学校代码：10286
分类号：TN47
密　级：公开
U D C：621.38
学　号：151228

东南大學
SOUTHEAST UNIVERSITY

硕士学位论文

宽电压时序推测型 SRAM 存储阵列的设计

研究生姓名：郭易辰
导 师 姓 名：杨军

申请学位类别 工学硕士　学位授予单位 东 南 大 学
一级学科名称 电子科学与技术　论文答辩日期 2018 年 6 月 9 日
二级学科名称 微电子学与固体电子学　学位授予日期 2018 年 月 日
答辩委员会主席 蔡跃明　评　阅　人 蔡跃明
吴建辉

2018 年 6 月 12 日

东南大学

硕士学位论文

宽电压时序推测型 SRAM 存储阵列的设计

专 业 名 称：微电子学与固体电子学

研究生姓名：郭易辰

导 师 姓 名：杨军

Design of Array for Wide-Voltage Timing Speculative Static Random Access Memory

A Thesis Submitted to

Southeast University

For the Academic Degree of Master of Engineering

BY

Guo Yichen

Supervised by

Prof. Yang Jun

School of Electronic Science and Engineering

Southeast University

June 2018

摘要

为满足片上系统（System on a Chip，SoC）的能效需求，低至近阈值区的宽电压静态随机存储器（Static Random Access Memory，SRAM）的设计在学术界引起了广泛的关注。存储阵列作为 SRAM 的关键模块，决定着 SRAM 的整体性能。随着电源电压降低，局部工艺波动导致电路需要的设计裕度越来越大，在近阈值区，过于悲观的设计裕度大大地增加了存储阵列的读出延时，SRAM 的性能因此严重退化。

时序推测方案能够在一定程度上降低过大的设计裕度对性能的影响，时序推测方案采用两次读出的方式，第一次读出为推测型读出，数据快速输出，用于降低存储阵列的延时，第二次读出为确认型读出，用于检错。现有的时序推测方案在近阈值区的检错延时过大，这限制了其在 SoC 芯片中的应用。本文提出了一种改进型的时序推测方案，该方案在推测型读出后通过快速调整灵敏放大器输入电压的极性实现快速检错，该方案可以大幅度降低存储阵列的读出延时，仿真结果表明：相比传统的读出方案，存储阵列的读出延时在低电压下（0.5V）和正常电压下（0.9V）分别降低了大约 50%和 10%。

本文以时序推测型存储阵列为核心，基于 TSMC 28nm 工艺，完成了一款容量为 256×32 的宽电压 SRAM 的设计，并完成后仿真的验证，仿真结果表明：相比于传统方案，本文 SRAM 的整体读出延时在 0.5V 和 0.9V 工作电压下分别降低了 36%和 2%。与传统的方案相比，本文方案收益的综合指标（Figure of Merit，FoM）提升了 1.96 倍。与其他的时序推测方案相比，本文方案收益的 FoM 提升了 1.75 倍。

关键词：静态随机存储器，存储阵列，近阈值，时序推测

Abstract

In order to meet the energy efficiency requirements for SoC, wide voltage SRAM design is an increasing concern in academic. As the critical part of SRAM, SRAM array have a decisive influence on the performance of SRAM. As the power supply voltage decreases, local process variation causes the increase of the design margin. In the near-threshold region, the pessimistic design margin greatly increases the read latency of SRAM array and severely degrades the performance of SRAM.

The timing speculative scheme can reduce the impact of the design margin on performance, it senses the bit-line twice, the first sensing is the speculative reading, the output data is quickly sent out after first sensing, which is used to reduce the read latency of SRAM array, the second sensing is the confirm reading, which is used for error detection. The existing timing speculative schemes have too large error detection delay in the near-threshold region, which limits its application in SoC design. This paper proposed an improved timing speculative scheme, which quickly adjusts the polarity of the input voltage of the sense amplifier after speculative reading to achieve rapid error detection. This scheme can significantly reduce the read latency of SRAM array. Compared with the conventional scheme, the read latency of the SRAM array is reduced by approximately 50% and 10% respectively at 0.5V and 0.9V.

Based on the TSMC 28nm process, this thesis designs a wide-voltage SRAM with a capacity of 256×32. The layout and post-layout simulation are completed. Compared with the conventional scheme, the proposed method achieves 36% reduction of read latency at 0.5V and 2% reduction of read latency at 0.9V. Compared with the conventional scheme, the FoM of gain is 1.96x. Compared with the other timing speculative scheme, the proposed scheme achieves 1.75x improvement in the FoM of gain.

Keywords: Static Random Access Memory, SRAM array, near-threshold region, timing speculative

目录

第一章 绪论

静态随机存储器（Static Random Access Memory，SRAM）是片上系统（System on a Chip，SoC）的关键模块之一，为了同时兼顾高性能和高能效的需求，宽电压 SRAM 的设计在学术界引起了广泛的关注。在近阈值区，SRAM 的性能退化相比于逻辑电路更加严重，针对此问题，本文深入研究了宽电压 SRAM 存储阵列的优化方案。

本章 1.1 节介绍了宽电压 SRAM 的研究背景及设计挑战；1.2 节介绍了国内外研究现状；1.3 节介绍了论文的主要工作；1.4 节介绍了论文的组织结构；1.5 节是本章小结。

1.1 宽电压 SRAM 研究背景及设计挑战

随着集成电路技术的快速发展，SoC 芯片的功能越来越强大，高性能和低功耗始终是 SoC 设计的两大追求目标，但是这两项指标是相互制约的，过度地追求低功耗会导致性能的严重退化。因此，能量效率成为了更重要的设计目标，它是指每次操作所需能量的倒数，亦即每瓦特能量能够完成的操作数，以 GOPS/Watt 为单位。随着电压的降低，电路的工作频率不断降低，而能效则呈现先增后减的趋势，能效在近阈值区达到最高，如图 1-1 所示。为了能够同时兼顾性能和能效需求，宽电压范围的 SoC 设计在学术界引起了广泛的关注[1-3]。SRAM 作为 SoC 芯片的关键模块，对 SoC 的整体指标起决定性的作用，因此低至近阈值区的宽电压 SRAM 设计得到了越来越多的关注[4-5]，但是近阈值区的 SRAM 性能急剧下降。

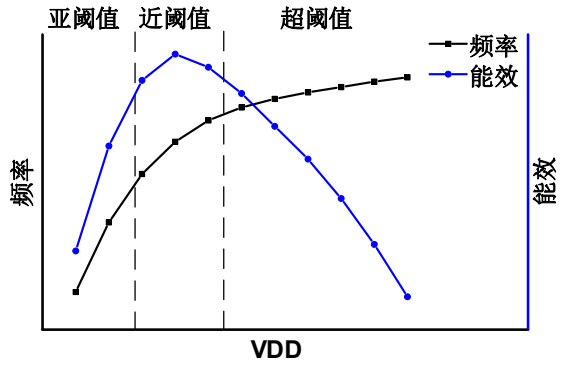


图 1-1 电源电压变化对频率和能效的影响

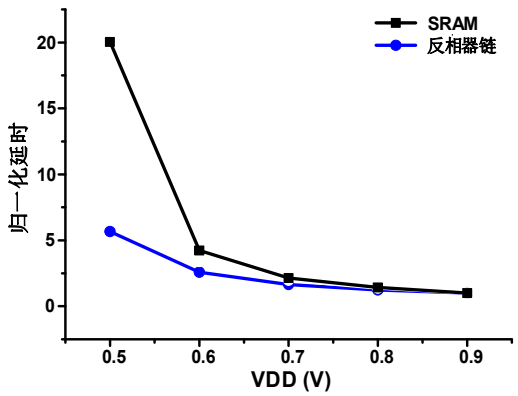


图 1-2 不同电压下 SRAM 和反相器链的归一化延时

采用 TSMC 28nm 工艺，图 1-2 给出了不同电压下 SRAM 和反相器链的归一化延时，反相器链代表了逻辑电路，由图可知，SRAM 和反相器链的延时随着电源电压的降低而变大，与逻辑电路相比，SRAM 的性能在近阈值区退化更加严重，这会限制 SoC 系统在近阈值区的最高工作频率。

SRAM 的性能取决于读操作。具体原因如下：读操作时，存储单元直接驱动位线上的负载电容，为了降低 SRAM 存储单元的静态功耗，提升存储密度，SRAM 存储单元通常采用高阈值，且违反设计规则的小尺寸器件，故存储单元的驱动能力较弱，因此存储单元的读出延时相对较大。写操作过程中，当字线开启时，存储单元内部交叉耦合反相器的正反馈作用会加速数据的写入，因此存储单元数据写入的延时相对较低，故读操作是 SRAM 性能的瓶颈。

SRAM 的读出路径如图 1-3 所示，下面做具体分析。SRAM 的读访问时间取决于三部分：字线驱动模块延时（T_{DEC}）、存储阵列的延时（T_{ARRAY}）及输出驱动延时（T_{OUT}）。

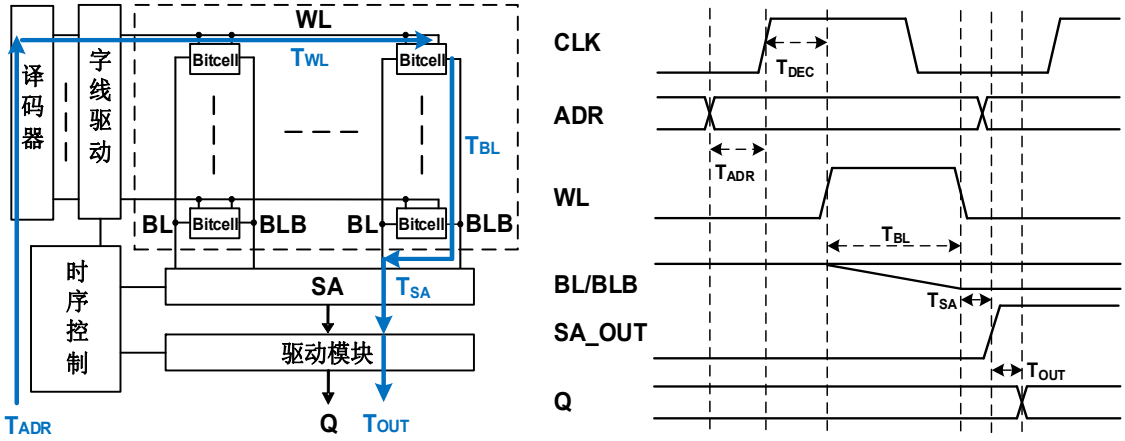


图 1-3 SRAM 读操作的关键路径

SRAM 读操作的过程为：地址信号 ADR 在时钟上升沿之前被送入 SRAM 内部，地址译码器开始工作，通常情况下，地址译码器在时钟上升沿之前结束工作，这个延时对应 T_{ADR}，此延时会影响 SoC 系统的工作频率，但是不会影响 SRAM 的读出延时。时钟上升沿之后，锁存器锁存地址信号，在控制信号的作用下，字线驱动模块启动，随后字线上升沿到来，时钟上升沿到字线的上升沿的延时记为 T_{DEC}，此延时会影响 SRAM 的读出延时。字线开启后，存储单元对位线放电，当所有的位线电压摆幅达到灵敏放大器的失调电压时，字线关断，此延时记为 T_{BL}，与此同时灵敏放大器开启，将位线的小摆幅信号进行放大，此延时记为 T_{SA}，故存储阵列的读出延时 $T_{ARRAY}=T_{BL}+T_{SA}$。灵敏放大器的输出经过驱动模块，相应的延时记为 T_{OUT}。故 SRAM 的读出延时主要由三个部分组成：

1) 字线驱动延时 T_{DEC}：时钟上升沿到字线上升沿的延时，主要由字线驱动模块决定。

2) 存储阵列的读出延时 T_{ARRAY}：存储单元的放电延时（字线上升沿到位线的摆幅达到灵敏放大器的失调电压所对应的延时）加上灵敏放大器的延时，相比于存储单元的放电延时，灵敏放大器的延时可以忽略不计。

3) 输出驱动延时 T_{OUT}：输出驱动模块对应的延时。

在整个读操作路径中，字线的驱动电路，输出驱动模块对应普通的逻辑电路，可以通过采用低阈值器件和增大晶体管尺寸等方法减小延时；而对于 T_{ARRAY}，延时的大小由存储单元的驱动能力决定，和逻辑电路相比，SRAM 存储单元的驱动能力相对较弱，因此其放电延时相对较大，图 1-4 给出了字线驱动延时、存储阵列延时和输出驱动延时在 SRAM 整体读出延时中的比例，存储阵列的读

出延时在 SRAM 整体读出延时的占比约为 80%，因此存储阵列的读出延时是 SRAM 整体读出延时的主要组成部分，降低存储阵列的读出延时将成为提升 SRAM 性能的关键。

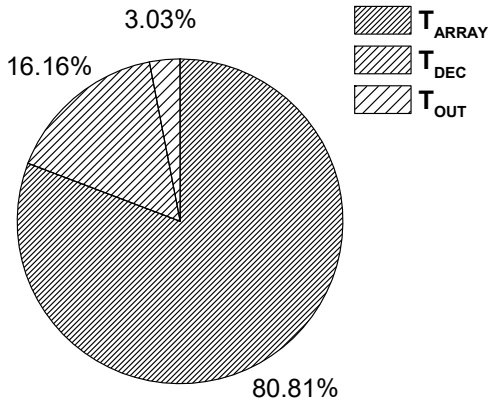


图 1-4 SRAM 读出延时的占比

下面对存储阵列的读出延时做具体的分析，存储阵列中的存储单元的驱动能力是不一致的，为了保证 SRAM 读操作的正确性，字线关断（灵敏放大器开启）时要保证存储阵列中所有存储单元的位线摆幅超过灵敏放大器的失调电压，因此存储阵列的读出延时取决于存储阵列中驱动最弱的存储单元，这就是存储阵列中的木桶效应，如图 1-5 所示。

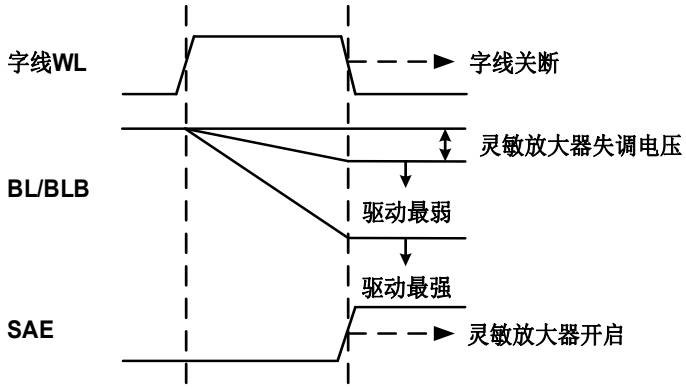


图 1-5 存储阵列中的木桶效应

存储单元驱动能力不一致的现象是由工艺波动造成的。为提高 SRAM 的存储密度，SRAM 存储单元通常采用违反设计规则的小尺寸器件，因此相比于数字逻辑电路，存储单元更容易受工艺波动的影响。工艺波动可以分为全局工艺波动和局部工艺波动[6-8]。在全局工艺波动的影响下，同一个芯片中的 NMOS 管和 PMOS 管的工艺参数的变化朝着同一个方向，不同芯片的工艺参数存在差异。局部工艺波动又称为片上偏差，在局部工艺波动的影响下，同一芯片中不同 NMOS 管和 PMOS 管之间的工艺参数存在差异，局部工艺波动由随机掺杂波动（Random Doping Fluctuation，RDF）和线边沿粗糙（Line Edge Roughness，LER）等随机因素引起。随着集成电路制造技术的发展，半导体工艺的特征尺寸不断减小，局部工艺波动的影响越来越显著。对于同一个 SRAM 存储阵列，全局工艺波动对存储单元的影响是一致的，局部工艺波动会造成 SRAM 存储单元驱动能力的不同，从而导致了存储单元的放电延时不一致，局部工艺波动的影响在近阈值区更加显著。图 1-6 给出了存储单元的放电延时（存储单元的位线摆幅达到灵敏放大器的失调电压对应的延时）的蒙特卡洛仿真结果，仿真

采用 TSMC 28nm 工艺，仿真条件分别为 0.9V，SSG 工艺角，0℃和 0.5V，SSG 工艺角，0℃。

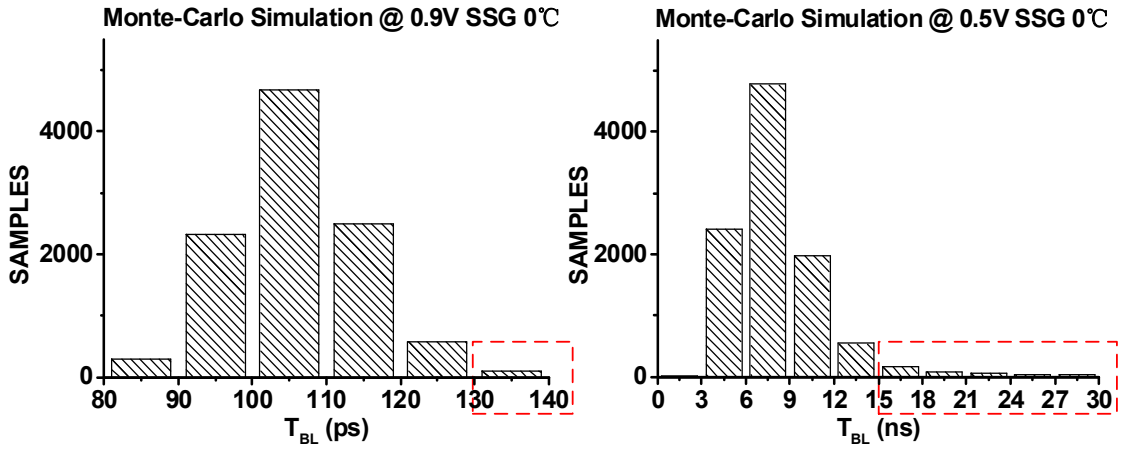


图 1-6 0.5V 和 0.9V 条件下存储单元放电延时的统计分布

为了保证 SRAM 的功能正确，设计中通常要预留一定的设计裕度。当电源电压降低至近阈值区时，局部工艺波动导致存储单元位线放电延时的统计分布存在明显的拖尾现象，绝大多数存储单元的放电延时分布在 15ns 以内，只有少数弱驱动存储单元的放电延时分布在 15ns 至 30ns 之间，弱驱动存储单元使得设计裕度随着电源电压的降低进一步变大，过于悲观的设计裕度使存储阵列的性能急剧下降，故 SRAM 的整体性能急剧降低。在 SRAM 六管存储单元结构中，为了保证读操作的稳定性，避免读操作过程中存储单元的数据发生翻转，近阈值区的 SRAM 无法采用字线电压提升技术降低局部工艺波动对存储单元放电延时的影响，如何克服局部工艺波动对存储阵列延时的影响成为了设计难点。

1.2 国内外研究现状

近些年来近阈值区 SRAM 的研究热点是提升 SRAM 存储单元的读写稳定性，因为读写稳定性决定了 SRAM 的最低工作电压，学术界的技术路线大体可以分为如下两类：1.采用超六管结构的存储单元[9-13]；2.采用存储单元的辅助优化技术[14-19]。超六管存储单元通过特殊的电路结构实现了读写稳定性的提升，缺点是面积开销大。存储单元的辅助优化技术通过动态地改变存储单元的晶体管驱动能力实现读写稳定性的提升。

为克服 SRAM 六管存储单元读写稳定性之间的矛盾，文献[9]提出了读写分离的 SRAM 八管存储单元结构，读写分离的八管存储单元从电路结构上保证了读位线与内部存储节点完全隔离，故读写分离的八管单元从根源上解决了读破坏问题，所以可以调整六管单元部分的晶体管尺寸，以实现写稳定性的增强。除了上述的读写分离的八管存储单元之外，学术界还出现了其他的超六管存储单元结构[10-13]。

存储单元的辅助优化技术可以分为读辅助优化技术和写辅助优化技术。读辅助优化技术是为了避免 SRAM 存储单元在读操作过程中发生读破坏现象，提升存储单元的读操作稳定性。其中，电源电压提升技术[14]能够提升 SRAM 存储单元中交叉耦合反相器的锁存能力，保证读操作过程中存储单元内部的数据不被破坏；字线电压降低技术[15-16]降低了存储单元传输管的驱动能力，降低了发生读破坏现象的概率。写辅助优化技术是为了保证数据正确写入 SRAM 存储单元，其中，电源电压降低技术[17]能够降低 SRAM 存储单元中交叉耦合反相器的锁存能力，使得数据能够更容易写入存储单

元；字线电压提升技术[18]提升了存储单元传输管的驱动能力，提升了存储单元的写稳定性；负位线技术[19]的原理与字线电压提升技术类似，皆是通过提升存储单元传输管的驱动能力提升存储单元的写稳定性。

上述优化技术皆可提升 SRAM 存储单元在低电压下的读写稳定性，能够降低 SRAM 的最低工作电压，但是不能提升 SRAM 的整体性能，无法解决 SRAM 在近阈值区性能退化严重的问题。时序推测优化技术[20-21]通过两次读出的方式降低存储阵列的读出延时，从而提升 SRAM 的整体性能，在该方案中灵敏放大器先后启动两次，第一次读出对应推测型读出，用于降低存储阵列的读出延时，第二次读出对应确认型读出，用于检错。如果推测型读出出错，系统通过纠错手段保证功能的正确性。但是上述的时序推测方案在近阈值区的检错延时过大，这限制了其在 SoC 中的应用。

1.3 研究内容

本文以降低 SRAM 存储阵列的读出延时为目标，以时序推测技术作为手段，论文在 TSMC 28nm 工艺下设计了一款容量为 256×32 的宽电压 SRAM（0.5V 至 0.9V），论文的主要工作如下：

1) 分析宽电压 SRAM 存储阵列的设计挑战，总结国内外的研究现状。

2) 阐述了时序推测技术的基本工作原理，分析并总结其在近阈值区的局限性。

3) 针对现有的时序推测技术在近阈值区的局限性，提出了改进型的时序推测方案，详细介绍了其工作原理和电路实现，并完成仿真分析。

4) 基于 TSMC 28nm 工艺，以时序推测型存储阵列为核心，完成了一款容量为 256×32 的宽电压 SRAM 的设计，包括了电路设计及版图的物理实现，并给出了仿真结果及对比分析。

1.4 论文组织结构

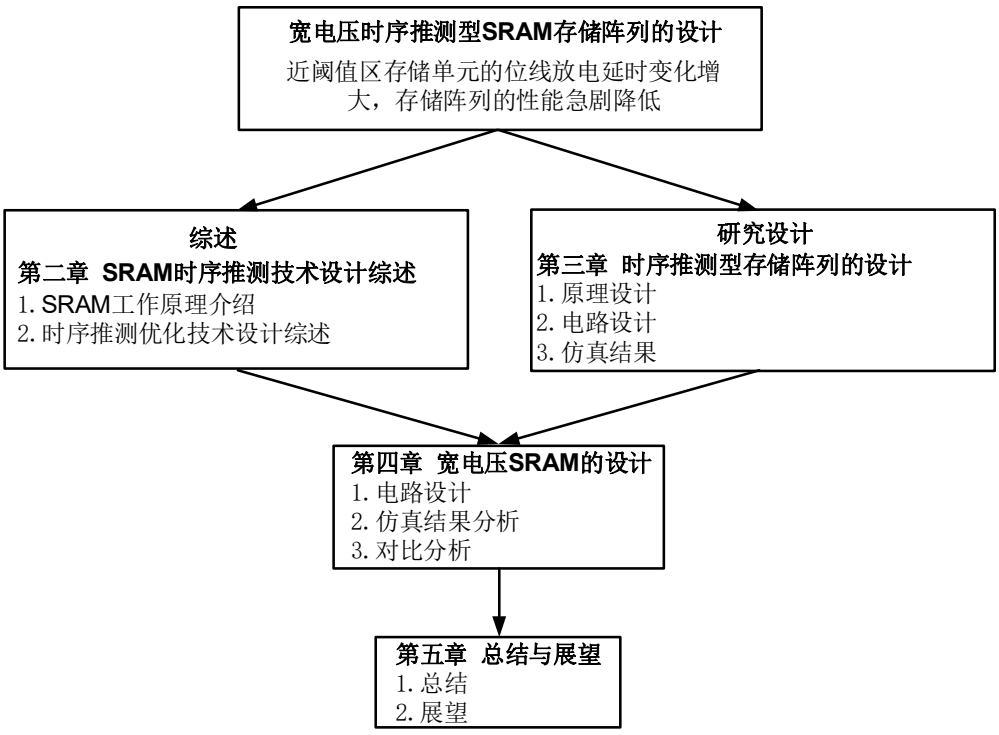


图 1-7 论文的组织结构

第一章为绪论内容，主要介绍了宽电压 SRAM 设计的重要意义，分析了近阈值区 SRAM 性能退化严重的原因，并给出了论文的主要工作及框架结构。

第二章是设计综述部分，介绍了 SRAM 的基本结构及存储单元的工作原理，并重点分析了时序推测技术的研究现状，最后指出现有的时序推测技术在近阈值区的局限性。

第三章提出了一种改进型的时序推测方案，首先介绍了改进型时序推测方案的原理及电路实现，并进行了仿真和分析。

第四章基于 TSMC 28nm 工艺，以存储阵列为核心设计了一款容量为 256×32 的宽电压（0.5V-0.9V）时序推测型 SRAM。首先，介绍宽电压 SRAM 的主要电路模块的设计及测试系统的设计；其次，给出了宽电压 SRAM 的仿真结果；最后给出了本文的时序推测方案和其他时序推测方案的对比。

第五章为总结与展望。首先对论文主要工作及创新点进行总结，然后对后续工作做进一步的展望。

1.5 本章小结

本章首先介绍了宽电压 SRAM 设计的重要意义，紧接着重点分析了近阈值区 SRAM 性能急剧下降的原因，随后概括了本文的主要工作，最后给出了论文的整体框架结构。

第二章 SRAM 时序推测技术设计综述

时序推测技术对降低存储阵列的读出延时有着重要的影响，因此研究时序推测技术有着重要的意义。本章给出了 SRAM 时序推测技术的设计综述。2.1 节介绍了 SRAM 的整体结构和 SRAM 存储单元的工作原理，这是全文的设计基础；2.2 节为时序推测技术的综述；2.3 节给出了本章小结。

