MORGAN & CLAYPOOL PUBLISHERS

General-Purpose Graphics Processor Architecture

Tor M. Aamodt
Wilson Wai Lun Fung
Timothy G. Rogers

Synthesis Lectures on Computer Architecture

Margaret Martonosi, *Series Editor*

MORGAN & CLAYPOOL PUBLISHERS

通用图形处理器架构

Tor M. Aamodt Wilson
Wai Lun Fung Timoth
y G. Rogers

SYNTHESIS LECTURES ON
COMPUTER ARCHITECTURE

Margaret Martonosi, *Series Editor*

General-Purpose Graphics Processor Architectures

通用图形处理器架构

Synthesis Lectures on Computer Architecture

Synthesis Lectures on Computer Architecture publishes 50- to 100-page publications on topics pertaining to the science and art of designing, analyzing, selecting and interconnecting hardware components to create computers that meet functional, performance and cost goals. The scope will largely follow the purview of premier computer architecture conferences, such as ISCA, HPCA, MICRO, and ASPLOS.

计算机体系结构综合讲座

编辑
玛格丽特·马托诺西, ***Princeton University***

名誉创始编辑 Mark D. Hill,

University of Wisconsin, Madison

Synthesis Lectures on Computer Architecture 出版 50 到 100 页的出版物，主题涉及设计、分析、选择和互连硬件组件的科学和艺术，以创建满足功能、性能和成本目标的计算机。范围将主要遵循顶级计算机架构会议的范围，例如 ISCA、HPCA、MICRO 和 ASPLOS。

通用图形处理器架构 Tor M. Aamodt、Wilson Wai Lun Fung 和 Timothy G. Rogers 2018

异构系统编译算法 Steven Bell、Jing Pu、James Hegarty 和 Mark Horowitz 2018

虚拟内存的架构和操作系统支持 Abhishek Bhattacharjee 和 Daniel Lustig 2017

计算机架构师的深度学习 Brandon Reagen、Robert Adolf、Paul Whatmough、Gu-Yeon Wei 和 David Brooks 2017

片上网络，第二版 Natalie Enright Jerger、Tushar Krishna 和 Li-Shiuan Peh 2017

使用时间神经网络进行时空计算 James E. Smith 2017

四

虚拟化的硬件和软件支持 Edouard Bugnion、Ja
son Nieh 和 Dan Tsafrir 2017

数据中心设计和管理：计算机架构师的视角 Benjamin C. Lee 2016

内存层次结构压缩入门 Somayeh Sardashti、Angelos Arelakis、
Per Stenström 和 David A. Wood 2015

硬件加速器研究基础设施 Yakun Sophia Shao 和
David Brooks 2015

分析 Rajesh Bordawekar、Bob Blainey 和 R
uchir Puri 2015

Customizable Computing Yu-Ting Chen, Jason Cong, Michael Gill, Gl
enn Reinman, and Bingjun Xiao 2015

Die-stacking Architecture
Yuan Xie and Jishen Zha
o 2015
单指令多数据执行 Christopher J. Hughes 2
015

节能计算机架构：最新进展 Magnus Själander、Margare
t Martonosi 和 Stefanos Kaxiras 2014

FPGA 加速计算机系统仿真 Hari Angepat、Derek Chio
u、Eric S. Chung 和 James C. Hoe 2014

硬件预取入门 Babak Falsafi 和 T
homas F. Wenisch 2014

On-Chip Photonic Interconnects: A Computer Architect's Perspective
Christopher J. Nitta, Matthew K. Farrens, and Venkatesh Akella
2013

Optimization and Mathematical Modeling in Computer Architecture
Tony Nowatzki, Michael Ferris, Karthikeyan Sankaralingam, Cristian Estan, Nilay Vaish, and David Wood
2013

Security Basics for Computer Architects
Ruby B. Lee
2013

The Datacenter as a Computer: An Introduction to the Design of Warehouse-Scale Machines, Second Edition
Luiz André Barroso, Jimmy Clidaras, and Urs Hölzle
2013

Shared-Memory Synchronization
Michael L. Scott
2013

Resilient Architecture Design for Voltage Variation
Vijay Janapa Reddi and Meeta Sharma Gupta
2013

Multithreading Architecture
Mario Nemirovsky and Dean M. Tullsen
2013

Performance Analysis and Tuning for General Purpose Graphics Processing Units (GPGPU)
Hyesoon Kim, Richard Vuduc, Sara Baghsorkhi, Jee Choi, and Wen-mei Hwu
2012

Automatic Parallelization: An Overview of Fundamental Compiler Techniques
Samuel P. Midkiff
2012

Phase Change Memory: From Devices to Systems
Moinuddin K. Qureshi, Sudhanva Gurumurthi, and Bipin Rajendran
2011

Multi-Core Cache Hierarchies
Rajeev Balasubramonian, Norman P. Jouppi, and Naveen Muralimanohar
2011

片上光子互连：计算机架构师的视角 Christopher J. Nitta、Matthe
w K. Farrens 和 Venkatesh Akella 2013 年

计算机架构中的优化和数学建模 Tony Nowatzki、Michael Ferris、Karthikeyan S
ankaralingam、Cristian Estan、Nilay Vaish 和 David Wood 2013

计算机架构师的安全基础知识 Ruby
B. Lee 2013

数据中心作为计算机：仓库规模机器设计简介，第二版 Luiz André Barroso、
Jimmy Clidaras 和 Urs Hölzle 2013 年

共享内存同步 Michael L. Scott
2013

电压变化的弹性架构设计 Vijay Janapa Reddi 和
Meeta Sharma Gupta 2013

多线程架构 Mario Nemirovsky 和
Dean M. Tullsen 2013

通用图形处理的性能分析与调优 单位
(GPGPU) Hyesoon Kim、Richard Vuduc、Sara Baghsorkhi、Jee Cho
i 和 Wen-mei Hwu 2012

自动并行化：基本编译器技术概述 Samuel P. Midkiffi 2012

相变存储器：从设备到系统 Moinuddin K. Qureshi、Sudhan
va Gurumurthi 和 Bipin Rajendran 2011

多核缓存层次结构 Rajeev Balasubramonian、Norman P. Jouppi
和 Naveen Muralimanohar 2011

A Primer on Memory Consistency and Cache Coherence
Daniel J. Sorin, Mark D. Hill, and David A. Wood
2011

Dynamic Binary Modification: Tools, Techniques, and Applications
Kim Hazelwood
2011

Quantum Computing for Computer Architects, Second Edition
Tzvetan S. Metodi, Arvin I. Faruque, and Frederic T. Chong
2011

High Performance Datacenter Networks: Architectures, Algorithms, and Opportunities
Dennis Abts and John Kim
2011

Processor Microarchitecture: An Implementation Perspective
Antonio González, Fernando Latorre, and Grigorios Magklis
2010

Transactional Memory, Second Edition
Tim Harris, James Larus, and Ravi Rajwar
2010

Computer Architecture Performance Evaluation Methods
Lieven Eeckhout
2010

Introduction to Reconfigurable Supercomputing
Marco Lanzagorta, Stephen Bique, and Robert Rosenberg
2009

On-Chip Networks
Natalie Enright Jerger and Li-Shiuan Peh
2009

The Memory System: You Can't Avoid It, You Can't Ignore It, You Can't Fake It
Bruce Jacob
2009

Fault Tolerant Computer Architecture
Daniel J. Sorin
2009

内存一致性和缓存一致性入门 Daniel J. Sorin、Mark D. Hill 和 David A. Wood 2011

动态二进制修改：工具、技术和应用 Kim Hazelwood 2011

计算机架构师的量子计算，第二版 Tzvetan S. Metodi、Arvin I. Faruque 和 Frederic T. Chong 2011

高性能数据中心网络：架构、算法和机遇 Dennis Abts 和 John Kim 2011

处理器微架构：实施视角 Antonio González、Fernando Latorre 和 Grigorios Magklis 2010

事务内存，第二版 Tim Harris、James Larus 和 Ravi Rajwar 2010

计算机体系结构性能评估方法 Lieven Eeckhout 2010

可重构超级计算简介 Marco Lanzagorta、Stephen Bique 和 Robert Rosenberg 2009

片上网络 Natalie Enright Jerger 和 Li-Shiuan Peh 2009

记忆系统：你无法回避它、无法忽略它、无法伪造它 Bruce Jacob 2009

容错计算机架构 Daniel J. Sorin 2009

The Datacenter as a Computer: An Introduction to the Design of Warehouse-Scale Machines
Luiz André Barroso and Urs Hölzle
2009

Computer Architecture Techniques for Power-Efficiency
Stefanos Kaxiras and Margaret Martonosi
2008

Chip Multiprocessor Architecture: Techniques to Improve Throughput and Latency
Kunle Olukotun, Lance Hammond, and James Laudon
2007

Transactional Memory
James R. Larus and Ravi Rajwar
2006

Quantum Computing for Computer Architects
Tzvetan S. Metodi and Frederic T. Chong
2006

数据中心作为计算机：仓库规模机器的设计简介

路易斯·安德烈·巴罗佐和乌尔斯·
霍尔泽尔 2009

计算机架构技术在节能方面的应用 Stefanos Kaxiras
和 Margaret Martonosi 2008

芯片多处理器架构：提高吞吐量和延迟的技术 Kunle Olukotun、Lance Hammo
nd 和 James Laudon 2007

事务性记忆 James R. Larus
和 Ravi Rajwar 2006

计算机架构师的量子计算 Tzvetan S. Metodi
和 Frederic T. Chong 2006

General-Purpose Graphics Processor Architectures

Tor M. Aamodt, Wilson Wai Lun Fung, and Timothy G. Rogers

www.morganclaypool.com

ISBN: 9781627059237 paperback
ISBN: 9781627056182 ebook
ISBN: 9781681733586 hardcover

DOI 10.2200/S00848ED1V01Y201804CAC044

A Publication in the Morgan & Claypool Publishers series
SYNTHESIS LECTURES ON COMPUTER ARCHITECTURE

Lecture #44
Series Editor: Margaret Martonosi, *Princeton University*
Founding Editor Emeritus: Mark D. Hill, *University of Wisconsin, Madison*
Series ISSN
Print 1935-3235 Electronic 1935-3243

通用图形处理器架构 Tor M. Aamodt、Wilson Wai Lun Fung 和 Timothy G. Rogers

www.morganclaypool.com

ISBN：9781627059237 平装本 ISBN：9781627056182 电子书 ISBN：9781681733586 精装本

DOI 10.2200/S00848ED1V01Y201804CAC044

Morgan & Claypool 出版社系列出版物
SYNTHESIS LECTURES ON COMPUTER ARCHITECTURE

第44讲
系列编辑：Margaret Martonosi，*Princeton University* 创始编辑 名誉编辑：Mark D. Hill，*University of Wisconsin, Madison* 系列 ISSN 印刷版 1935-3235 电子版 1935-3243

General-Purpose Graphics Processor Architectures

Tor M. Aamodt
University of British Columbia

Wilson Wai Lun Fung
Samsung Electronics

Timothy G. Rogers
Purdue University

SYNTHESIS LECTURES ON COMPUTER ARCHITECTURE #44

MORGAN & CLAYPOOL PUBLISHERS

通用图形处理器架构

Tor M. Aamodt
University of British Columbia
冯伟麟
Samsung Electronics
蒂莫西·G·罗杰斯
Purdue University

SYNTHESIS LECTURES ON COMPUTER ARCHITECTURE #44

MORGAN & CLAYPOOL PUBLISHERS

ABSTRACT

Originally developed to support video games, graphics processor units (GPUs) are now increasingly used for general-purpose (non-graphics) applications ranging from machine learning to mining of cryptographic currencies. GPUs can achieve improved performance and efficiency versus central processing units (CPUs) by dedicating a larger fraction of hardware resources to computation. In addition, their general-purpose programmability makes contemporary GPUs appealing to software developers in comparison to domain-specific accelerators. This book provides an introduction to those interested in studying the architecture of GPUs that support general-purpose computing. It collects together information currently only found among a wide range of disparate sources. The authors led development of the GPGPU-Sim simulator widely used in academic research on GPU architectures.

