一种高速低功耗多端口寄存器堆的设计*

丛高建† 齐家月

（清华大学微电子学研究所，北京 100084）

摘要：通过使用特殊的存储单元，减小工作电流，设计了一种 32×32bit 的 1 写 8 读 9 端口寄存器堆，读操作位线和写操作位线都实现了低摆幅，结合使用自复位地址译码电路、门限时钟和优化的时序控制电路等，实现了高速和低功耗的目标，并用 SMIC 0.18μm 工艺设计了全定制版图.在 1.8V 工作电压下用 Hspice 进行版图后仿真结果显示，写入时间为 1.7ns，读取时间为 1.32ns，时钟频率为 500MHz 时，9 个端口同时工作的最大功耗为 70 mW.

关键词：高速；低功耗；多端口；SRAM；寄存器堆
EEACC：2570D；1265D
中图分类号：TN432 **文献标识码**：A **文章编号**：0253-4177(2007)04-0614-05

1 引言

超标量微处理器中的重命名寄存器，具有众多读写端口，其存取速度直接影响着处理器的性能.寄存器的重命名在高性能微处理器中用于提高指令并行执行能力，广泛应用于各种通用微处理器中，有众多的实现形式[1].重命名寄存器堆的端口十分众多，如 SPARC 的重命名寄存器堆就具有 4 个写端口和 10 个读端口[2].处理器的并行处理能力越强，要求寄存器堆的端口越多.端口的增加不但增加了功耗，而且加大了读写延时，限制了微处理器时钟频率的提高.一些微处理器实例中存储电路的功耗占整个电路功耗的 40%以上[3].为了减轻端口太多引起的负面影响，Power2，Power3 和 Alpha 21264 中采用了复制寄存器堆的办法来减少每个寄存器堆的端口[1]，文献[4]中提出了一种新的重命名寄存器结构，每个执行单元对应一个寄存器分区，这样每个寄存器分区的写端口只需一个就行了.在此基础上我们用 SMIC 0.18μm 工艺实现了一个读写均采用低摆幅位线的 1 写 8 读 32×32bit 大小的寄存器堆[1].

高速和低功耗是多端口寄存器堆的设计目标，两者相互制约.寄存器堆电路的功耗主要来自：地址译码电路、存储阵列、灵敏放大器和外围接口电路[3].延时主要来自译码电路和灵敏放大器.电路模拟表明，译码电路延时约占总延时的 50%，灵敏放大器延时约占总延时的 30%[5].可以采用多种办法减小功耗和延时，比如在地址译码电路中使用 SCL (source coupled logic)电路提高工作速度[5]；使用电流工作方式的灵敏放大器[6,7]；使用单端灵敏放大器[2,6,8]；使用电平转换电路[9]等.

我们在寄存器堆的设计中综合应用了多种方法实现了 1.7ns 写入、1.32ns 读出的速度和 0.14mW/MHz 的功耗.

2 电路结构

如图 1 所示，电路主要分为地址锁存译码、存储阵列、读写时序控制和灵敏放大几部分.存储阵列为 32×32 个存储单元，每个单元有 1 写 8 读 9 个独立端口，每个读端口都有各自的地址锁存译码和灵敏放大电路.为了加快读出速度，我们使用了双位线差分读出的办法.位线预充电至 V_{cc}.

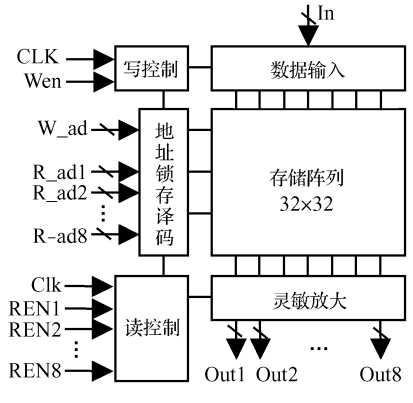

图 1 多端口 SRAM 电路结构

Fig.1 Block diagram of the multi-port SRAM

* INTEL 大学合作计划资助项目

† 通信作者.Email:conggj00@mails.tsinghua.edu.cn

2006-09-14 收到，2006-10-21 定稿

3　各部分电路详细介绍

电路中应用的高速和低功耗设计有：

(1)位线低摆幅.读操作位线低摆幅已经很常见,为了进一步减小功耗,写操作位线同样使用了低摆幅.由于时序控制方面的原因,写操作需更长时间.

