信息科学技术学术著作丛书

面向片上缓存子系统的功耗优化方法

何炎祥　沈凡凡　著

科学出版社

信息科学技术学术著作丛书

面向片上缓存子系统的功耗优化方法

何炎祥　沈凡凡　著

科学出版社

北京

内 容 简 介

缓存作为计算机存储体系结构中的重要组成部分，对系统功耗和性能非常关键。本书全面系统地介绍缓存优化方法及其关键技术，从存储体系结构的角度出发，解决缓存的静态功耗和动态功耗问题，从而保证系统整体功耗的降低。同时，本书还重点阐述新型非易失性存储技术在架构缓存中的应用与实践。本书的主要内容是作者近年来在该领域的最新研究成果，具有较强的原创性。

本书可作为高等院校和科研院所计算机科学与技术、计算机系统结构、计算机应用技术等相关专业的高年级本科生或研究生用书，也可供软件优化等相关领域的研究人员学习和参考。

图书在版编目(CIP)数据

面向片上缓存子系统的功耗优化方法/何炎祥，沈凡凡著. —北京：科学出版社，2018.1

（信息科学技术学术著作丛书）

ISBN 978-7-03-056477-1

Ⅰ.①面… Ⅱ.①何… ②沈… Ⅲ.①网络服务器 Ⅳ.①TP368.5

中国版本图书馆 CIP 数据核字(2018)第 020002 号

责任编辑：魏英杰 / 责任校对：桂伟利

责任印制：张 伟 / 封面设计：陈 敬

科 学 出 版 社 出版

北京东黄城根北街 16 号

邮政编码：100717

http://www.sciencep.com

北京中石油彩色印刷有限责任公司 印刷

科学出版社发行 各地新华书店经销

*

2018 年 1 月第 一 版 开本：720×1000 1/16

2018 年 1 月第一次印刷 印张：13 1/2

字数：270 000

定价：95.00 元

（如有印装质量问题，我社负责调换）

《信息科学技术学术著作丛书》序

21世纪是信息科学技术发生深刻变革的时代，一场以网络科学、高性能计算和仿真、智能科学、计算思维为特征的信息科学革命正在兴起。信息科学技术正在逐步融入各个应用领域并与生物、纳米、认知等交织在一起，悄然改变着我们的生活方式。信息科学技术已经成为人类社会进步过程中发展最快、交叉渗透性最强、应用面最广的关键技术。

如何进一步推动我国信息科学技术的研究与发展；如何将信息技术发展的新理论、新方法与研究成果转化为社会发展的推动力；如何抓住信息技术深刻发展变革的机遇，提升我国自主创新和可持续发展的能力？这些问题的解答都离不开我国科技工作者和工程技术人员的求索和艰辛付出。为这些科技工作者和工程技术人员提供一个良好的出版环境和平台，将这些科技成就迅速转化为智力成果，将对我国信息科学技术的发展起到重要的推动作用。

《信息科学技术学术著作丛书》是科学出版社在广泛征求专家意见的基础上，经过长期考察、反复论证之后组织出版的。这套丛书旨在传播网络科学和未来网络技术，微电子、光电子和量子信息技术、超级计算机、软件和信息存储技术、数据知识化和基于知识处理的未来信息服务业、低成本信息化和用信息技术提升传统产业，智能与认知科学、生物信息学、社会信息学等前沿交叉科学，信息科学基础理论，信息安全等几个未来信息科学技术重点发展领域的优秀科研成果。丛书力争起点高、内容新、导向性强，具有一定的原创性，体现出科学出版社“高层次、高水平、高质量”的特色和“严肃、严密、严格”的优良作风。

希望这套丛书的出版，能为我国信息科学技术的发展、创新和突破带来一些启迪和帮助。同时，欢迎广大读者提出好的建议，以促进和完善丛书的出版工作。

中国工程院院士

原中国科学院计算技术研究所所长

此为试读，需要完整PDF请访问：www.ertongbook.com

前　言

近年来，随着科技的迅猛发展，智能手机、电脑、可穿戴设备、智能家居设备和无人机等电子产品已被广泛使用并将逐步普及，这给人们的生活带来极大的便利。然而，这些产品续航能力不足的问题也渐渐地凸显出来。以低功耗设计为优化目标的产品是当今绿色智能电子设备发展的必由之路。