2.1 SRAM 简介

2.1.1 SRAM 整体结构

SRAM 的整体结构如图 2-1 所示，包括了如下几个部分：

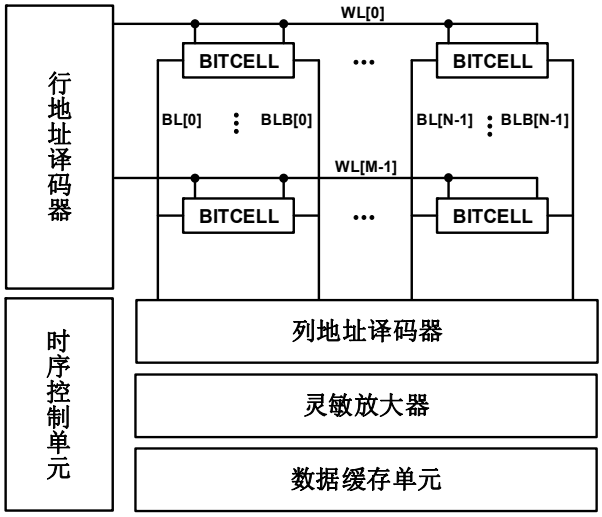


图 2-1 SRAM 的整体结构

1) 存储单元：存储单元是 SRAM 的核心部件，每个存储单元存储一位二进制数据，交叉耦合的反相器构成了锁存器，在不掉电的情况下，锁存器处在保持状态，数据保持稳定，存储单元按照行和列的形式进行排列，行和列分别对应着 SRAM 中的字线 WL 和位线 BL/BLB。

2) 地址译码器：要对 SRAM 中的某个存储单元进行读/写操作，必须通过地址译码器选中要操作的存储单元。地址译码器的作用就是对存储单元进行选择。对于大容量的存储器一般都是二维译码，通过行地址译码器选择行，再通过列地址译码器进行列选择，行地址和列地址交叉处的单元就是选中的存储单元。同时译码器还要负责驱动字线等大负载信号线。

3) 时序控制电路[22-25]：SRAM 的时序控制单元以复制位线技术为核心产生读写控制信号，复制位线技术用于跟踪 SRAM 读操作关键路径，并用于控制灵敏放大器的开启，以实现 SRAM 的高速低功耗操作，相比于传统的反相器链式的跟踪方式，复制位线技术跟踪工艺、电压及温度（Process、Voltage、Temperature，PVT）变化的能力更强。在近阈值区，受局部工艺波动的影响，传统复制位线技术的延时变化增大，这大大地增加了 SRAM 的读出延时，因此抗局部工艺波动的复制位线技术成为了学术界的研究热点。

4) 灵敏放大器[26-29]：读操作时，存储单元直接驱动位线的负载电容，在一些大容量存储阵列中，位线很长，因此位线的负载电容非常大，并且存储单元的驱动能力非常弱，因此位线电压达到

逻辑电路所要求的逻辑低电平“0”或是逻辑高电平“1”的标准所消耗的时间很长，这限制了 SRAM 的读性能，因此 SRAM 采用灵敏放大器放大位线上的小摆幅电压差并输出外围逻辑电路所需要的逻辑电平。

5) 数据缓冲单元：数据缓冲单元是 SRAM 和外部交换信息的接口电路，数据缓冲单元包括输入缓冲单元和输出缓冲单元。输入缓冲单元用于锁存输入信号（地址信号，输入数据，读写使能及片选使能）并负责驱动位线，输出缓冲单元用于提供输出驱动能力。

2.1.2 SRAM 存储单元的工作原理

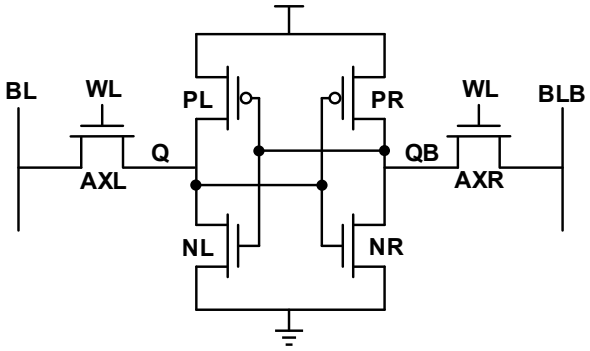


图 2-2 SRAM 六管存储单元结构

SRAM 六管存储单元[30-31]的结构如图 2-2 所示，六管单元由两个 PMOS 管，四个 NMOS 管及字线 WL 和位线 BL/BLB 构成，PL 和 NL 构成的反相器和 PR 和 NR 构成的反相器交叉耦合构成锁存结构，AXL 和 AXR 是两个传输管，字线 WL 控制传输管的开启和关断，位线 BL/BLB 通过传输管 AXL/AXR 分别和内部存储节点 Q 和 QB 连通。控制电路通过对字线 WL 和位线 BL/BLB 的控制完成对 SRAM 存储单元的三个基本操作，分别对应写入操作、读出操作和保持操作。

1) 写入操作

假定此时存储单元存“0”，即左边的存储节点 Q=“0”，右边的存储节点 QB=“1”，此时写入数据“1”，字线 WL 开启之前，BL 预充电至高电平，BLB 放电至低电平。写操作开始时，字线 WL 变成高电平，传输管 AXL 和 AXR 导通，位线 BL/BLB 与存储单元的内部节点连通，存储节点 QB 放电至低电平，存储节点 Q 充电至高电平，交叉耦合反相器的正反馈作用会加速数据的写入。

2) 读出操作

假定此时存储单元存“0”，即左边的存储节点 Q=“0”，右边的存储节点 QB=“1”。字线 WL 开启之前，位线 BL/BLB 通过预充电电路预充至高电平 VDD，读操作开始时，字线 WL 开启，位线 BL 开始放电，位线 BLB 保持在高电平 VDD，当位线的摆幅 $V_{SW}=V_{BL}-V_{BLB}$ 达到灵敏放大器的失调电压时，字线关断，复制位线控制灵敏放大器开启，数据全摆幅输出。在读操作过程中，放电路径存在分压问题，即存储节点 Q 的电压由零升至 ΔV，当 ΔV 的大小超过反相器 PR-NR 的翻转阈值电压时，存储单元内部存储的数据发生改变，这个现象称为读破坏现象。

3) 保持操作

数据保持时存储单元的字线 WL 关断，存储节点与外界隔离，交叉耦合反相器的正反馈作用维持存储单元中的数据，只要系统不断电，存储信息可以长期维持。

传统的 SRAM 六管单元电路简单，面积开销小，被广泛地应用在 SoC 芯片中，但是在近阈值

区，局部工艺波动造成的晶体管失配导致了读写过程中的竞争现象，因此六管存储单元的读写稳定性急剧降低，为克服传统六管存储单元读稳定性与写稳定性之间的矛盾，文献[9]提出了读写分离的八管 SRAM 存储单元结构。如图 2-3 所示。

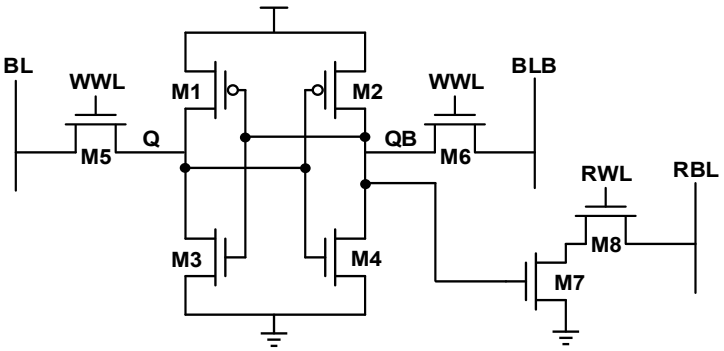


图 2-3 读写分离的八管 SRAM 存储单元

读写分离的八管单元是在六管单元的基础上改造形成的，与六管单元的电路结构相比较，八管单元多出了读放电通路。六管单元部分用于写操作和数据保持，读操作则通过读放电支路完成。

读写分离的八管存储单元从电路结构上保证了读位线 RBL 与内部存储节点 Q/QB 完全隔离，故读写分离的八管单元完全消除了读破坏现象，所以可以通过调整六管单元部分的晶体管尺寸实现写稳定性的增强。读写分离的八管单元的缺点是面积开销过大，并且采用单端读方式，不利于灵敏放大器的检测。

2.2 时序推测优化技术设计综述

近阈值区的 SRAM 读写辅助优化技术皆在提升 SRAM 存储单元的稳定性，可以降低 SRAM 的最低工作电压，但是无法提升 SRAM 存储阵列的性能。时序推测技术与传统的读写辅助技术有所不同，时序推测优化技术可以降低存储阵列中驱动最弱的存储单元对存储阵列延时的影响，从而可以大幅度地降低存储阵列的读出延时。

文献[32]中首先提出了时序推测的概念，文献采用了如图 2-4 所示的监测单元来判断 SoC 系统中关键路径的时序是否出错，监控单元包括了 D 触发器、影子锁存器和一个二输入的异或门，它除了具备普通触发器的功能，还能用于检测时序错误。D 触发器用于采样输入数据 DATA，影子锁存器用于确认 D 触发器采样的数据是否正确。现分如下两种情况进行讨论：

1) 未发生建立时间违规

数据 DATA 的翻转发生在时钟上升沿之前，故 D 触发器和影子锁存器均正确采样，异或门输出低电平。

2) 发生建立时间违规

数据 DATA 的翻转发生在时钟上升沿之后，此时路径发生了建立时间违规，故 D 触发器采集到错误数据，而锁存器由于时钟高电平透明，可以采样到数据翻转，从而 D 触发器的输出和锁存器的输出异或运算结果为高电平，此时 D 触发器采集到错误的数据，系统通过纠错手段保证功能的正确性。

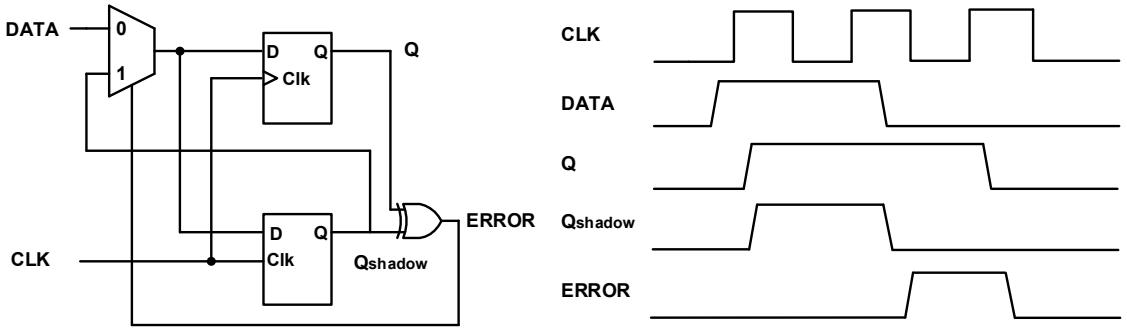


图 2-4 时序推测技术在数字逻辑电路中的应用

这种时序推测的思想也被应用于 SRAM 中，文献[20]提出了一种应用在动态电压调节（Dynamic Voltage Scaling，DVS）系统中的 SRAM。文献[20]通过主灵敏放大器和影子灵敏放大器两次采样位线的方式降低 PVT 偏差带来的电压余量，如图 2-5 所示。该 SRAM 的工作原理如下：主灵敏放大器在时钟信号的下降沿使能，将输出数据送入 SRAM 后接的组合逻辑，位线继续保持放电状态，经过一定的延时，位线的摆幅进一步放大，影子灵敏放大器使能（设计要保证影子灵敏放大器的输出正确），其输出用于判断主灵敏放大器输出的数据是否准确，若异或门输出为低电平，则主灵敏放大器输出的结果是正确的，若异或门输出为高电平，则主灵敏放大器输出的结果是错误的。如果系统判断主灵敏放大器的输出错误，多路选择器输出影子灵敏放大器的结果，系统通过错误纠正机制保证功能的正确性。该 SRAM 应用在 DVS 系统中，当系统的电压降低时，存储单元的驱动电流降低，在主灵敏放大器使能时刻（时钟下降沿），位线的摆幅达不到灵敏放大器的失调电压，主灵敏放大器的输出出错，影子灵敏放大器开启时，位线的摆幅足够大，影子灵敏放大器输出正确，异或门输出高电平。在降低电压的同时，系统统计异或门输出高电平的概率，当出错率达到系统设置的阈值时，系统停止降低电源电压，此方式可以大约节省 35%的动态功耗。该方案也可以用于提升 SRAM 的性能，主灵敏放大器的输出直接送入 SRAM 后接的组合逻辑对提升性能有一定的帮助。

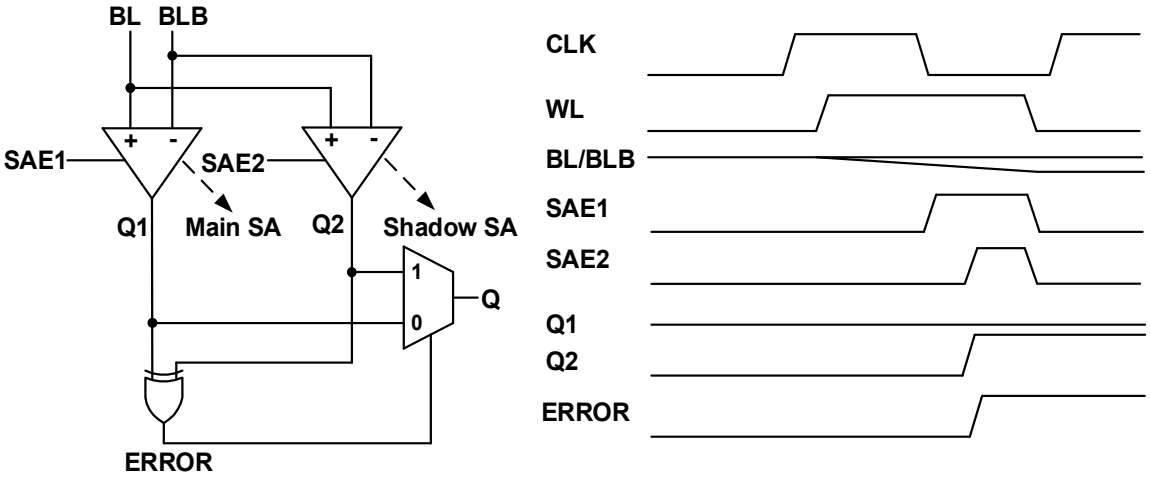


图 2-5 时序推测型 SRAM 的电路结构及读操作波形

文献[21]将文献[20]中的思想用于降低 SRAM 的读出延时。低电压下存储单元放电延时的拖尾现象导致了过于悲观的设计裕度，这使得 SRAM 的整体性能在近阈值区退化严重，传统的读出方式在字线的关断时刻开启灵敏放大器，为了降低 SRAM 的读出延时，文献[21]采用了如图 2-6 所示的电路结构，该方案在字线开启的中间时刻开启灵敏放大器，如图 2-7 所示，灵敏放大器的输出结果送入 SRAM 后接的组合逻辑电路，系统将灵敏放大器的第一次输出结果存入寄存器中，位线继续放

电，在字线关断时刻，灵敏放大器第二次启动（设计时保证灵敏放大器第二次输出的结果正确），异或门通过对比寄存器中的结果和灵敏放大器第二次的输出结果判断灵敏放大器第一次的输出结果是否正确，如果异或门输出高电平，多路选择器输出灵敏放大器第二次的输出结果，系统通过错误纠正机制保证功能的正确性。在灵敏放大器第一次启动时，绝大多数存储单元的位线摆幅已经超过了灵敏放大器的失调电压，只有极少数的弱驱动存储单元对应的位线摆幅低于灵敏放大器的失调电压，在第一次的读出结果中，出错率较低，纠错带来的额外代价可以忽略不计，采用本文的方案可以在一定程度上消除过度悲观的设计裕度带来的性能损失，故 SRAM 的工作频率大幅度提升，文献指出在低电压下 SRAM 的最高工作频率可以近似地提升两倍。

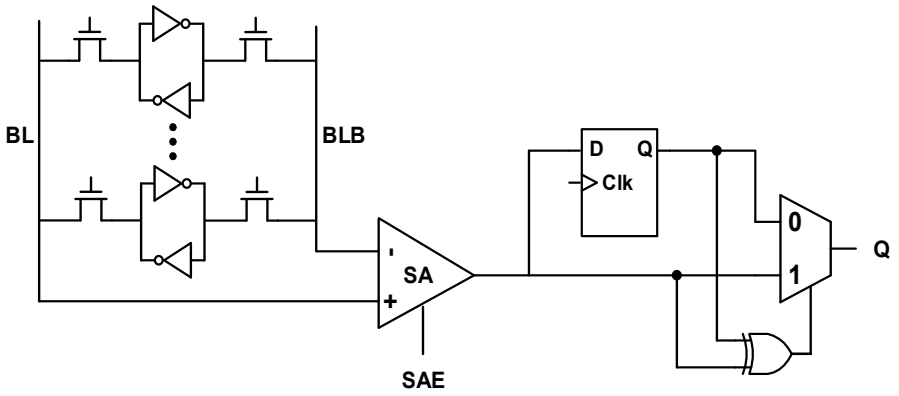


图 2-6 时序推测型 SRAM 的电路结构

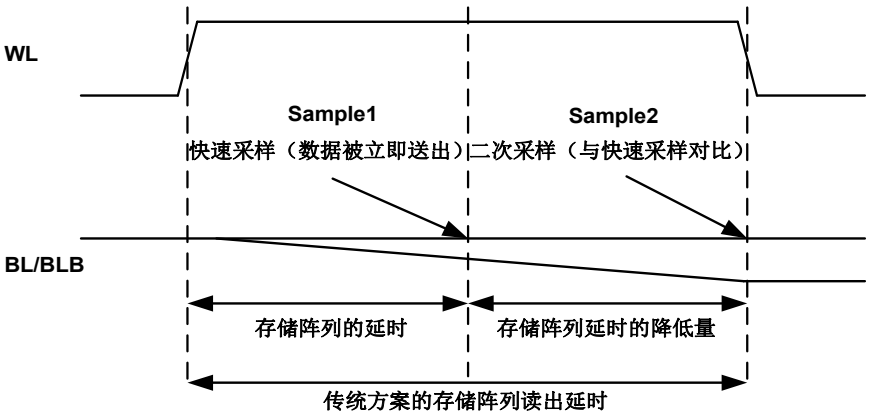


图 2-7 时序推测型 SRAM 的读出方式

下面对时序推测方案做总结与分析。时序推测方案通过降低存储阵列的读出延时提升 SRAM 的性能，与传统的读出方式不同，时序推测方案包含了检错和纠错过程，在读操作过程中，灵敏放大器两次启动，如图 2-8 所示，灵敏放大器的第一次启动对应推测型读出，多路选择器输出灵敏放大器第一次的读出结果，数据直接送入 SRAM 后接的组合逻辑，推测型读出用于提升 SRAM 的整体性能，此时并非所有的存储单元的位线摆幅均超过灵敏放大器的失调电压，数据存在出错的可能。灵敏放大器的第二次启动对应确认型读出，设计要保证确认型读出的结果一定正确，检错电路通过对比确认型读出和推测型读出的结果判断推测型读出的结果是否出错，如果检错电路判断推测型读出的结果错误，多路选择器输出灵敏放大器第二次的读出结果，系统通过纠错的方式保证功能的正确性。一般来说，推测型读出出错的概率较低，纠错对系统的功耗及吞吐率的影响可以忽略不计。推测型读出和确认型读出的时间间隔记为 ΔT，在近阈值区，由于存储单元位线放电延时的统计分布存

在较长的拖尾现象，为了降低弱驱动存储单元对 SRAM 存储阵列延时的影响，推测型读出的时刻会比较早，推测型读出和确认型读出的时间间隔 ΔT 会比较长。

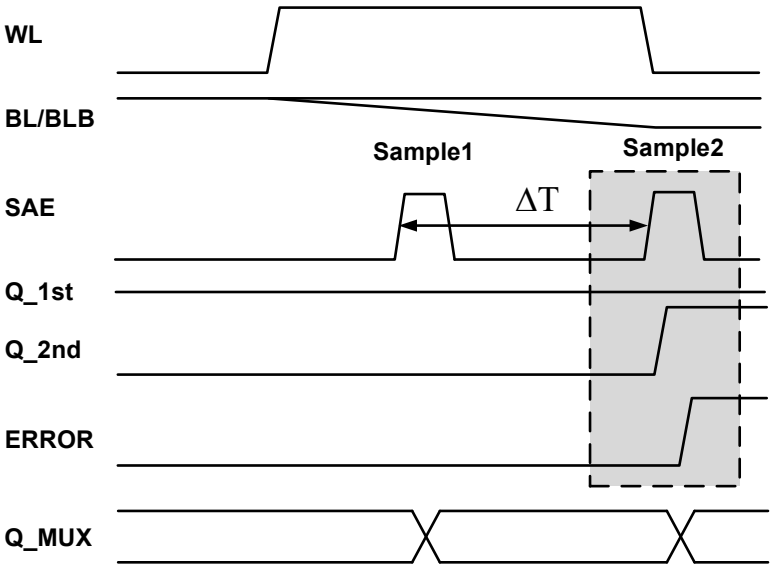


图 2-8 推测型读出的波形示意图

考虑时序推测技术在 SoC 系统中的应用，图 2-9 展示了 SoC 系统中的关键路径，关键路径由 SRAM 和组合逻辑电路组成，SRAM 对应两个读出延时，如果推测型读出正确，读出延时记为 T_{CQ1}，如果推测型读出错误，读出延时记为 T_{CQ2}，组合逻辑的延时记为 T_{COM}，现在做出如下定义，如果关键路径中的组合逻辑延时的比例较大，则称此关键路径为组合逻辑占主导的关键路径，如果关键路径中的 SRAM 的读出延时比例较大，则称此关键路径为 SRAM 占主导的关键路径。现对上述两种情况做具体分析。

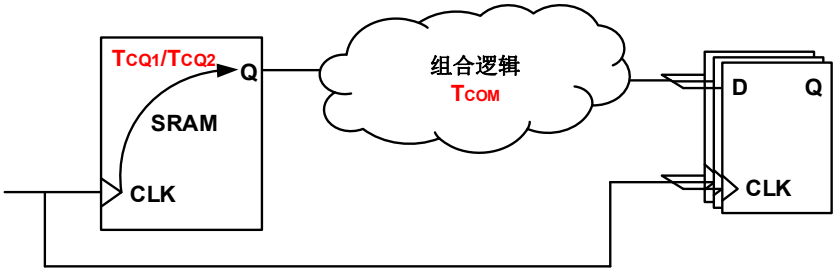


图 2-9 SoC 系统中的关键路径

1) 场景 1:$T_{CQ1}+T_{COM}>T_{CQ1}+\Delta T$（组合逻辑占主导的关键路径）

在场景 1 中，在推测型读出时刻，灵敏放大器输出 Q_1st，多路选择器输出 Q_1st，相应的延时记为 T_{CQ1}，紧接着位线继续放电，在确认型读出时刻，灵敏放大器输出 Q_2nd，如果 Q_2nd 不等于 Q_1st，ERROR 信号由低变高，推测型读出出错，多路选择器输出正确结果 Q_2nd，此时 SRAM 的读出延时更长，对应 T_{CQ2}，此时系统的关键路径会发生建立时间违规，由于系统在第二个时钟上升沿之前被告知推测型读出的结果是错误的，因此系统可以采用时钟门控[33-34]的方式对系统时钟做二分频以保证关键路径末端的寄存器采到正确的数据，如图 2-10 所示。

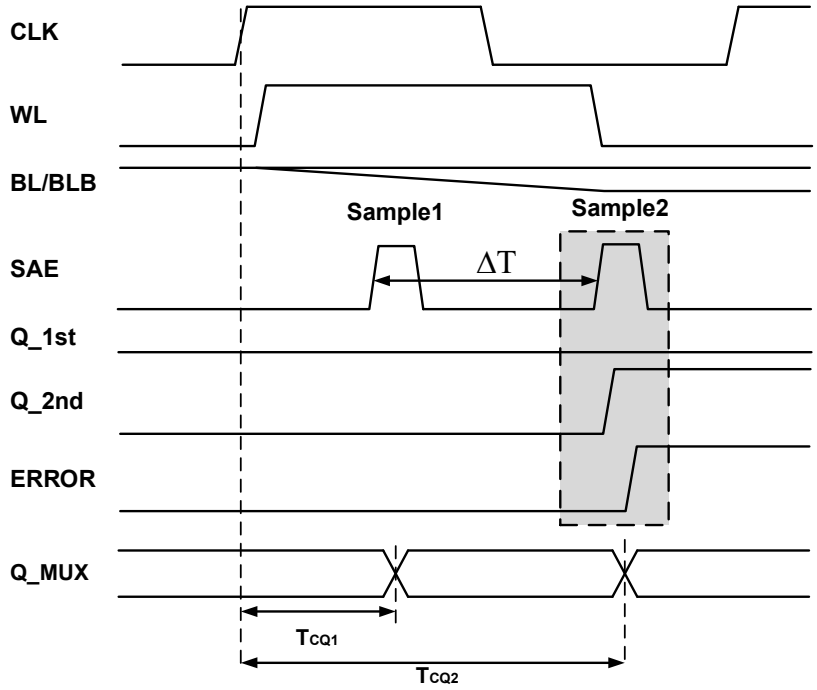


图 2-10 组合逻辑占主导地位的关键路径的读操作示意图

1) 场景 2:$T_{CQ1}+T_{COM1}<T_{CQ1}+\Delta T$（SRAM 占主导的关键路径）

在场景 2 中，在推测型读出时刻，灵敏放大器输出 Q_1st，多路选择器输出 Q_1st，相应的延时记为 T_{CQ1}，紧接着位线继续放电，在确认型读出时刻，灵敏放大器输出 Q_2nd，如果 Q_2nd 不等于 Q_1st，ERROR 信号由低变高，推测型读出出错，此时多路选择器输出正确的结果 Q_2nd，由于 SRAM 后接的组合逻辑延时过小，检错电路在第二个时钟上升沿后触发 ERROR 信号，故数据的读出和确认需要两个时钟周期，如图 2-11 所示。