The first chapter of this book describes the basic hardware structure of GPUs and provides a brief overview of their history. Chapter 2 provides a summary of GPU programming models relevant to the rest of the book. Chapter 3 explores the architecture of GPU compute cores. Chapter 4 explores the architecture of the GPU memory system. After describing the architecture of existing systems, Chapters 3 and 4 provide an overview of related research. Chapter 5 summarizes cross-cutting research impacting both the compute core and memory system.

This book should provide a valuable resource for those wishing to understand the architecture of graphics processor units (GPUs) used for acceleration of general-purpose applications and to those who want to obtain an introduction to the rapidly growing body of research exploring how to improve the architecture of these GPUs.

KEYWORDS

GPGPU, computer architecture

抽象的

图形处理器 (GPU) 最初是为支持视频游戏而开发的，现在越来越多地用于从机器学习到加密货币挖掘等通用（非图形）应用。与中央处理单元 (CPU) 相比，GPU 可以通过将更大比例的硬件资源用于计算来实现更高的性能和效率。此外，与领域特定加速器相比，当代 GPU 的通用可编程性使软件开发人员对它更具吸引力。本书为那些有兴趣研究支持通用计算的 GPU 架构的读者提供了入门知识。它收集了目前仅在各种不同来源中找到的信息。作者领导了 GPGPU-Sim 模拟器的开发，该模拟器广泛用于 GPU 架构的学术研究。

本书第一章介绍了 GPU 的基本硬件结构并简要概述了其历史。第 2 章总结了与本书其余部分相关的 GPU 编程模型。第 3 章探讨了 GPU 计算核心的架构。第 4 章探讨了 GPU 内存系统的架构。在描述现有系统的架构之后，第 3 章和第 4 章概述了相关研究。第 5 章总结了影响计算核心和内存系统的跨领域研究。

本书为希望了解用于通用应用程序加速的图形处理器单元 (GPU) 架构的人士，以及想要了解快速发展的有关如何改进这些 GPU 架构的研究的人士提供了宝贵的资源。

关键词

GPGPU、计算机架构

Contents

内容

Preface

This book is intended for those wishing to understand the architecture of graphics processor units (GPUs) and to obtain an introduction to the growing body of research exploring how to improve their design. It is assumed readers have a familiarity with computer architecture concepts such as pipelining and caches and are interested in undertaking research and/or development related to the architecture of GPUs. Such work tends to focus on trade-offs between different designs, and thus this book is written with a view to providing insights into such trade-offs so that the reader can avoid having to learn by trial and error what is already known to experienced designers.

To help achieve this, the book collects together into one resource many relevant bits of information currently found among a wide range of disparate sources such as patents, product documents, and research papers. It is our hope this will help reduce the time it takes for a student or practitioner just starting to do their own research to become productive.

While this book covers aspects of current GPU designs, it also attempts to "synthesize" published research. This is partly due to necessity, as very little has been said by vendors on the microarchitecture of specific GPU products. In describing a "baseline" GPGPU architecture, this book relies both upon published product descriptions (journal papers, whitepapers, manuals) and, in some cases, descriptions in patents. The details found in patents may differ substantially from the microarchitecture of actual products. In some cases, microbenchmark studies have clarified for researchers some details, but in others our baseline represents our "best guess" based upon publicly available information. Nonetheless, we believe this will be helpful as our focus is understanding architecture trade-offs that have already been studied or might be interesting to explore in future research.

Several portions of this book focus on summarizing the many recent research papers on the topic of improving GPU architectures. As this topic has grown significantly in popularity in recent years, there is too much to cover in this book. As such, we have had to make difficult choices about what to cover and what to leave out.

Tor M. Aamodt, Wilson Wai Lun Fung, and Timothy G. Rogers
April 2018

Preface

本书面向希望了解图形处理器单元 (GPU) 架构并了解日益增多的改进设计研究的读者。本书假定读者熟悉计算机架构概念（例如流水线和缓存），并有兴趣进行与 GPU 架构相关的研究和/或开发。此类工作往往侧重于不同设计之间的权衡，因此本书旨在提供对此类权衡的见解，以便读者避免通过反复试验来学习经验丰富的设计师已知的知识。

为了实现这一目标，本书将目前在专利、产品文档和研究论文等各种不同来源中找到的许多相关信息汇集到一个资源中。我们希望这将有助于减少刚开始进行自己的研究的学生或从业者取得成果所需的时间。

本书虽然涵盖了当前 GPU 设计的各个方面，但也试图“综合”已发表的研究成果。这部分是出于必要，因为供应商很少谈论特定 GPU 产品的微架构。在描述“基线”GPGPU 架构时，本书既依赖已发布的产品描述（期刊论文、白皮书、手册），有时也依赖专利中的描述。专利中的细节可能与实际产品的微架构大不相同。在某些情况下，微基准研究为研究人员澄清了一些细节，但在其他情况下，我们的基线代表了我们基于公开信息的“最佳猜测”。尽管如此，我们相信这会有所帮助，因为我们的重点是了解已经研究过的或可能在未来研究中值得探索的架构权衡。

本书的几个部分重点总结了最近关于改进 GPU 架构的许多研究论文。由于这个主题近年来越来越受欢迎，本书要涵盖的内容太多了。因此，我们不得不做出艰难的选择，决定要涵盖什么，要省略什么。

Tor M. Aamodt、Wilson Wai Lun Fung 和 Timothy G. Rogers 2
018 年 4 月

Acknowledgments

We would like to thank our families for their support while writing this book. Moreover, we thank our publisher, Michael Morgan and editor, Margaret Martonosi, for the extreme patience they have shown while this book came together. We also thank Carole-Jean Wu, Andreas Moshovos, Yash Ukidave, Aamir Raihan, and Amruth Sandhupatla for providing detailed feedback on early drafts of this book. Finally, we thank Mark Hill for sharing his thoughts on strategies for writing Synthesis Lectures and specific suggestions for this book.

Tor M. Aamodt, Wilson Wai Lun Fung, and Timothy G. Rogers
April 2018

致谢

我们要感谢家人在撰写本书期间给予的支持。此外，我们还要感谢我们的出版商 Michael Morgan 和编辑 Margaret Martonosi，感谢他们在本书完成过程中表现出的极大耐心。我们还感谢 Carole-Jean Wu、Andreas Moshovos、Yash Ukidave、Aamir Raihan 和 Amruth Sandhupatla 对本书初稿提供的详细反馈。最后，我们感谢 Mark Hill 分享他对撰写综合讲座的策略的想法以及对本书的具体建议。

Tor M. Aamodt、Wilson Wai Lun Fung 和 Timothy G. Rogers 2
018 年 4 月

CHAPTER 1

Introduction

This book explores the hardware design of graphics processor units (GPUs). GPUs were initially introduced to enable real-time rendering with a focus on video games. Today GPUs are found everywhere from inside smartphones, laptops, datacenters, and all the way to supercomputers. Indeed, an analysis of the Apple A8 application processor shows that it devotes more die area to its integrated GPU than to central processor unit (CPU) cores [A8H]. The demand for ever more realistic graphics rendering was the initial driver of innovation for GPUs [Montrym and Moreton, 2005]. While graphics acceleration continues to be their primary purpose, GPUs increasingly support non-graphics computing. One prominent example of this receiving attention today is the growing use GPUs to develop and deploying machine learning systems [NVIDIA Corp., 2017]. Thus, the emphasis of this book is on features relevant to improving the performance and energy efficiency of non-graphics applications.

This introductory chapter provides a brief overview of GPUs. We start in Section 1.1 by considering the motivation for the broader category of computation accelerators to understand how GPUs compare to other options. Then, in Section 1.2, we provide a quick overview of contemporary GPU hardware. Finally, Section 1.4 provides a roadmap to the rest of this book.

1.1 THE LANDSCAPE OF COMPUTATION ACCELERATORS

For many decades, succeeding generations of computing systems showed exponential increasing performance per dollar. The underlying cause was a combination of reduced transistor sizes, improvements in hardware architecture, improvements in compiler technology, and algorithms. By some estimates half of those performance gains were due to reductions in transistor size that lead to devices that operate faster [Hennessy and Patterson, 2011]. However, since about 2005, the scaling of transistors has failed to follow the classical rules now known as Dennard Scaling [Dennard et al., 1974]. One key consequence is that clock frequencies now improve much more slowly as devices become smaller. To improve performance requires finding more efficient hardware architectures.