智能电子设备中最重要的组成部分是处理器和存储器。这两部分通常也是功耗开销的主要部分。随着半导体工艺的进步，处理器的运行速度越来越快，而主存的访问速度则相对缓慢，它们之间的性能差距逐渐增大，“存储墙”问题也日益严峻，片上缓存能在一定程度上缓解访问速度不匹配的问题，因此被广泛地使用在各种计算设备上。传统的片上缓存通常采用 SRAM 架构，因为它有访问速度快和使用寿命长等优点。然而，随着半导体特征尺寸的进一步降低，基于传统 CMOS 工艺的 SRAM 片上缓存的漏电功耗（静态功耗）将急剧增加，并逐渐占据主导地位。对于大容量的缓存，SRAM 存储单元将耗费大量的芯片面积。基于 SRAM 设计的片上缓存已经无法满足现代计算设备对低功耗和高性能的要求。

新型非易失性存储器（NVM）的出现为计算机存储技术提供了新的解决方案。NVM 很有希望替代传统存储技术，因为它具有漏电功耗低、存储密度高和非易失性等优良特点。为了充分利用 NVM 的这些优点，近年来有研究者提出使用 NVM 技术架构片上缓存。然而，新型存储器件的制造工艺和设计原理与 SRAM 不同，NVM 通常都有相同的缺点，即写操作的功耗相对较高、写操作的延迟相对较长和存储单元

的写寿命有限等。传统的缓存优化方法已不适应新技术的发展。那么运用新型非易失性存储技术架构片上缓存，需要在尽可能利用 NVM 优点的同时克服其写操作代价大的问题。

本书从存储体系结构的角度出发，分别采用分区技术、反馈学习、磨损均衡技术、数据分配技术、周期性学习和编译技术等方法优化系统的功耗。全书共 9 章，各章的主要内容组织如下。

第 1 章为绪论。首先，介绍传统缓存技术和新型非易失性存储技术的研究背景。然后，从功耗的角度讨论国内外相关领域的研究现状。最后，详细地介绍本书的主要研究内容和创新点，以及全书章节的组织结构。

第 2 章讨论缓存技术的研究现状。首先，介绍传统缓存技术的静态功耗和动态功耗优化方法。然后，重点讨论新型缓存技术的研究现状。

第 3 章讨论基于分区技术的缓存功耗优化。首先，讨论现有缓存存在功耗优化的不足。其次，分析缓存分区技术和消除死写块的潜在好处，并以此为研究动机提出相应的方法。再次，详细地介绍复用局部性感知的缓存分区方法。最后，给出实验评估方法及所提方法的实验效果。

第 4 章讨论基于反馈学习的非易失性缓存功耗优化。首先，讨论 STT-RAM 架构缓存尚存在的问题。其次，通过例子分析了死写终止的好处，并以此为研究动机提出相应的方法。再次，详细地介绍基于反馈学习的死写终止方法。最后，给出实验评估方法及所提方法的实验效果。

第 5 章讨论基于磨损均衡技术的非易失性缓存功耗优化。首先，讨论非易失性缓存及现有优化方法尚存在的问题。其次，通过实验分析缓存组内组间的写操作压力，并以此为研究动机提出相应的方法。再次，详细介绍磨损均衡技术指导缓存数据分配。最后，给出评估所提

方法的实验效果。

第6章讨论基于数据分配技术的混合缓存优化方法。首先，讨论混合缓存及现有优化方法尚存在的问题。其次，通过实验分析混合缓存的动态功耗和写操作问题，并以此为研究动机提出相应的方法。再次，详细地介绍缓存访问的统计行为指导数据分配。最后，给出评估所提方法的实验效果。

第7章讨论基于周期性学习的多级非易失性缓存功耗优化。首先，讨论多级STT-RAM缓存及现有优化方法尚存在的问题，分析多级STT-RAM缓存优化数据分配的好处，并以此为研究动机提出相应的方法。其次，形式化定义多级STT-RAM缓存功耗优化问题，同时给出了贪心算法的解决思路。再次，介绍离线分析缓存访问行为，并通过周期性学习这些行为来指导缓存数据分配。最后，给出评估所提方法的实验效果。

第8章讨论基于编译技术的PCM功耗优化。首先，讨论PCM及其现有优化方法存在的问题。其次，通过探索MLC PCM的写延迟和数据保留时间之间的关系，以分析写指令适合的写模式，并以此为研究动机提出相应的方法。再次，详细地介绍编译技术指导的双重写方法，包括构造控制流程图、存储器地址分析、定义可达性分析、最坏情形生命期分析和代码注入。最后，给出评估所提方法的实验效果。