(2)使用高速低功耗的灵敏放大器.

(3)使用特殊的存储单元电路.限制了位线接地电流,不但减小了短路电流功耗,同时也配合实现了位线的低摆幅.

(4)使用自复位动态与逻辑地址译码电路.

(5)门限时钟,由读写使能控制,在不工作时杜绝内部电路无效翻转.这会增加一些门延时,但节约功耗的效果比较明显.

(6)使用 SCL 电路进行地址锁存,并产生双相缓冲信号,这两个信号最多只有一个门的延时.

3.1　灵敏放大器电路

电流模式的灵敏放大器工作时,位线电平几乎保持不变,因而几乎消除了位线电容充放电引起的功耗,但代之以短路电流功耗,如文献[6]和[7]中使用了电流工作模式的灵敏放大器,位线的预充电电路是常通的,如图 2 所示.当进行读操作时,由预充电电路和存储单元的下拉电路、灵敏放大器形成接地直流通路,这是电流工作模式所必须的.如果不对接地电流加以限制,虽然达到了令人满意的读取速度,却大大抵消了抑制位线电容充放电带来的功耗节省,甚至功耗反而增大了,小规模的 SRAM 电路更容易出现这种情况.

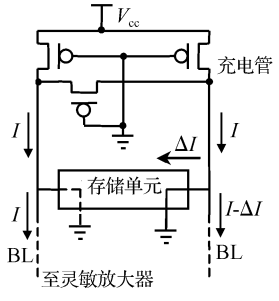

图 2　充电管和下拉电路形成直流通路

Fig.2　Short circuit current caused by precharging and pulling down transistors

我们采用了文献[10]中提出的一种电压模式灵敏放大器,如图 3 所示.由于从电源至地只有三个管子串联,可以使用更低的电源电压,而且与 CL (current latch) SA 和传统 SA 相比功耗延时积分别减小 14% 和 63%[10].

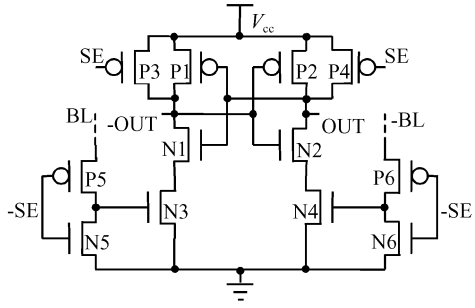

图 3　高速低电压灵敏放大器

Fig.3　High-speed low-voltage SA

当 SE 信号为低电平时,P3,P4,N5,N6 开启,N3,N4,P5,P6 截止,电路处于复位状态,两个输出端均为高电平;当 SE 信号是高电平时,P3,P4,N5,N6 截止,位线电平通过 P5,P6 连在 N3,N4 的栅上,N3,N4 相当于共源差分放大器,将栅上的差分电压放大后在输出端输出.输出端的电压差经过 P1,P2,N1,N2 组成的正反馈回路,快速增大至满摆幅输出.电路模拟表明,在电源电压 1.8V 的情况下,输入位线差分电压达 0.05V 即可获得良好的工作性能,如图 4 所示.

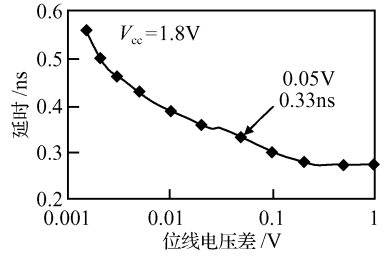

图 4　灵敏放大器性能

Fig.4　Performance of the SA

3.2　存储单元电路

传统的多端口存储单元如图 5 所示,由于每增加一个端口,都要增加横向和纵向的字线和位线,存储单元的面积和端口数的平方成正比[11],多端口的存储单元比普通 SRAM 的六管单元要大得多.由于负载非常大,在进行读操作时,存储单元输出端与位线直接连接有可能造成存储单元的错误反转,所以存储单元最好与位线隔绝.

为了实现写操作位线的低摆幅,我们采用了灵敏放大的写过程.新的存储单元电路如图 6 所示,由灵敏放大器加上字线、位线的连接电路组成.灵敏放大器的 SE 端连接写地址译码的字线.由于受字线位线布线的限制,使用新的单元电路与传统电路相比对面积造成的影响不大.