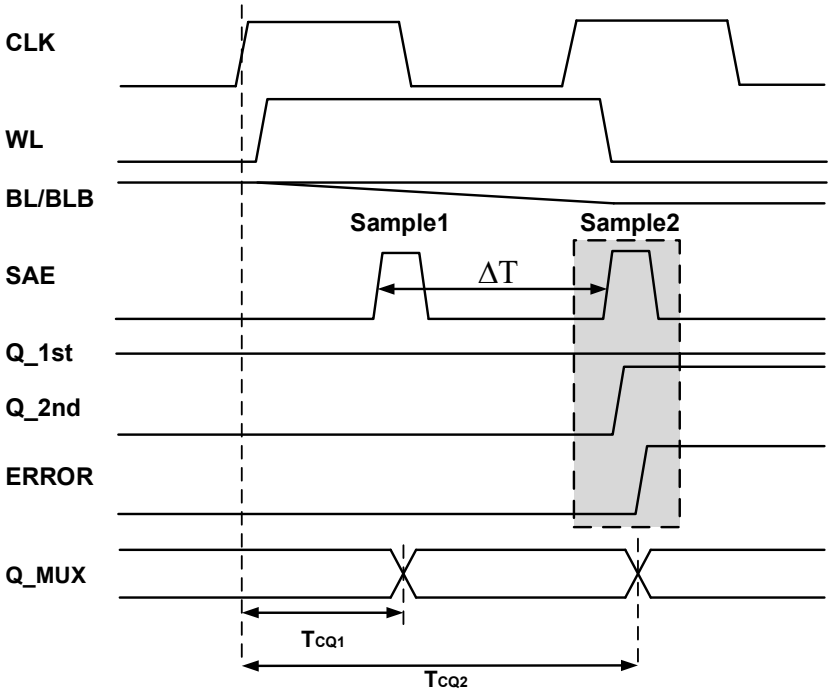


图 2-11 SRAM 占主导地位的关键路径的读操作示意图

在该场景下，时序推测技术会存在如下的缺点：

- 由于 ERROR 信号是在第二个时钟上升沿后触发，故系统是在第二个时钟上升沿之后被告

知推测型输出的结果是错误的，系统无法通过时钟门控的方式避免关键路径末端的寄存器采到错误的数据，系统需要复杂的纠错机制才能够保证功能的正确性。

- 在第二个时钟周期，由于确认型读出的需要，位线要接着第一个时钟周期继续保持放电状态，若系统在第二个时钟周期对 SRAM 其他地址的存储单元进行读写，这会导致存储阵列的位线访问冲突，故系统在第二个时钟周期内无法对 SRAM 其他地址的存储单元进行读写，这限制了该技术在 SoC 系统中的应用。

综上所述，上述的两种的时序推测方案在近阈值区的检错延时会比较长，这会限制该技术在 SoC 系统中的应用。除了上述的缺点之外，在文献[20]中，影子灵敏放大器、异或门和多路选择器构成了检错电路，而在文献[21]中，寄存器、异或门和多路选择器构成了检错电路，影子灵敏放大器和寄存器的面积较大，复杂的检错逻辑使得上述两种方案的面积开销相对较大，尤其是在一些性能要求较高的场合（性能要求高的场合位线上的存储单元数目少，检错电路的面积占比会比较高）。

2.3 本章小结

本章给出了 SRAM 时序推测技术的设计综述，首先介绍了 SRAM 的基本结构和工作原理，包括 SRAM 的模块组成及各个模块的功能；然后介绍了 SRAM 六管存储单元的工作原理，这是本文设计的基础；其次给出了时序推测技术的设计综述，时序推测技术是一种能够降低存储阵列中弱驱动存储单元对存储阵列延时影响的优化技术，该技术能够提升 SRAM 的整体性能；本章在最后对时序推测技术做了总结，重点分析了现有的时序推测技术在近阈值区的局限性，同时指出现有的时序推测技术其面积开销较大的问题。

第三章 时序推测型存储阵列的设计

针对现有的时序推测方案在近阈值区检错延时过大的缺点，本章提出了一种改进型时序推测方案，该方案在推测型读出后通过快速地调整灵敏放大器输入电压的极性实现快速检错，本章以时序推测方案为核心完成了时序推测型 SRAM 存储阵列的设计。3.1 节介绍了改进型的时序推测方案的原理；3.2 节介绍了时序推测方案的电路设计与实现，即时序推测型存储阵列的电路设计；3.3 节介绍了电路中的一些非理想因素对本文时序推测方案的影响；3.4 节介绍了本文的改进型时序推测方案的仿真结果。

3.1 时序推测方案的原理设计

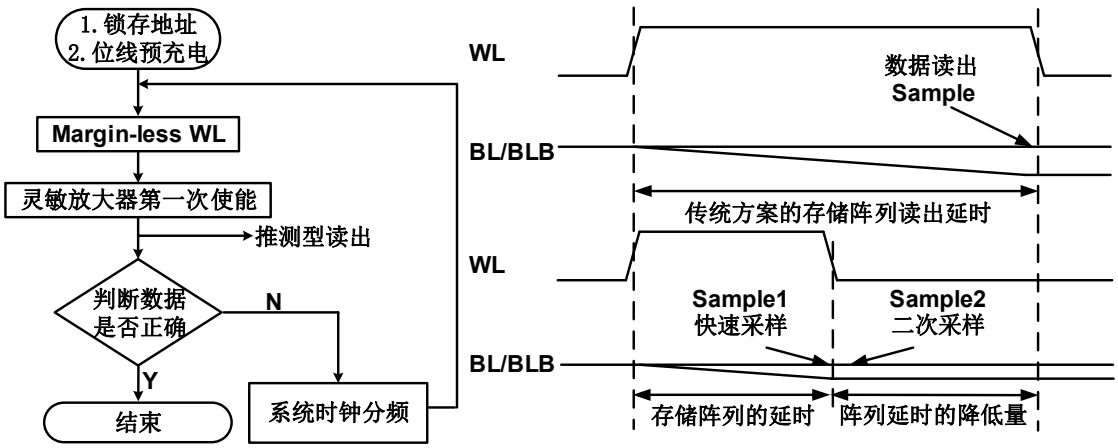


图 3-1 本文的时序推测方案的设计思路

本文提出的改进型时序推测方案是为了降低存储阵列中延时最大的存储单元对整体阵列延时的影响，且追求较低的检错延时。具体的设计思路如图 3-1 所示：第一步，字线开启，位线放电，当大部分存储单元的位线摆幅超过灵敏放大器的失调电压时，字线关断，灵敏放大器启动，数据快速输出至 SRAM 后接的组合逻辑电路。第二步，检错电路判断灵敏放大器的输出是否正确，如果出错，系统通过时钟门控的方式对系统时钟做分频，位线继续放电，位线的摆幅进一步放大，灵敏放大器再次启动，依次循环直至数据正确读出。时序推测方案的设计关键是检错方案的设计，本文的检错方案的设计思路如下：在灵敏放大器快速启动后（推测型读出），检错电路动态地调节灵敏放大器的输入电压和失调电压之间的关系，之后灵敏放大器二次启动（确认型读出），检错电路通过对比灵敏放大器两次的输出结果实现检错目的。

由于检错方案涉及到灵敏放大器的输入电压和失调电压对输出的影响，为了便于下文的讨论和分析，首先介绍灵敏放大器的输入电压和失调电压对输出的影响。

电压型灵敏放大器被广泛地应用在 SRAM 设计中，电压型灵敏放大器采用正反馈结构实现对小信号的快速放大，以高阻输入电压型灵敏放大器为例说明灵敏放大器的输入电压和失调电压对输出的影响[35-36]，高阻输入电压型灵敏放大器如图 3-2 所示。

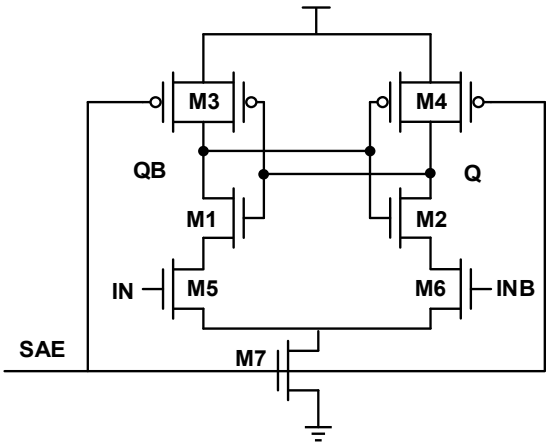


图 3-2 高阻输入电压型灵敏放大器

首先分析其工作原理（暂不考虑局部工艺波动造成的晶体管失配）：假定此时灵敏放大器的输入电压 V_{IN}=VDD，V_{INB}<VDD，$V_{INPUT}=V_{IN}-V_{INB}$>0。当 SAE 信号为低电平时，输出节点 Q 和 QB 保持 VDD，当 SAE 变为高电平时，正反馈建立，Q/QB 节点与地存在直流放电通路，由于晶体管 M5 的过驱动电压大于晶体管 M6，因此 QB 到地的放电电流大于 Q 到地的放电电流，故 QB 节点电压降低的速度比 Q 节点要快，与此同时 PMOS 管 M3 和 M4 导通，电源对 Q/QB 节点存在充电通路，且电源对 Q 节点的充电电流大于电源对 QB 节点的充电电流，正反馈作用的最终结果是 V_Q=VDD，V_{QB}=0。同理可得，当 $V_{INPUT}=V_{IN}-V_{INB}$<0 时，正反馈作用的最终结果是 V_Q=0，V_{QB}=VDD。

通过上文的分析可以得出如下结论：在不考虑局部工艺波动的理想情况下，当 V_{INPUT}>0 时，对应的输出 V_Q=VDD；当 V_{INPUT}<0 时，对应的输出 V_Q=0。随着半导体工艺特征尺寸的不断缩小，局部工艺波动造成的晶体管失配会对灵敏放大器的输出造成一定的影响，现考虑如下两种情况：

1) 局部工艺波动导致放电通路 M1-M5 的驱动能力弱于 M2-M6 的驱动能力。

当 V_{INPUT}>0 时，灵敏放大器的输出会受到局部工艺波动的影响，当 0<V_{INPUT}<ΔV 时，灵敏放大器的输出 V_Q=0，与理想情况不一致，当 V_{INPUT}>ΔV 时，灵敏放大器的输出 V_Q=VDD，与理想情况一致。当 V_{INPUT}<0 时，灵敏放大器的输出不受到局部工艺波动的影响，相应的 V_Q=0，在这种情况下灵敏放大器的失调电压 V_{OFFSET}=ΔV。

2) 局部工艺波动导致放电通路 M1-M5 的驱动能力强于 M2-M6 的驱动能力。

当 V_{INPUT}<0 时，灵敏放大器的输出会受到局部工艺波动的影响，当-ΔV<V_{INPUT}<0 时，灵敏放大器的输出 V_Q=VDD，与理想情况不一致，当 V_{INPUT}<-ΔV 时，灵敏放大器的输出 V_Q=0，与理想情况一致。当 V_{INPUT}>0 时，灵敏放大器的输出不受到局部工艺波动的影响，相应的 V_Q=VDD，在这种情况下灵敏放大器的失调电压 V_{OFFSET}=-ΔV。

通过上文的分析，可以得出如下的结论：当|V_{INPUT}|>|V_{OFFSET}|时，灵敏放大器的输出一定正确，故在 SRAM 的设计中，灵敏放大器的启动时刻要求位线的摆幅超过灵敏放大器的失调电压（更准确地来说是位线的摆幅超过灵敏放大器失调电压的绝对值）。表格 3.1 和表格 3.2 对上述的分析做了总结。

表 3.1 灵敏放大器的输入电压和失调电压对灵敏放大器输出的影响（$V_{OFFSET}>0$）

	实际输出	期望输出
$V_{INPUT}<0$	0	0
$0<V_{INPUT}<V_{OFFSET}$	0	1
$V_{INPUT}>V_{OFFSET}$	1	1

表 3.2 灵敏放大器的输入电压和失调电压对灵敏放大器输出的影响（$V_{OFFSET}<0$）

	实际输出	期望输出
$V_{INPUT}<V_{OFFSET}$	0	0
$V_{OFFSET}<V_{INPUT}<0$	1	0
$V_{INPUT}>0$	1	1

上文分析了灵敏放大器的输入电压和失调电压对输出的影响，本文的检错方案通过动态地调节灵敏放大器的输入电压和失调电压之间的关系以达到检错的目的，本文检错方案的流程如图 3-3 所示。

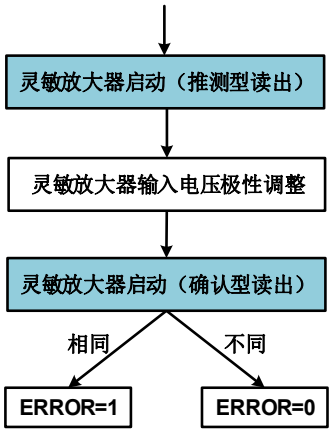


图 3-3 检错原理

灵敏放大器第一次启动（推测型读出）之后，检错电路迅速地调整灵敏放大器的输入电压极性，接着灵敏放大器再次启动（确认型读出），如果两次输出结果相反，ERROR 信号置“0”，如果两次输出的结果相同，ERROR 信号置“1”。图 3-4 展示了灵敏放大器的输入电压在极性调整前后对应的统计分布。假定灵敏放大器的失调电压 V_{OFFSET} 大于零。

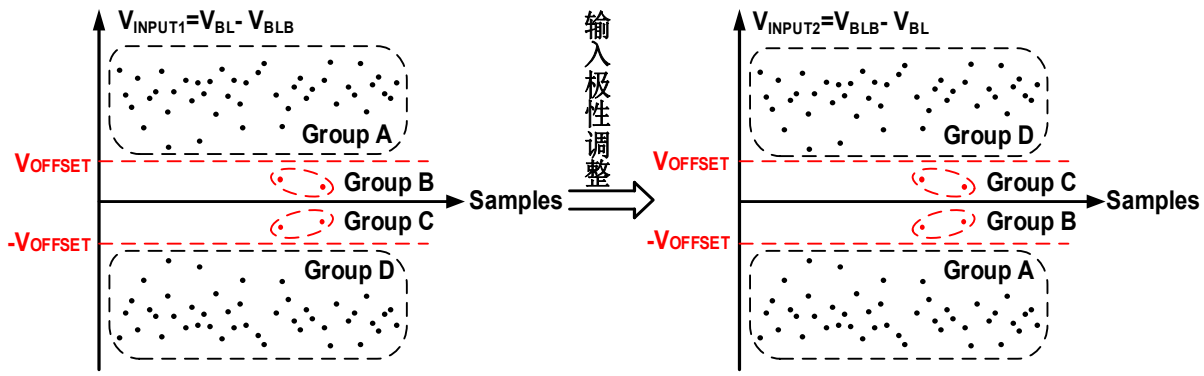


图 3-4 灵敏放大器的输入电压的摆幅分布

如果推测型读出的结果是“1”，则灵敏放大器的输入电压 $V_{INPUT1}=V_{SW}=V_{BL}-V_{BLB}>V_{OFFSET}$，即对应着图 3-4 中的 A 组，检错电路调整灵敏放大器的输入电压极性后，$V_{INPUT2}=-V_{INPUT1}<V_{OFFSET}$，确认型的读出结果为“0”，即当存储单元的位线摆幅 $V_{SW}>V_{OFFSET}$ 时，推测型读出的结果和确认型读出的结果相反。

如果推测型读出的结果为“0”，则灵敏放大器的输入电压 $V_{INPUT1}=V_{SW}<V_{OFFSET}$，对应着图 3-4 中的 B 组、C 组或 D 组。若 V_{INPUT1} 处在 B 组，检错电路调整灵敏放大器的输入电压极性后，$V_{INPUT2}=-V_{INPUT1}<V_{OFFSET}$，确认型读出的结果仍为“0”。若 V_{INPUT1} 处在 C 组，检错电路调整灵敏放大器的输入电压极性后，$V_{INPUT2}=-V_{INPUT1}<V_{OFFSET}$，确认型读出的结果仍为“0”。即当存储单元的位线摆幅 $|V_{SW}|<|V_{OFFSET}|$ 时，确认型读出和推测型读出一致。若 V_{INPUT1} 处在 D 组，经过极性调整，$V_{INPUT2}=-V_{INPUT1}>V_{OFFSET}$，确认型读出结果为“1”，即当存储单元的位线摆幅 $V_{SW}<-V_{OFFSET}$ 时，确认型读出结果和推测型读出结果相反。通过上述的分析，将结论总结如表 3.3 所示。

表 3.3 推测型读出结果和确认型读出结果与存储单元位线摆幅之间的关系

	期望输出	推测型读出（Q_1st）	确认型读出（Q_2nd）	Q_1st NXOR Q_2nd
$V_{SW}>V_{OFFSET}$	1	1	0	0
$0<V_{SW}<V_{OFFSET}$	1	0	0	1
$-V_{OFFSET}<V_{SW}<0$	0	0	0	1
$V_{SW}<-V_{OFFSET}$	0	0	1	0

综上所述，当存储单元的位线摆幅 $|V_{SW}|>|V_{OFFSET}|$ 时，位线的摆幅足够大，推测型读出结果正确，并且确认型读出与推测型读出的结果相反，当存储单元的位线摆幅 $|V_{SW}|<|V_{OFFSET}|$ 时，位线摆幅的绝对值相对较小，灵敏放大器可能会导致误操作，推测型读出可能出错，并且确认型读出与推测型读出的结果一致。因此检错方案可以以推测型读出和确认型读出的结果相同作为 ERROR 信号置“1”的触发条件，即本文的检错标准如式（3.1）所示。

$$ERROR = Q\_1st \odot Q\_2nd \tag{3.1}$$

上述的检错标准不会存在漏判，即检错电路将错误的推测型读出判对，当 $0<V_{INPUT1}<V_{OFFSET}$ 时，期望输出“1”，推测型读出的结果为“0”，推测型读出错误，经过灵敏放大器输入电压极性的调整，确认型读出的结果一定为“0”，ERROR 信号置高。上述的检错标准会存在误判，即检错电路将正确的推测型读出判错，当 $-V_{OFFSET}<V_{INPUT1}<0$ 时，期望输出“0”，推测型读出的结果为“0”，推测型读出正确，经过灵敏放大器输入电压极性的调整，确认型读出的结果为“0”，ERROR 置高，即确认型读出与推测型读出的结果一致是推测型读出结果出错的必要不充分条件。总而言之，当存储单元的位线摆幅满足 $-V_{OFFSET}<V_{SW}<V_{OFFSET}$ 时，检错方案无法判断推测型读出是否正确，检错方案带来的误判会对本文时序推测方案的收益造成一定的影响，这将在后文进行分析。

3.2 时序推测方案的电路设计

3.2.1 时序推测型存储阵列的整体结构

时序推测型存储阵列总体的结构如图 3-5 所示，包括了 M×N 个存储单元（一条字线连接 N 个

存储单元，一条位线连接 M 个存储单元）、切换开关、灵敏放大器、锁存器及总线检测单元。切换开关连接位线和灵敏放大器的输入端，用于调整灵敏放大器输入电压的极性。时序推测方案需要对比灵敏放大器前后两次的输出结果，需要记忆单元存储灵敏放大器第一次的输出结果，相比于寄存器，锁存器的面积开销更小，因此本文采用锁存器存储灵敏放大器第一次的输出结果。总线检测单元用于对比推测型读出和确认型读出的结果，并触发 ERROR 信号。切换开关、锁存器和总线检测单元构成了检测逻辑。

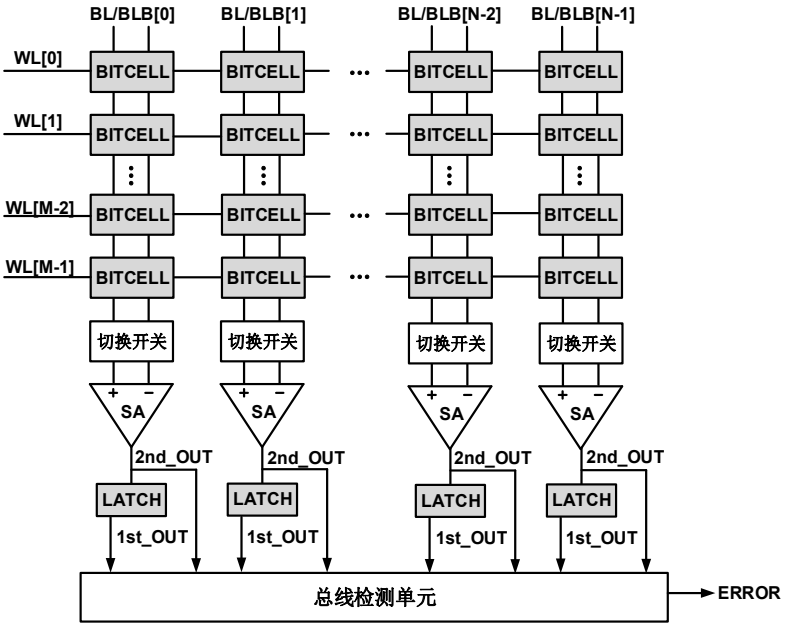
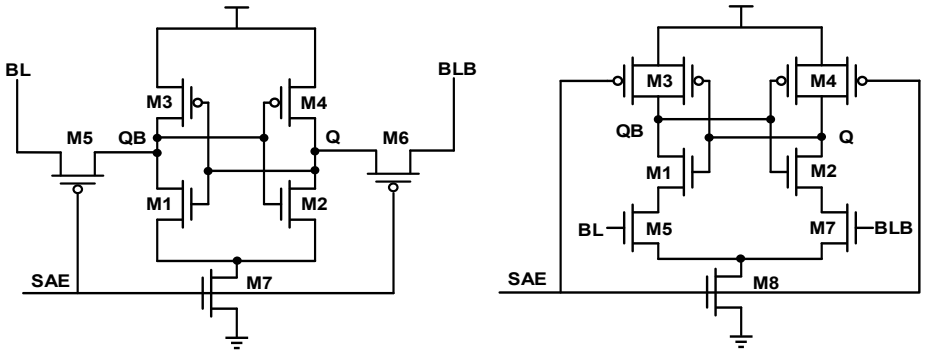


图 3-5 时序推测型存储阵列的整体结构

3.2.2 灵敏放大器的设计



(a)解耦型灵敏放大器　　(b)高阻输入型灵敏放大器

图 3-6 电压型灵敏放大器

为了降低 SRAM 的读出延时，SRAM 的设计中通常采用灵敏放大器，电压型灵敏放大器由于电路结构简单被广泛地应用在 SRAM 的设计中，电压型灵敏放大器依靠内部交叉耦合反相器的正反馈作用实现对小信号的快速放大。电压型灵敏放大器可以分为：解耦型灵敏放大器和高阻输入型灵敏放大器，如图 3-6 所示。本文的检错方案通过动态地调整灵敏放大器的输入电压与失调电压之间的关系实现快速检错，对灵敏放大器的设计要求如下：在推测型读出和确认型读出之间，位线和灵敏放大器之间不能存在泄漏电流，泄漏电流会影响位线的摆幅，从而对灵敏放大器的输入电压造成一

定的影响，使得检错结果受到一定影响。对于解耦型灵敏放大器，位线 BL/BLB 和灵敏放大器的内部节点通过传输管连接，当传输管导通的时候，位线与内部节点存在漏电路径，泄漏电流会对检错结果造成影响，导致 SRAM 功能错误，如图 3-7 所示。

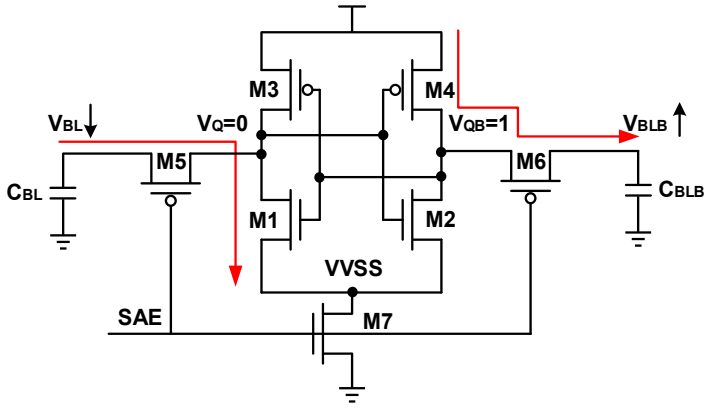


图 3-7 解耦型灵敏放大器中的泄漏电流

假定此时存储单元存储数据“1”，字线关断后位线的摆幅相对较小，此时灵敏放大器的输入电压 $V_{INPUT1}=V_{BL}-V_{BLB}<V_{OFFSET}$，SAE 信号变高后，在正反馈的作用下，灵敏放大器输出 Q=“0”，即推测型读出是错误的。在推测型读出与确认型读出的间隙，SAE 信号保持为低电压，由于晶体管 M5 和 M6 导通，BL 通过 M5 和 M1 对内部节点 VVSS 充电，电源通过 M4 和 M6 对 BLB 充电，泄漏电流对位线的摆幅 V_{SW} 造成了影响，当灵敏放大器二次启动时，可能会出现 $V_{INPUT2}>V_{OFFSET}$ 的情形，此时确认型读出的结果为“1”，即确认型读出结果和推测型读出的结果相反，根据前文对检错原理的描述，此时检错电路认定推测型读出正确，故泄漏电流可能会导致检错电路发生漏判，从而造成了系统功能性错误。

对于高阻型灵敏放大器，位线 BL/BLB 和 MOS 管的栅极直接相连，位线和灵敏放大器的栅极之间存在栅极泄漏电流，相比于 MOS 管的亚阈值电流，栅极泄漏电流可以忽略不计，故高阻型灵敏放大器是本设计的最佳选择。

3.2.3 切换开关的设计

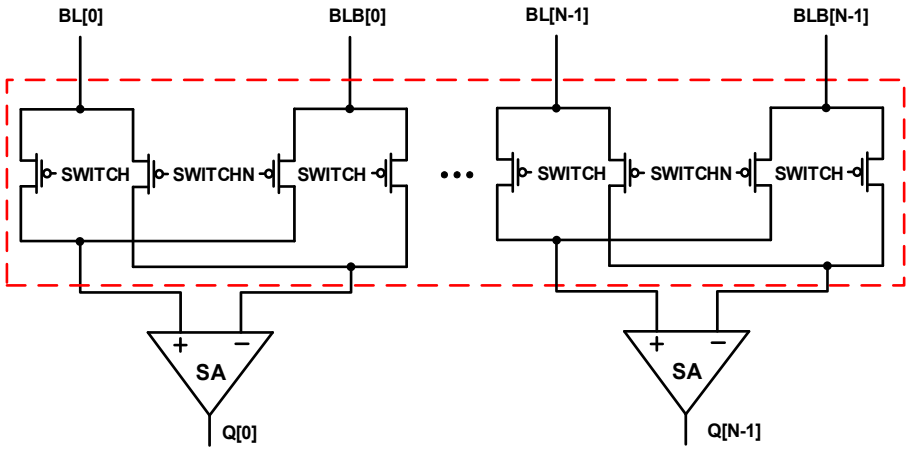


图 3-8 PMOS 切换开关

切换开关可以采用 CMOS 传输门结构，也可以采用 MOS 传输管结构。CMOS 传输门在传输低电平和传输高电平时都不会存在阈值损失，NMOS 传输管在传输低电平时无阈值损失，但是在传输