By exploiting hardware specialization it is possible to improve energy efficiency by as much as 500× [Hameed et al., 2010]. As shown by Hameed et al., there are several key aspects to attaining such gains in efficiency. Moving to vector hardware, such as that found in GPUs, yields about a 10× gain in efficiency by eliminating overheads of instruction processing. A large part of the remaining gains of hardware specialization are a result of minimizing data movement which

CHAPTER 1

介绍

本书探讨了图形处理器单元 (GPU) 的硬件设计。GPU 最初是为了实现实时渲染而引入的，主要应用于视频游戏。如今，从智能手机、笔记本电脑、数据中心到超级计算机，GPU 随处可见。事实上，对 Apple A8 应用处理器的分析表明，它为集成 GPU 分配的芯片面积比中央处理器单元 (CPU) 内核的芯片面积更大 [A8H]。对更逼真的图形渲染的需求是 GPU 创新的最初驱动力 [Montrym and Moreton，2005]。虽然图形加速仍然是其主要用途，但 GPU 越来越多地支持非图形计算。如今，一个备受关注的突出例子是越来越多地使用 GPU 来开发和部署机器学习系统 [NVIDIA Corp.，2017]。因此，本书的重点是与提高非图形应用程序的性能和能源效率相关的功能。

本章为入门章节，简要概述了 GPU。我们从第 1.1 节开始，探讨了更广泛类别的计算加速器的发展动机，以了解 GPU 与其他选项的比较。然后，在第 1.2 节中，我们简要概述了当代 GPU 硬件。最后，第 1.4 节提供了本书其余部分的路线图。

1.1 计算加速器的概况

几十年来，一代又一代的计算系统都表现出了价格比呈指数级增长的趋势。其根本原因是晶体管尺寸减小、硬件架构改进、编译器技术和算法改进等因素的共同作用。据估计，这些性能提升中有一半是由于晶体管尺寸减小，从而导致设备运行速度更快 [Hennessy and Patterson，2011]。然而，自 2005 年左右以来，晶体管的缩放比例已不再遵循现在称为 Dennard Scaling 的经典规则 [Dennard et al.，1974]。一个关键的后果是，随着设备尺寸的减小，时钟频率的提高速度现在要慢得多。要提高性能，就需要找到更高效的硬件架构。

通过利用硬件专业化，可以将能源效率提高多达 500× [Hameed 等人，2010]。正如 Hameed 等人所指出的，实现这种效率提升有几个关键方面。转向矢量硬件（例如 GPU 中的矢量硬件）通过消除指令处理的开销，可使效率提高约 10×。硬件专业化其余收益的很大一部分是最小化数据移动的结果，这

can be achieved by introducing complex operations that perform multiple arithmetic operations while avoiding accesses to large memory arrays such as register files.

A key challenge for computer architects today is finding better ways to balance the gains in efficiency that can be obtained by using specialized hardware with the need for flexibility required to support a wide range of programs. In the absence of architectures only algorithms that can be used for a large number of applications will run efficiently. An emerging example is hardware specialized for supporting deep neural networks such as Google's Tensor Processing Unit [Jouppi et al., 2017]. While machine learning appears likely to occupy a very large fraction of computing hardware resources, and these may migrate to specialized hardware, we argue there will remain a need for efficiently supporting computation expressed as software written in traditional programming languages.

One reason for the strong interest in GPU computing outside of the use of GPUs for machine learning is that modern GPUs support a Turing Complete programming model. By Turing Complete, we mean that any computation can be run given enough time and memory. Relative to special-purpose accelerators, modern GPUs are flexible. For software that can make full use of GPU hardware, GPUs can be an order of magnitude more efficient than CPUs [Lee et al., 2010]. This combination of flexibility and efficiency is highly desirable. As a consequence many of the top supercomputers, both in terms of peak performance and energy efficiency now employ GPUs [top]. Over succeeding generations of products, GPU manufacturer's have refined the GPU architecture and programming model to increase flexibility while simultaneously improving energy efficiency.

1.2 GPU HARDWARE BASICS

Often those encountering GPUs for the first time ask whether they might eventually replace CPUs entirely. This seems unlikely. In present systems GPUs are not stand-alone computing devices. Rather, they are combined with a CPU either on a single chip or by inserting an add-in card containing only a GPU into a system containing a CPU. The CPU is responsible for initiating computation on the GPU and transferring data to and from the GPU. One reason for this division of labor between CPU and GPU is that the beginning and end of the computation typically require access to input/output (I/O) devices. While there are ongoing efforts to develop application programming interfaces (APIs) providing I/O services directly on the GPU, so far these all assume the existence of a nearby CPU [Kim et al., 2014, Silberstein et al., 2013]. These APIs function by providing convenient interfaces that hide the complexity of managing communication between the CPU and GPU rather than eliminating the need for a CPU entirely. Why not eliminate the CPU? The software used to access I/O devices and otherwise provide operating system services would appear to lack features, such as massive parallelism, that would make them suitable to run on the GPU. Thus, we start off by considering the interaction of the CPU and GPU.

可以通过引入执行多个算术运算的复杂操作来实现，同时避免访问大型存储器阵列（例如寄存器文件）。

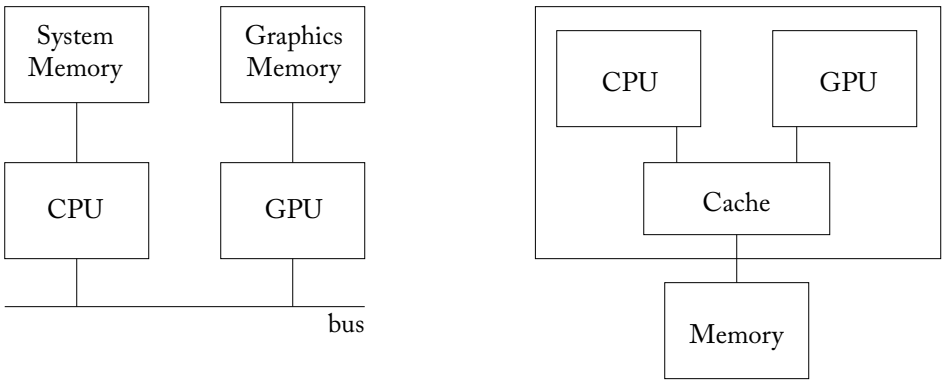
当今计算机架构师面临的一个关键挑战是找到更好的方法来平衡使用专用硬件所能获得的效率提升与支持各种程序所需的灵活性需求。在没有架构的情况下，只有可用于大量应用程序的算法才能高效运行。一个新兴的例子是专门用于支持深度神经网络的硬件，例如 Google 的张量处理单元 [Jouppi et al., 2017]。虽然机器学习似乎可能会占用很大一部分计算硬件资源，并且这些资源可能会迁移到专用硬件，但我们认为，仍然需要有效地支持以传统编程语言编写的软件来表示的计算。

除了将 GPU 用于机器学习之外，人们对 GPU 计算产生浓厚兴趣的原因之一是现代 GPU 支持图灵完备编程模型。图灵完备的意思是，只要有足够的时间和内存，就可以运行任何计算。相对于专用加速器，现代 GPU 非常灵活。对于可以充分利用 GPU 硬件的软件，GPU 的效率可以比 CPU 高出一个数量级 [Lee et al., 2010]。这种灵活性和效率的结合非常可取。因此，许多顶级超级计算机（无论是峰值性能还是能效）现在都采用了 GPU [top]。在后续几代产品中，GPU 制造商已经改进了 GPU 架构和编程模型，以提高灵活性，同时提高能效。

1.2 GPU 硬件基础知识

那些第一次接触 GPU 的人经常会问，GPU 最终是否会完全取代 CPU。这似乎不太可能。在目前的系统中，GPU 并不是独立的计算设备。相反，它们与 CPU 组合在单个芯片上，或者通过将仅包含 GPU 的附加卡插入包含 CPU 的系统中来实现。CPU 负责在 GPU 上启动计算并将数据传输到 GPU 和从 GPU 传输数据。CPU 和 GPU 之间进行这种分工的原因之一是计算的开始和结束通常需要访问输入/输出 (I/O) 设备。虽然人们一直在努力开发应用程序编程接口 (API)，以直接在 GPU 上提供 I/O 服务，但到目前为止，这些都假设存在附近的 CPU [Kim et al., 2014, Silberstein et al., 2013]。这些 API 的功能是提供方便的接口，隐藏管理 CPU 和 GPU 之间通信的复杂性，而不是完全消除对 CPU 的需求。为什么不消除 CPU？用于访问 I/O 设备并以其他方式提供操作系统服务的软件似乎缺乏适合在 GPU 上运行的功能（例如大规模并行性）。因此，我们首先考虑 CPU 和 GPU 之间的交互。

An abstract diagram showing a typical system containing a CPU and GPU is shown in Figure 1.1. On the left is a typical discrete GPU setup including a bus connecting the CPU and GPU (e.g., PCIe) for architectures such as NVIDIA's Volta GPU, and on the right is a logical diagram of a typical integrated CPU and GPU such as AMD's Bristol Ridge APU or a mobile GPU. Notice that systems including discrete GPUs have separate DRAM memory spaces for the CPU (often called system memory) and the GPU (often called device memory). The DRAM technology used for these memories is often different (DDR for CPU vs. GDDR for GPU). The CPU DRAM is typically optimized for low latency access whereas the GPU DRAM is optimized for high throughput. In contrast, systems with integrated GPUs have a single DRAM memory space and therefore necessarily use the same memory technology. As integrated CPUs and GPUs are often found on low-power mobile devices the shared DRAM memory is often optimized for low power (e.g., LPDDR).



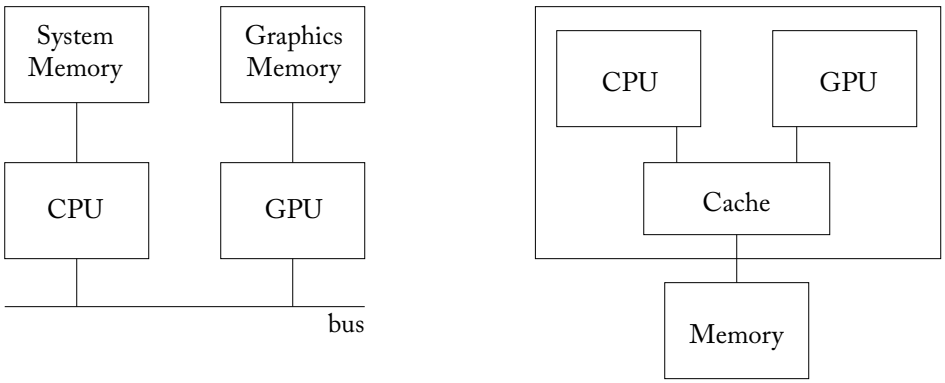
(a) System with discrete GPU

(b) Integrated CPU and GPU

Figure 1.1: GPU computing systems include CPUs.