图示的位线连接方法保证了读操作不会引起存

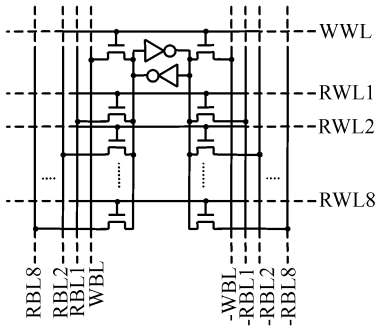

图 5 传统的多端口存储单元

Fig. 5 Traditional multi-port SRAM cell

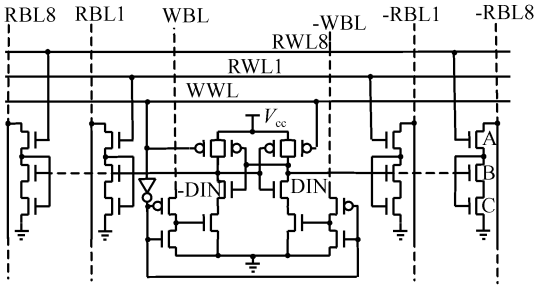

图 6 新的存储单元电路

Fig. 6 Sense-amplifying memory cell

储单元的错误翻转. 各读写位线的下拉电路由三个 nMOS 管串联而成，加入了一个限制电流的管子. 目的是在保证字线和存储单元负载最小的情况下，减小下拉电流. 如果通过改变管子宽长比来减小电流会增大管子面积，增加字线和存储单元的电容负载. 图 6 中当 RWL8 为高电平时，假设存储单元右端输出为高电平，则由 A，B，C 组成的下拉电路导通，A，C 管处于饱和区，B 管处于线性区，各管均为最小尺寸，有

$$I_{dsA} = \frac{1}{2}\mu_n c_{ox}\left(\frac{W}{L}\right)_A (V_{cc} - V_{gC} - V_{th})^2$$

$$I_{dsC} = \frac{1}{2}\mu_n c_{ox}\left(\frac{W}{L}\right)_C (V_{gC} - V_{th})^2$$

得到 $V_{gC} = \frac{1}{2}V_{cc}$. 不使用限流管时下拉电流为

$$I_{ds} = \frac{1}{2} \times \frac{1}{2}\mu_n c_{ox} \frac{W}{L}(V_{cc} - V_{th})^2$$

当 $V_{cc} = 1.8\text{V}$，$V_{th} = 0.45\text{V}$ 时，两种情况的下拉电流比值为

$$\frac{\left(\frac{1}{2}V_{cc} - V_{th}\right)^2}{\frac{1}{2}(V_{cc} - V_{th})^2} = \frac{2}{9}$$

小规模 SRAM 电路位线电容不是很大，为了减小位线摆幅，限制接地电流是十分必要的. 这样，由于灵敏放大器中没有位线接地通路，同时存储单元的接地电流受到了限制，因此与使用电流模式灵敏放大器和传统下拉电路相比，本电路在进行读操作时工作电流大大减小.

3.3 地址锁存和译码电路

为了减小延时，我们使用了图 7 所示的 SCL (source-couple-logic) 电路[5] 来进行地址锁存以及产生用于地址译码的正反相信号. SCL 电路的优点是，不论正相还是反相信号最大延时都只有一个门延时，比一般锁存器快，锁存后输入高电平降低对输出没有影响. 缺点是这种锁存在时钟为高电平时有效，时钟为低电平时处于复位状态，而且锁存后 AD 由低到高变化会使输出反转，如果前级使用动态逻辑则没有这个问题.

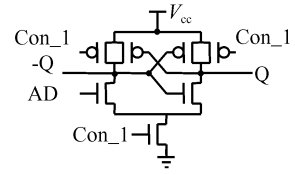

图 7 SCL 电路用作地址锁存

Fig. 7 SCL circuit used for address latching

出于减小功耗的目的，我们选择了动态与逻辑地址译码电路. 读地址译码电路在时钟为低电平时，电路处于复位状态，字线输出是低电平，在时钟上升沿，充电管关闭，电路开始取值. 字线的输出脉冲宽度为半个时钟周期. 由于读地址译码结果控制着位线下拉电路的通断，因此地址译码输出的脉冲宽度只要保证灵敏放大器能正确取值即可，通过限制地址译码的脉冲宽度可以减少一些不必要的功耗.