高电平时存在阈值损失，与之相反，PMOS 传输管在传输高电平时无阈值损失，但是在传输低电平时存在阈值损失。为了降低面积开销，本文采用 PMOS 管作为切换开关，相比传输门结构节约了大约 50%的面积开销。切换开关的电路结构如图 3-8 所示，阈值损失不会对本文的设计造成任何影响，具体分析如下：

根据 PMOS 传输管的传输特性，可以得到灵敏放大器的输入电压和位线摆幅 V_{SW} 之间的关系如下：

1) 当$|V_{SW}|<VDD-|V_{THP}|$时：

$$\begin{cases}|V_{INPUT1}|=|V_{SW}| \\ V_{INPUT2}=-V_{INPUT1}\end{cases} \tag{3.2}$$

2) 当$|V_{SW}|>VDD-|V_{THP}|$时：

$$\begin{cases}|V_{INPUT1}|=VDD-|V_{THP}| \\ V_{INPUT2}=-V_{INPUT1}\end{cases} \tag{3.3}$$

从式（3.2）和式（3.3）可以看出，只有当位线摆幅的绝对值$|V_{SW}|$大于 $VDD-|V_{THP}|$时才会发生阈值损失，相应的灵敏放大器的输入电压被钳位在 $VDD-|V_{THP}|$。在本文的设计中，PMOS 传输开关采用 TSMC 28nm 工艺的超低阈值型 PMOS 管，相应的阈值电压大约为 150mV，在 0.5V 和 0.9V 条件下，当位线的摆幅超过 350mV 和 750mV 时，相应的灵敏放大器的输入电压被钳位在 350mV 和 750mV。在 SRAM 的实际工作过程中，位线的摆幅通常较小，受局部工艺偏差的影响，少部分存储单元的位线摆幅会超过上述值，并且此电压值远大于灵敏放大器的失调电压，故阈值损失造成的电压钳位现象不会对灵敏放大器的输出造成任何影响。

3.2.4 锁存器的设计

锁存器可以分为两大类：静态锁存器和动态锁存器，如图 3-9 所示。

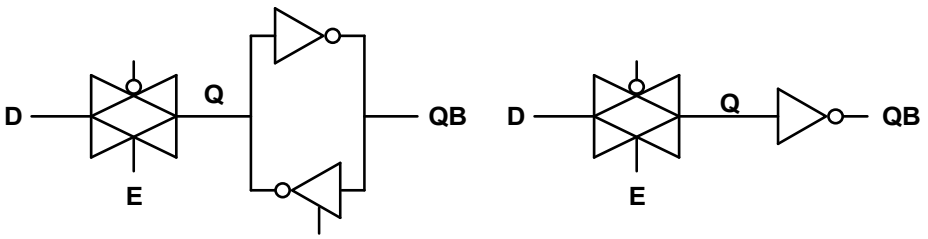


图 3-9 静态锁存器和动态锁存器

静态锁存器由传输门、反相器及三态反相器构成，数据写入时，三态反相器关断，数据通过传输门送入内部；数据保持时，传输门关断，三态反相器使能，交叉耦合反相器的正反馈作用保证数据的稳定。动态锁存器由传输门和反相器构成，数据写入时，数据经过传输门直接送入内部；数据保持时，传输门关断，锁存器依靠内部的寄生电容存储信息。

静态锁存器的优点是稳定性好，缺点是面积开销大。动态锁存器的优点是面积开销小，缺点是稳定性差。动态锁存器依靠内部的寄生电容存储信息，但是内部节点的电荷易受到泄漏电流及噪声的影响，电荷容易丢失，锁存器内部存储的信息易发生改变，因此动态锁存器只能短暂的锁存数据。

在本文的设计中，灵敏放大器的第一次输出结果存在锁存器中，此结果不仅用于和灵敏放大器的第二次输出结果作比较，还要送入 SRAM 后接的组合逻辑电路，为了避免存储信息的改变对后接

组合逻辑电路的影响，本文在动态锁存器的基础上增加了泄漏电流补偿管 M1 和 M2，如图 3-10 所示，若内部节点 Q 存“1”，PMOS 管 M1 对 Q 点充电，故 Q 点可以维持在高电平；若内部节点 Q 存“0”，NMOS 管 M2 对 Q 点放电，故 Q 点可以维持在低电平。为了避免数据写入的过程中传输门与 M1/M2 的竞争现象，M1 和 M2 采用超高阈值型晶体管。相比于传统的静态锁存器，本文的锁存器面积大约降低了 30%。

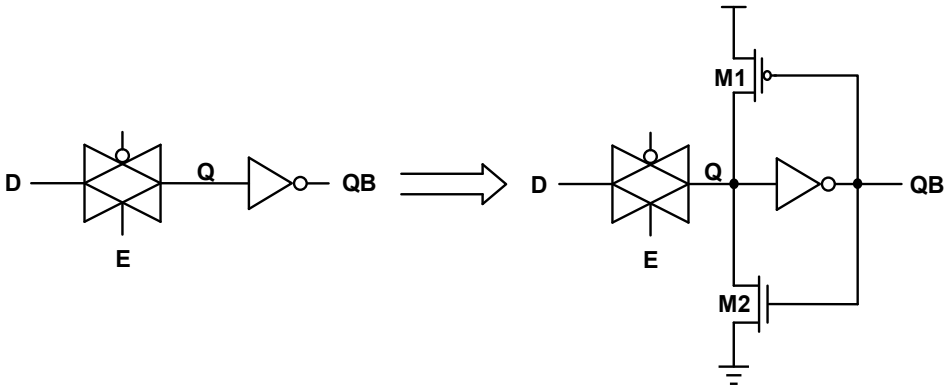
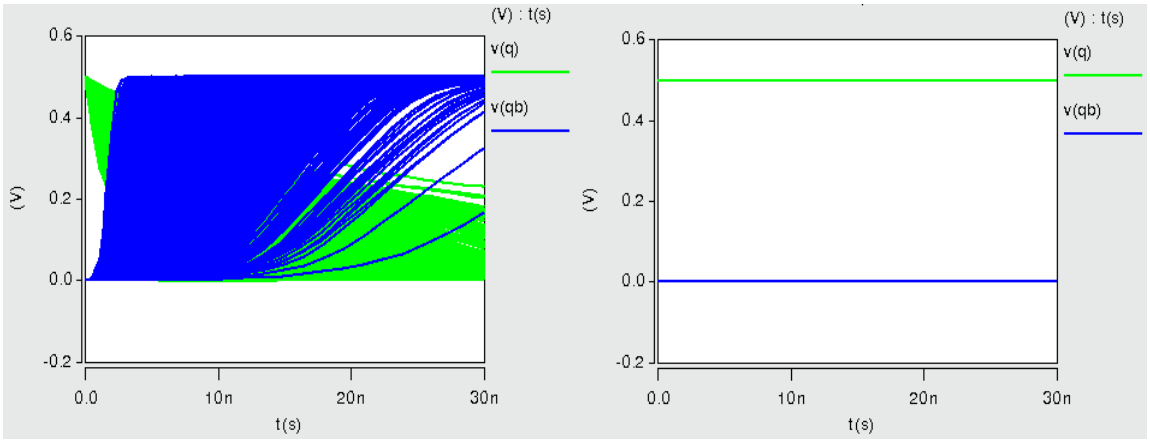


图 3-10 带漏电补偿晶体管的锁存器

图 3-11 是未采用泄漏电流补偿管和采用了泄漏电流补偿管的蒙特卡洛仿真对应的波形，仿真条件为 0.5V，FFG 工艺角，70℃。仿真结果表明：泄漏电流补偿管消除了泄漏电流对存储节点的影响，存储信息可以长久维持。



（a）未采用泄漏电流补偿管的锁存器　　（b）采用泄漏电流补偿管的锁存器

图 3-11 锁存器的蒙特卡洛仿真波形

3.2.5 总线检测电路的设计

总线检测电路是用来对比推测型读出和确认型读出的结果，并用于触发 ERROR 信号。根据前文对检错原理的分析，在 N 个灵敏放大器中，若所有灵敏放大器的推测型输出与确认型输出的结果相反，则 ERROR 信号置“0”；若至少存在一个灵敏放大器，其对应的推测型输出和确认型输出结果一致，则 ERROR 信号置“1”。其对应的逻辑表达式如下：

$$ERROR = Q[0]\_1st \odot Q[0]\_2nd + \cdots + Q[N-1]\_1st \odot Q[N-1]\_2nd \quad (3.4)$$

将上述逻辑表达式翻译成电路如图 3-12 所示。

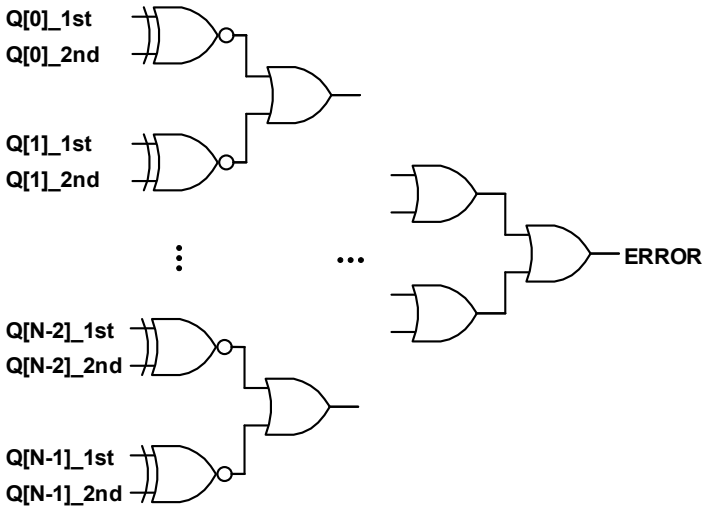


图 3-12 电路实现

上述电路采用静态门结构，由 N 个同或门及若干或门组成，这大大地增加了面积开销，并且增加了检错电路的延时和功耗。本文提出了一种总线型的检测单元，该检测单元具有面积小，延时低，且功耗低的特点，总线型检测单元的电路如图 3-13 所示。

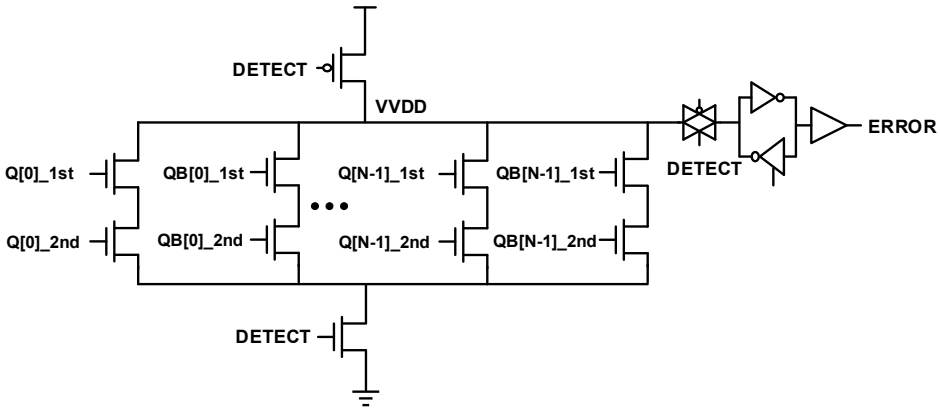


图 3-13 总线检测的电路实现

总线型检测单元采用动态门结构，其工作原理如下：当 DETECT 信号为低电平时，电路进入预充阶段，内部节点 VVDD 被预充至 VDD；当 DETECT 信号变为高电平时，电路进入求值阶段，当所有灵敏放大器的推测型输出与确认型输出的结果相反时，内部节点 VVDD 至 VSS 不存在直流通路，VVDD 保持高电平，ERROR 信号输出“0”。在 N 个灵敏放大器中，若至少存在一个灵敏放大器，其对应的推测型输出和确认型输出结果一致，则内部节点 VVDD 至 VSS 存在直流通路，VVDD 被拉低，ERROR 信号输出高电平。

总线型检测单元的设计需要注意如下问题：由于内部节点 VVDD 的负载很大，为了加速电路求值的速度，NMOS 下拉通路通常会选择低阈值型晶体管，低阈值型晶体管速度快，但是泄漏电流较高，假定所有的灵敏放大器的推测型输出与确认型输出的结果相反，即在动态门结构中不存在直流通路，只存在泄漏电流通路，泄漏电流可能会将 VVDD 放电至低电平，使得 ERROR 信号翻转为高电平，导致误操作，如图 3-14 所示。

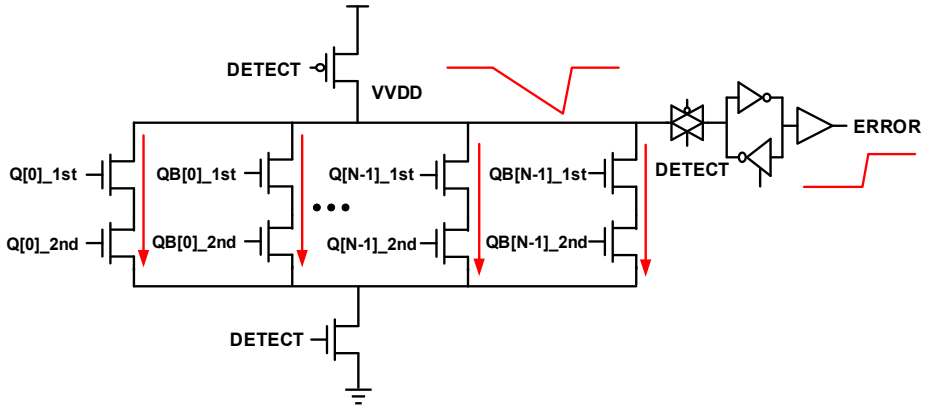


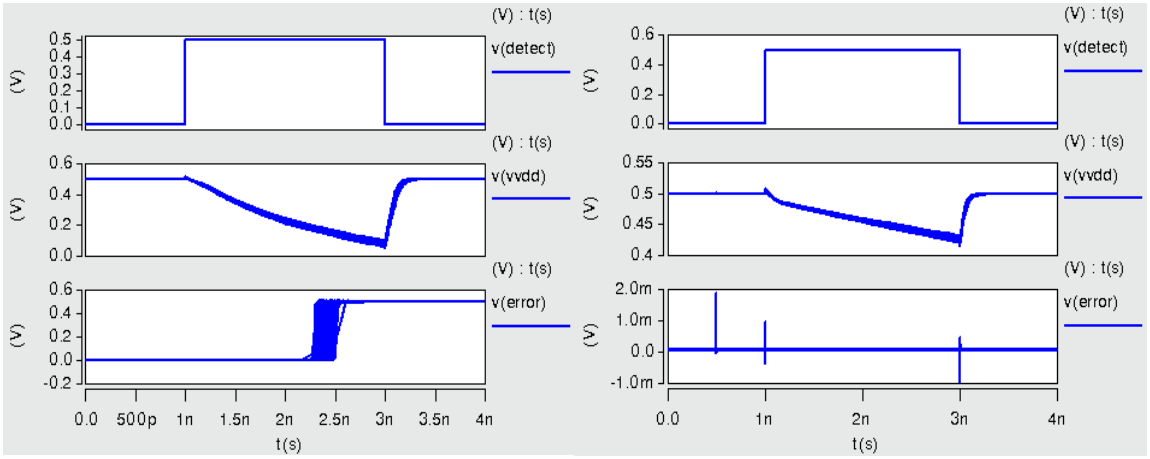
图 3-14 总线检测单元中的泄漏电流

MOS 管的泄漏电流[37]的公式如下：

$$I_{leakage} = I_0 \exp\left(\frac{V_{GS} - V_{TH}}{nV_t}\right) \tag{3.5}$$

在式（3.5）中，I_0 代表 $V_{GS}=V_{TH}$ 时的 MOS 管电流，V_{GS} 代表 MOS 管的过驱动电压，V_{TH} 代表 MOS 管的阈值电压，$V_t=kt/q$ 叫做热电压。

MOS 管的泄漏电流和阈值电压 V_{TH} 呈指数关系，阈值电压 V_{TH} 的变化会对泄漏电流的大小造成显著的影响。在先进工艺下，MOS 管存在短沟道效应（Short-Channel Effect，SCE）[38-39]，MOS 晶体管沟道越短，源漏区 PN 结耗尽层电荷在总的沟道区耗尽层电荷中的比例越大，使实际由栅极电压控制的耗尽层电荷减少，造成了 MOS 管的阈值电压随着沟道长度的减小而降低，即 MOS 管的阈值电压会随着沟道长度的增加而增加，MOS 管的沟道长度稍做增加，其泄漏电流可以大幅度降低。图 3-15 给出了不同沟道长度的 MOS 管对输出的影响（假定 VVDD 到 VSS 不存在直流通路，只存在漏电通路），仿真条件为 0.5V，FFG 工艺角，70℃。



（a）L=30nm　　（b）L=40nm

图 3-15 泄漏电流对总线检测单元输出的影响

当 MOS 管沟道长度 L=30nm 时，此时泄漏电流较大，泄漏电流使得检测单元的内部节点 VVDD 的电压大幅度降低，ERROR 信号变成高电平，造成了误操作。当 MOS 管的沟道长度调整成为 40nm 时，泄漏电流大幅度降低，节点 VVDD 的压降只有几十个毫伏，ERROR 信号仍保持在低电平。

栅长的增加会降低 MOS 管的宽长比，导致 MOS 管饱和电流的降低，使得总线检测单元的延时增加，表 3.4 给出了沟道长度 L=30nm 和 L=40nm 时，不同 PVT 条件下的总线检测单元的延时。

表 3.4 总线检测单元的延时与 MOS 管沟道长度的关系

	L=30nm	L=40nm
0.5V FFG 工艺角 70℃	506ps	554ps
0.5V TTG 工艺角 25℃	762ps	824ps
0.5V SSG 工艺角 0℃	1154ps	1237ps

仿真结果表明：沟道长度的增加会导致总线检测单元的延时大约增加 10%，和存储阵列的读出延时相比较，总线检测单元延时的增加可以忽略不计。

3.2.6 时序推测方案的工作过程

本文提出的改进型时序推测方案的读操作的完整过程如下：字线 WL 开启，位线 BL/BLB 放电，当绝大多数存储单元的位线摆幅超过灵敏放大器的失调电压时，字线关断，灵敏放大器在 SAE 信号的控制下启动（推测型读出），数据快速输出至 SRAM 的后接组合逻辑电路，锁存器在 LATCH 信号的控制下锁存灵敏放大器的输出，紧接着通过切换开关调整灵敏放大器输入电压的极性，灵敏放大器二次启动（确认型读出），当 DETECT 信号为高时，总线检测单元对比推测型读出和确认型读出的结果，并触发 ERROR 信号。由于推测型读出的数据被立即送到 SRAM 外部的组合逻辑电路，灵敏放大器输入电压极性调整的时间和确认型读出的时间不会影响存储阵列的读出延时，故该方案可以在一定程度上降低存储阵列的读出延时。当位线的摆幅较大时，读操作的波形对应图 3-16，此时推测型读出的结果与确认型读出的结果恰好相反，推测型读出的结果正确，ERROR 信号置“0”，此时系统不需要采用任何纠错手段。

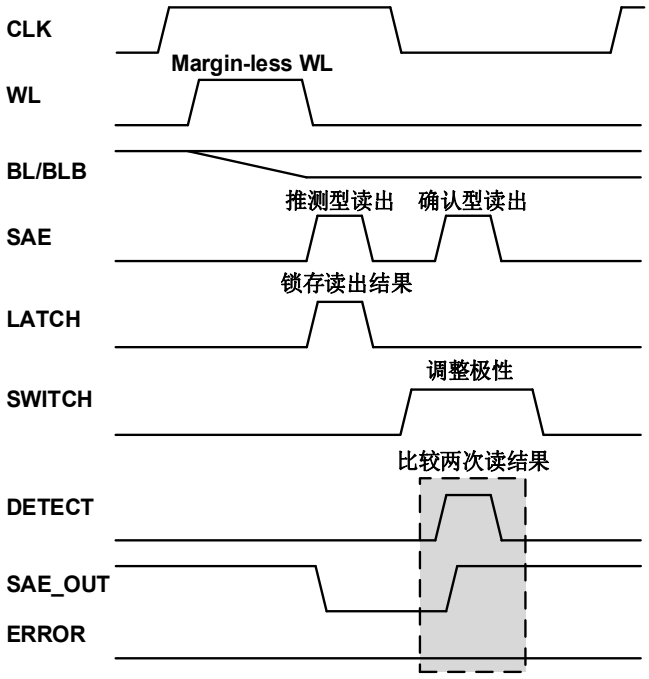


图 3-16 一个周期读

当位线的摆幅较低时，读操作的波形如图 3-17 所示，此时推测型输出和确认型输出的结果相同，ERROR 信号置“1”，此时检错电路无法判断推测型读出的结果是否正确，字线需要进一步开启以保证建立足够的位线摆幅，此时存储阵列的读出延时更长，系统通过时钟门控的方式关断下个时钟周期的上升沿或下降沿，系统时钟的分频保证了多个时钟周期读取当前地址的数据，避免了关键路径

末端的触发器采样错误数据。

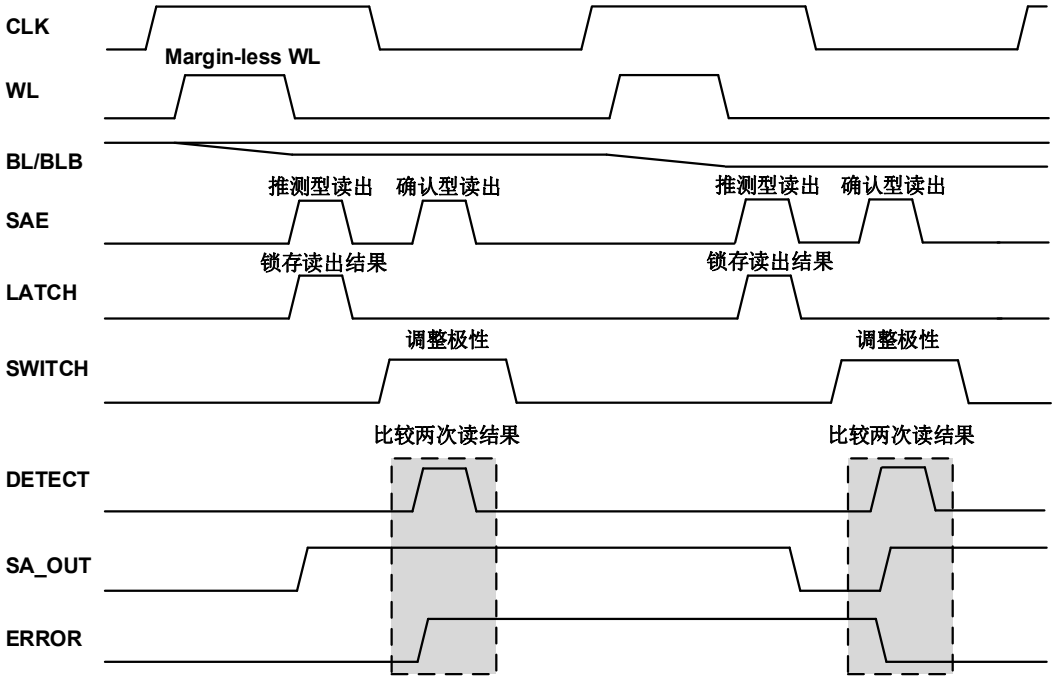


图 3-17 多个时钟周期读

3.3 噪声分析

在前文对时序推测方案的原理描述中，我们假定灵敏放大器的输入电压在极性调整后仅仅是极性发生了变化，电压的摆幅大小并没有发生变化，这种假设过于理想，电路中的一些非理想的因素会导致灵敏放大器的输入电压的摆幅发生变化，这些非理想因素来源于四个方面：1. 灵敏放大器的泄漏电流；2. 存储单元的泄漏电流；3. 串扰；4. 电荷共享。上述的非理想噪声因素可能会给本文的时序推测方案造成一定的影响，需要做具体的分析。

3.3.1 灵敏放大器的泄漏电流对灵敏放大器输入电压的影响

上文已经对该问题做了具体的分析，本文采用高阻输入型灵敏放大器，高阻输入型灵敏放大器的栅极直接与位线相连接，栅极的泄漏电流对位线电压的影响可以忽略不计，因此灵敏放大器输入电压的摆幅大小不会受到影响，如图 3-18 所示。

