doi:10.3772/j.issn.1002-0470.2014.02.010

采用钳位二极管的新型低功耗 SRAM 的设计①

张立军②* 吴 晨** 王子欧* 毛凌锋*

(*苏州大学城市轨道交通学院 苏州 215006)

(**苏州秉亮科技有限公司 苏州 215021)

摘 要 给出了一种设计低功耗静态随机存储器(SRAM)的技术,实现了在电路级与架构级层次上同时降低漏电流与动态功耗。该技术采用源极偏压结构控制漏电流,将一个钳位二极管与 NMOS 管并联插入 GND 与 SRAM 单元的源极之间,当 NMOS 打开时 SRAM 进行正常的读写操作,而 NMOS 关闭则会将源极电压抬高至钳位电压,降低漏电流的同时保证了数据的稳定性;对 SRAM 结构进行独特的布局,引入 Z 译码电路,极大地减少每次操作时激活的存储单元数量,明显降低动态功耗;将 power-gating 技术与高阈值(high-V_{th})器件相结合的低功耗设计应用于外围电路,进一步降低漏电流。基于 UMC 55nm SP CMOS 工艺制造了包含多个 SRAM 实例(instance)的测试芯片,测试结果证明了该技术的有效性与可靠性。

关键词 静态随机存储器(SRAM),低功耗,钳位二极管,漏电流

0 引 言

静态随机存储器(static random access memory,SRAM)是高性能微处理器和便携式设备中非常关键的一部分。降低 SRAM 的功耗将会有效地提高整个系统的性能、可靠性和成本。低功耗、高性能的 SRAM 设计已经引起了业界的广泛关注。

位线充放电等引起的动态功耗一度主导了 SRAM 的能量消耗。随着工艺的不断进步,阈值电压与栅氧厚度的不断减小使漏电流急剧增大,静态功耗所占的比重也随之上升。目前业界已提出多种降低 SRAM 漏电流的技术[1-5]。其中,文献[1]通过抬高未选中行单元的 GND 电压来控制漏电流,其控制信号即为字线选择信号,但浮动的 GND 电压会造成静态噪声容限(static noise margin,SNM)下降,数据易受噪声影响。考虑到位于存储阵列不同位置的单元有着不同的读写延时,文献[3]对整个阵列进行了优化,在不影响器件性能和面积的前提下,采用双阈值(dual-V_{th})和双栅氧厚度(dual-T_{ox})的方法降低漏电流,但这种方法需要更复杂的工艺技术。

本文提出了一套低功耗 SRAM 设计技术,它能够在电路级与架构级层次上同时降低 SRAM 漏电流与动态功耗。首先,基于源极偏压结构,将一个钳位二极管与 n 沟道金属氧化物半导体(NMOS)管并联插入 GND 与 SRAM 单元的源极之间,降低漏电流的同时保证了数据的稳定性。其次,还对 SRAM 结构进行了独特的布局,结合层次化字线与位线技术,极大地减少了每次操作时激活的存储单元数量,明显降低了动态功耗。同时,对外围电路也提出了低功耗设计方案。用 UMC 55nm SP CMOS 工艺制造的测试芯片的测试结果证明了该技术的有效性与可靠性。

1 钳位二极管电路设计

1.1 电路原理

我们采用源极偏压结构来控制存储单元的漏电流,如图 1 所示,在 SRAM 单元的 GND 与 V_{SL}之间连接高阈值 NMOS 晶体管 M7。在工作状态时,M7 打

① 国家自然科学基金(61272105,61076102)资助项目。

② 男,1971 年生,博士,研究员;研究方向:器件可靠性,纳米尺度 SRAM 低功耗、稳定性及可测性技术和设计;联系人,E-mail:zhanglijun@suda.edu.cn

(收稿日期:2013-07-16)