A GPU computing application starts running on the CPU. Typically, the CPU portion of the application will allocate and initialize some data structures. On older discrete GPUs from both NVIDIA and AMD the CPU portion of the GPU Computing application typically allocates space for data structures in both CPU and GPU memory. For these GPUs, the CPU portion of the application must orchestrate the movement of data from CPU memory to GPU memory. More recent discrete GPUs (e.g., NVIDIA's Pascal architecture) have software and hardware support to automatically transfer data from CPU memory to GPU memory. This can be achieved by leveraging virtual memory support [Gelado et al., 2010], both on the CPU and GPU. NVIDIA calls this "unified memory." On systems in which the CPU and GPU are integrated onto the same chip and share the same memory, no programmer controlled copying from CPU memory to GPU memory is necessary. However, because CPUs and GPUs use caches and

图 1.1 显示了包含 CPU 和 GPU 的典型系统的抽象图。左侧是典型的独立 GPU 设置，包括连接 CPU 和 GPU 的总线（例如 PCIe），用于 NVIDIA 的 Volta GPU 等架构，右侧是典型的集成 CPU 和 GPU 的逻辑图，例如 AMD 的 Bristol Ridge APU 或移动 GPU。请注意，包含独立 GPU 的系统为 CPU（通常称为系统内存）和 GPU（通常称为设备内存）提供了单独的 DRAM 内存空间。用于这些内存的 DRAM 技术通常不同（CPU 使用 DDR，GPU 使用 GDDR）。CPU DRAM 通常针对低延迟访问进行了优化，而 GPU DRAM 针对高吞吐量进行了优化。相比之下，具有集成 GPU 的系统具有单个 DRAM 内存空间，因此必然使用相同的内存技术。由于集成 CPU 和 GPU 通常出现在低功耗移动设备上，因此共享 DRAM 内存通常针对低功耗进行了优化（例如 LPDDR）。



(a) System with discrete GPU

(b) Integrated CPU and GPU

图 1.1：GPU 计算系统包括 CPU。

GPU 计算应用程序开始在 CPU 上运行。通常，应用程序的 CPU 部分将分配并初始化一些数据结构。在 NVIDIA 和 AMD 的旧款独立 GPU 上，GPU 计算应用程序的 CPU 部分通常会在 CPU 和 GPU 内存中为数据结构分配空间。对于这些 GPU，应用程序的 CPU 部分必须协调数据从 CPU 内存到 GPU 内存的移动。较新的独立 GPU（例如 NVIDIA 的 Pascal 架构）具有软件和硬件支持，可自动将数据从 CPU 内存传输到 GPU 内存。这可以通过利用 CPU 和 GPU 上的虚拟内存支持 [Gelado 等，2010] 来实现。NVIDIA 称之为“统一内存”。在 CPU 和 GPU 集成在同一芯片上并共享相同内存的系统中，不需要程序员控制从 CPU 内存到 GPU 内存的复制。但是，由于 CPU 和 GPU 使用缓存和

some of these caches may be private, there can be a cache-coherence problem, which hardware developers need to address [Power et al., 2013b].

At some point, the CPU must initiate computation on the GPU. In current systems this is done with the help of a driver running on the CPU. Before launching computation on the GPU, a GPU computing application specifies which code should run on the GPU. This code is commonly referred to as a kernel (more details in Chapter 2). At the same time the CPU portion of the GPU computing application also specifies how many threads should run and where these threads should look for input data. The kernel to run, number of threads, and data location are conveyed to the GPU hardware via the driver running on the CPU. The driver will translate the information and place it memory accessible by the GPU at a location where the GPU is configured to look for it. The driver then signals the GPU that it has new computations it should run.

A modern GPU is composed of many cores, as shown in Figure 1.2. NVIDIA calls these cores *streaming multiprocessors* and AMD calls them *compute units*. Each GPU core executes a single-instruction multiple-thread (SIMT) program corresponding to the kernel that has been launched to run on the GPU. Each core on a GPU can typically run on the order of a thousand threads. The threads executing on a single core can communicate through a scratchpad memory and synchronize using fast barrier operations. Each core also typically contains first-level instruction and data caches. These act as bandwidth filters to reduce the amount of traffic sent to lower levels of the memory system. The large number of threads running on a core are used to hide the latency to access memory when data is not found in the first-level caches.

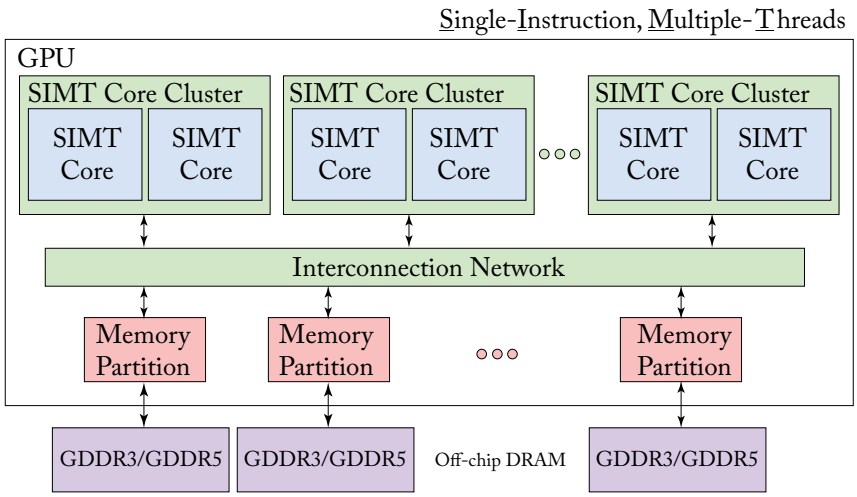


Figure 1.2: A generic modern GPU architecture.

To sustain high computation throughput it is necessary to balance high computational throughput with high memory bandwidth. This in turn requires parallelism in the memory sys-

其中一些缓存可能是私有的，可能存在缓存一致性问题，硬件开发人员需要解决这个问题[Power et al., 2013b]。

在某个时刻，CPU 必须在 GPU 上启动计算。在当前系统中，这是在 CPU 上运行的驱动程序的帮助下完成的。在 GPU 上启动计算之前，GPU 计算应用程序会指定应在 GPU 上运行哪些代码。此代码通常称为内核（第 2 章中有更多详细信息）。同时，GPU 计算应用程序的 CPU 部分还指定应运行多少个线程以及这些线程应在何处查找输入数据。要运行的内核、线程数和数据位置通过 CPU 上运行的驱动程序传达给 GPU 硬件。驱动程序将转换信息并将其放置在 GPU 可访问的内存中，GPU 已配置为查找它的位置。然后，驱动程序向 GPU 发出信号，告知它有新的计算需要运行。

现代 GPU 由许多核心组成，如图 1.2 所示。NVIDIA 将这些核心称为 *streaming multiprocessors*，AMD 将它们称为 *compute units*。每个 GPU 核心执行与已启动以在 GPU 上运行的内核相对应的单指令多线程 (SIMT) 程序。GPU 上的每个核心通常可以运行大约一千个线程。在单个核心上执行的线程可以通过暂存器进行通信并使用快速屏障操作进行同步。每个核心通常还包含第一级指令和数据缓存。它们充当带宽过滤器，以减少发送到内存系统较低级别的流量。当在第一级缓存中找不到数据时，核心上运行的大量线程用于隐藏访问内存的延迟。

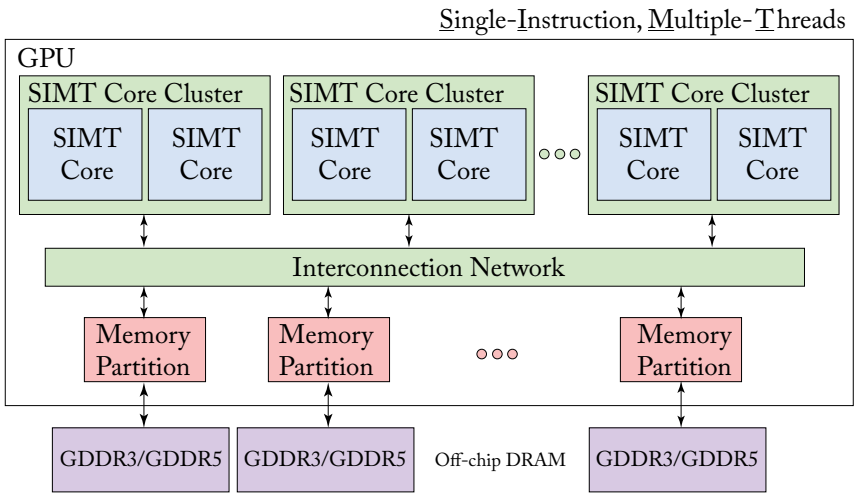


图 1.2：通用的现代 GPU 架构。

为了维持高计算吞吐量，必须在高计算吞吐量和高内存带宽之间取得平衡。这又要求内存系统中具有并行性。

tem. In GPUs this parallelism is provided by including multiple memory channels. Often, each memory channel has associated with it a portion of last-level cache in a memory partition. The GPU cores and memory partitions are connected via an on-chip interconnection network such as a crossbar. Alternative organizations are possible. For example, the Intel Xeon Phi, which directly competes with GPUs in the supercomputing market, distributes the last-level cache with the cores.

GPUs can obtain improved performance per unit area vs. superscalar out-of-order CPUs on highly parallel workloads by dedicating a larger fraction of their die area to arithmetic logic units and correspondingly less area to control logic. To develop intuition into the tradeoffs between CPU and GPU architectures, Guz et al. [2009] developed an insightful analytical model showing how performance varies with number of threads. To keep their model simple, they assume a simple cache model in which threads do not share data and infinite off-chip memory bandwidth. Figure 1.3 which reproduces a figure from their paper, illustrates an interesting trade-off they found with their model. When a large cache is shared among a small number of threads (as is the case in multicore CPUs), performance increases with the number of threads. However, if the number of threads increases to the point that the cache cannot hold the entire working set, performance decreases. As the number of threads increases further, performance increases with the ability of multithreading to hide long off-chip latency. GPUs architectures are represented by the right-hand side of this figure. GPUs are designed to tolerate frequent cache misses by employing multithreading.

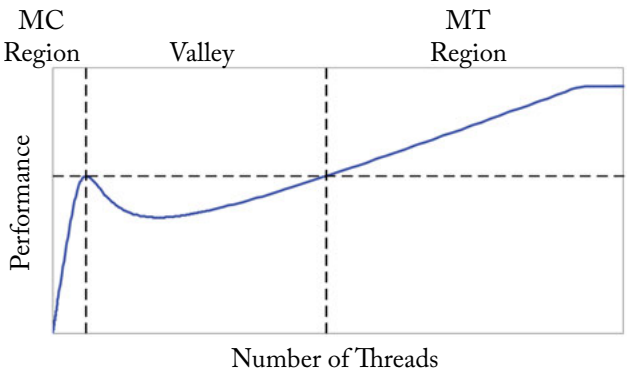


Figure 1.3: An analytical model-based analysis of the performance tradeoff between multicore (MC) CPU architectures and multithreaded (MT) architectures such as GPUs shows a "performance valley" may occur if the number of threads is insufficient to cover off-chip memory access latency (based on Figure 1 from Guz et al. [2009]).