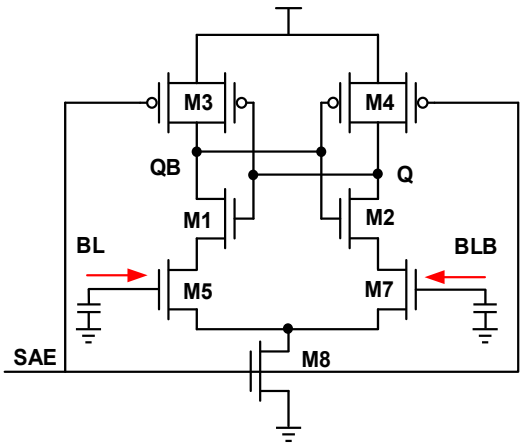


图 3-18 位线与高阻型灵敏放大器间的栅极泄漏电流

3.3.2 存储单元的泄漏电流对灵敏放大器输入电压的影响

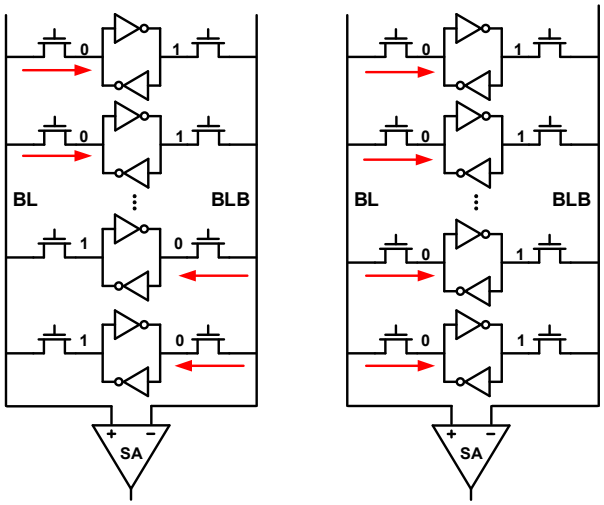
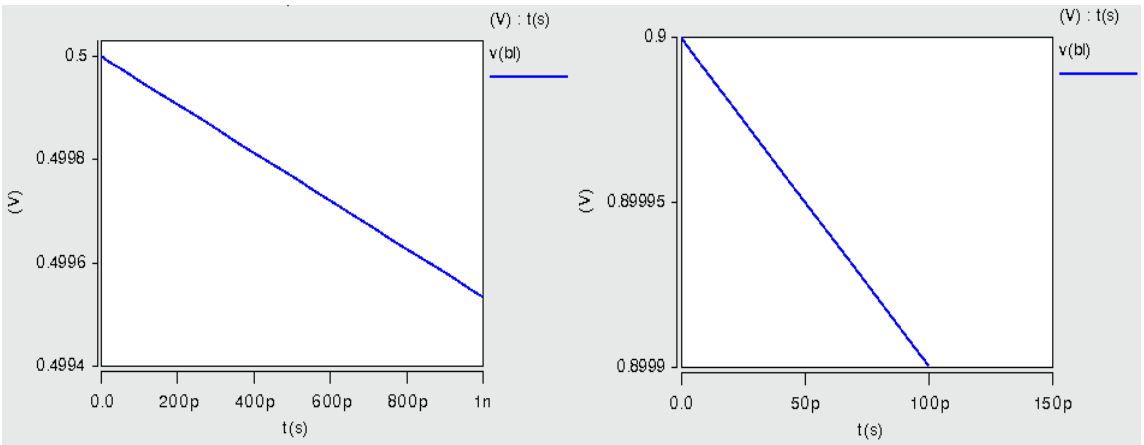


图 3-19 存储单元的泄漏电流

在灵敏放大器两次开启的间隔内，字线处在关断状态，但是存储单元的泄漏电流可能会影响位线的摆幅。一般情况下，同一条位线上存“0”和存“1”的存储单元数目各占 50%，两条位线上的泄漏电流几乎一致，因此位线摆幅不会受到影响。现考虑一种极端情况：假定位线上的 M 个存储单元全部存“0”，这样位线 BL 上存在泄漏电流，而位线 BLB 上不存在泄漏电流，因此位线的摆幅会受到泄漏电流的影响，如图 3-19 所示。图 3-20 给出了泄漏电流最大的 PVT 条件下（0.5V，FF 工艺角，70℃/0.9V，FF 工艺角，70℃）存储单元的泄漏电流对位线电压影响的仿真结果（仿真只需要考虑单个存储单元，单个存储单元的位线负载电容为 C_{BL}/M，C_{BL} 代表整条位线的负载电容，M 为一根位线上存储单元的数目）。



（a）0.5V，FF 工艺角，70℃　　（b）0.9V，FF 工艺角，70℃

图 3-20 存储单元的泄漏电流对位线摆幅的影响

仿真结果表明：在 FF 工艺角，70℃条件下，0.5V 和 0.9V 下位线的电压变化仅有 0.08%和 0.011%，故存储单元的泄漏电流对灵敏放大器输入电压的影响可以忽略不计，具体原因如下：1.为了降低存储单元的静态功耗，存储单元选用高阈值的晶体管，泄漏电流相对较小；2.灵敏放大器两次启动的时间间隔相对较短，泄漏电流的影响较小。

3.3.3 串扰对灵敏放大器输入电压的影响

串扰[40-41]是指集成电路中两个节点之间的耦合干扰，由电路中的耦合电容引起，随着半导体制

造工艺的特征尺寸不断减小，MOS 器件的物理尺寸和互连线的间距不断减小，导致了串扰的影响不断增强。在本文的设计中，PMOS 切换开关的栅源/栅漏耦合电容会对位线的电压造成一定的影响，具体如图 3-21 所示，在极性调整的过程中，SWITCH 信号由低变高，耦合电容会使得位线 BL 的电压增加，SWITCHN 信号由高变低，耦合电容会使得位线 BL 的电压降低，位线电压的正向变化及负向变化相互抵消，故串扰的影响相互抵消，位线的电平不会发生改变，故串扰对灵敏放大器输入电压的影响可以忽略不计。

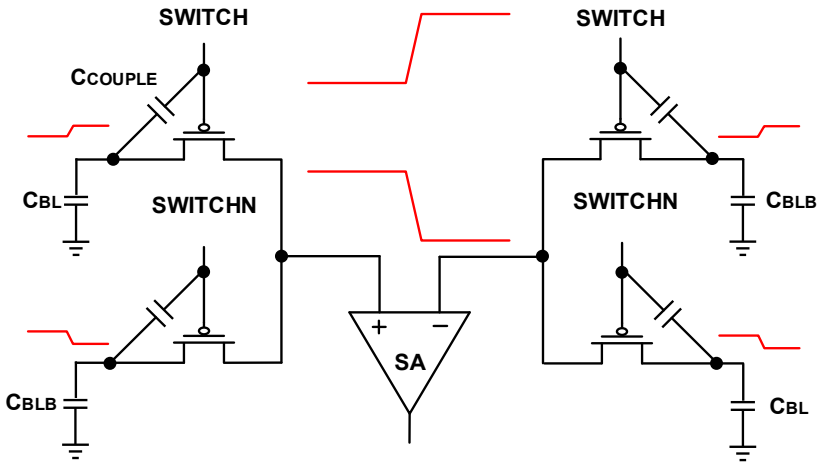


图 3-21 开关切换过程中的串扰现象

3.3.4 电荷共享对灵敏放大器输入电压的影响

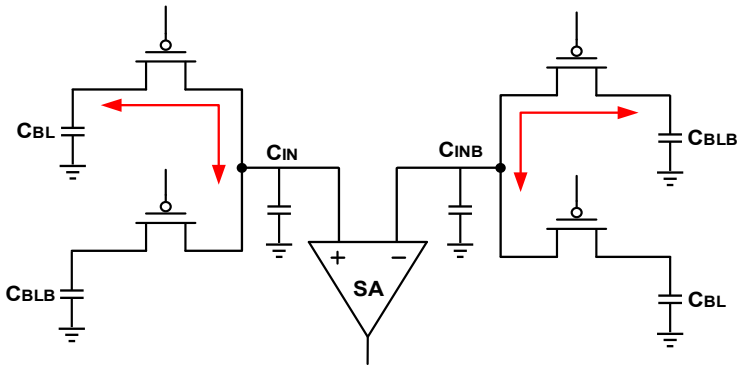


图 3-22 开关切换过程中的电荷共享

当位线摆幅 V_{SW}<VDD-|V_{THP}|时，PMOS 可以看成是理想开关，在开关切换的过程中，位线负载电容 C_{BL}/C_{BLB} 和 C_{IN}/C_{INB} 发生电荷的重新分配，如图 3-22 所示，电荷共享会使得灵敏放大器的输入电压受到影响，表格 3.5 给出了存储阵列深度分别为 128、256 和 512 时对应的仿真结果。

表 3.5 电荷共享对灵敏放大器输入电压摆幅的影响

	M=128	M=256	M=512
$\|V_{INPUT2}\|/\|V_{INPUT1}\|$	88%	92%	96%

仿真结果表明：电荷共享会对灵敏放大器的输入电压造成一定的影响，灵敏放大器第二次的输入电压的摆幅大约是第一次输入电压摆幅的 90%。现假定灵敏放大器输入电压的 V_{INPUT1}<-V_{OFFSET}<0，此时推测型读出的结果正确，极性调整之后，若不考虑电荷共享的影响，V_{INPUT2}>V_{OFFSET}，此时确认型读出的结果与推测型读出的结果相反，ERROR 信号保持在低电平；若考虑极性调整过程中电荷共

享造成的摆幅损失，$V_{INPUT2}<V_{OFFSET}$，此时确认型读出的结果与推测型读出的结果一致，ERROR 信号置高，如图 3-23 所示。因此电荷共享会增加检错电路的误判率，从而导致了存储阵列延时收益的降低。

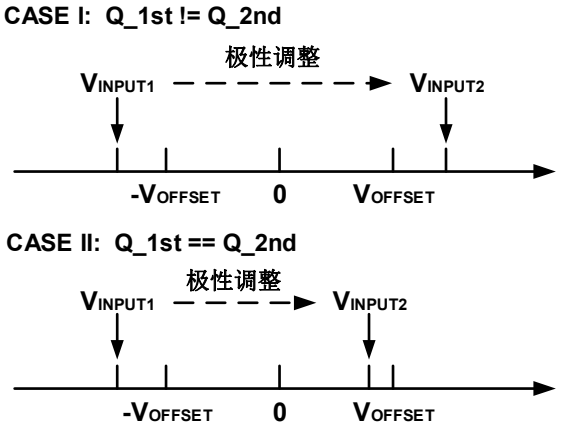


图 3-23 电荷共享对检测方案的影响

现对式（3.2）及式（3.3）做出如下的修正。C_{IN}/C_{INB} 的大小是由 PMOS 管的源漏扩散电容，灵敏放大器的栅极输入电容及互连线寄生电容决定的。假定此时存储单元存储数据“1”，字线关断时，相应的位线摆幅 $V_{SW}=V_{BL}-V_{BLB}<VDD-|V_{THP}|$，且 $C_{IN}=C_{INB}$，$C_{BL}=C_{BLB}$。根据电荷共享的原理可以做出如下推导：

$$\begin{aligned} V_{INPUT2} &= \frac{C_{BLB}V_{BLB}+C_{IN}V_{BL}}{C_{IN}+C_{BLB}} - \frac{C_{BL}V_{BL}+C_{INB}V_{BLB}}{C_{INB}+C_{BL}} \\ &= \frac{(C_{BL}-C_{IN})V_{BLB}-(C_{BL}-C_{IN})V_{BL}}{C_{IN}+C_{BL}} \\ &= -\frac{C_{BL}-C_{IN}}{C_{BL}+C_{IN}}V_{INPUT1} \end{aligned} \tag{3.6}$$

经过修正，灵敏放大器的输入电压和位线摆幅 V_{SW} 之间的关系如式（3.7）和式（3.8）：

1) 当$|V_{SW}|<VDD-|V_{THP}|$时：

$$\begin{cases} |V_{INPUT1}| = |V_{SW}| \\ V_{INPUT2} = -\dfrac{C_{BL}-C_{IN}}{C_{BL}+C_{IN}}V_{INPUT1} \end{cases} \tag{3.7}$$

2) 当$|V_{SW}|>VDD-|V_{THP}|$时：

$$\begin{cases} |V_{INPUT1}| = VDD-|V_{THP}| \\ V_{INPUT2} = -V_{INPUT1} \end{cases} \tag{3.8}$$

式（3.7）和式（3.8）将用于下一小节存储阵列延时的收益计算。

3.4 仿真结果

3.4.1 HSPICE-MATLAB 混合仿真方法

在计算本文时序推测方案的收益之前，首先做出如下的定义：

$T_{ARRAY-CONV}$ 代表传统模式下的存储阵列读出延时，如式（3.9），其中 $T_{WL-CONV}$ 是能够保障所有存储单元的位线摆幅超过灵敏放大器失调电压的字线使能时间，T_{SA} 代表灵敏放大器的读出延时。

$$T_{ARRAY-CONV} = T_{WL-CONV} + T_{SA} \tag{3.9}$$

T_{ARRAY} 代表了本文时序推测方案的存储阵列读出延时，对应加权平均延时，对应式（3.10）。时序推测方案的存储阵列读出延时取决于字线的使能时间。ERROR 信号置“0/1”的概率取决于字线使能时间，字线的使能时间越长，推测型读出的正确率就越高，ERROR 信号置“1”的概率也就越低。

$$T_{ARRAY} = T_{WL} \times P(ERROR=0) + (N-1) \times T_{WL} \times P(ERROR=1) + T_{SA} \tag{3.10}$$

其中 N 等于式（3.11）

$$N = \frac{T_{WL-CONV}}{T_{WL}} \tag{3.11}$$

在此定义存储阵列延时的收益如式（3.12）：

$$Gain = \frac{T_{ARRAY-CONV} - T_{ARRAY}}{T_{ARRAY-CONV}} \tag{3.12}$$

通过上文的分析，本文的时序推测方案的收益与字线的使能时间有关，时序推测方案的核心是降低局部工艺波动带来的设计裕度对 SRAM 存储阵列的性能的影响，时序推测方案的收益需要通过蒙特卡洛仿真才能得到，由于电路的规模相对较大，考虑局部工艺偏差的 HSPICE 蒙特卡洛仿真的时间成本太大，不利于得到方案的最佳字线使能时间及对应的收益，为了快速计算出最佳的字线使能时间，本文开发了一种 HSPICE-MATLAB 混合型仿真方法，该方法可以快速并准确地得到最佳的字线使能时间及对应的时序推测方案的收益，该 HSPICE-MATLAB 混合型仿真流程如图 3-24 所示。

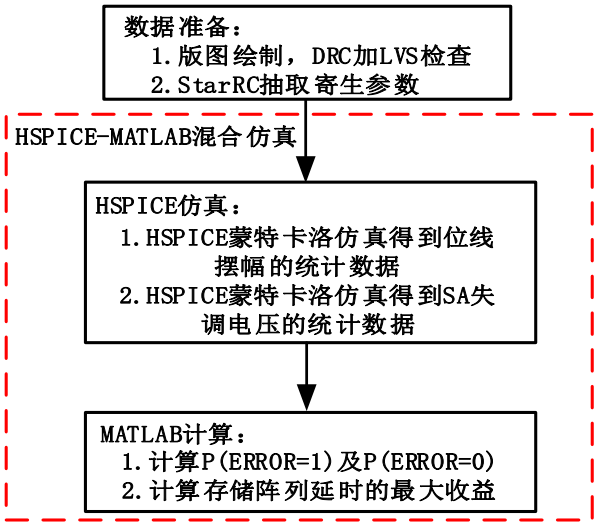


图 3-24 HSPICE-MATLAB 混合仿真流程

1) 数据准备阶段：

➢ 做出一列存储单元的版图，通过 Calibre 软件对其进行设计规则检查（Design Rule Check，DRC）和版图原理图一致性检查（Layout Versus Schematics，LVS）。其中，DRC 用于检查版图是否违反了工艺厂家提供的设计规则，LVS 用于检查从版图中提取出的电路图是否和电路原理图一致。

➢ DRC 和 LVS 通过后，采用 StarRC 软件对版图进行寄生参数的抽取，得到带寄生参数信息

的标准寄生（Standard Parasitic Format，SPF）文件。

2) HSPICE 仿真：

➢ 导入带寄生参数的 SPF 文件，通过蒙特卡洛仿真得到 $T_{WL\text{-}CONV}$，并将字线使能时间 $T_{WL\text{-}CONV}$ 分 k 个档位，单位的字线使能时间 $T=T_{WL\text{-}CONV}/k$。

➢ 通过蒙特卡洛仿真得到存储单元的位线摆幅 V_{SW} 并存储在二维矩阵中，矩阵的大小为 m×k，其中 m 代表蒙特卡洛仿真的次数，k 代表字线的档位数。

$$V_{SW}=\begin{bmatrix} VSW_{1,1} & VSW_{1,2} & .... & VSW_{1,k} \\ VSW_{2,1} & VSW_{2,2} & .... & VSW_{2,k} \\ .... & .... & .... & .... \\ VSW_{m,1} & VSW_{m,2} & .... & VSW_{m,k} \end{bmatrix} \tag{3.13}$$

➢ 通过蒙特卡洛仿真得到灵敏放大器的失调电压并存入矩阵中，矩阵的大小为 m×1，m 代表蒙特卡洛仿真的次数。

$$V_{OFFSET}=\begin{bmatrix} VOFFSET_1 \\ VOFFSET_2 \\ .... \\ VOFFSET_m \end{bmatrix} \tag{3.14}$$

3) MATLAB 仿真：

➢ 将 HSPICE 的仿真结果导入 MATLAB 中，根据式（3.7）和式（3.8）计算灵敏放大器的输入电压 V_{INPUT1} 和 V_{INPUT2}，并与灵敏放大器的失调电压 V_{OFFSET} 作对比，并计算不同字线档位下 ERROR 置“1”和 ERROR 置“0”的概率，MATLAB 计算过程的伪代码如下：

ERROR 信号置“0/1”概率计算的伪代码

Input：位线摆幅 V_{SW} 和灵敏放大器失调电压 V_{OFFSET} 的 HSPICE 仿真数据

Output：P(ERROR=1)和 P(ERROR=0)

```
for i=1:k     //k 代表字线的档位数
  begin
    for j=1:mc_num     //mc_num 代表蒙特卡洛仿真次数
      begin
        SW_index=random(mc_num);   SA_index=random(mc_num);
        V_SW=V_SW(SW_index,i);   //随机挑选一个存储单元
        V_OFFSET=V_OFFSET(SA_index);        //随机挑选一个灵敏放大器
        Calculate V_INPUT1 and V_INPUT2;   //根据公式计算 V_INPUT1 和 V_INPUT2
        if((V_INPUT1-V_OFFSET)×(V_INPUT2-V_OFFSET)>0) //推测型和确认型读出相同
              flag=flag+1;
      end
      P1(i)=(flag/mc_num)^N;     //计算 ERROR 信号置高的概率，N 为位宽
      P0(i)=1-P1(i);     //计算 ERROR 信号为低的概率
  end
```

➢ 根据式(3.10)、式(3.11)及式(3.12)计算不同字线档位下的存储阵列加权平均延时 T_{ARRAY}，并得到最佳的字线使能时间及对应的时序推测方案的最大收益。

为了验证本文提出的 HSPICE-MATLAB 混合型快速仿真方法的准确性，以不同字线使能时间下 ERROR 信号置“0”的概率作为参照，HSPICE 的仿真结果和 HSPICE-MATLAB 混合仿真结果的对比如图 3-25 所示，假定此时存储阵列的宽度 N=32，位线 BL/BLB 的负载为 50fF，仿真条件为 0.5V，TTG 工艺角，25℃。

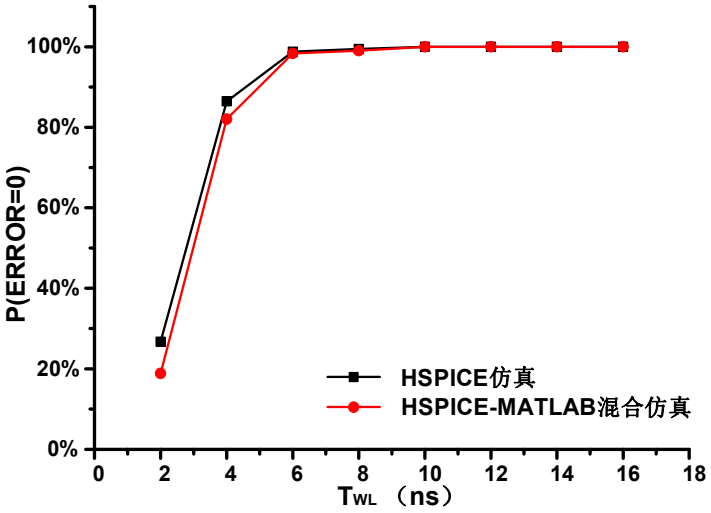


图 3-25 HSPICE-MATLAB 混合仿真与 HSPICE 仿真的对比

仿真结果表明：本文提出的 HSPICE-MATLAB 混合型快速仿真方法的仿真结果与 HSPICE 的仿真结果高度吻合，因此可以通过该仿真方法进行时序推测方案的收益分析。

3.4.2 仿真结果

假定此时存储阵列的深度 M=256，宽度 N=32。根据上文的仿真流程，可以得到不同电压下的存储阵列延时的最大收益及取得最大收益所对应的 ERROR 信号为零的概率(推测型读出正确的概率)，仿真结果如图 3-26 及 3-27 所示。

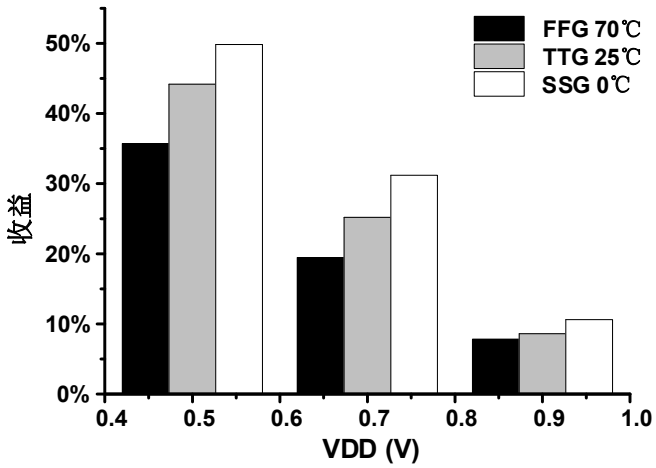


图 3-26 本文的时序推测方案在不同电压下的最大收益

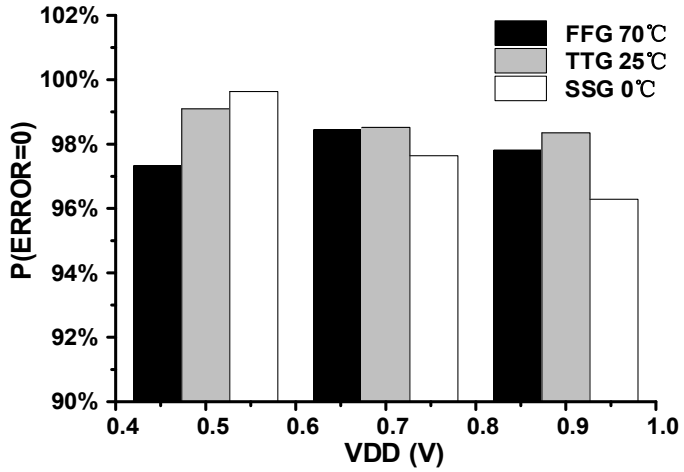


图 3-27 不同电压下推测型读出正确的概率

由于本文的时序推测方案存在误判，图 3-28 给出了不同电压下对应的误判率。

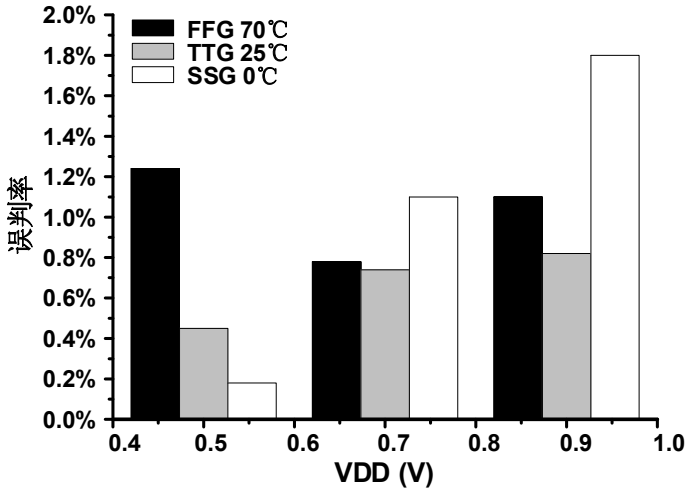


图 3-28 时序推测方案在不同电压下的误判率

由仿真结果可以得出如下结论：

1) 时序推测技术可以克服存储阵列中弱驱动的存储单元对存储阵列延时的影响，极大地提升存储阵列的性能。

2) 时序推测技术的收益随着电源电压的升高逐渐降低，这是因为局部工艺波动对存储单元的影响随着电压的升高逐渐降低。

3) 在不同的电压下，最坏情况下（SSG 工艺角，0℃）的收益最高，最好情况下（FFG 工艺角，70℃）的收益最低，典型情况下（TTG 工艺角，25℃）的收益处在中间。这是因为在 SSG 工艺角，0℃条件下，晶体管的阈值电压偏高，局部工艺波动对存储单元的影响较高；在 FFG 工艺角，70℃条件下，晶体管的阈值电压偏低，局部工艺波动对存储单元的影响较低；在 TTG 工艺角，25℃条件下，局部工艺波动对存储单元的影响处在中间。

4) 时序推测方案在取得最佳收益时，相应的 ERROR 信号置高的概率较低，即推测型读出的结果出错的概率较低，故纠错带来的代价可以忽略不计，故存储阵列的读出延时 T_{ARRAY} 近似等于推测型读出的延时。

5) 本文的检错方案虽然存在误判，但是误判率较低，存储阵列的收益几乎不受影响。

为了全面地评估本文的时序推测方案的收益，还需要考虑容量对收益的影响，即还需要考虑存储阵列的深度和宽度对收益的影响。假定存储阵列的宽度 N 为 32，令存储阵列的深度 M 分别为 128，256 和 512，可以得到 0.5V，0.7V 和 0.9V 条件下的存储阵列延时的最大收益如图 3-29 所示，仿真是在 TTG 工艺角，25℃下进行的。

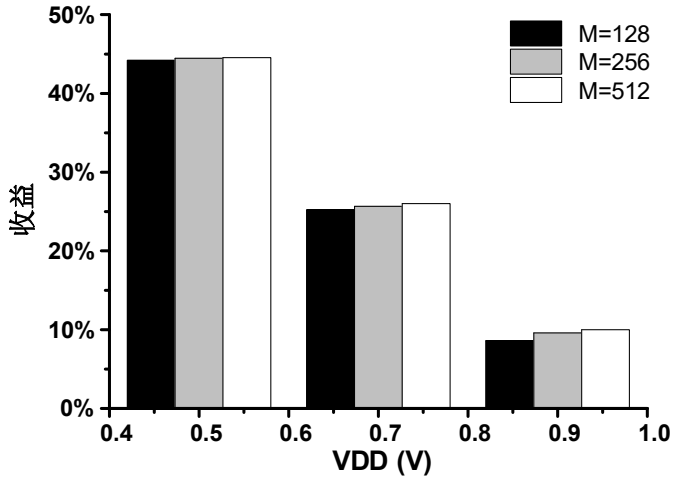


图 3-29 存储阵列的深度对收益的影响

假定存储阵列的深度 M 为 128，令存储阵列的宽度 N 分别为 32，64 和 128，可以得到 0.5V，0.7V 和 0.9V 条件下的存储阵列延时的最大收益如图 3-30 所示，仿真条件同上。