写地址译码输出连接存储单元的 SE 端，存储单元的复位只需很窄的脉冲，所以需要额外的电路来限制写地址译码的脉冲宽度. 电路单元如图 8 所示.

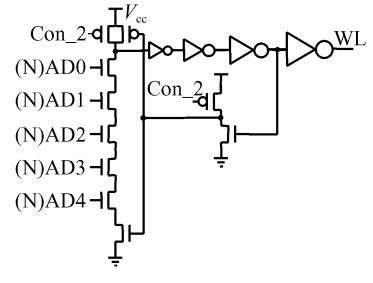

图 8 写译码电路单元

Fig. 8 Write address decoder cell

3.4 时序控制电路

读控制电路主要由门限时钟电路和 SE 脉冲电路两部分组成，控制整个电路的工作时序，如图 9 所

示.写控制电路与此类似,不同的是不需要产生 SE 脉冲.

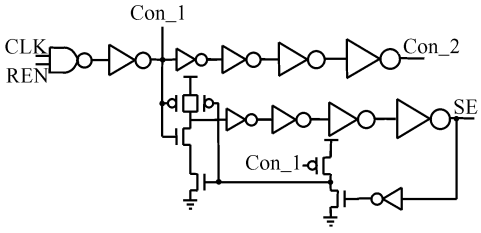

图 9　读操作时序控制电路

Fig.9　Timing control circuit for read operations

时序控制电路和地址译码电路的输出端多是长连线,其寄生电容大大增加了负载,如果忽视将导致最终电路性能与设计预期相差很远,设计电路的时候必须保证输出有足够的驱动能力.如连线采用分布 rc 模型,其驱动延时为

$$t_p = 0.69R_sC_w + 0.38R_wC_w$$

式中　R_s 是驱动管电阻;R_w 是连线电阻;C_w 是连线电容,本电路中最长连线只有几百微米,$R_w \ll R_s$,长连线可以近似为纯电容负载.可以通过插入尺寸优化的缓冲链来实现足够的驱动能力和最优化的延时,如图 9 所示,但缓冲链在减小延时的同时也引入了额外的功耗.

4　版图后仿真结果

我们用 SMIC 0.18μm 工艺设计了全定制版图.存储单元大小为 10.08μm×14.04μm,整个电路大小为 392.45μm × 612.87μm,共 7.5 万个 MOS 管.在工作电压为 1.8V、时钟频率为 500MHz 时,用 Hspice 进行版图后仿真波形,如图 10 所示.写入时间为 1.7ns,读出时间为 1.32ns(从时钟上升沿中点至 SA 取值至 90%),9 个端口同时工作的总功耗为 70mW.由于读操作在时钟周期开始后的 0.76ns 才通过字线选通存储单元,而且写操作在 1.32ns 时存储单元已处于求值阶段,求值过程可以在读时钟周期刚开始的这段时间继续完成,因此电路可以正常工作在 1.32ns 的时钟周期上.

电路总延时中,门限时钟和地址锁存缓冲延时约占 36%,译码电路及字线驱动延时约占 21%,灵敏放大器延时约占 28%.由于存储单元和灵敏放大器工作电流比较小,地址译码路径中一系列缓冲管的功耗在整个电路的功耗中占相当大比例.模拟显示,地址锁存、缓冲、译码及字线驱动电路的功耗约占总功耗的 57%.表 1 是本文电路及文献中一些多端口寄存器堆性能的总结,可以看出,本文电路的设计基本满足了高速和低功耗的要求.