With the end of Dennard Scaling [Horowitz et al., 2005], increasing energy efficiency has become a primary driver of innovation in computer architecture research. A key observation is that accessing large memory structures can consume as much or more energy as computation.

tem。在 GPU 中，这种并行性是通过包含多个内存通道来实现的。通常，每个内存通道都与内存分区中的最后一级缓存的一部分相关联。GPU 内核和内存分区通过片上互连网络（如交叉开关）连接。其他组织方式也是可能的。例如，在超级计算市场上直接与 GPU 竞争的 Intel Xeon Phi 将最后一级缓存与内核一起分配。

在高度并行的工作负载上，GPU 可以通过将更大比例的芯片面积分配给算术逻辑单元并相应地减少控制逻辑面积来获得比超标量乱序 CPU 更高的单位面积性能。为了直观地了解 CPU 和 GPU 架构之间的权衡，Guz 等人 [2009] 开发了一个富有洞察力的分析模型，展示了性能如何随线程数而变化。为了使模型简单化，他们假设了一个简单的缓存模型，其中线程不共享数据和无限的片外内存带宽。图 1.3 重现了他们论文中的一张图，说明了他们在模型中发现的一个有趣的权衡。当少数线程共享一个大缓存时（多核 CPU 就是这种情况），性能会随着线程数量的增加而提高。但是，如果线程数量增加到缓存无法容纳整个工作集的程度，性能就会下降。随着线程数进一步增加，性能也随之提高，多线程能够隐藏较长的片外延迟。GPU 架构由该图的右侧表示。GPU 旨在通过采用多线程来容忍频繁的缓存未命中。

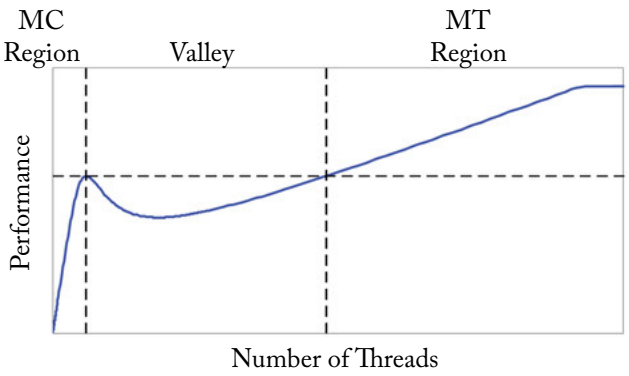


图 1.3：基于分析模型对多核 (MC) CPU 架构和多线程 (MT) 架构（如 GPU）之间的性能权衡的分析表明，如果线程数量不足以覆盖片外内存访问延迟，则可能出现“ 性能低谷 ”（基于 Guz 等人 [2009] 的图 1）。

随着 Dennard Scaling [Horowitz 等，2005] 的终结，提高能源效率已成为计算机架构研究创新的主要驱动力。一个关键的观察结果是，访问大型内存结构所消耗的能量可能与计算一样多或更多。

For example, Table 1.1 provides data on the energy for various operations in a 45 nm process technology [Han et al., 2016]. When proposing novel GPU architecture designs it is important to take energy consumption into account. To aid with this, recent GPGPU architecure simulators such as GPGPU-Sim [Bakhoda et al., 2009] incorporate energy models [Leng et al., 2013].

Table 1.1: Energy consumption of various operations for a 45 nm process technology (based on Table 1 in Han et al. [2016])

Operation	Energy [pJ]	Relative Cost
32 bit int ADD	0.1	1
32 bit float ADD	0.9	9
32 bit int MULT	3.1	31
32 bit float MULT	3.7	37
32 bit 32KB SRAM	5	50
32 bit DRAM	640	6400

1.3 A BRIEF HISTORY OF GPUS

This section briefly describes the history of graphics processing units. Computer graphics emerged in the 1960s with projects such as Ivan Sutherland's Sketchpad [Sutherland, 1963]. From its earliest days computer graphics have been integral to off-line rendering for animation in films and in parallel the development of real-time rendering for use in video games. Early video cards started with the IBM Monochrome Display Adapter (MDA) in 1981 which only supported text. Later, video cards introduced 2D and then 3D acceleration. In addition to video games 3D accelerators targeted computer-aided design. Early 3D graphics processors such as the NVIDIA GeForce 256 were relatively fixed-function. NVIDIA introduced programmability to the GPU in the form of vertex shaders [Lindholm et al., 2001] and pixel shaders in the GeForce 3 introduced in 2001. Researchers quickly learned how to implement linear algebra using these early GPUs by mapping matrix data into into textures and applying shaders [Krüger and Westermann, 2003] and academic work at mapping general-purpose computing onto GPUs such that the programmer did not need to know graphics soon followed [Buck et al., 2004]. These efforts inspired GPU manufacturers to directly support general-purpose computing in addition to graphics. The first commercial product to do so was the NVIDIA GeForce 8 Series. The GeForce 8 Series introduced several innovations including ability to write to arbitrary memory addresses from a shader and scratchpad memory to limit off-chip bandwidth, which had been lacking in earlier GPUs. The next innovation was enabling caching of read-write data with NVIDIA's Fermi architecture. Subsequent refinements include AMD's Fusion architecture which integrated CPU and GPU on the same die and dynamic parallelism that enables

例如，表 1.1 提供了 45 nm 工艺技术中各种操作的能量数据 [Han et al., 2016]。在提出新颖的 GPU 架构设计时，重要的是要考虑能耗。为了帮助实现这一点，最近的 GPGPU 架构模拟器（如 GPGPU-Sim [Bakhoda et al., 2009]）采用了能量模型 [Leng et al., 2013]。

表 1.1：45 nm 工艺技术各项操作的能耗（基于 Han et al. [2016] 中的表 1）

Operation	Energy [pJ]	Relative Cost
32 bit int ADD	0.1	1
32 bit float ADD	0.9	9
32 bit int MULT	3.1	31
32 bit float MULT	3.7	37
32 bit 32KB SRAM	5	50
32 bit DRAM	640	6400

1.3 GPU 简史

本节简要介绍图形处理单元的历史。计算机图形学出现于 20 世纪 60 年代，当时出现了 Ivan Sutherland 的 Sketchpad [Sutherland, 1963] 等项目。从早期开始，计算机图形学就成为电影动画离线渲染不可或缺的一部分，同时还用于视频游戏的实时渲染开发。早期的视频卡始于 1981 年的 IBM 单色显示适配器 (MDA)，它仅支持文本。后来，视频卡引入了 2D 和 3D 加速。除了视频游戏之外，3D 加速器还针对计算机辅助设计。早期的 3D 图形处理器（如 NVIDIA GeForce 256）功能相对固定。 NVIDIA 在 2001 年推出的 GeForce 3 中以顶点着色器 [Lindholm 等人，2001] 和像素着色器的形式为 GPU 引入了可编程性。研究人员很快学会了如何使用这些早期的 GPU 通过将矩阵数据映射到纹理中并应用着色器 [Krüger 和 Westermann，2003] 来实现线性代数，而将通用计算映射到 GPU 上以使程序员不需要了解图形的学术工作也随之而来 [Buck 等人，2004]。这些努力启发了 GPU 制造商除了支持图形之外，还直接支持通用计算。第一款这样做的商业产品是 NVIDIA GeForce 8 系列。GeForce 8 系列引入了多项创新，包括从着色器和暂存器内存写入任意内存地址的能力，以限制片外带宽，而这些是早期 GPU 所缺乏的。下一个创新是使用 NVIDIA 的 Fermi 架构实现读写数据缓存。后续改进包括 AMD 的 Fusion 架构，该架构将 CPU 和 GPU 集成在同一芯片上，并实现动态并行性，从而实现

launching of threads from the GPU itself. Most recently, NVIDIA's Volta introduces features such as Tensor Cores that are targeted specifically at machine learning acceleration.

1.4 BOOK OUTLINE

The rest of this book is organized as follows.

When designing hardware it is important to consider the software that it will support. Thus, in Chapter 2, we provide a brief summary of the programming model, code development process, and compilation flow.

In Chapter 3, we explore the architecture of individual GPU cores that support execution of thousands of threads. We incrementally build up an increasingly detailed understanding of the trade-offs involved in supporting high throughput and a flexible programming model. The chapter finishes up by summarizing recent research related to the architecture of GPU cores to help quickly bring up to speed those new to the field.

In Chapter 4, we explore the memory system including both the first-level caches found within the GPU cores, and the internal organization of the memory partitions. It is important to understand the memory system of GPUs as computations that run on GPUs are often limited by off-chip memory bandwidth. The chapter finishes up by summarizing recent research related to GPU memory system architecture.

Finally, Chapter 5 gives an overview of additional research on GPU computing architectures that does not neatly fit into Chapter 3 or 4.

从 GPU 本身启动线程。最近，NVIDIA 的 Volta 推出了 Tensor Cores 等功能，专门用于机器学习加速。

1.4 本书大纲

本书的其余部分组织如下。

在设计硬件时，考虑硬件支持的软件非常重要。因此，在第 2 章中，我们简要概述了编程模型、代码开发过程和编译流程。

在第 3 章中，我们探讨了支持执行数千个线程的单个 GPU 核心的架构。我们逐渐建立了对支持高吞吐量和灵活编程模型所涉及的权衡的越来越详细的理解。本章最后总结了与 GPU 核心架构相关的最新研究，以帮助新进入该领域的人员快速掌握最新知识。

在第 4 章中，我们将探讨内存系统，包括 GPU 核心内的一级缓存以及内存分区的内部组织。了解 GPU 的内存系统非常重要，因为在 GPU 上运行的计算通常受到片外内存带宽的限制。本章最后总结了与 GPU 内存系统架构相关的最新研究。

最后，第 5 章概述了有关 GPU 计算架构的额外研究，这些研究并不完全适合第 3 章或第 4 章。

CHAPTER 2

Programming Model

The goal of this chapter is to provide enough context about how GPUs are programmed for non-graphics computing so that those who have no prior experience with GPUs can follow the discussion in later chapters. We focus here on essential material, leaving more in-depth coverage to other references (e.g., [Kirk and Wen-Mei, 2016]). Many GPU computing benchmark suites exist which can be employed for architecture research. Learning how GPUs are programmed is relevant to a computer architect interested in GPU computing to gain a better understanding of the hardware/software interface but it becomes essential if you want to explore making changes to the hardware/software interface as part of your research. In the latter case, existing benchmarks may not exist and so will may need to be created, perhaps by modifying the source code of existing GPU computing applications. For example, research exploring the introduction of transactional memory (TM) on GPUs required this because current GPUs do not support TM (see Section 5.3).