开,其电阻很小,电流从电源电压流至接近实际的 GND,存储单元进行正常的读写操作。在空闲状态时,M7 关闭,源极电压 V_{SL} 升高,可以同时降低亚阈值漏电流与栅漏电流,这两种漏电流在目前的互补型金属氧化物半导体(CMOS)电路中占据着主导地位。

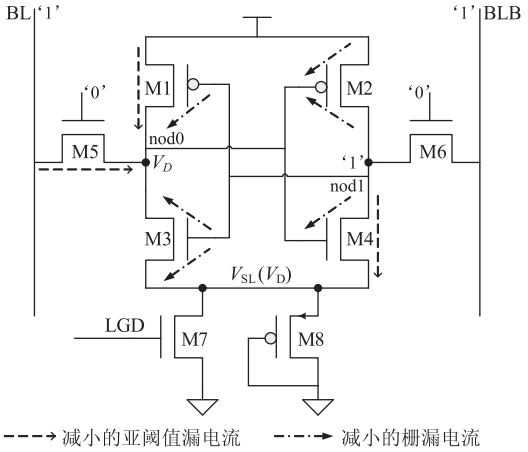

图 1 钳位二极管电路结构

这种结构的缺点是下拉路径中的 M7 会减慢速度,增大面积,同时还会增加动态功耗。因此晶体管 M7 的尺寸主要受三个因素影响:读操作时的噪声容限,读速度以及当存储单元从空闲状态跳转至工作状态时源极电压回到 GND 的时间。为了尽量减小存储器性能损失,M7 的尺寸必须设计得比较大。然而更重要的是,当 M7 关闭时,V_{SL} 被微弱的漏电流拉至一个浮动的正电压,存"0"的节点 V_{nod0} 也随之浮动。此时若存在噪声信号,数据的稳定性将会受到严峻的挑战。为了解决这一问题,如图 1 所示,将一个由 p 沟道金属氧化物半导体(PMOS)晶体管 M8 构成的钳位二极管与 M7 并联,组成钳位单元,使空闲状态下的 V_{SL} 固定在一个稳定的电压值 V_D。V_D 的值由二极管的尺寸和阈值电压决定,它必须低于反向器的翻转电压以保证数据的稳定性,同时在漏电流降低与数据保持特性之间进行折中选择。若存在噪声信号使 V_{SL} 电压升高,此时二极管打开,将 V_{SL} 钳位在 V_D。当存储单元被选中时,M7 打开,V_{SL} 与 V_{nod0} 从 V_D 回到 GND。

在目前的 CMOS 电路中,将 NMOS 的栅漏短接或者将 PMOS 的栅极接至 GND 均可构成 MOS 二极管[6]。文献[7]表明:如果 NMOS 与 PMOS 的尺寸相同,PMOS 电阻更大,它对漏电流的抑制能力更强。此外,仿真结果表明 PMOS 二极管对于压力、体积、温度(PVT)变化表现出更好的特性。因此在本文中采用的是由 PMOS 构成的二极管。

1.2 漏电流降低

在目前的集成电路中,主要有三种漏电流:亚阈值漏电流,栅漏电流和结漏电流,其中前两者占的比例较大。

亚阈值漏电流是指当晶体管的栅源电压小于阈值电压时流过漏—源的漏电流。钳位二极管将空闲状态时的 V_{SL} 升至 V_D,反相器中电压差($V_{DD}-V_D$)减小,同时由于衬底电压不变,抬高源极电压相当于产生衬偏效应,阈值电压增大,亚阈值漏电流减小。此外,源漏电压差减小减弱了漏场感应势垒降低(DIBL)效应,也降低了亚阈值漏电流。

栅漏电流主要包括三个部分:栅 - 源/漏交叠区,栅 - 沟道及栅 - 衬底,它们分别与相应的电压差有关。其中,处于打开状态的 NMOS 晶体管的栅 - 沟道漏电流最关键。由于 V_{nod0} 随 V_{SL} 由"0"升至 V_D,SRAM 单元中大部分晶体管的栅源、栅漏及栅至衬底的电压差都减小,因此图 1 中 M1、M2、M3 与 M4 的栅漏电流均有大幅下降。