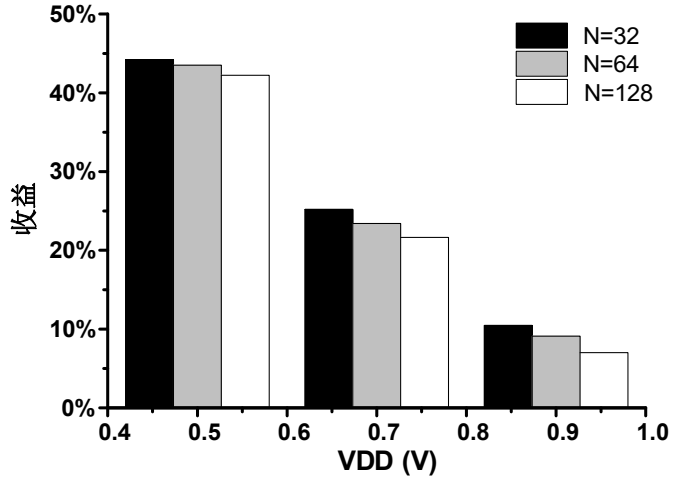


图 3-30 存储阵列宽度对收益的影响

仿真结果表明：SRAM 存储阵列的深度和宽度会对收益造成一定的影响。当存储阵列的深度变大时，位线负载电容变大，根据式（3.6），电荷共享对灵敏放大器的输入电压的摆幅影响随着位线负载电容的增加而降低，故存储阵列延时的最大收益随着存储阵列深度的增加而增加。当存储阵列的位宽增加时，一个字中包含弱驱动存储单元的概率就会增加，这增加了推测型读出出错的概率，故存储阵列延时的最大收益随着存储阵列宽度的增加而降低。总体而言，存储阵列的深度和宽度对存储阵列延时的最大收益的影响较小。

3.5 本章小结

时序推测技术能够在一定程度上降低存储阵列中弱驱动存储单元对存储阵列整体延时的影响，

实现存储阵列性能的提升。针对现有的时序推测技术在低电压条件下的缺点，本文提出了一种改进型的时序推测方案。3.1 节介绍了本文的时序推测方案的思路，并重点介绍了检错方案的设计，本文通过动态地调节灵敏放大器输入电压的极性实现在低电压条件下的快速检错。3.2 节具体地介绍了本文提出的时序推测方案的电路实现，即时序推测型存储阵列的设计与实现。3.3 节具体分析了电路中的一些非理想因素（泄漏电流/串扰/电荷共享）对检错方案的影响。3.4 节介绍了仿真流程及仿真结果，仿真结果表明：本文的时序推测方案在 0.5V 条件下能够降低大约 50%的存储阵列延时，在 0.9V 条件下能够降低大约 10%的存储阵列延时。

第四章 宽电压 SRAM 的设计

本章以时序推测型存储阵列为主体，基于 TSMC 28nm CMOS 工艺，完成了一款容量为 256×32 的宽电压 SRAM（工作电压范围为 0.5V 至 0.9V）的设计，并完成后仿真验证工作。4.1 节介绍了 SRAM 整体的电路结构和版图布局；4.2 节为后仿真结果；4.3 节为时序推测方案的对比分析；4.4 节为本章总结。

4.1 电路设计

4.1.1 电路结构

本文设计的容量为 256×32 的 SRAM 的结构如图 4-1 所示，整个 SRAM 的深度为 256，宽度为 32，即位线连接 256 个存储单元，字线连接 32 个存储单元，SRAM 的整体结构包括存储阵列（SRAM 的主体），输入电路模块（负责输入数据的锁存及位线的驱动），译码器（负责地址信号的译码），译码器驱动级（负责驱动字线并负责字线电压的调节），时序控制电路（负责提供 SRAM 读写控制信号），输出驱动电路（为输出提供驱动）。

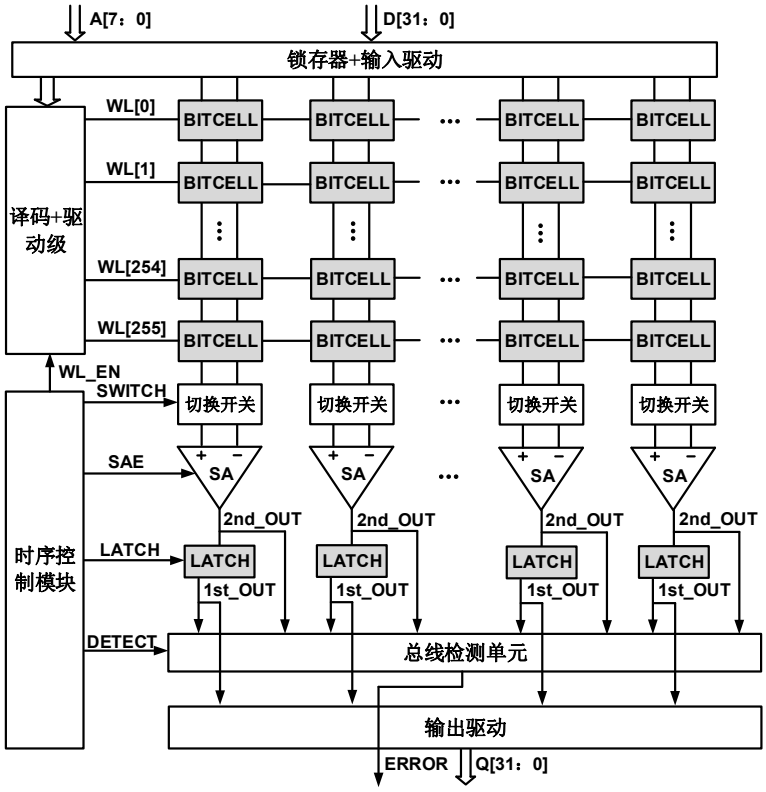


图 4-1 SRAM 的整体结构

4.1.1.1 时序推测型存储阵列

存储阵列是 SRAM 的设计主体，同样也是本文的重点研究对象。存储阵列包括了存储单元、切换开关、灵敏放大器、锁存器和总线检测单元，如图 4-2 所示。上述电路模块的设计在第三章已经做了详细的描述。

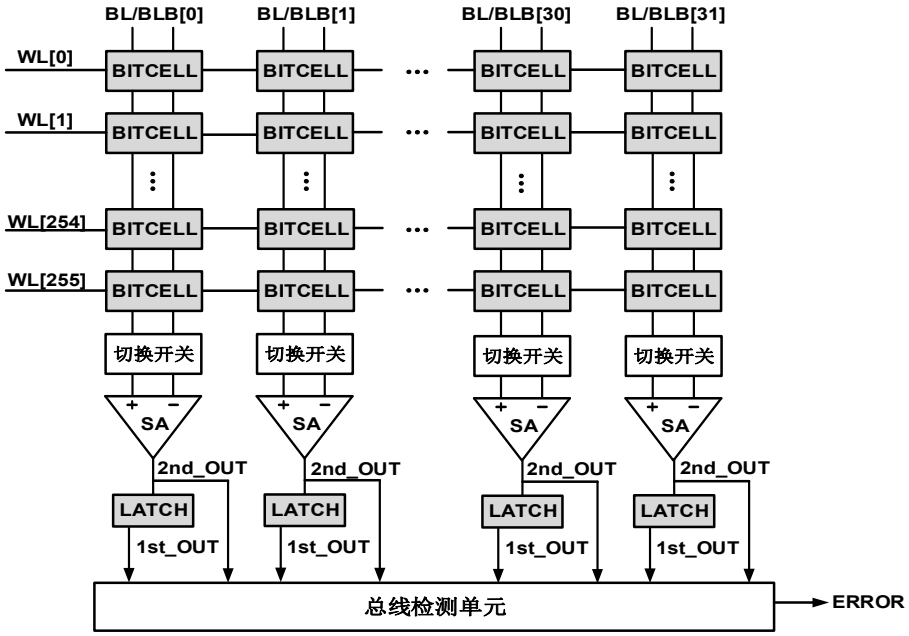


图 4-2 时序推测型存储阵列的设计

4.1.1.2 字线驱动级—字线电压调节模块的设计

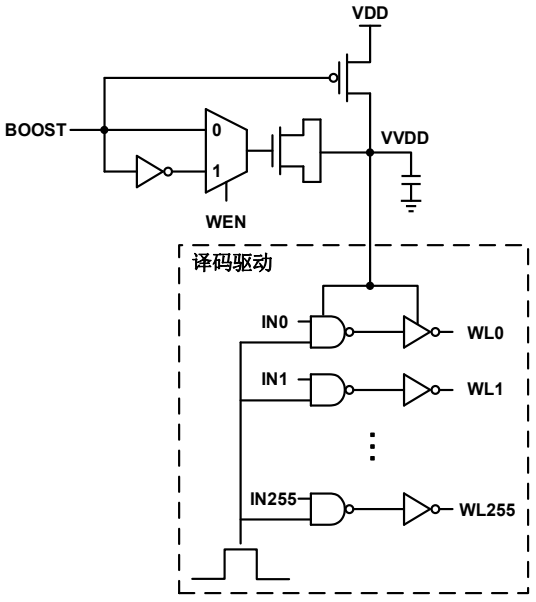


图 4-3 字线电压调节模块

为了保证 SRAM 存储单元在低电压下的读写稳定性，本文采用字线电压调节技术。字线电压调节技术通过电容耦合的原理动态地调节字线电压，本文采用 MOS 电容实现耦合作用，该方案可以产生高于电源电压的输出电压，也可以产生低于电源电压的输出电压。字线电压调节技术的电路结构如图 4-3 所示，WEN(读写使能信号)控制多路选择器的选通，写操作时（WEN=0），字线电压提升可以保证写操作的稳定性，读操作时（WEN=1），字线电压降低可以保证读操作的稳定性。

4.1.1.3 锁存器及输入驱动器

锁存器及输入驱动器的电路结构如图 4-4 所示。锁存器用于锁存地址和数据信号，锁存器采用第三章中的带泄漏电流补偿管的锁存器结构，输入驱动器用于驱动位线，写操作时，WEN 信号为低

电平，传输门保持导通，数据被送入位线；读操作时，WEN 信号为高电平，在位线预充电阶段，传输门导通，位线被预充电至 VDD，之后传输门截止，位线处在高阻态，读操作开始。

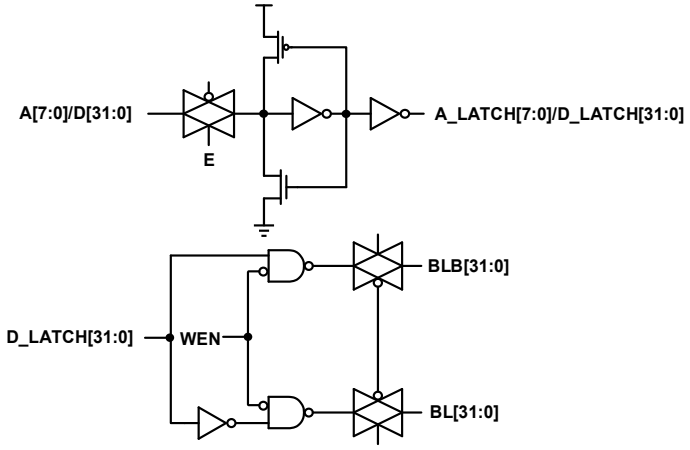


图 4-4 锁存器及输入驱动器

4.1.1.4 译码器

译码电路中的逻辑门可以采用 CMOS 静态门或动态门结构。相比于静态电路，动态电路的速度更快，动态电路的输出依靠内部寄生电容存储的电荷，电荷容易丢失，动态电路的输出易受到干扰，故动态电路的设计需要格外谨慎。本文的设计采用 CMOS 静态门结构。在 SRAM 中，译码器通过二进制地址信息选择相应的存储单元，若地址译码器采用单级译码方式，对于一个 N 位地址输入的译码器，一共需要 2^N 个 N 输入的与门，在本文的设计中地址输入为 8 位，因此译码器就需要 256 个 8 输入与门，这大大的增加了译码器的面积与延时，所以在 SRAM 中译码器通常采用多级译码[42]。本文译码器的设计采用二级译码，两个 4-16 译码器组成了第一级译码器，两个 4-16 译码器的输出组合形成第二级译码器，第二级译码器的输出连接译码驱动模块，如图 4-5 所示。

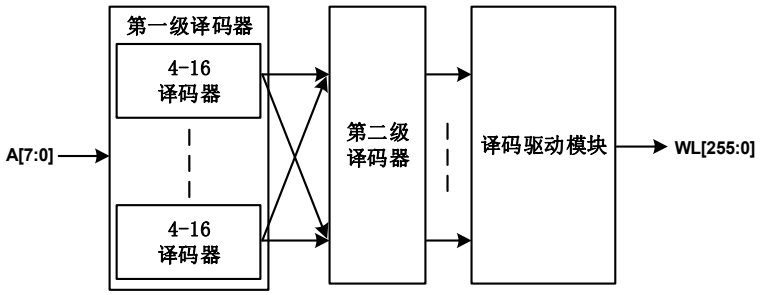


图 4-5 译码器的设计

4.1.1.5 输出驱动模块

输出驱动模块是由一排驱动单元构成，连接灵敏放大器的输出，用于驱动 SRAM 外部的负载，如图 4-6 所示。

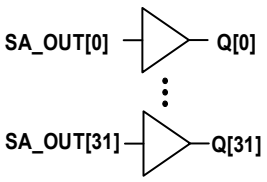


图 4-6 输出驱动模块

4.1.1.6 时序控制模块

时序控制模块的核心是复制位线技术，SRAM 采用复制位线技术跟踪读操作的关键路径，与反相器链的跟踪方式相比，复制位线技术的 PVT 跟踪能力更好。在近阈值区，局部工艺波动对复制位线的影响较大，会导致 SRAM 性能的退化，本文采用了数字化的复制位线技术，如图 4-7 所示。该技术能够很好的克服局部工艺波动的影响。数字化复制位线技术通过多个存储单元对位线 RBL 和 RBLB 交替放电的方式产生时钟脉冲，以该时钟脉冲作为基础，时序控制模块产生 SRAM 读写控制信号。

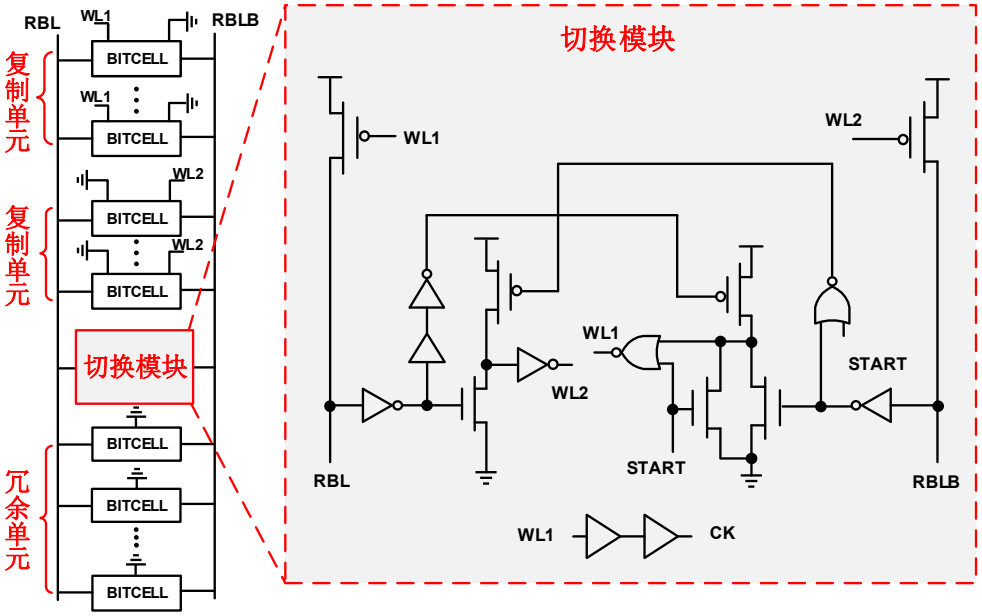


图 4-7 数字化复制位线技术

4.1.2 SRAM 版图的设计

版图设计作为从电路设计到生产制造的桥梁[43]，是集成电路技术中的一项重要环节，其不仅关系到集成电路的功能是否正确，而且还会对芯片生产制造的良率产生一定的影响。同时不同的版图实现形式会带来不同的寄生参数，这在一定程度上会影响集成电路的性能和功耗等指标。版图设计中最重要的环节就是版图的布局，合理的布局可以让后续的手动连线工作变得快捷，同时也可以减小信号线的长度，从而降低信号线的寄生负载，提升速度，并且降低了功耗。SRAM 版图的布局要充分考虑各个模块之间的连线问题。为了保证字线的连线长度最短，译码驱动级的输出连线与存储单元的字线要处在同一条垂直线上。同理，对于输入驱动电路和灵敏放大器，为了保证位线的连线长度最短，输入驱动电路及灵敏放大器要紧靠存储单元，且输入驱动电路及灵敏放大器的位线与存储单元的位线处在同一水平线上。在手动布线的过程中，要注意串扰的影响。为了使串扰最小，不能使导线之间的耦合电容过大，对于同一层的平行导线，增加导线的间距能够减少串扰，必要时可以在两条信号线之间增加电源屏蔽线。SRAM 宏单元的版图如图 4-8 所示，TSMC 28nm 工艺的设计规则要求存储单元的栅极垂直走向，存储单元的版图如图 4-9 所示，故 SRAM 宏单元版图中的字线是垂直走向，位线是水平走向。

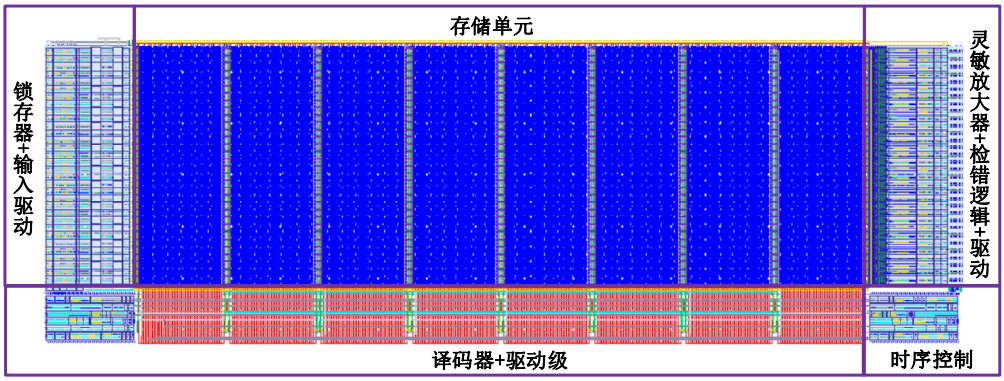


图 4-8 SRAM 的整体版图布局

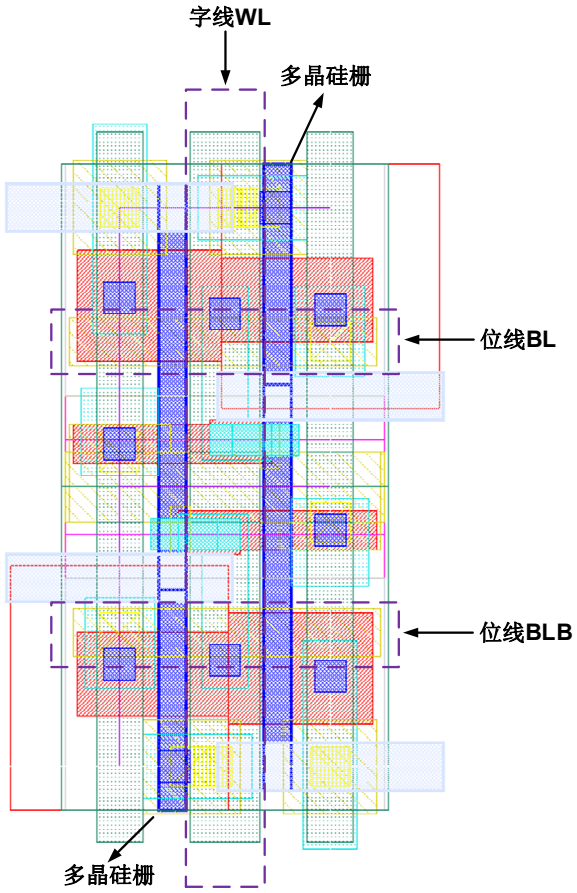


图 4-9 SRAM 六管单元版图

4.1.3 SRAM 测试模块的设计

为了验证时序推测型 SRAM 在系统中的应用，搭建如图 4-10 所示的测试系统，测试系统包括了 SRAM 宏单元，读写控制器及比较器。读写控制器用于控制片选信号、读写使能信号、地址信号及数据信号的产生，比较器接收来自 SRAM 宏单元的输出，并判断 SRAM 输出的结果是否正确，比较器内部包含计数模块，如果出现读错误，SRAM 的错误统计信号加一。若 SRAM 内部的检错电路判推测型读出出错，则 ERROR 信号置高，为了保证系统功能的正确性，通过时钟门控的方式对系统时钟做分频处理，保证比较器能够接收正确数据。

本文设计的测试方案的工作流程如下：

1) 顺序遍历地址，完成写 5 操作。

2) 顺序遍历地址，读取 SRAM 的数据，将数据送入比较器做比较，如果 SRAM 输出错误，比较器内部的统计模块统计错误数。

3) 顺序遍历地址，完成写 A 操作。

4) 顺序遍历地址，读取 SRAM 的数据，将数据送入比较器做比较，如果 SRAM 输出错误，比较器内部的统计模块统计错误数。

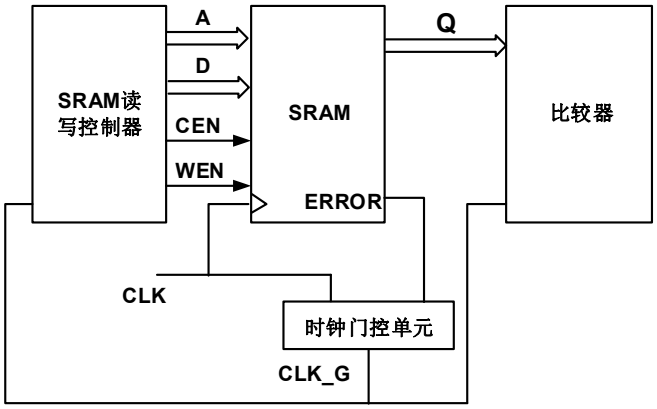


图 4-10 测试系统的设计

4.2 仿真结果

4.2.1 稳定性仿真

随着电源电压下降，局部工艺波动导致了 SRAM 六管存储单元的晶体管失配，晶体管失配导致了存储单元读写操作过程中的竞争现象，从而使 SRAM 存储单元的读写稳定性在近阈值区急剧降低。存储单元的读写稳定性指标决定了 SRAM 的最低工作电压，本文通过字线电压调节技术提升存储单元的读写稳定性（提升字线电压可以保证写操作的稳定性，降低字线电压可以保证读操作的稳定性）。由于存储单元的失效属于小概率事件，采用 HSPICE 的蒙特卡洛仿真时间成本大，为了验证设计的可靠性，本文采用 ProPlus 公司的专业良率分析工具 NanoYield 对存储单元的读写稳定性良率进行仿真分析。图 4-11 给出了低电压区间（0.5V~0.6V）的良率仿真结果（包括了一次读操作和一次写操作）。

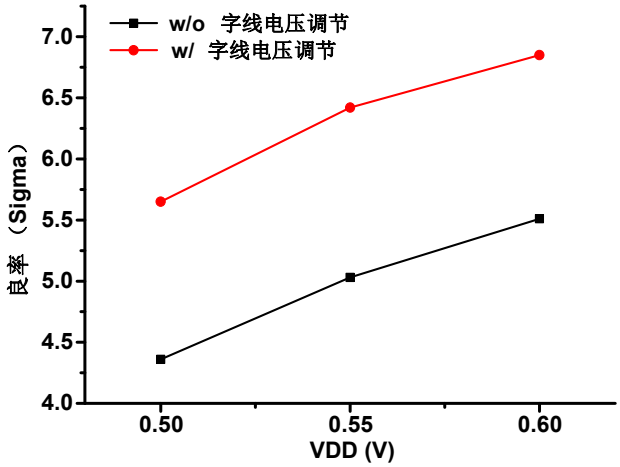


图 4-11 低电压区间的稳定性良率仿真结果

为了保证 SRAM 存储单元在近阈值区的良率，本文采用了字线电压调节技术。仿真结果证明字

线电压调节技术能够提升存储单元的读写稳定性，通过字线电压调节技术，存储单元的稳定性良率能够达到 5 至 6 个 Sigma 的设计要求。当电源电压为 0.5V 时，存储单元的稳定性良率大约为 5.5 个 Sigma，对应的概率为 99.9999962%，当 SRAM 存储阵列的容量达到 Mb 级别时，存储阵列对应的良率为 0.999999962^(1024×1024)=96%，能够满足设计的需求。

4.2.2 功能仿真

为了验证时序推测型 SRAM 在系统中的应用，采用 VCS-HSIM 混合仿真方法对测试系统做功能性的仿真。VCS-HSIM 混合仿真采用 VCS 数字仿真器和 HSIM 模拟仿真器实现数模混合信号仿真，使用直接内核接口（Direct Kernel Interface，DKI）在 VCS 数字仿真器和 HSIM 模拟仿真器之间交换信息。VCS-HSIM 联合仿真支持两种电路网表：一种是 Verilog 网表，另一种是 SPICE 网表。本文中的 SRAM 读写控制器和比较器实例由 Verilog 描述，SRAM 宏单元由 SPICE 描述。仿真分为两种情况，仿真波形如图 4-12 和 4-13 所示。

1) 系统未访问弱驱动存储单元（推测型读出正确，ERROR 信号一直保持在低电平）

读写控制器对 SRAM 的前 4 位地址写入数据 32’hAAAAAAAA，紧接着读取前 4 位地址的数据，由于系统未访问弱驱动存储单元，故推测型读出为 32’hAAAAAAAA，在读操作过程中，ERROR 信号一直保持低电平，错误统计信号 COUNT 也一直保持为零。