Modern GPUs employ wide SIMD hardware to exploit the data-level parallel in GPU applications. Instead of exposing this SIMD hardware directly to the programmer, GPU computing APIs, such as CUDA and OpenCL, feature a MIMD-like programming model that allows the programmer to launch a large array of scalar threads onto the GPU. Each of these scalar threads can follow its unique execution path and may access arbitrary memory locations. At runtime, the GPU hardware executes groups of scalar threads, called *warps* (or *wavefronts* in AMD terminology), in lockstep on SIMD hardware to exploit their regularities and spatial localities. This execution model is called single-instruction, multiple-thread (SIMT) [Lindholm et al., 2008a, Nickolls and Reusch, 1993].

The rest of this chapter expands upon this discussion and is organized as follows. In Section 2.1, we explore the conceptual execution model used by recent GPU programming models and provide a concise summary of the execution model for typical GPUs released in the past decade. In Section 2.2, we explore the compilation process for GPU computing applications and take a brief look at GPU instruction set architectures.

2.1 EXECUTION MODEL

A GPU computing application starts execution on a CPU. For discrete GPUs, the CPU portion of the application will typically allocate memory for use in the computation on the GPU and then initiate transfer of input data into GPU memory, and finally launch a computational kernel on the GPU. For integrated GPUs only the last step is necessary. A computational kernel is

CHAPTER 2

编程模型

本章的目的是提供足够的背景信息，说明如何对 GPU 进行非图形计算编程，以便那些之前没有使用过 GPU 的人能够理解后面章节的讨论。我们在这里重点介绍基本材料，将更深入的介绍留给其他参考资料（例如，[Kirk and Wen-Mei, 2016]）。有许多 GPU 计算基准套件可用于架构研究。了解 GPU 的编程方式对于对 GPU 计算感兴趣的计算机架构师来说很重要，这样可以更好地理解硬件/软件接口，但如果您想在研究中探索对硬件/软件接口进行更改，这一点就变得至关重要。在后一种情况下，现有的基准可能不存在，因此可能需要创建基准，也许可以通过修改现有 GPU 计算应用程序的源代码来创建。例如，探索在 GPU 上引入事务内存 (TM) 的研究需要这样做，因为当前的 GPU 不支持 TM（参见第 5.3 节）。

现代 GPU 采用宽 SIMD 硬件来利用 GPU 应用程序中的数据级并行。GPU 计算 API（例如 CUDA 和 OpenCL）不直接向程序员公开这种 SIMD 硬件，而是采用类似 MIMD 的编程模型，允许程序员在 GPU 上启动大量标量线程。这些标量线程中的每一个都可以遵循其独特的执行路径，并可以访问任意内存位置。在运行时，GPU 硬件在 SIMD 硬件上同步执行标量线程组（在 AMD 术语中称为 *warps*（或 *wavefronts*，即））），以利用它们的规律性和空间局部性。这种执行模型称为单指令多线程 (SIMT) [Lindholm 等，2008a，Nickolls 和 Reusch，1993]。

本章的其余部分将在此讨论的基础上展开，内容安排如下。在第 2.1 节中，我们探讨了最近的 GPU 编程模型所使用的概念执行模型，并对过去十年发布的典型 GPU 的执行模型进行了简要总结。在第 2.2 节中，我们探讨了 GPU 计算应用程序的编译过程，并简要介绍了 GPU 指令集架构。

2.1 执行模型

GPU 计算应用程序开始在 CPU 上执行。对于独立 GPU，应用程序的 CPU 部分通常会分配用于 GPU 计算的内存，然后启动将输入数据传输到 GPU 内存的传输，最后在 GPU 上启动计算内核。对于集成 GPU，只需要最后一步。计算内核是

composed of (typically) thousands of threads. Each thread executes the same program, but may follow a different control flow through that program depending upon the outcome of the computation. Below we consider this flow in detail using a specific code example written in CUDA. In the following section we look at the execution model at the assembly level. Our discussion does not dwell on performance aspects of GPU programming models. However, one interesting observation made by Seo et al. [2011] in the context of OpenCL (a programming model similar to CUDA which can be compiled to many architectures) is that code carefully optimized for one architecture (e.g., a GPU) may perform poorly on another (e.g., a CPU).

Figure 2.1 provides C code for a CPU implementation of the well-known operation *single-precision scalar value A times vector value X plus vector value Y*, known as SAXPY. SAXPY is part of the well-known Basic Linear Algebra Software (BLAS) library [Lawson et al., 1979], and is useful for implementing higher level matrix operations such as Gaussian elimination [McCool et al., 2012]. Given its simplicity and utility, it is often used as an example when teaching computer architecture [Hennessy and Patterson, 2011]. Figure 2.2 provides a corresponding CUDA version of SAXPY that splits execution across a CPU and GPU.

The example in Figure 2.2 demonstrates the abstraction provided by CUDA and related programming models (e.g., OpenCL [Kaeli et al., 2015]). The code begins execution with the function `main()`. To keep the example focused on details specific to computation on the GPU we omit details of allocating and initializing the arrays `x` and `y`. Next, the function `saxpy_serial` is called. This function takes as input arguments the number of elements in the vectors `x` and `y` in parameter `n`, the scalar value in parameter `a`, and pointers to arrays used to represent the vectors `x` and `y`. The function iterates over each element of the arrays `x` and `y`. In each iteration the code on line 4 reads the values `x[i]` and `y[i]` using the loop variable `i`, multiplies `x[i]` by `a` then adds `y[i]`, and then updates `x[i]` with the result. For simplicity, we omit details of how the CPU uses the result of the function call.

Next, we consider a CUDA version of SAXPY. Similar to a traditional C or C++ program the code in Figure 2.2 begins execution by running the function `main()` on the CPU. Rather than walking through this code line by line, we will first highlight aspects specific to GPU execution.

Threads that execute on the GPU are part of a compute *kernel* specified by a function. In the CUDA version of SAXPY, shown in Figure 2.2, the CUDA keyword `__global__` on line 1 indicates the kernel function `saxpy` will run on the GPU. In the example in Figure 2.2 we have parallelized the "`for`" loop from Figure 2.1. Specifically, each iteration of the "`for`" loop on line 4 in the original CPU-only C code from Figure 2.1 is translated into an individual thread running the code on lines 3–5 in Figure 2.2.

A compute kernel typically consists of thousands of threads, each of which starts by running the same function. In our example the CPU starts computation on the GPU on line 17 using CUDA's kernel configuration syntax. The kernel configuration syntax looks much like a function call in C with some additional information specifying the number of threads contained

通常由数千个线程组成。每个线程执行相同的程序，但根据计算结果，可能遵循该程序的不同控制流。下面我们使用用 CUDA 编写的特定代码示例详细考虑此流程。在下一节中，我们将从汇编级别研究执行模型。我们的讨论并不涉及 GPU 编程模型的性能方面。然而，Seo 等人 [2011] 在 OpenCL（一种类似于 CUDA 的编程模型，可以编译为多种架构）的背景下提出了一个有趣的观察，即针对一种架构（例如 GPU）精心优化的代码可能在另一种架构（例如 CPU）上表现不佳。

图 2.1 提供了众所周知的 *single-precision scalar value A times vector value X plus vector value Y* 运算的 CPU 实现的 C 代码，即 SAXPY。SAXPY 是著名的基本线性代数软件 (BLAS) 库 [Lawson 等，1979] 的一部分，可用于实现高斯消元法等更高级的矩阵运算 [McCool 等，2012]。鉴于其简单性和实用性，它经常在教授计算机体系结构时用作示例 [Hennessy 和 Patterson，2011]。图 2.2 提供了相应的 CUDA 版本的 SAXPY，它将执行分为 CPU 和 GPU。

图 2.2 中的示例演示了 CUDA 和相关编程模型（例如 OpenCL [Kaeli et al., 2015]）提供的抽象。代码从函数 `main()` 开始执行。为了使示例专注于 GPU 上的计算特定细节，我们省略了分配和初始化数组 `x` 和 `y` 的细节。接下来，调用函数 `saxpy_serial`。此函数将参数 `n` 中向量 `x` 和 `y` 中的元素数量、参数 `a` 中的标量值以及用于表示向量 `x` 和 `y` 的数组指针作为输入参数。该函数对数组 `x` 和 `y` 的每个元素进行迭代。在每次迭代中，第 4 行的代码使用循环变量 `i` 读取值 `x[i]` 和 `y[i]`，将 `x[i]` 乘以 `a`，然后加上 `y[i]`，最后用结果更新 `x[i]`。为简单起见，我们省略了 CPU 如何使用函数调用结果的细节。

接下来，我们考虑 CUDA 版本的 SAXPY。与传统的 C 或 C++ 程序类似，图 2.2 中的代码通过在 CPU 上运行函数 `main()` 开始执行。我们不会逐行介绍此代码，而是首先重点介绍 GPU 执行的特定方面。

在 GPU 上执行的线程是函数指定的计算 *kernel* 的一部分。在图 2.2 所示的 CUDA 版本的 SAXPY 中，第 1 行上的 CUDA 关键字 `__global__` 表示核函数 `saxpy` 将在 GPU 上运行。在图 2.2 的示例中，我们已将图 2.1 中的“ `for` ”循环并行化。具体而言，图 2.1 中原始的仅 CPU C 代码中第 4 行上的“ `for` ”循环的每次迭代都转换为运行图 2.2 中第 3-5 行代码的单独线程。

计算内核通常由数千个线程组成，每个线程都通过运行相同的函数来启动。在我们的示例中，CPU 使用 CUDA 的内核配置语法在第 17 行开始在 GPU 上进行计算。内核配置语法看起来很像 C 中的函数调用，但包含一些附加信息来指定所包含的线程数

```
void saxpy_serial(int n, float a, float *x, float *y)
{
  for (int i = 0; i < n; ++i)
    y[i] = a*x[i] + y[i];
}
main() {
  float *x, *y;
  int n;
  // omitted: allocate CPU memory for x and y and initialize contents
  saxpy_serial(n, 2.0, x, y); // Invoke serial SAXPY kernel
  // omitted: use y on CPU, free memory pointed to by x and y
}
```

Figure 2.1: Traditional CPU code (based on Harris [2012]).

```
__global__ void saxpy(int n, float a, float *x, float *y)
{
   int i = blockIdx.x*blockDim.x + threadIdx.x;
   if(i<n)
     y[i] = a*x[i] + y[i];
}
int main() {
  float *h_x, *h_y;
  int n;
  // omitted: allocate CPU memory for h_x and h_y and initialize contents
  float *d_x, *d_y;
  int nblocks = (n + 255) / 256;
  cudaMalloc( &d_x, n * sizeof(float) );
  cudaMalloc( &d_y, n * sizeof(float) );
  cudaMemcpy( d_x, h_x, n * sizeof(float), cudaMemcpyHostToDevice );
  cudaMemcpy( d_y, h_y, n * sizeof(float), cudaMemcpyHostToDevice );
  saxpy<<<nblocks, 256>>>(n, 2.0, d_x, d_y);
  cudaMemcpy( h_x, d_x, n * sizeof(float), cudaMemcpyDeviceToHost );
  // omitted: use h_y on CPU, free memory pointed to by h_x, h_y, d_x, and d_y
}
```