1.3 数据保持特性

由于空闲状态下电源电压与源极电压之间的电压差($V_{DD}-V_D$)减小,噪声容限随之降低,但其依然维持在一个可以接受的水平,数据稳定性不会受到影响。在不同温度下对钳位二极管电路进行仿真模拟,存"1"节点 nod1 的电压一直维持在 V_{DD},如图 2

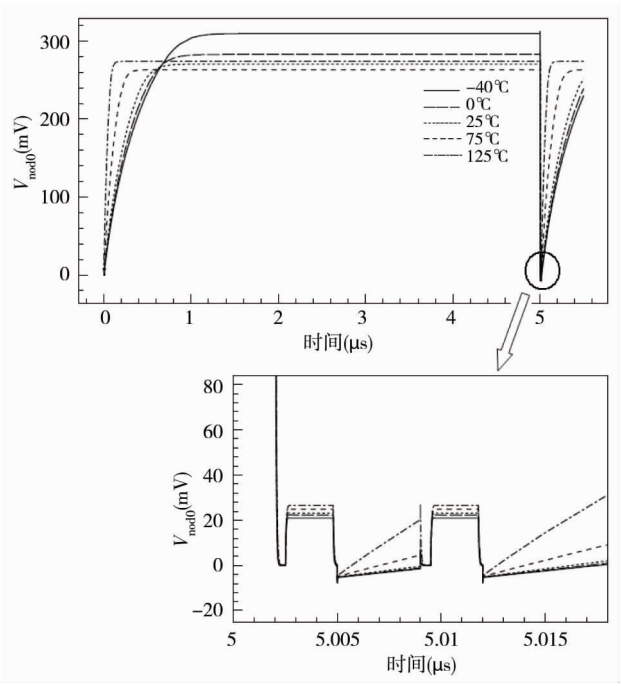

图 2 不同温度下的数据稳定性分析

所示为存“0”节点 nod0 的电压波形。当 NMOS 关闭时,源极电压 V_{SL} 升高,V_{nod0} 也随之抬高。随着温度的升高,漏电流增大,V_{nod0} 被更快地抬高至钳位电压 V_D。当存储单元被选中时,V_{SL} 接至 GND,V_{nod0} 被快速地拉回至“0”进行读写操作(图 2 中的局部放大图),这表明了存储器在不同温度下都能维持数据的稳定性。

2 整体结构布局

本文所设计的 SRAM 是单端口低功耗存储器,在满足一定规则的前提下,可以根据不同的需要产生不同组合的字(word)数,数(bit)位,字节(byte)数与列复用(column multiplexing)数目。

对存储阵列进行分块布局,可以将存储器的工作范围限制在某一子阵列内,显著减少了每次操作时激活的存储单元数量,提高了存储器的读写速度,而且还缩短了字线(WL)和位线(BL)的长度,减小了字线和位线的负载电容,从而降低功耗。但是,考虑到各子阵列的外围译码控制电路等也需消耗功耗,并增加芯片面积,因此需对整个存储阵列进行合适的布局划分。下面以 512kb SRAM(16384 × 32 × 1CM16)为例详细描述其结构布局(图 3)。

