就是说，只有 B 和 D 两种结构可以用来替换原始的 A 结构.在此次设计中，B 和 D 结构中的厚栅氧化层厚度采用 1.6nm.

具体的替换过程如下：首先，找到 SRAM 中最慢的读操作和写操作延时.由于 B 结构具有最小的泄漏功耗消耗，所以应首先考虑使用 B 结构进行替换，之后再考虑 D 结构.如果 B 结构自身的访问延时没有比最慢的 SRAM 单元的访问延时更长，那么这个单元就可以使用 B 结构来代替；如果 B 结构的延时无法满足，而 D 结构的访问延时小于最慢单元的访问延时，就使用 D 结构进行替换，否则，不进行替换.如此反复的进行下去，直到所有的 SRAM 单元替换完为止.为了加速替换的过程，我们并没有对所有的 SRAM 单元都一一进行比较替换，而是采取了分块替换的方法，即只要某备选块中最慢的单元都可以满足延时的要求，那么整个模块就会被替换掉.最终，在 SRAM 单元阵列中，A，B 和 D 结构所占的比例分别为 10%，37%以及 53%.由于增加栅氧化层的厚度不会造成单元版图上任何端口位置的变化，因此，这种替换的方法可以在不影响版图布局布线的情况下轻松地实现.

5.2.2 SRAM 的整体仿真

单元替换完成以后，仍然通过 HSPICE 采用 45nm 体硅工艺 BSIM4 模型在供电电压为 1.0V、温度为 110℃的条件下进行整体仿真.由于在 45nm 的工艺条件下，互连线延时已经不可以被忽略，因此所有的局部和全局互连，包括位线、字线以及解码器的连线等，都使用了分布式的 RC 电路对其进行了模型化.通过仿真得到，SRAM 整体泄漏功耗的消

耗与原始结构相比降低了 54.8%.

6　结论

提出了一种在 45nm 体硅工艺下使用双-栅氧化层厚度技术实现低泄漏功耗 SRAM 设计的方法，这种方法可以在不降低性能的前提下通过增加部分 SRAM 单元内关键晶体管的栅氧化层厚度来实现泄漏功耗的降低.提出了三种新型的 SRAM 单元结构，分别基于泄漏功耗的降低、延时的增加以及静态噪声边界的改善等性能指标进行了仿真和分析比较.最后，设计了一个 1GHz、32kb 的 SRAM，并使用新结构进行了适当的替换，仿真结果表明，SRAM 的整体泄漏功耗减小了 54.8%左右.未来的研究重点包括 SRAM 单元替换算法的改进以及在实际的 45nm 工艺线上完成文中所设计的双-栅氧化层厚度的 SRAM 的流片，进行芯片测试等.

参考文献

[1] http://www.itrs.net/Links/2006Update/2006UpdateFinal.htm

[2] Azizi N, Najm F N, Moshovos A. Low-leakage asymmetric-cell SRAM. IEEE Trans Very Large Scale Integr Syst, 2003, 11(4): 701

[3] Nii K, Tsukamoto Y, Yoshizawa T, et al. A 90-nm low-power 32kB embedded SRAM with gate leakage suppression circuit for mobile applications. IEEE J Solid-State Circuits, 2004, 39(4): 684

[4] Takeyama Y, Otake H, Hirabayashi O, et al. A low leakage SRAM macro with replica cell biasing scheme. IEEE J Solid-State Circuits, 2006, 41(4): 815

[5] Hodges D A, Jackson H G, Saleh R A. Jiang Anping, Wang Xin'an, Chen Zili, et al. Translation. Analysis and design of digital integrated circuits. In: Deep Submicron Technology. Third Edition. Beijing: Publishing House of Electronics Industry, 2005(in Chinese)[Hodges D A, Jackson H G, Salehd R A. 蒋安平，王新安，陈自力. 等译. 数字集成电路分析与设计——深亚微米工艺. 第三版. 北京：电子工业出版社，2005]

[6] http://www-device.eecs.berkeley.edu/~bsim3/bsim4.html

[7] Mukhopadhyay S, Neau C, Cakici R T, et al. Gate leakage reduction for scaled devices using transistor stacking. IEEE Trans Very Large Scale Integr Syst, 2003, 11(4): 716

[8] Mohanty S P, Velagapudi R, Kougianos E. Dual-K versus dual-T technique for gate leakage reduction: a comparative perspective. ISQED, 2006: 564

[9] http://synopsys.com/products/mixedsignal/hspice/hspice.html

Reducing Leakage of SRAM Using Dual-Gate-Oxide-Thickness Transistors in 45nm Bulk Technology

Yang Song[1,2,†], Wang Hong[1], and Yang Zhijia[1]

(1 *Shenyang Institute of Automation, Chinese Academy of Sciences, Shenyang* 110016, *China*)
(2 *Graduate University of Chinese Academy of Sciences, Beijing* 100049, *China*)

Abstract: This paper presents a method based on dual-gate-oxide-thickness assignment to reduce the total leakage power dissipation of SRAM in 45nm bulk technology. The proposed technique incurs neither area nor delay overhead and can improve the static noise margin. In addition, it results in a slight change in the SRAM design flow. Three novel SRAM cell configurations are proposed. Simulation results demonstrate that this technique can reduce the total leakage power dissipation of 32kb of SRAM with these configurations by more than 50%.

Key words: gate leakage current; SRAM; gate-oxide-thickness; SNM
EEACC: 2560F; 2570D
Article ID: 0253-4177(2007)05-0745-05

† Corresponding author. Email: yangsong@sia.cn

Received 13 November 2006, revised manuscript received 16 January 2007