Figure 2.2: CUDA code (based on Harris [2012]).

```
void saxpy_serial(int n, float a, float *x, float *y)
{
  for (int i = 0; i < n; ++i)
    y[i] = a*x[i] + y[i];
}
main() {
  float *x, *y;
  int n;
  // omitted: allocate CPU memory for x and y and initialize contents
  saxpy_serial(n, 2.0, x, y); // Invoke serial SAXPY kernel
  // omitted: use y on CPU, free memory pointed to by x and y
}
```

图 2.1：传统 CPU 代码（基于 Harris [201 2]）。

```
__global__ void saxpy(int n, float a, float *x, float *y)
{
   int i = blockIdx.x*blockDim.x + threadIdx.x;
   if(i<n)
     y[i] = a*x[i] + y[i];
}
int main() {
  float *h_x, *h_y;
  int n;
  // omitted: allocate CPU memory for h_x and h_y and initialize contents
  float *d_x, *d_y;
  int nblocks = (n + 255) / 256;
  cudaMalloc( &d_x, n * sizeof(float) );
  cudaMalloc( &d_y, n * sizeof(float) );
  cudaMemcpy( d_x, h_x, n * sizeof(float), cudaMemcpyHostToDevice );
  cudaMemcpy( d_y, h_y, n * sizeof(float), cudaMemcpyHostToDevice );
  saxpy<<<nblocks, 256>>>(n, 2.0, d_x, d_y);
  cudaMemcpy( h_x, d_x, n * sizeof(float), cudaMemcpyDeviceToHost );
  // omitted: use h_y on CPU, free memory pointed to by h_x, h_y, d_x, and d_y
}
```

图 2.2：CUDA 代码（基于 Harris [2012]）。

between triple angle brackets (<<<>>>). The threads that make up a compute kernel are organized into a hierarchy composed of a *grid* of *thread blocks* consisting of *warps*. In the CUDA programming model, individual threads execute instructions whose operands are scalar values (e.g., 32-bit floating-point). To improve efficiency typical GPU hardware executes groups of threads together in lock-step. These groups are called *warps* by NVIDIA and *wavefronts* by AMD. NVIDIA warps consists of 32 threads while AMD wavefronts consist of 64 threads. Warps are grouped into a larger unit called a cooperative thread array (CTA) or thread block by NVIDIA. Line 17 indicates the compute kernel should launch a single grid consisting of `nblocks` thread blocks where each thread block contains 256 threads. The arguments passed by the CPU code to the kernel configuration statement are distributed to each instance of a running thread on the GPU.

Many of today's mobile device system-on-chips integrate a CPU and a GPU into a single chip as do processors found on today's laptop and desktop computers. However, traditionally, GPUs had their own DRAM memory and this continues today for GPUs found inside data centers used for machine learning. We note that NVIDIA has introduced Unified Memory, which transparently updates GPU memory from CPU memory and CPU memory from GPU memory. In systems enabled with Unified Memory the runtime and hardware are responsible for performing copies on the programmer's behalf. Given the increasing interest in machine learning and as the goal in this book is to understand hardware, in our example we consider the general case of separate GPU and CPU memories managed by the programmer.

Following the style used in many NVIDIA CUDA examples, we use the prefix `h_` in naming pointer variables for memory allocated in CPU memory and `d_` for pointers for memory allocated in GPU memory. On line 13 the CPU calls the CUDA library function `cudaMalloc`. This function invokes the GPU driver and asks it to allocate memory on the GPU for use by the program. The call to `cudaMalloc` sets `d_x` to point to a region of GPU memory containing enough space to hold `n` 32-bit floating-point values. On line 15 the CPU calls the CUDA library function `cudaMemcpy`. This function invokes the GPU driver and asks it to copy the contents of the array in CPU memory pointed to by `h_x` to the array in GPU memory pointed to by `d_x`.

Let us finally focus on the execution of threads on the GPU. A common strategy employed in parallel programming is to assign each thread a portion of the data. To facilitate this strategy, each thread on the GPU can lookup its own identity within the grid of blocks of threads. The mechanism for doing this in CUDA employs grid, block and thread identifiers. In CUDA, grids and thread blocks have x, y, and z dimensions. While it executes, each thread has a fixed, unique combination of non-negative integer x, y, and z coordinates within the grid and thread block. Each thread block has x, y, and z coordinates within a grid. Similarly, each thread has x, y, and z coordinates within a thread block. The extents of these coordinates are set by the kernel configuration syntax (line 17). In our example, y and z dimensions are not specified and so all threads have zero values for their y and z thread block and thread coordinates. On line 3 the value of `threadIdx.x` identifies the x coordinate of the thread within its thread block and `blockIdx.x`

在三角括号 (<<<>>>) 之间。组成计算内核的线程被组织成一个层次结构，该层次结构由 *grid* 和 *thread blocks* 组成，后者由 *warps* 组成。在 CUDA 编程模型中，各个线程执行操作数为标量值（例如 32 位浮点数）的指令。为了提高效率，典型的 GPU 硬件会同步执行线程组。NVIDIA 将这些组称为 *warps*，AMD 将这些组称为 *wavefronts*。NVIDIA 的 Warp 由 32 个线程组成，而 AMD 的 Wavefront 由 64 个线程组成。Warp 被分组为一个更大的单元，NVIDIA 将其称为协作线程阵列 (CTA) 或线程块。第 17 行表示计算内核应启动由 `nblocks` 线程块组成的单个网格，其中每个线程块包含 256 个线程。CPU 代码传递给内核配置语句的参数分发给 GPU 上正在运行的线程的每个实例。

当今许多移动设备片上系统都将 CPU 和 GPU 集成到单个芯片中，就像当今笔记本电脑和台式电脑上的处理器一样。但是，传统上，GPU 有自己的 DRAM 内存，而如今用于机器学习的数据中心内的 GPU 仍然保留着这种做法。我们注意到，NVIDIA 推出了统一内存，它可以透明地从 CPU 内存更新 GPU 内存，从 GPU 内存更新 CPU 内存。在启用统一内存的系统中，运行时和硬件负责代表程序员执行复制。鉴于人们对机器学习的兴趣日益浓厚，并且本书的目标是了解硬件，在我们的示例中，我们考虑程序员管理单独的 GPU 和 CPU 内存的一般情况。

按照许多 NVIDIA CUDA 示例中使用的样式，我们使用前缀 `h_` 来命名在 CPU 内存中分配的内存的指针变量，使用 `d_` 来命名在 GPU 内存中分配的内存的指针。在第 13 行，CPU 调用 CUDA 库函数 `cudaMalloc`。此函数调用 GPU 驱动程序并要求它在 GPU 上分配内存供程序使用。对 `cudaMalloc` 的调用将 `d_x` 设置为指向 GPU 内存的某个区域，该区域包含足够的空间来保存 `n` 32 位浮点值。在第 15 行，CPU 调用 CUDA 库函数 `cudaMemcpy`。此函数调用 GPU 驱动程序并要求它将 `h_x` 指向的 CPU 内存中数组的内容复制到 `d_x` 指向的 GPU 内存中数组。

最后，让我们关注 GPU 上的线程执行。并行编程中采用的一种常见策略是为每个线程分配一部分数据。为了促进这一策略，GPU 上的每个线程都可以在线程块网格中查找自己的身份。在 CUDA 中执行此操作的机制采用网格、块和线程标识符。在 CUDA 中，网格和线程块具有 x、y 和 z 维度。在执行时，每个线程在网格和线程块内都有一个固定的、唯一的非负整数 x、y 和 z 坐标组合。每个线程块在网格内都有 x、y 和 z 坐标。同样，每个线程在线程块内都有 x、y 和 z 坐标。这些坐标的范围由内核配置语法（第 17 行）设置。在我们的示例中，y 和 z 维度未指定，因此所有线程的 y 和 z 线程块和线程坐标都为零值。在第 3 行中，`threadIdx.x` 的值标识了线程块内线程的 x 坐标，而 `blockIdx.x`

indicates the x coordinate of the thread block within its grid. The value `blockDim.x` indicates the maximum number of threads in the x-dimension. In our example, `blockDim.x` would evaluate to 256 since this is the value specified on line 17. The expression `blockIdx.x*blockDim.x + threadIdx.x` is used to compute an offset `i` for use when accessing the arrays `x` and `y`. As we will see, using index `i` we have assigned each thread a unique element of `x` and `y`.

To a significant extent, the combination of compiler and hardware enables the programmer to remain oblivious to the lock-step nature of thread execution in a warp. The compiler and hardware enable the appearance of each thread within a warp executing independently. On line 4 in Figure 2.2 we compare the value of index `i` with `n`, the size of the arrays `x` and `y`. Threads for which `i` is less than `n` execute line 5. Line 5 in Figure 2.2 performs one iteration of the original loop in Figure 2.1. After all threads in the grid are completed, the compute kernel returns control to the CPU after line 17. On line 18 the CPU invokes the GPU driver to copy the array pointed to by `d_y` from GPU memory back to CPU memory.

Some additional details of the CUDA programming model that are not illustrated by the SAXPY example, but which we will discuss later, are as follows.

Threads within a CTA can communicate with each other efficiently via a per compute core scratchpad memory. This scrathpad is called *shared memory* by NVIDIA. Each streaming multiprocessor (SM) contains a single shared memory. The space in the shared memory is divided up among all CTAs running on that SM. AMD's Graphics Core Next (GCN) architecture [AMD, 2012] includes a similar scratchpad memory that AMD calls the *local data store* (LDS). These scratchpad memories are small, ranging from 16–64 KB per SM, and exposed to programmers as different memory spaces. Programmers allocate memory into scratchpad memory using special keywords in their source code (e.g., "`__shared__`" in CUDA). The scratchpad memory acts as a software controlled cache. While GPUs also contain hardware managed caches, accessing data through such caches can lead to frequent cache misses. Applications benefit from using scratchpad memory when the programmer can identify data that is reused frequently and in a predictable manner. Unlike GPUs from NVIDIA, AMD's GCN GPUs also includes a *global data store* (GDS) scratchpad memory shared by all cores on the GPU. Scratchpad memories are used in graphics applications to pass results between different graphics shaders. For example, LDS is used for passing of parameter values betwen vertex and pixel shaders in GCN [AMD, 2012].