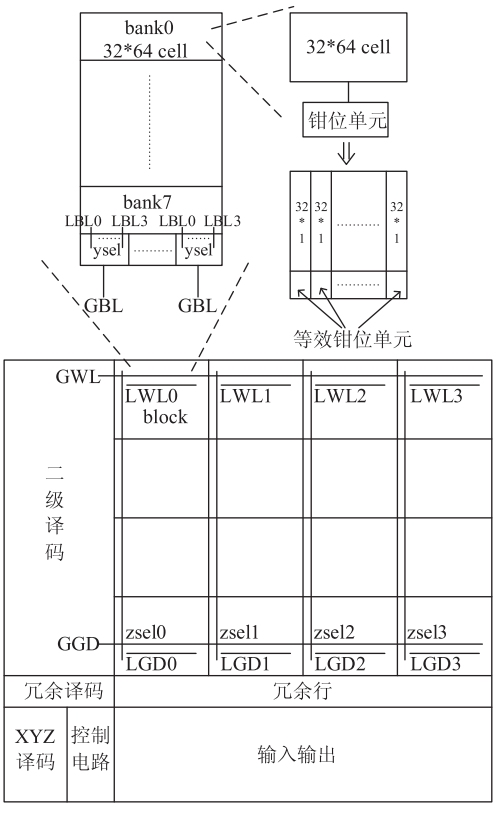

图 3 512kb SRAM 的整体结构布局(右半部分)

图 3 所示为 512kb 存储器整体结构布局(以右半部分为例)。相对于传统的横向(X)与纵向(Y)译码,本文增加了 Z 译码控制。首先将存储器阵列分成四大行,四大列。由 X 译码选择一大行,Z 译码选择一大列,交叉选择一块(block)存储器。每一块单元又可分为 8 个 bank(即 bank0 ~ bank7),亦是由 X 方向译码控制,每个 bank 包含 32 × 64 个基本单元,如图 3 中的左上部分所示。Y 方向译码则是通过列复用电路从每一大列中选择存取的列单元。由此可知,Y 译码与 Z 译码都是用于控制列单元选择的。X、Y、Z 译码器结合选中存取单元。

与架构布局相应的层次化字线与位线技术也应用在本文中。如图 3 所示,字线与位线均分为全局的和局部的。Z 译码信号 zsel0 – zsel3 从四大列单元中选中一列,即将全局字线(GWL)接至局部字线 LWL0 – LWL3 中的一个。Y 译码信号 ysel 在每一大列电路中进行局部位线选择,即将全局位线(GBL)信号传输至选中的局部位线(LBL)。

通过增加 Z 译码电路,结合层次化字线与位线技术,每次读写操作时选中的字线数目降为原来的 1/4,极大地减少了对位线充放电的存储单元的数目,不仅在很大程度上降低了动态功耗,还降低了由于过多单元漏电流引起的数据错误翻转的可能性。

为了节省面积,一个 bank 共用一个钳位单元。同时为防止升高的源极电压 V_D 随着存储器的容量而变化,以使钳位单元可以更方便地应用到不同容量的 SRAM 中,将每个钳位单元等效成 64 等份并联而成,即每一列(32 × 1)单元共用一小份,如图 3 中右上部分所示。二极管的尺寸和阈值电压决定了 V_D 的值,而该值不会随着位数的变化而波动。钳位单元的选择与局部字线选择类似,也是由 Z 译码信号 zsel0 – zsel3 从全局信号(GGD)选择局部信号(LGD)。由于在该结构中,一个 bank 中所有等效钳位单元的栅极控制都连至 LGD,由 X、Z 译码器结合选中一个 bank 即选中其钳位单元。

为了提高 SRAM 的可靠性,我们还在其中加入了冗余行(redundant rows),若存储器中发现错误,则将冗余行代替错误行,提高芯片良率。

3 外围电路设计

为了进一步降低 SRAM 的功耗,我们对整个存储器结构进行了仿真,发现除了存储阵列外,字线驱动电路占了总漏电流中比较大的比例。因此对于外

围电路设计,我们着重考虑字线驱动电路的静态功耗优化。

对字线驱动电路进行仿真,发现当电路处于空闲状态(WL = 0)时,最后两级反向器占了总漏电流的主要部分。如图 4(a)所示,此时 XP1 与 XM2 关闭,漏电流主要由这两个晶体管的亚阈值漏电流与栅漏电流构成。因此我们所采用的低功耗技术也是主要针对它们的。