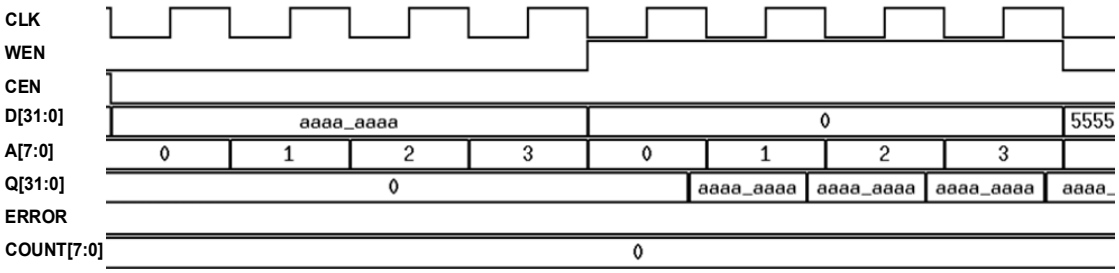


图 4-12 功能仿真波形（系统未访问弱驱动存储单元）

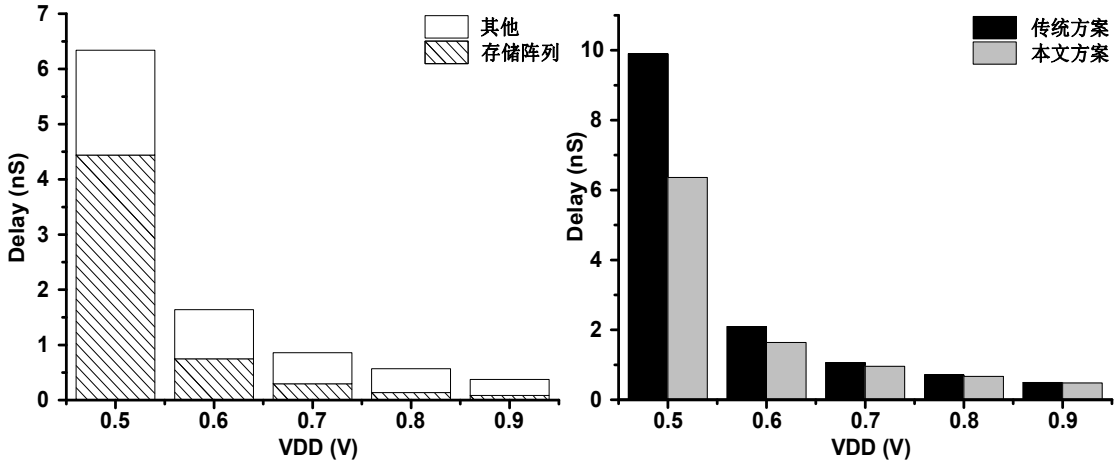
2) 系统访问弱驱动存储单元（推测型读出出错，ERROR 信号置高）：

和第一种情况一致，读写控制器对 SRAM 的前 4 位地址写入数据 32’hAAAAAAAA，紧接着读取前 4 位地址的数据。在读操作过程中，当地址信号 A=8’b00000010 时，由于系统访问弱驱动存储单元，数据输出 32’hAAAAAAAB，即推测型读出出错，检错电路触发 ERROR 信号置高，时钟门控单元将读写控制器的时钟下降沿关断（读写控制器时钟下降沿发送数据），从而保证了两个时钟周期的时间读取当前地址数据，由于比较器接收正确数据，在读操作过程中错误统计信号 COUNT 一直保持为 0。

图 4-13 功能仿真波形（系统访问弱驱动存储单元）

4.2.3 性能仿真

图 4-14 给出了不同电压下 SRAM 的延时仿真结果。图中阴影部分代表了存储阵列的读出延时（存储阵列的读出延时由字线的使能时间决定，字线的最佳使能时间按照第三章介绍的方法得到）。在低电压下（0.5V）和正常电压（0.9V）下，存储阵列的读出延时占据 SRAM 整体读出延时的比例分别为 70%和 17.6%，所以存储阵列的读出延时对 SRAM 的性能有着重要的影响。和传统模式相比，本文的 SRAM 采用了时序推测方案，时序推测方案可以在一定程度上降低设计裕度对存储阵列读出延时的影响，故 SRAM 的整体延时在低电压条件下（0.5V）和正常电压条件下（0.9V）分别降低了约 36%和 2%。



（a）存储阵列读出延时与 SRAM 读出延时　　（b）本文的方案与传统方案的对比

图 4-14 SRAM 在不同电压下的延时

图 4-15 给出不同电压下传统 SRAM、时序推测型 SRAM 和反相器链的归一化延时。反相器链代表了数字逻辑电路。随着电源电压的降低，SRAM 和逻辑电路的性能不断降低，相比于逻辑电路，SRAM 的性能降低程度更加严重，采用时序推测方案可以在一定程度上减缓电压的降低对 SRAM 性能的影响。

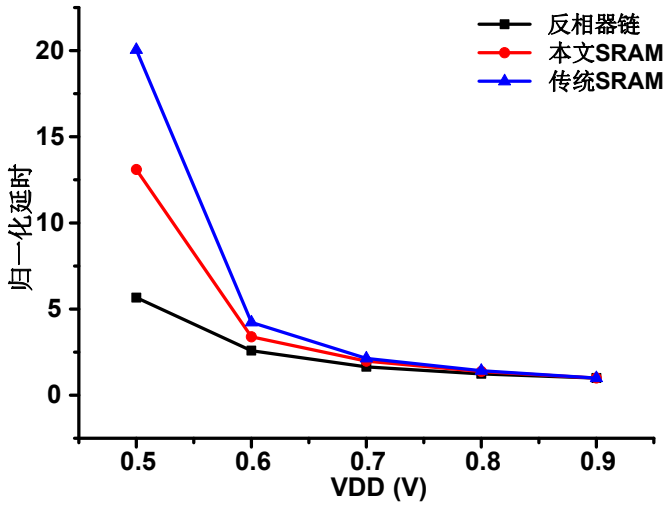


图 4-15 不同电压下反相器链、传统 SRAM 和时序推测型 SRAM 的归一化延时

4.3 时序推测方案的对比分析

4.3.1 工作模式对比

图 4-16 总结了不同的时序推测方案的读操作波形。在传统的 SRAM 存储阵列中，当所有存储单元的位线摆幅超过灵敏放大器的失调电压时，灵敏放大器开启，对应的读出延时为 $T_{ARRAY-CONV}$，时序推测技术在较早的时刻启动灵敏放大器，检错和纠错保证了系统功能的正确性。时序推测型 SRAM 存储阵列包含两个性能参数，一个是存储阵列的读出延时 T_{ARRAY}，由于推测型读出错误的概率较低，因此可以认为存储阵列的读出延时 T_{ARRAY} 近似等于推测型读出的延时，另一个是检错延时 T_{ERROR}，如图 4-16 所示。

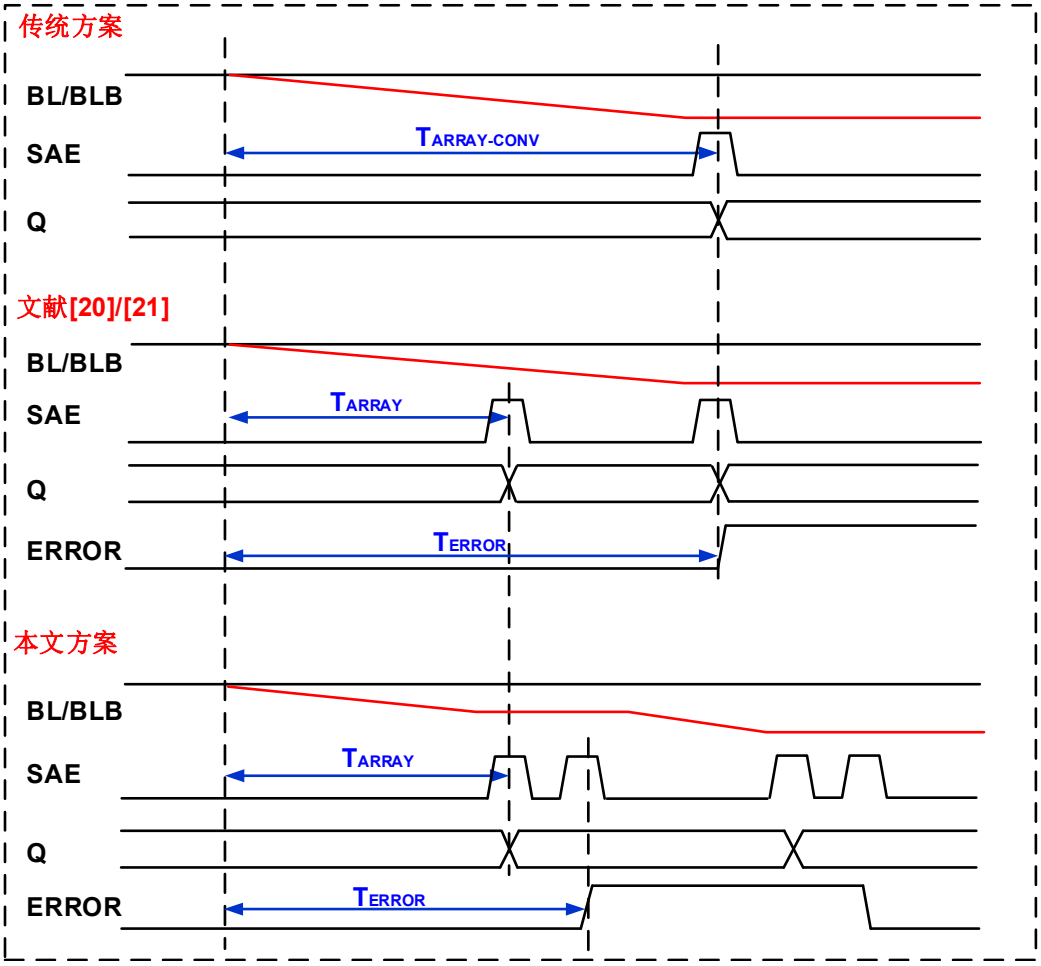


图 4-16 不同时序推测方案的读出波形

时序推测方案包括了检错和纠错的过程，首先对比三种时序推测方案的检错方式。文献[20]的方案在较早的时刻使能灵敏放大器（推测型读出），数据快速输出，灵敏放大器第二次启动（确认型读出）时要保证所有存储单元的位线电压摆幅超过灵敏放大器的失调电压，在近阈值区，存储单元放电延时的统计分布存在拖尾现象，因此检错延时较高，考虑在 SoC 中的应用，该方案仅适用于组合逻辑延时占主导地位的关键路径，而不适用于 SRAM 延时占主导地位的关键路径。文献[21]的方案采用了和文献[20]相类似的检错方式，过高的检错延时使该方案不适用于 SRAM 延时占主导地位的关键路径。在本文的方案中，在灵敏放大器第一次启动后（推测型读出），检错电路迅速地调整灵敏

放大器的输入电压极性，之后灵敏放大器二次启动（确认型读出），检错电路通过对比推测型输出和确认型输出的结果实现检错，灵敏放大器输入电压的极性调整实际上是一个电荷共享的过程，速度相对较快，故本文方案的检错延时相对较低。

下面讨论三种时序推测方案的纠错方式。在文献[20]和文献[21]中，当检错电路判断推测型读出出错时，由于确认型读出的结果一定正确，多路选择器直接输出确认型读出的结果。在本文的方案中，当检错电路判断出推测型读出出错时，由于本文的检错方案存在误判，即 ERROR 信号置高时，检错电路无法判断出推测型读出的结果是否正确，为了得到正确的输出，字线再次开启，位线继续放电，当位线的摆幅足够大时，灵敏放大器再次启动并输出正确的结果。在上述三种方案中，当 ERROR 信号置高时，SRAM 的读出延时更长，系统的关键路径会出现建立时间违规，系统采用时钟门控的方式对时钟做二分频以保证关键路径末端的寄存器采集到正确的数据。

4.3.2 性能能耗面积对比

本节将从性能、能耗及面积等三个角度对时序推测方案做全方位的对比，由于存储阵列是 SRAM 的主体，对 SRAM 的性能、能耗以及面积有着重要的影响，因此本节的对比均是从存储阵列的角度出发。图 4-17 展示了三种不同类型的存储阵列，其中传统的存储阵列作为对比基准（存储单元加灵敏放大器），本小节将从性能、能耗及面积等三个角度全方面对比不同类型的存储阵列。不同的工艺和存储阵列的容量均会对对比产生影响，为了公平的对比，基于 TSMC 28nm 工艺，本文对文献[20]和文献[21]的时序推测方案做了复现。存储阵列的深度和宽度会对收益造成一定的影响，仿真结果表明这种影响可以忽略不计，但是存储阵列的深度会对面积开销和能耗开销的对比造成一定影响，为了全方位的对比，本文对比的存储阵列容量大小分为三种：128×32、256×32 及 512×32。由于本文的时序推测方案是针对现有的时序推测方案在近阈值区的缺点提出的，因此本节的对比均是在低电压区间进行，对比的 PVT 条件为 0.5V，TTG 工艺角，25℃。

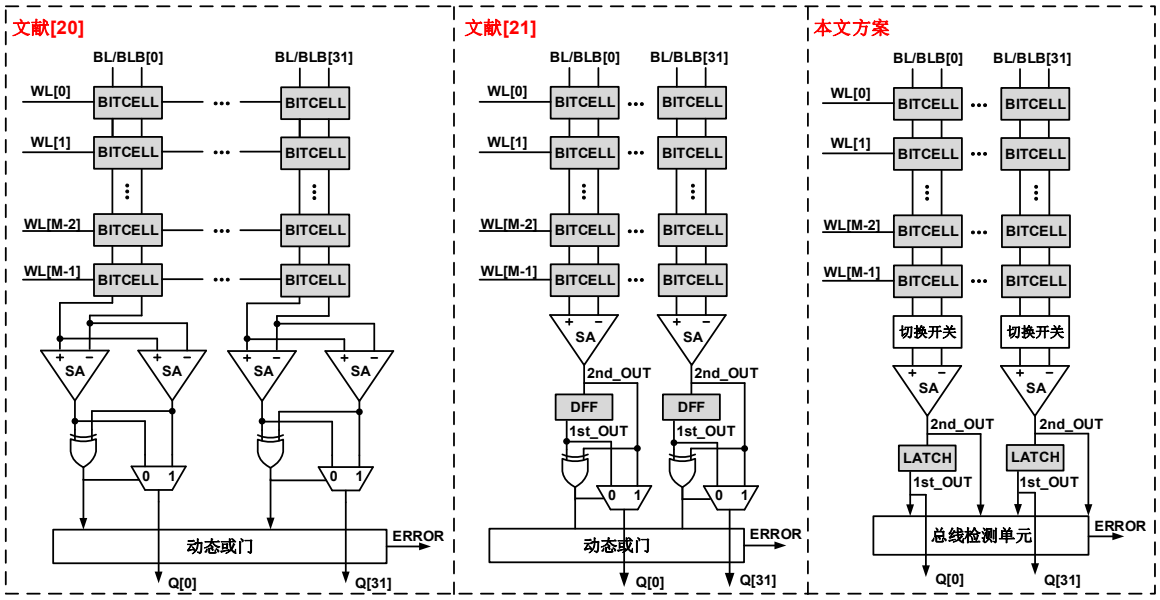


图 4-17 不同时序推测方案的电路实现

4.3.2.1 性能对比

时序推测型的 SRAM 存储阵列包含了两个延时指标 T_{ARRAY} 和 T_{ERROR}，T_{ARRAY} 代表存储阵列的读出延时，T_{ERROR} 代表了存储阵列的检错延时。通过 HSPICE-MATLAB 混合仿真方法可以得到另外两种方案的存储阵列延时的最大收益，仿真结果如图 4-18 和 4-19 所示，M 代表存储阵列的深度。在文献[20]和文献[21]中，检错原理是类似的，这使得两个方案具有几乎一致的 T_{ARRAY} 和 T_{ERROR}。与传统的读出方案相比，T_{ARRAY} 几乎减少了 44%，而 T_{ERROR} 就等于传统方案存储阵列的读出延时。

与传统的读出方案相比，在三种不同容量的存储阵列中，本文方案的存储阵列读出延时 T_{ARRAY} 分别减少了 42%、43.3%和 43.9%，与文献[20]和文献[21]的方案相比，本文提出的方案在 T_{ARRAY} 上的收益略低，这是因为本文的检错方案会存在一定的误判，即检错电路将正确的推测型读出结果判错，由于误判率较低，故可以近似认为本文方案的收益和上述两种方案一致。本文的时序推测方案通过灵敏放大器输入电压极性的调整实现检错，此方案大大地降低了检错延时，相比于前两种方案的检错延时，在三种不同容量的存储阵列中，本文的检错延时 T_{ERROR} 分别减少了 37.32%、40.9%和 42.76%。减小检错延时对提升 SoC 系统的吞吐率有着重要的作用，这将在下一小节做具体分析。

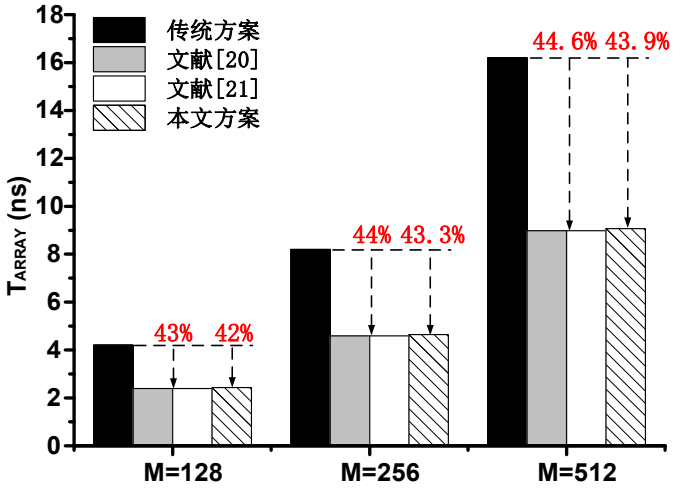


图 4-18 T_{ARRAY} 对比

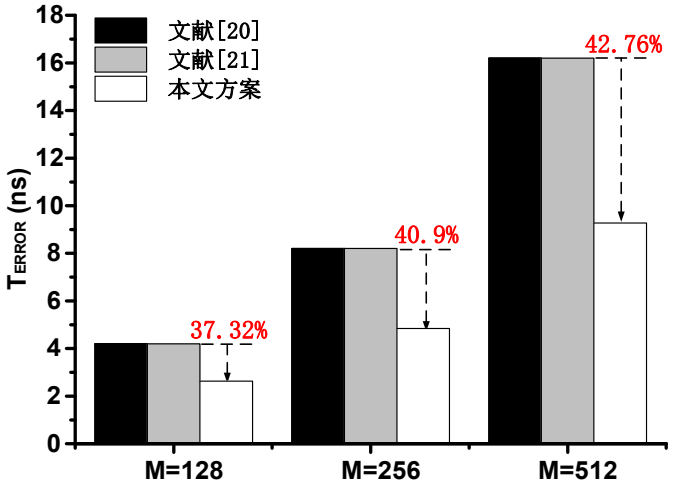


图 4-19 T_{ERROR} 对比

4.3.2.2 能耗对比

能耗包括动态能耗及静态能耗，时序推测方案的检错逻辑带来的额外静态能耗可以通过电源门控的方式加以抑制，故本小节不会将静态能耗作为对比指标。动态能耗包括写动态能耗及读动态功能耗，时序推测方案仅仅是针对读操作，没有对写操作做任何优化，故本节将以读能耗作为对比指标。传统的 SRAM 存储阵列的读能耗包含了位线预充电阶段所消耗的能量及灵敏放大器消耗的能量，时序推测型 SRAM 存储阵列的能耗不仅包括位线预充电所消耗的能量和灵敏放大器消耗的能量，还包括了检错电路消耗的能量。位线的摆幅会显著地影响存储阵列的预充电能耗，由于局部工艺波动的原因，即便是相同的字线使能时间，位线摆幅会出现一定的差异，为了准确地反映出存储阵列的能耗水平，本节采用蒙特卡洛仿真得到存储阵列读操作的平均能耗，仿真的方法是对一列存储单元进行蒙特卡洛仿真得到列的平均能耗，列平均能耗乘 32 得到整个存储阵列的能耗。在三种时序推测方案中，由于推测型读出正确的概率比较高，因此纠错带来的能耗代价可以忽略不计，故纠错能耗不在本节的对比范围之内，对比结果如图 4-20 所示，M 代表存储阵列的深度。

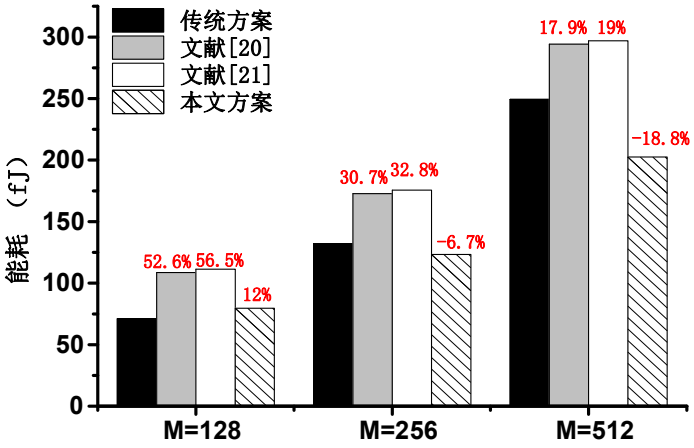


图 4-20 能耗对比

在传统方案中，在 128×32、256×32 和 512×32 三种阵列结构中，读操作消耗的能量分别为 71fJ，132fJ 和 246fJ，其中位线预充电消耗的能量在总体能耗中的所占比例分别为 82.47%、91%和 95.2%，其余部分为灵敏放大器所消耗的能量。由于在读操作的过程中，若干根位线同时使能，故存储阵列的能量消耗大部分集中在位线的预充电阶段。

在文献[20]中，在三种不同容量的存储阵列中，读操作消耗的能量分别为 108fJ、172fJ 和 294fJ。影子灵敏放大器、异或门和多路选择器等检错逻辑消耗了额外的能量。与传统的读出方案相比，在三种不同容量的存储阵列中，能耗分别增加了 52.6%，30.7%和 17.9%。文献[21]的推测型读出原理与文献[20]相似，所以其对应的能耗开销几乎相同。

在本文的方案中，由于字线使能时间较短，因此位线平均摆幅相对较低，相比于传统的方案，位线预充电阶段消耗的能量会有一定程度的降低，这能够在一定程度上弥补检错逻辑带来的额外能量消耗。与传统的读出方案相比，在三种不同容量的存储阵列中，能耗分别增加了 12%，-6.7%和-18.8%，当 SRAM 存储阵列的深度达到 256 和 512 时，对应的位线预充电能耗在总体能耗中的比例相对较高，节省的位线预充电能耗超过了检错逻辑带来的额外能量消耗，因此总体能耗有所降低。

4.3.2.3 面积对比

图 4-21 为不同的时序推测方案的版图对比，通过版图可以看出影子灵敏放大器和寄存器的面积开销较大，本文检错电路的结构经过精心设计，面积开销最小，面积对比的结果如图 4-22 所示。

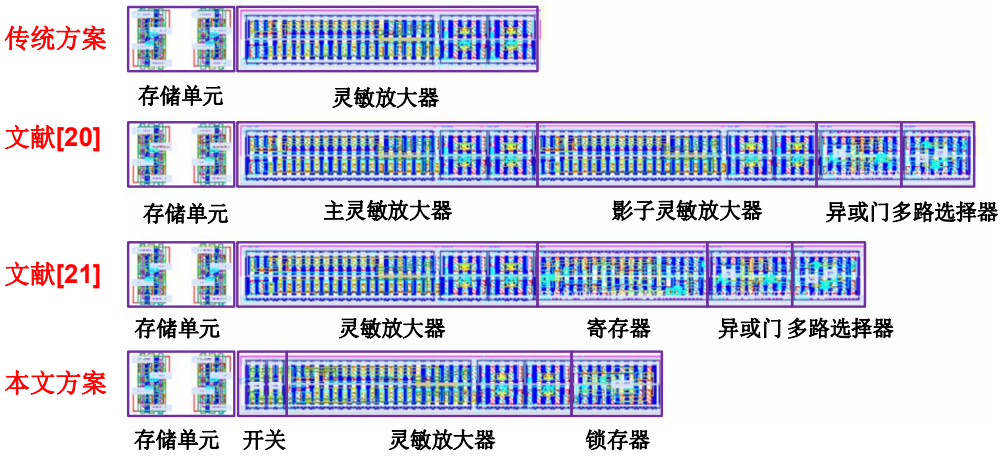


图 4-21 不同时序推测技术的版图对比

在文献[20]中，影子灵敏放大器、异或门、多路选择器及动态或门构成了检错逻辑。与传统方案相比较，在 128×32、256×32 和 512×32 阵列中，面积分别增加了 17.6%、9.4%和 4.8%。大部分的面积开销是由影子灵敏放大器引起的。

在文献[21]中，寄存器、异或门、多路选择器及动态或门构成了检错逻辑。在 128×32、256×32 及 512×32 的阵列结构中，面积分别增加了 13.2%、7%和 3.6%。大部分的面积开销是寄存器造成的。

本文提出的方案中，MOS 开关、锁存器及总线检测单元构成本文的检错逻辑，与传统方案相比较，在不同容量的存储阵列中，面积分别增加了为 6.4%、3.4%和 1.8%。