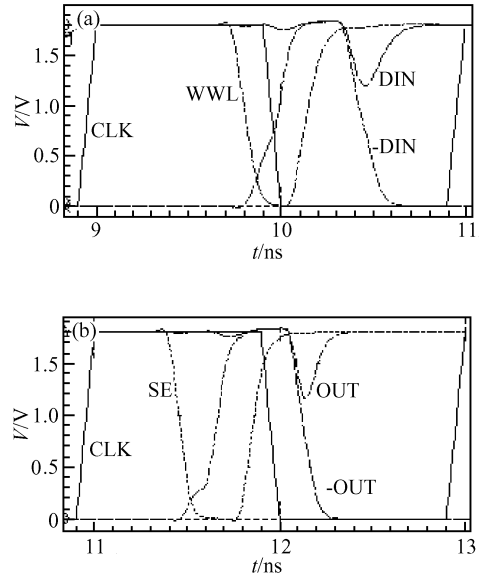

图 10　版图后仿真波形　(a) 写操作波形;(b) 读操作波形

Fig.10　Post-layout simulation waveforms　(a) Write operations;(b) Read operations

表 1　电路性能比较

Table 1　Result comparison of the circuit properties

规模大小 /bit	端口数目	特征尺寸 /μm	电压 /V	功耗 /(mW@500M)	最短周期 /ns
32×32	1w/8r	0.18	1.8	70	1.32
34×64[6]	6w/10r	0.11	1.2	133	1.78
32×64[12]	2w/6r	0.5	2.5	1200	1.9
32×32[7]	2w/3r	0.35	3.3	640	1.85
16×32[8]	6w/10r	0.18	1.8	46	1.7

5　小结

通过大大减小工作电流,我们设计了读写操作均采用低摆幅位线的 1 写 8 读 32×32bit 寄存器堆,加上电路时序的优化、门限时钟的使用等措施,实现了高速与低功耗的设计目标,并用 SMIC 0.18μm 工艺设计了版图.用 Hspice 进行版图后仿真结果显示,工作电压 1.8V 时,写入时间为 1.7ns,读出时间为 1.32ns,时钟频率为 500MHz 时,9 个端口总功耗为 70mW.

参考文献

[1] Sima D. The design space of register renaming techniques. IEEE Micro,2000,5(20):70

[2] Asato C. A 14-port 3.8ns 116-word 64b read-renaming register file. IEEE J of Sold-State Circuits,1995,30(11):1254

[3] Margala M. Low-power SRAM circuit design. IEEE Int Workshop on Memory Technology, Design, and Testing,

1999:115
[4] Liu Zhenyu, Qi Jiayue. A novel rename register architecture and performance analysis. Asia-Pacific Computer Systems Architecture Conference, 2004: 503
[5] Nambu H, Kanetani K, Yamasaki K, et al. A 1.8-ns access, 550-MHz, 4.5-Mb CMOS SRAM. IEEE J Solid-State Circuits,1998,33(11):1650
[6] Tzartzanis N, Walker W W. A differential current-mode sensing method for high-noise-immunity, single-ended register files. Digest of Technical Papers, ISSCC IEEE International Solid-State Circuits Conference,2004,1:506
[7] Wang Jiajing, Zhang Qianling. A 500-MHz low-power five-port CMOS register file. Design Automation Conference, Proceedings of the ASP-DAC, Asia and South Pacific, 2003:511
[8] Yu Qian, Wang Donghui, Zhang Tiejun, et al. A design of 500MHz 10-read 6-write register file. 6th International Conference on ASIC, ASICON,2005,1:311
[9] Tzartzanis N, Walker W W. A transparent voltage conversion method and its application to a dual-supply-voltage register file. Computer Design, Proceedings, 21st International Conference,2003:107
[10] Chow Hwang-Cherng, Chang Shu-Hsien. High performance sense amplifier circuit for low power SRAM applications. Circuits and Systems. ISCAS Proceedings of the International Symposium,2004,2:II-741-4
[11] Wallace S, Bagherzadeh N. A scalable register file architecture for superscalar processors. Microprocessors and Microsystems,1998,22(1):49
[12] Hwang W, Joshi R V, Henkels W H. A 500-MHz 32×64 eight-port self-resetting CMOS register file. IEEE J Solid-State Circuits,1999,34(1):56

Design of a High-Speed Low-Power 9-Port Register File*

Cong Gaojian† and Qi Jiayue

(*Institute of Microelectronics, Tsinghua University, Beijing* 100084, *China*)

Abstract: A 1-write-port 8-read-port 32×32-bit register file has been designed in 1.8V 0.18μm CMOS technology. Low-swing bit-lines are used for both read and write operations. Together with the use of novel memory cells, high speed sensor amplifiers, self-reset address decoders, SCL circuits, clock gating, and delicate time control circuits, it has achieved both high speed and low power. Post-layout simulations in 1.8V with HSPICE indicate a write time of 1.7ns and a read time of 1.32ns. The power dissipation is 70mW for all 9 ports at 500MHz.

Key words: high speed; low power; multi-port; SRAM; register file
EEACC: 2570D; 1265D
Article ID: 0253-4177(2007)04-0614-05

* Project supported by the Joint Program of INTEL University
† Corresponding author. Email: conggj00@mails.tsinghua.edu.cn
Received 14 September 2006, revised manuscript received 21 October 2006