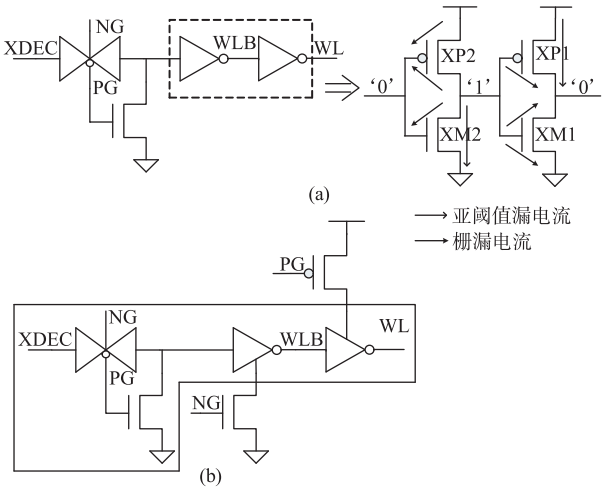

图 4 字线驱动电路漏电流分布(a)和 power gating 降漏电流方法(b)

本文主要采用两种方法来降低其漏电流:一,将关闭晶体管 XP1 与 XM2 的沟道长度 L 从 60nm 增大至 65nm,提高其阈值电压,以将亚阈值漏电流降低 35%;二,如图 4(b)所示,对 XP1 与 XM2 进行 power gating,即在空闲状态时 NG = 0、PG = 1,在 XP1 的源极与 V_{DD} 之间连接关闭的 PMOS,在 XM2 的源极与 GND 之间连接关闭的 NMOS,产生类似堆叠效应(stacking effect)[8] 的效果,漏电流可以减小 75%。

对于其它外围电路,我们也采用增大晶体管沟道长度的方法并结合高阈值器件来降低功耗,这种做法对存储器的性能没有太大影响。

4 测试结果

为了验证该技术的可靠性与有效性,我们采用 UMC 55nm SP CMOS 工艺制造了 58 片测试芯片,每片包含 11 个不同容量的 SRAM 实例。512kb SRAM 电路由于增加了横向的钳位单元与纵向的 Z 译码电路,面积为 $779.4\text{m} \times 633.6\text{m} = 0.49\text{mm}^2$,相对于现有的 SRAM 结构增大了 23%。存储阵列在横向与纵向均分成 4 份,之间则是局部字线与位线的驱动电路。下面叙述芯片电压稳定性与功耗测试的结果。

4.1 最低工作电压

在常温下对测试芯片进行工作电压扫描,通过 SRAM 中数据的翻转特性确定其最低工作电压(VCC_min),测试结果如图 5 所示,证明了 SRAM 的可靠性。

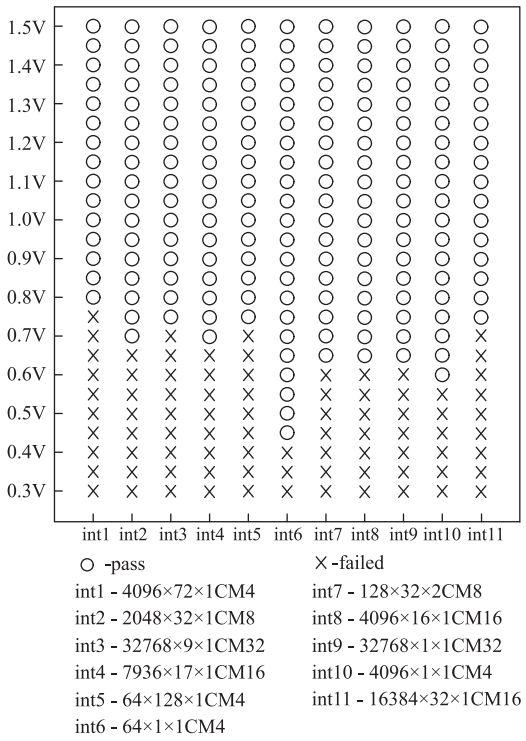

图 5 测试芯片良率

4.2 功耗降低

如前文所述,电压差($V_{DD} - V_D$)的减小和 NMOS 的衬偏效应降低了亚阈值漏电流。同时升高源极电压减小了晶体管的栅源和栅漏电压差,降低了栅漏电流。虽然钳位单元也会引起新的漏电流,但由于多个 SRAM 存储单元共用一个钳位单元,且其尺寸比 6T - SRAM 中的晶体管大得多,因此相比于减小的漏电流,它可以忽略不计。