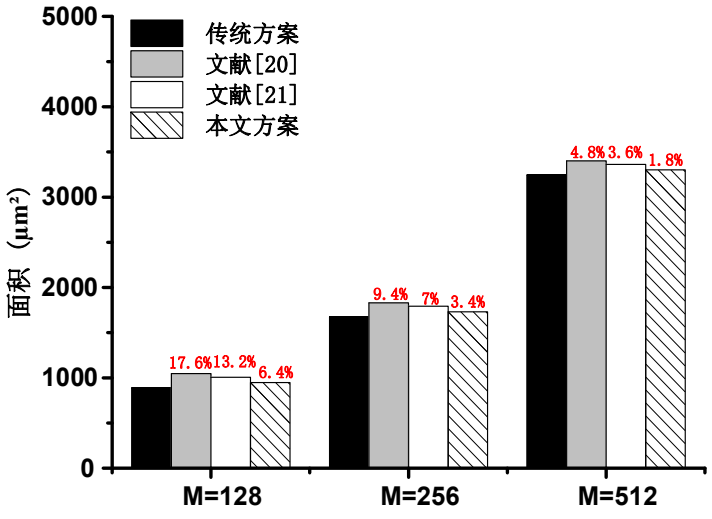


图 4-22 不同时序推测方案的面积对比

通过上述对比，可以得出以下的结论：

1) 不同方案的面积开销会随着存储阵列深度的增加而降低。

2) 本文提出的方案面积开销大约是其他方案的一半，当存储阵列的深度较小时，本文的时序推测方案在面积上的优势会更加明显。

4.3.3 吞吐率对比

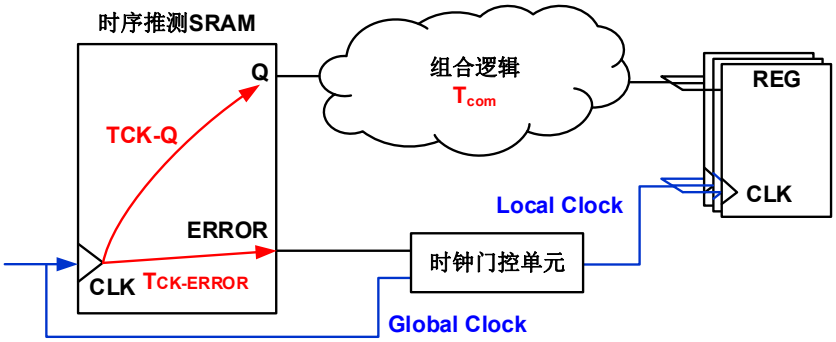


图 4-23 时序推测系统的关键路径

时序推测系统中的关键路径图 4-23 所示，与传统的关键路径有所不同，SRAM 的 ERROR 信号控制着时钟门控单元，当检错电路判断推测型读出的结果是错误的时候，ERROR 信号置高，下级寄存器的时钟信号被时钟门控单元控制，以等待 SRAM 输出正确的数据。时序推测型 SRAM 有两个参数一个是 $T_{CK\text{-}Q}$，用于衡量 SRAM 的整体延时，另一个是 $T_{CK\text{-}ERROR}$，用于衡量 SRAM 宏单元的检错延时，分别等于式（4.1）和式（4.2）：

$$T_{CK-Q} = T_{DEC} + T_{ARRAY} + T_{OUT} \tag{4.1}$$

$$T_{CK-ERROR} = T_{DEC} + T_{ERROR} \tag{4.2}$$

在式（4.1）和式（4.2）中，T_{DEC} 代表时钟上升沿至字线上升沿的延时，T_{ARRAY} 代表存储阵列的读出延时，T_{ERROR} 代表存储阵列的检错延时，T_{OUT} 代表输出驱动模块的延时。

为了保证时序推测系统的正确功能，时序推测系统必须满足如下的两项约束条件：

1) 约束项 1 如式（4.3），其中 $T_{CK\text{-}Q}$ 代表 SRAM 的数据读出延时，T_{COM} 代表组合逻辑的延时，$T_{SETUP\text{-}REG}$ 代表关键路径末端的寄存器的建立时间，T_{CYCLE} 代表时钟周期。

$$T_{CK-Q} + T_{COM} + T_{SETUP-REG} < T_{CYCLE} \tag{4.3}$$

2) 约束项 2 如式（4.4），其中 $T_{CK\text{-}ERROR}$ 代表 SRAM 的检错延时，$T_{SETUP\text{-}ICG}$ 代表时钟门控单元的建立时间，T_{CYCLE} 代表时钟周期。

$$T_{CK-ERROR} + T_{SETUP-ICG} < T_{CYCLE} \tag{4.4}$$

时序推测系统的吞吐率收益如式（4.5）所示（为了便于分析，忽略关键路径末端触发器和时钟门控单元的建立时间的大小）：

$$Gain = \frac{T_{CK\text{-}Q\text{-}CONV} + T_{COM}}{max\left\{T_{CK\text{-}Q} + T_{COM}, T_{CK\text{-}ERROR}\right\}} \tag{4.5}$$

系统的最高吞吐率不仅仅与 SRAM 的读出延时 $T_{CK\text{-}Q}$ 有关，还与检错延时 $T_{CK\text{-}ERROR}$ 有关，现分两种情况进行讨论：

1) 第一种情况：$T_{CK\text{-}ERROR} > T_{CK\text{-}Q} + T_{COM}$

$$Gain=\frac{T_{CK-Q-CONV}+T_{COM}}{T_{CK-ERROR}} < \frac{T_{CK-Q-CONV}+T_{CK-ERROR}-T_{CK-Q}}{T_{CK-ERROR}}=1+\frac{T_{CK-Q-CONV}-T_{CK-Q}}{T_{CK-ERROR}} \quad (4.6)$$

2) 第二种情况：$T_{CK-ERROR}<T_{CK-Q}+T_{COM}$

$$\begin{aligned} Gain&=\frac{T_{CK-Q-CONV}+T_{COM}}{T_{CK-Q}+T_{COM}} \\ &=\frac{T_{CK-Q-CONV}-T_{CK-Q}+T_{CK-Q}+T_{COM}}{T_{CK-Q}+T_{COM}} \\ &=1+\frac{T_{CK-Q-CONV}-T_{CK-Q}}{T_{CK-Q}+T_{COM}} \\ &<1+\frac{T_{CK-Q-CONV}-T_{CK-Q}}{T_{CK-ERROR}} \end{aligned} \quad (4.7)$$

经过上述的分析，可以得到时序推测系统的最高吞吐率收益如式（4.8）所示，时序推测系统的最高吞吐率收益不仅与时序推测型 SRAM 的读出延时有关，还与系统的检错延时有关，检错延时越低系统的吞吐率就越高。

$$Gain_{MAX}=1+\frac{T_{CK-Q-CONV}-T_{CK-Q}}{T_{CK-ERROR}} \quad (4.8)$$

根据式（4.8）计算出三种不同的时序推测方案的系统最高吞吐率收益如图 4-24 所示。

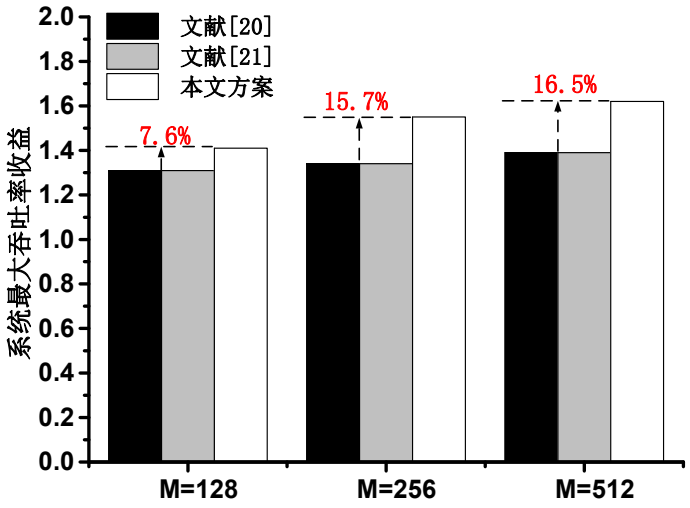


图 4-24 时序推测系统的最大吞吐率收益对比

三种时序推测方案有近似一致的读出延时，但是与其他两种的时序推测方案相比，本文的时序推测方案通过灵敏放大器输入电压极性的调整实现快速检错，此方案可以大幅度地提升时序推测系统的最大吞吐率收益。文献[20]和文献[21]采用了相似的检错方式，故其最大的系统吞吐率收益几乎一致，本文的方案在最大吞吐率收益方面分别提升了 7.6%，15.7%和 16.5%。

4.3.4 时序推测方案的对比总结

根据上文的分析与讨论，表 4.1 对三种时序推测方案的对比进行了总结，为了公平对比，三种时

序推测方案均采用 TSMC 28nm 工艺，SRAM 的存储阵列的容量大小分为以下三种：128×32、256×32 和 512×32。文献[20]的时序推测方案，文献[21]的时序推测方案及本文的时序推测方案均能够在一定程度上降低弱驱动存储单元对存储阵列性能的影响，取得了近似一致的存储阵列延时收益，相比于其他两种方案，本文的时序推测方案的面积开销及功耗开销最小。由于本文的时序推测方案采用了灵敏放大器输入电压极性调整的方式实现检错，本文的时序推测方案的检错延时最小，故本文的方案在吞吐率方面更具优势。

综合考虑上述三种方案的性能、能耗及面积，在此定义时序推测型存储阵列收益的综合指标（Figure of Merit，FoM）如式（4.9）所示。

$$FoM_{GAIN} = \frac{Gain_{throughput}}{Area_{overhead} \times Power_{overhead}} \tag{4.9}$$

以文献[20]中容量为 128×32 大小的 SRAM 为例，对应的 FoM 为 1.31（系统最大吞吐率收益）/1.176（面积开销）×1.526（能耗开销）=0.73。与传统的方案相比，本文方案收益的 FoM 提升了 1.96 倍。与文献[20]和文献[21]中的时序推测方案相比，本文方案收益的 FoM 分别提升了 1.75 倍和 1.73 倍，故综合考虑性能，能耗和面积，本文的方案是最佳选择。

表 4.1 时序推测方案的对比总结

		文献[20]			文献[21]			本文方案		
工艺		TSMC 28nm			TSMC 28nm			TSMC 28nm		
工作电压		0.5V			0.5V			0.5V		
存储阵列容量		128×32	256×32	512×32	128×32	256×32	512×32	128×32	256×32	512×32
检错方式		灵敏放大器两次使能			灵敏放大器两次使能			灵敏放大器两次使能+输入电压极性调整		
纠错方式		多路选择器+时钟门控			多路选择器+时钟门控			位线继续放电+时钟门控		
适用条件		组合逻辑主导的关键路径			组合逻辑主导的关键路径			组合逻辑/SRAM 主导的关键路径		
面积开销		17.6%	9.4%	4.8%	13.2%	7%	3.6%	6.4%	3.4%	1.8%
能耗开销		52.6%	30.7%	17.9%	56.5%	32.8%	19%	12%	-6.7%	-18.8%
性能	T_{CONV}/T_{ARRAY}	1.75	1.78	1.80	1.75	1.78	1.80	1.73	1.77	1.79
	T_{CONV}/T_{ERROR}	~1	~1	~1	~1	~1	~1	1.6	1.69	1.75
最大吞吐率收益		1.31	1.34	1.39	1.31	1.34	1.39	1.41	1.55	1.62
阵列收益的 FoM		0.73	0.94	1.12	0.74	0.94	1.13	1.18	1.61	1.96

*注：由于本文的时序推测方案是针对现有的时序推测方案在近阈值区的缺点提出的，故本节的对比是在低电压条件下进行的。

4.4 本章小结

本章以存储阵列为主体，设计了一款容量为 256×32 的宽电压时序推测型 SRAM，并完成版图的绘制和后仿真的验证，仿真结果表明：相比于传统的方案，基于时序推测方案的 SRAM 在 0.5V 和 0.9V 条件下的读出延时分别降低了 36%和 2%，采用时序推测方案可以在一定程度上减缓电压降低

对 SRAM 性能的影响。与传统的方案相比，本文方案收益的 FoM 提升了 1.96 倍。与文献[20]和文献[21]中的时序推测方案相比，本文方案收益的 FoM 分别提升了 1.75 倍和 1.73 倍，故综合考虑性能、能耗和面积，本文的方案是最佳选择。

第五章 总结与展望

随着集成电路的飞速发展，SoC 系统对 SRAM 提出了越来越严格的性能和能效要求。低至近阈值区的宽电压 SRAM 能够兼顾高性能和高能效需求，宽电压 SRAM 的设计在学术界引起了广泛的关注。但是近阈值区 SRAM 的性能退化严重，这会限制 SoC 芯片在近阈值区的最高工作频率。

5.1 总结

本文以降低存储阵列的读出延时为目标，以时序推测技术为技术手段。针对现有的时序推测方案在近阈值区的检错延时过大的缺点，本文提出了一种改进型的时序推测方案，本文的时序推测方案通过调节灵敏放大器的输入电压的极性实现快速检错。本文以时序推测型存储阵列为核心，基于 TSMC 28nm 工艺设计了一款容量为 256×32 的宽电压 SRAM，并完成后仿真验证。本文设计的 SRAM 的整体的概况如表 5.1 所示

表 5.1 SRAM 宏单元总体概况

宽电压时序推测型 SRAM 的设计		
工艺	TSMC 28nm CMOS	
存储单元类型	Foundry 6T	
容量	8 Kbits（256×32 bit）	
面积	2726μm²	
工作电压范围	0.5V-0.9V	
性能	6.34ns@0.5V	0.47ns@0.9V

本文的主要工作如下：

1) 宽电压 SRAM 的研究背景

第一章主要介绍了宽电压 SRAM 的研究背景与意义，并指出 SRAM 的性能在近阈值区严重退化，然后重点地分析了 SRAM 的读操作关键路径，并指出近阈值区 SRAM 性能急剧下降的原因。

2) SRAM 时序推测技术的设计综述

第二章为 SRAM 时序推测技术的设计综述，首先介绍了 SRAM 的基本结构和工作原理，包括 SRAM 的模块组成及各个模块的功能，然后介绍了 SRAM 六管存储单元的工作原理，这是本文设计的基础，其次给出了时序推测技术的设计综述，时序推测技术是一种能够克服存储阵列中弱驱动存储单元对存储阵列延时影响的优化技术，该技术能够提升 SRAM 的整体性能，本章在最后对时序推测技术做了总结，重点分析了现有的时序推测技术在近阈值区的局限性，同时指出现有的时序推测技术面积开销相对较大的问题。

3) 时序推测型存储阵列的设计

第三章介绍了本文提出的一种改进型的时序推测方案，本章首先介绍了改进型的时序推测方案的原理设计及电路实现，紧接着给出了 HSPICE-MATLAB 混合仿真方法，最后给出了仿真结果，仿真结果表明：本文的时序推测方案能够在一定程度上减小设计裕度，从而提升了存储阵列的整体性

能，相比传统的读出方案，存储阵列的读出延时在低电压下（0.5V）和正常电压下（0.9V）分别降低了大约 50%和 10%。

4) 宽电压 SRAM 的设计

第四章以时序推测型存储阵列为核心，基于 TSMC 28nm 工艺完成了一款容量为 256×32 的宽电压（0.5V-0.9V）SRAM 宏单元的设计。本章详细地介绍了 SRAM 的各部分的电路结构，版图设计及测试系统。仿真结果表明在低电压下（0.5V），相比于传统的设计方案，SRAM 的整体读出延时降低了 36%，在正常电压下（0.9V），相比于传统的设计方案，SRAM 的读出延时降低了 2%，故时序推测方案是消除设计裕度带来的性能损失的有效手段，采用时序推测方案可以在一定程度上减缓电压降低对 SRAM 性能的影响。与传统的方案相比，本文方案收益的 FoM 提升了 1.96 倍。与文献[20]和文献[21]中的时序推测方案相比，本文方案收益的 FoM 分别提升了 1.75 倍和 1.73 倍，故综合考虑性能、能耗和面积，本文的方案是最佳选择。

5.2 展望

1) 灵敏放大器的失调电压是影响存储阵列性能的一个重要方面，本文仅仅从时序推测的角度优化了存储阵列的性能，将时序推测技术与降低 SRAM 灵敏放大器失调电压[44-46]相结合的方案可以进一步提升存储阵列的性能。

2) 位线分级结构[47-49]经常使用在对性能要求高的场合，可以将位线分级技术和时序推测技术进一步结合，并建立位线分级的数学模型，实现对存储阵列的进一步优化。

3) 本文提出的时序推测方案正在流片，截止到目前，无法得到实测数据验证方案的合理性。

致谢

我的论文是在我的导师杨军教授的指导下完成的，杨老师为此付出了大量的心血，在此向杨老师表达最诚挚的感谢！杨老师勇于探索，不断创新的精神永远是我的学习目标。

感谢东南大学国家专用集成电路工程技术研究中心的所有老师，是他们为我提供了良好的硬件环境和研究氛围，感谢他们在项目、研究工作和生活上对我的指导和帮助。特别感谢已经毕业的丁瑞、高帅、浦浩、叶沐阳师兄，是他们指导我进行了 SRAM 基础知识的学习。

感谢吉昊、朱吉喆、周陶梅三位同学在论文撰写期间的启发与帮助。三年来，我们团结协作，共同努力，一起克服了工作中的各种困难。感谢他们的陪伴。

感谢我的师弟和师妹们：顾东志、刘炎、曹政坤、许逸波、李晓敏，他们为了课题组营造了活力，很怀念和他们在一起的日子，他们是我身后强大的精神支柱。祝愿他们前程似锦！

感谢商新超、孔羽尧，郭静静和周永亮四位博士，他们为实验室营造了良好的科研氛围，从他们的身上我学习到了不畏艰难，迎难而上的科研精神。

感谢我的父母，感谢父母的养育之恩，感谢父母无微不至的关怀，父母永远是我前进的动力。

最后，谨以此文献给所有关心、爱护、帮助过我的师长，家人和朋友们！

参考文献

[1] Alioto M. Ultra-low power VLSI circuit design demystified and explained: A tutorial[J]. IEEE Transactions on Circuits and Systems I: Regular Papers, 2012, 59(1): 3-29

[2] Damaraju S, George V, Jahagirdar S, et al. A 22nm IA multi-CPU and GPU system-on-chip[C]. In: IEEE International Solid-State Circuits Conference, Digest of Technical Papers. 2012. 56-57

[3] Lin B, Mallik A, Dinda P A, et al. Power reduction through measurement and modeling of users and cpus: Summary[C]. In: ACM SIGMETRICS Performance Evaluation Review. 2007. 363-364

[4] Dreslinski R G, Wieckowski M, Blaauw D, et al. Near-threshold computing: Reclaiming moore's law through energy efficient integrated circuits[J]. Proceedings of the IEEE, 2010, 98(2): 253-266

[5] Dreslinski R G, Chen G K, Mudge T, et al. Reconfigurable energy efficient near threshold cache architectures[C]. In: Symposium on Microarchitecture. 2008. 459-470

[6] Chen Q. A comparative study of threshold variations in symmetric and asymmetric undoped double-gate MOSFETs[C]. In: IEEE SOI Conference. 2002. 30-31

[7] Nikolic B, Park J H, Kwak J, et al. Technology variability from a design perspective[J]. IEEE Transactions on Circuits and Systems I: Regular Papers, 2011, 58(9): 1996-2009

[8] Kuhn K J, Giles M D, Becher D, et al. Process technology variation[J]. IEEE Transactions on Electron Devices, 2011, 58(8): 2197-2208

[9] Chang L, Montoye R K, Nakamura Y, et al. An 8T-SRAM for variability tolerance and low-voltage operation in high-performance caches[J]. IEEE Journal of Solid-State Circuits, 2008, 43(4): 956-963

[10] Takeda K, Ikeda H, Hagihara Y, et al. Redefinition of write margin for next-generation SRAM and write-margin monitoring circuit[C]. In: IEEE International Solid-State Circuits Conference, Digest of Technical Papers. 2006. 2602-2611

[11] Chang M F, Wu J J, Chen K T, et al. A differential data-aware power-supplied (DAP) 8T SRAM cell with expanded write/read stabilities for lower VDDmin applications[J]. IEEE Journal of Solid-State Circuits, 2010, 45(6): 1234-1245

[12] Yabuuchi M, Nii K, Tsukamoto Y, et al. A 45nm 0.6 V cross-point 8T SRAM with negative biased read/write assist[C]. In: IEEE Symposium on VLSI Circuits. 2009. 158-159

[13] Liu Z, Kursun V. Characterization of a novel nine-transistor SRAM cell[J]. IEEE transactions on very large scale integration (VLSI) systems, 2008, 16(4): 488-492

[14] Zhang K, Bhattacharya U, Chen Z, et al. A 3-GHz 70-Mb SRAM in 65-nm CMOS technology with integrated column-based dynamic power supply[J]. IEEE Journal of Solid-State Circuits, 2006, 41(1): 146-151

[15] Ohbayashi S, Yabuuchi M, Nii K, et al. A 65-nm SoC embedded 6T-SRAM designed for manufacturability with read and write operation stabilizing circuits[J]. IEEE Journal of Solid-State Circuits, 2007, 42(4): 820-829

[16] Yabuuchi M, Nii K, Tsukamoto Y, et al. A 45nm low-standby-power embedded SRAM with improved immunity against process and temperature variations[C]. In: IEEE International Solid-State Circuits Conference, Digest of Technical Papers. 2007. 326-606

[17] Sharma V, Cosemans S, Ashouei M, et al. A 4.4 pJ/Access 80 MHz, 128 kbit variability resilient SRAM with multi-sized sense amplifier redundancy[J]. IEEE Journal of Solid-State Circuits, 2011, 46(10): 2416-2430

[18] Sinangil M E, Mair H, Chandrakasan A P. A 28nm high-density 6T SRAM with optimized peripheral-assist circuits for operation down to 0.6V[C]. In: IEEE International Solid-State Circuits Conference, Digest of Technical Papers. 2011. 260-262

[19] Nii K, Yabuuchi M, Tsukamoto Y, et al. A 45-nm single-port and dual-port SRAM family with robust read/write stabilizing circuitry under DVFS environment[C]. In: IEEE Symposium on VLSI Circuits. 2008. 212-213

[20] Karl E, Sylvester D, Blaauw D. Timing error correction techniques for voltage-scalable on-chip memories[C]. In: IEEE International Symposium on Circuits and Systems. 2005. 3563-3566

[21] Khayatzadeh M, Saligane M, Wang J, et al. 17.3 A reconfigurable dual-port memory with error detection and correction in 28nm FDSOI[C]. In: IEEE International Solid-State Circuits Conference, Digest of Technical Papers. 2016. 310-312

[22] Houle R M. Simple Statistical Analysis Techniques to Determine Optimum Sense Amp Set Times [J]. IEEE Journal of Solid-State Circuits, 2008, 43 (8): 1816-1825

[23] Niki Y, Kawasumi A, Suzuki A, et al. A Digitized Replica Bitline Delay Technique for Random-Variation-Tolerant Timing Generation of SRAM Sense Amplifiers[J]. IEEE Journal of Solid-State Circuits, 2011, 46(11): 2545-2551

[24] Lin Z, Wu X, Li Z, et al. A Pipeline Replica Bitline Technique for Suppressing Timing Variation of SRAM Sense Amplifiers in a 28-nm CMOS Process[J]. IEEE Journal of Solid-State Circuits, 2017, 52(3): 669-677

[25] Komatsu S, Yamaoka M, Morimoto M, et al. A 40-nm low-power SRAM with multi-stage replica-bitline technique for reducing timing variation[C]. In: IEEE Custom Integrated Circuits Conference. 2009. 701-704

[26] Do A T, Kong Z H, Yeo K S. 0.9 V current-mode sense amplifier using concurrent bit-and data-line tracking and sensing techniques[J]. Electronics Letters, 2007, 43(25): 1421-1422

[27] Kawashima S, Mori T, Sasagawa R, et al. A charge-transfer amplifier and an encoded-bus architecture for low-power SRAM's[J]. IEEE Journal of Solid-State Circuits, 1998, 33(5): 793-799

[28] Pileggi L, Keskin G, Li X, et al. Mismatch analysis and statistical design at 65 nm and below[C]. In: IEEE Custom Integrated Circuits Conference. 2008. 9-12

[29] Wicht B, T Nirschl, D Schmitt-Landsiedel. Yield and speed optimization of a latch-type voltage sense amplifier [J]. IEEE Journal of Solid-State Circuits, 2004, 39(7): 1148-1158

[30] 关立军. 基于28nm工艺低电压SRAM单元电路设计[D]. 合肥：安徽大学，2017

[31] 吴秋雷. 低功耗 SRAM 存储单元关键技术研究及电路设计[D]. 合肥：安徽大学，2016

[32] Dan E, Kim N S, Das S, et al. Razor:A Low-Power Pipeline Based on Circuit-Level Timing Speculation[C]. In: International Symposium on Microarchitecture. 2003. 7-18

[33] 张永新, 陆生礼, 茆邦琴. 门控时钟的低功耗设计技术[J]. 微电子学与计算机, 2004, 21(1): 23-26

[34] Wu Q, Pedram M, Wu X. Clock-gating and its application to low power design of sequential circuits[C]. IEEE Custom Integrated Circuits Conference. 2000. 479-482.

[35] 高帅. 宽电压SRAM灵敏放大器的研究与实现[D]. 南京：东南大学，2016

[36] 张钿钿. 低电压 SRAM 存储单元及灵敏放大器设计[D]. 南京：东南大学，2015

[37] Roy K, Mukhopadhyay S, Mahmoodimeimand H. Leakage current mechanisms and leakage reduction techniques in deep-submicrometer CMOS circuits[J]. Proceedings of the IEEE, 2003, 91(2): 303-304

[38] Zhao Y, Parke S, Burke F. Modeling and characterization of deep-submicron MOSFET with short-channel effect based on BSIMTM[J]. Acta Electronica Sinica, 2004, 32(5): 841-844

[39] Kim T H, Keane J, Eom H, et al. Utilizing Reverse Short-Channel Effect for Optimal Subthreshold Circuit Design[J]. IEEE Transactions on Very Large Scale Integration Systems, 2007, 15(7): 821-829

[40] 常晓夏. 超大规模集成电路串扰问题的研究[D]. 北京邮电大学, 2006
[41] 马剑武, 陈书明, 孙永节. 深亚微米集成电路设计中串扰分析与解决方法[J]. 计算机工程与科学, 2005, 27(4): 102-104
[42] 朱贾峰，低电压 SRAM 关键技术研究与实现[D]. 南京:东南大学，2013
[43] 陆学斌. 集成电路版图设计[M]. 北京大学出版社, 2012
[44] Kawasumi A, Takeyama Y, Hirabayashi O, et al. A Low-Supply-Voltage-Operation SRAM With HCI Trimmed Sense Amplifiers [J]. IEEE Journal of Solid-State Circuits, 2010, 45 (11): 2341-2347
[45] Verma N, Chandrakasan A P. A 65nm 8T sub-Vt SRAM employing sense-amplifier redundancy[C]. In: IEEE International Solid-State Circuits Conference, Digest of Technical Papers. 2007. 328-606
[46] Sinangil Y, Chandrakasan A P. A 128 kbit SRAM with an embedded energy monitoring circuit and sense-amplifier offset compensation using body biasing[J]. IEEE Journal of Solid-State Circuits, 2014, 49(11): 2730-2739
[47] Karandikar A, Parhi K K. Low power SRAM design using hierarchical divided bit-line approach[C]. In: IEEE International Conference on Computer Design: VLSI in Computers and Processors. 1998. 82-88
[48] Yang B D, Kim L S. A low-power SRAM using hierarchical bit line and local sense amplifiers[J]. IEEE Journal of Solid-State Circuits, 2005, 40(6): 1366-1376
[49] Cosemans S, Dehaene W, Catthoor F. A low-power embedded SRAM for wireless applications[J]. IEEE journal of solid-state circuits, 2007, 42(7): 1607-1617