第9章总结本书的主要研究工作，展望后续的研究工作。

本书是武汉大学计算机学院众多科研人员多年学习、研究和工程实践沉淀的成果总结。参与相关研究的人员包括吴伟、陈勇、徐超、江南、喻涛、张军、陈木朝、汪吕蒙、孙发军、吴炳廉、刘子俊、闫国昌、唐洪峰、张晓瞳、刘瑞、沈云飞、周一泓等。本书第8章主要由李清安参与撰写，其余章节主要由沈凡凡参与撰写。何炎祥具体规划和设计全书的内容，并对全书进行统稿，李清安对本书的初稿提出很多建设性意见。在此，对他们的积极参与和热心帮助表示衷心的感谢。

本书是专门针对缓存存储体系结构研究的学术著作，对相关领域的研究人员具有一定的借鉴意义和参考价值。本书的出版得到国家自然科学基金青年科学基金“基于编译的 PCM 内存损耗均衡方法研究”（项目编号：61502346）、国家自然科学基金青年科学基金“面向众核处理器的非易失性缓存低功耗技术的研究”（项目编号：61402145）、国家自然科学基金“基于线程调度的通用图形处理器性能优化方法研究”（项目编号：61662002）、国家自然科学基金“基于编译的嵌入式软件可靠性加强方法研究”（项目编号：61640220）、湖北省自然科学基金青年基金“面向嵌入式片上存储的低功耗编译优化方法”（项目编号：2015CFB338）、安徽省自然科学基金青年基金“基于 NVM 的高性能低功耗缓存系统研究”（项目编号：1508085QF138）、南京审计大学人才引进项目资助，以及江西省教育厅科技项目“通用图形处理器线程调度优化方法研究”（项目编号：GJJ150605）等项目的资助，在此一并表示感谢。

缓存技术及其应用是当前处于科学前沿的研究课题之一，相关的理论和技术还在发展中，许多新的思想、理论和方法还需要进一步完善和验证。限于作者的水平和经验，书中不妥之处在所难免，恳请读者批评指正，共同推进缓存技术研究的进步和发展。

2017 年 10 月于武汉

目　录

此为试读,需要完整PDF请访问: www.ertongbook.com

第1章 绪 论

本章首先介绍研究背景和存储技术的发展形势，讨论国内外相关领域的研究现状，然后总结全书的主要工作和创新点，最后介绍全书的组织结构。

1.1 研究背景

随着半导体工艺和集成电路技术的飞速发展，处理器主频因功耗问题无法进一步提升，研究者逐渐转向多核心处理器设计。多核技术的日趋成熟使计算机系统性能大幅度提升。为了平衡核心数增加带来的数据访问压力，需要容量更大的片上缓存，其缓存功耗也随之上升，逐步成为处理器功耗预算中的重要部分。当前计算机存储架构中大多采用传统存储技术，如 SRAM、DRAM 和 Flash 技术等，它们已经无法适应集成电路技术的新发展。近年来，新型非易失性存储器（non-volatile memory，NVM）得到学术界和工业界的高度关注，NVM 技术为计算机存储架构提供了新的解决方案。研究者提出使用 NVM 存储技术取代传统存储技术，以适应新工艺和新技术的发展。

1.1.1 传统缓存技术所面临的问题

近四五十年来，随着半导体工艺技术的提高和计算机体系结构的优化，多线程并行计算技术的广泛应用，现代集成电路技术的迅猛发展，处理器的性能得到飞跃式的提升。图 1.1 展示了处理器近 35 年的

发展趋势，处理器的评价指标包括单位面积的晶体管数量（即集成度）、CPU 的性能、CPU 的时钟频率、CPU 的功耗和 CPU 的核心数目[1]。可以看出，直到 2005 年，CPU 的集成度越来越高，性能、时钟频率和功耗等均逐渐上升，提升的主要原因是晶体管的工艺尺寸（technology node）在逐渐缩小。这有多个好处：一是芯片可以增加更多的功能；二是根据摩尔定律[2]，芯片集成度的提升将大大降低产品的成本。另外，从理论上说，晶体管缩小可以降低单个晶体管的功耗，因为工艺缩小原理要求降低芯片的整体供电电压，进而降低功耗。然而，实际上并非如此，从物理原理上讲，单位面积的功耗并不会降低，因此功耗随着集成度的提高而提高。温度过高会影响芯片的性能，甚至影响其正常工作。因此，在 2005 年以后，CPU 的时钟频率不再增长，性能的提升逐渐转向多核架构（图 1.1）。“功耗墙”问题减缓了处理器发展的速度。事实上，功耗问题在当今移动设备和大数据计算中心等领域都是一个非常重要的问题，逐渐成为制约处理器发展的瓶颈。