此外,由于本文在存储器结构布局方面增加了 Z 译码电路,减少了每次操作时半选择(half-selected)存储单元的数量,因此极大地降低了位线充放电功耗。

表 1 与表 2 所示分别为常温下测得的部分 SRAM 的漏电流与动态电流,并与现有 SRAM 结构

进行比较,降低比率表明该技术的有效性。

表 1　SRAM 漏电流比较

Instance	漏电流(μA)		降低比率(%)
	本文设计	现有 SRAM 结构	
7936 × 17 × 1CM16	136.644	357.997	61.8
16384 × 32 × 1CM16	396.600	1126.300	64.8
32768 × 1 × 1CM32	109.435	344.370	68.2
4096 × 16 × 1CM16	59.150	177.120	66.6
2048 × 32 × 1CM8	63.020	201.220	68.9
32768 × 9 × 1CM32	282.200	673.320	58.1
128 × 32 × 2CM8	49.120	86.860	43.4

表 2　SRAM 动态电流比较

Instance	动态电流(mA/MHz)		降低比率(%)
	本文设计	现有 SRAM 结构	
7936 × 17 × 1CM16	0.010	0.016	37.5
16384 × 32 × 1CM16	0.019	0.030	36.7
32768 × 1 × 1CM32	0.005	0.008	37.5
4096 × 16 × 1CM16	0.009	0.015	40
2048 × 32 × 1CM8	0.013	0.017	23.5
32768 × 9 × 1CM32	0.014	0.017	17.6
128 × 32 × 2CM8	0.018	0.022	18.2

4.3　读写性能

由于钳位单元在不同的状态之间进行转换会增加延时,因此存储单元的读写操作性能会受到一定影响。表 3 所示为相对于现有结构,不同设计 SRAM 的读写延时增加比较。

表 3　读写延时增加比较

操作	延时增加		
	本文设计	IWL-VC SRAM[9]	A-SRAM[10]
读操作	3%	2.42%	4%
写操作	0	4.4%	4%

对于读操作而言,读取时间主要决定于两条位线之间产生足够电压差(如 100mV)的时间。如图 1 所示,M7 的存在会增大 BL 放电的时间,其读取时间相对于现有 SRAM 单元增加了 3%。

对于写操作而言,若将“0”写入 nod1,则需将位线 BLB 拉至低电平,此时电源电压通过 M2 与 M6 对 BLB 放电直至 V_{nod1} 使对面的反相器翻转为止。由此可知,钳位单元对写操作的影响比较小,写入时间几乎维持不变。

文献[9]提出的 IWL-VC SRAM 采用了两个传输晶体管对 GND 电压进行偏置,并且采用 PMOS 降低字线电压,增大了字线的有效电阻,读写延时分别增加了 2.42% 和 4.4%。文献[10]提出的不对称 SRAM 结构在存储单元的反相器中插入了一个 NMOS 以降低高电平电压,达到减小存“0”单元漏电流的目的,但同时也增大了传导路径中的有效电阻,因此降低了存取速度。通过将本文设计与其他两种结构的 SRAM 进行比较,可知我们所提出的 SRAM 结构最大的优点在于降低漏电流与动态功耗的同时,把存储器的性能损失控制在了 3% 以内,而且该技术能够切实应用于实际工程中。