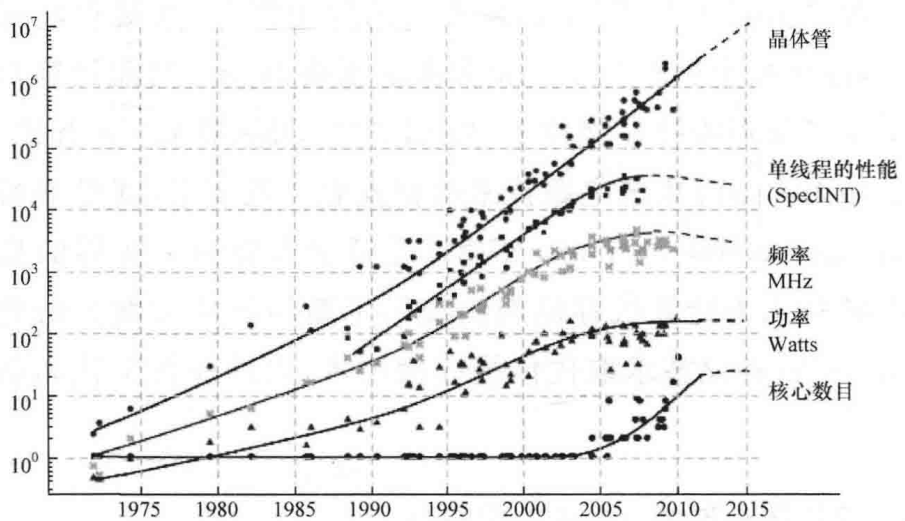

图 1.1 处理器近 35 年的发展趋势

功耗的增大将影响处理器的散热、封装及稳定性，处理器和应用程序的设计和维护成本也会相应的增加，这一因素限制了处理器性能的

进一步提高。通过分析处理器的体系结构可知，高性能的处理器高度依赖片上缓存(on-chip cache)。在处理器结构中，缓存占据大部分芯片面积，同时消耗了大量的功耗，占处理器功耗的30%～60%[3,4]。随着处理器负载的增加，缓存的功耗还会进一步上升。

现代计算机系统通常采用SRAM架构片上缓存，因为SRAM具有良好的读写性能，同时它的使用寿命长达10^{18}，特别适合用于靠近CPU的片上存储部分，如一级、二级和三级高速缓存。然而，SRAM存储单元是CMOS工艺制成的，其单元大小在120～200F^2[5]，因此其存储密度较低。随着半导体工艺尺寸的进一步缩小，基于传统CMOS工艺的SRAM所消耗的漏电功耗(静态功耗)会急剧增加，占据的比例逐步上升并成为主导因素。对于大容量缓存，它消耗的功耗更为严重，例如Intel Haswell-EP Xeon E5-2699 v3系列处理器的缓存大小为45MB，其平均功耗为145W[6]。

由此可见，基于SRAM的传统缓存技术存在存储密度不够高，漏电功耗相对较大的问题，因此不能适应半导体技术的发展趋势。优化缓存功耗问题成为当前处理器技术的重要研究方向。

1.1.2 新型非易失性存储技术带来的机遇

为了解决传统存储技术的不足，研究者探索了许多新型非易失性存储技术，新型NVM最有潜力取代传统存储技术。因为NVM具有访问速度快、漏电功耗低、集成度高和非易失性等优点。当前比较典型的NVM有自旋转移力矩存储器(spin-transfer torque RAM，STT-RAM)[7,8]、相变存储器(phase change memory，PCM)[9]、阻变存储器(resistive RAM，RRAM)[10]和赛道存储器(domain-wall memory或racetrack memory，DWM)[11,12]等。随着技术的革新和快速发展，这些技术已逐步从产品原型阶段走向产品产业化阶段。例如，SAMSUNG公司研制的512MB PCM存储芯片已经在手机存储卡中使用，它使用

此为试读，需要完整PDF请访问：www.ertongbook.com