5　结　论

本文提出了一种在电路级与架构级层次上同时降低 SRAM 漏电流与动态功耗的技术。在电路层次上,采用源极偏压结构控制漏电流,将钳位单元插入 GND 与 SRAM 单元的源极之间,降低漏电流的同时保证数据的稳定性。在架构层次上,将整个存储阵列进行分块布局,引入 Z 译码电路,结合层次化字线与位线结构,减少每次操作时的半选择单元数量进而降低动态功耗。

另外,本文还对外围电路提出了低功耗设计方案。并将新的 SRAM 设计在 UMC 55nm SP CMOS 工艺下进行流片,通过测试不同容量 SRAM 实例结果表明,该技术最多可以降低 68.9% 的漏电流与 40% 的动态功耗,且该设计在实现低功耗目的的同时能够将器件(cell)性能损失控制在 3% 以内。

参考文献

[1] Agarwal A, Li H, Roy K. A single-Vt low-leakage gated-ground cache for deep submicron. *IEEE Journal of Solid-State Circuits*, 2003, 38(2): 319-328

[2] Zhang L, Wu C, Ma Y, et al. Leakage power reduction techniques of 55nm SRAM cells. *IETE Technical Review*, 2011, 29(2): 135-145

[3] Amelifard B, Fallah F, Pedram M. Leakage minimization of SRAM cells in a dual-Vt and dual-Tox technology. *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, 2008, 16(7): 851-859

[4] Islam A, Hasan M. Leakage characterization of 10T SRAM cell. *IEEE Transactions on Electron Devices*, 2012, 59(3):631-638

[5] Calimera A, Macii A, Macii E, et al. Design techniques and architectures for low-leakage SRAMs. *IEEE Transactions on Circuits and Systems—I*, 2012, 59(9):1992-2007

[6] Levacq D, Dessard V, Flandre D. Low leakage SOI CMOS static memory cell with ultra-low power diode. *IEEE Journal of Solid-State Circuits*, 2007, 42(3):689-702

[7] Hua C, Cheng T, Hwang W. Distributed data-retention power gating techniques for column and row co-controlled embedded SRAM. In: Proceedings of IEEE International Workshop on Memory Technology, Design, and Testing, Taipei, China, 2005. 129-134

[8] Roy K, Mukhopadhyay S, Mahmoodi-meimand H. Leakage current mechanisms and leakage reduction techniques in deep sub-micrometer CMOS circuits. *Proceedings of the IEEE*, 2003, 91(2):305-327

[9] Razavipour G, Afzali-Kusha A, Pedram M. Design and analysis of two low-power SRAM cell structures. *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, 2009, 17(10):1551-1555

[10] Azizi N, Najm F N. An asymmetric SRAM cell to lower gate leakage. In: Proceedings of the 5th International Symposium Quality Electronic Design, San Jose, USA, 2004. 534-539

Design of a new low power SRAM with clamping diode

Zhang Lijun*, Wu Chen**, Wang Ziou*, Mao Lingfeng*

(* School of Urban Rail Transportation, Soochow University, Suzhou 215006)

(** Aicestar Technology Corp., Suzhou 215021)

Abstract

A new technique for design of a low power static random access memory (SRAM) was proposed to realize the simultaneous reduction of the leakage current and the dynamic power-consumption in the levels of circuit and architecture. The technique adopts a source biasing scheme (a NMOS transistor is inserted between the ground line and the SRAM cell) to reduce the leakage current. It requires an extra clamping diode in parallel with the NMOS transistor to avoid the floating virtual ground voltage and obtain the data retention capability. The SRAM is in an active mode when the NMOS transistor is turned on. Turning off the NMOS transistor can raise the source voltage and lead to a large reduction in the leakage current. Besides, the memory architecture is uniquely partitioned to decrease the number of half-selected SRAM cells and thus reducing the dynamic power. Power-gating techniques combined with high-V_{th} devices are applied to low power periphery circuits. Test chips with kinds of embedded SRAM instances were fabricated in the UMC 55nm SP CMOS process and the measurement results proved the effectiveness and reliability of the proposed technique.

Key words: static random access memory (SRAM), low power, clamping diode, leakage current