Threads within a CTA can synchronize efficiently using hardware-supported barrier instructions. Threads in different CTAs can communicate but do so through a global address space that is accessible to all threads. Access to this global address space is typically more expensive than access to shared memory both in terms of time and energy.

NVIDIA introduced CUDA Dynamic Parallelism (CDP) with the Kepler generation of GPUs [NVIDIA Corporation, a]. CDP is motivated by the observation that data intensive irregular applications can result in load imbalance across threads running on a GPU, leading the

表示线程块在其网格中的 *x* 坐标。值 `blockDim.x` 表示 *x* 维度中的最大线程数。在我们的示例中，`blockDim.x` 的计算结果为 256，因为这是第 17 行指定的值。表达式 `blockIdx.x*blockDim.x + threadIdx.x` 用于计算偏移量 `i`，以便在访问数组 `x` 和 `y` 时使用。我们将看到，使用索引 `i`，我们为每个线程分配了唯一的元素 `x` 和 `y`。

在很大程度上，编译器和硬件的组合使程序员能够忽略 warp 中线程执行的锁步特性。编译器和硬件使得 warp 中的每个线程看起来都是独立执行的。在图 2.2 中的第 4 行，我们将索引 `i` 的值与 `n`、数组 `x` 和 `y` 的大小进行比较。`i` 小于 `n` 的线程执行第 5 行。图 2.2 中的第 5 行执行图 2.1 中原始循环的一次迭代。在网格中的所有线程完成后，计算内核在第 17 行之后将控制权返回给 CPU。在第 18 行，CPU 调用 GPU 驱动程序将 `d_y` 指向的数组从 GPU 内存复制回 CPU 内存。

SAXPY 示例未说明的 CUDA 编程模型的一些其他细节，但我们将在后面讨论，如下所示。

CTA 内的线程可以通过每个计算核心的暂存器有效地相互通信。NVIDIA 将此暂存器称为 *shared memory*。每个流式多处理器 (SM) 包含一个共享内存。共享内存中的空间由在该 SM 上运行的所有 CTA 分配。AMD 的 Graphics Core Next (GCN) 架构 [AMD, 2012] 包含一个类似的暂存器，AMD 将其称为 *local data store* (LDS)。这些暂存器很小，每个 SM 的范围为 16-64 KB，并作为不同的内存空间暴露给程序员。程序员使用源代码中的特殊关键字（例如 CUDA 中的“ `__shared__` ”）将内存分配到暂存器中。暂存器充当软件控制的缓存。虽然 GPU 也包含硬件管理的缓存，但通过此类缓存访问数据可能会导致频繁的缓存未命中。当程序员能够识别出经常以可预测的方式重复使用的数据时，应用程序可以从使用暂存器中受益。与 NVIDIA 的 GPU 不同，AMD 的 GCN GPU 还包括一个由 GPU 上的所有核心共享的 *global data store* (GDS) 暂存器。暂存器用于图形应用程序在不同的图形着色器之间传递结果。例如，LDS 用于在 GCN 中的顶点和像素着色器之间传递参数值 [AMD，2012]。

CTA 中的线程可以使用硬件支持的屏障指令高效同步。不同 CTA 中的线程可以通信，但需要通过所有线程都可以访问的全局地址空间进行通信。访问此全局地址空间通常比访问共享内存更耗时，无论是在时间上还是在能源上。

NVIDIA 在 Kepler 一代 GPU 中引入了 CUDA 动态并行 (CDP) [NVIDIA Corporation, a]。CDP 的动机是观察到数据密集型不规则应用程序可能导致 GPU 上运行的线程之间的负载不平衡，从而导致

GPU hardware to be underutilized. In many ways, the motivation is similar to that of Dynamic Warp Formation (DWF) [Fung et al., 2007] and related approaches discussed in Section 3.4.

2.2 GPU INSTRUCTION SET ARCHITECTURES

In this section, we briefly discuss the translation of compute kernels from a high-level language such as CUDA and OpenCL to the assembly level executed by the GPU hardware and the form of current GPU instruction sets. An interesting aspect of GPU architectures that is somewhat different from CPU architectures is the way the GPU ecosystem has evolved to support instruction set evolution. For example, x86 microprocessors are backwards compatible to the Intel 8086 released in 1976. Backward compatibility means that a program compiled for a prior generation architecture will run on the next generation architecture without any changes. Thus, software compiled 40 years ago for an Intel 8086 can in theory run on any of today's x86 processors.

2.2.1 NVIDIA GPU INSTRUCTION SET ARCHITECTURES

Given the, at times, large number of vendors offering GPU hardware (each with their own hardware designs), a level of instruction set virtualization, via OpenGL Shading Language (OGSL) and Microsoft's High-Level Shading Language (HLSL), became common as early GPUs became programmable. When NVIDIA introduced CUDA in early 2007, they decided to follow a similar path and introduced their own high-level virtual instruction set architecture for GPU computing called the Parallel Thread Execution ISA, or PTX [NVI, 2017]. NVIDIA fully documents this virtual instruction set architecture with each release of CUDA to the point that it was easy for the authors of this book to develop the GPGPU-Sim simulator to support PTX [Bakhoda et al., 2009]. PTX is many ways similar to a standard reduced instruction set computer (RISC) instruction set architecture like ARM, MIPS, SPARC, or ALPHA. It also shares a similarity to the intermediate representations used within optimizing compilers. One such example is the use of a limitless set of virtual registers. Figure 2.3 illustrates a PTX version of the SAXPY program from Figure 2.2.

Before running PTX code on the GPU it is necessary to compile PTX down to the actual instruction set architecture supported by the hardware. NVIDIA calls this level SASS which is short for "Streaming ASSembler" [Cabral, 2016]. The process of converting from PTX to SASS can be accomplished either by the GPU driver or a stand-alone program called `ptxas` provided with NVIDIA's CUDA Toolkit. NVIDIA does not fully document SASS. While this makes it harder for academic researchers to develop architecture simulators that capture all compiler optimization effects, it frees NVIDIA from customer demands to provide backwards compatibility at the hardware level enabling complete redesign of the instruction set architecture from one generation to the next. Inevitably, developers who wished to understand performance at a low level started to create their own tools to disassemble SASS. The first such effort due to Wladimir Jasper van der Laan and named "decuda" [van der Lann], arrived in late 2007 for NVIDIA's GeForce 8 Series (G80), within about a year of the first release of CUDA-enabled hardware.

GPU 硬件未得到充分利用。在许多方面，其动机与动态扭曲形成 (DWF) [Fung et al., 2007] 以及第 3.4 节中讨论的相关方法类似。

2.2 GPU 指令集架构

在本节中，我们简要讨论了计算内核从 CUDA 和 OpenCL 等高级语言到 GPU 硬件执行的汇编语言的转换，以及当前 GPU 指令集的形式。GPU 架构与 CPU 架构略有不同的一个有趣方面是 GPU 生态系统的发展方式，以支持指令集的发展。例如，x86 微处理器向后兼容 1976 年发布的 Intel 8086。向后兼容性意味着为上一代架构编译的程序将在下一代架构上运行而无需任何更改。因此，40 年前为 Intel 8086 编译的软件理论上可以在当今的任何 x86 处理器上运行。

2.2.1 NVIDIA GPU 指令集架构

鉴于提供 GPU 硬件的供应商数量众多（每家都有自己的硬件设计），随着早期 GPU 变得可编程，通过 OpenGL 着色语言 (OGSL) 和 Microsoft 的高级着色语言 (HLSL) 进行一定程度的指令集虚拟化变得十分普遍。当 NVIDIA 于 2007 年初推出 CUDA 时，他们决定走一条类似的道路，并推出了自己的用于 GPU 计算的高级虚拟指令集架构，称为并行线程执行 ISA 或 PTX [NVI，2017]。NVIDIA 会在每次发布 CUDA 时完整记录此虚拟指令集架构，以至于本书的作者可以轻松开发 GPGPU-Sim 模拟器来支持 PTX [Bakhoda et al.，2009]。PTX 在许多方面类似于标准精简指令集计算机 (RISC) 指令集架构，如 ARM、MIPS、SPARC 或 ALPHA。它还与优化编译器中使用的中间表示有相似之处。其中一个例子是使用一组无限的虚拟寄存器。图 2.3 展示了图 2.2 中的 SAXPY 程序的 PTX 版本。

在 GPU 上运行 PTX 代码之前，需要将 PTX 编译为硬件支持的实际指令集架构。NVIDIA 将此级别称为 SASS，是“ Streaming ASSembler ”的缩写 [Cabral, 2016]。从 PTX 到 SASS 的转换过程可以通过 GPU 驱动程序或 NVIDIA CUDA 工具包提供的独立程序 `ptxas` 来完成。NVIDIA 没有完整记录 SASS。虽然这使得学术研究人员更难开发出捕捉所有编译器优化效果的架构模拟器，但它使 NVIDIA 摆脱了客户需求，即在硬件级别提供向后兼容性，从而能够从一代到下一代完全重新设计指令集架构。不可避免的是，希望从低层次了解性能的开发人员开始创建自己的工具来反汇编 SASS。第一个此类成果由 Wladimir Jasper van der Laan 完成，名为“ decuda ”[van der Lann]，于 2007 年底为 NVIDIA 的 GeForce 8 系列 (G80) 推出，距首次发布支持 CUDA 的硬件不到一年时间。

```
.visible .entry _Z5saxpyifPfS_(
.param .u32 _Z5saxpyifPfS__param_0,
.param .f32 _Z5saxpyifPfS__param_1,
.param .u64 _Z5saxpyifPfS__param_2,
.param .u64 _Z5saxpyifPfS__param_3
)
{
.reg .pred %p<2>;
.reg .f32 %f<5>;
.reg .b32 %r<6>;
.reg .b64 %rd<8>;


ld.param.u32 %r2, [_Z5saxpyifPfS__param_0];
ld.param.f32 %f1, [_Z5saxpyifPfS__param_1];
ld.param.u64 %rd1, [_Z5saxpyifPfS__param_2];
ld.param.u64 %rd2, [_Z5saxpyifPfS__param_3];
mov.u32 %r3, %ctaid.x;
mov.u32 %r4, %ntid.x;
mov.u32 %r5, %tid.x;
mad.lo.s32 %r1, %r4, %r3, %r5;
setp.ge.s32 %p1, %r1, %r2;
@%p1 bra BB0_2;

cvta.to.global.u64 %rd3, %rd2;
cvta.to.global.u64 %rd4, %rd1;
mul.wide.s32 %rd5, %r1, 4;
add.s64 %rd6, %rd4, %rd5;
ld.global.f32 %f2, [%rd6];
add.s64 %rd7, %rd3, %rd5;
ld.global.f32 %f3, [%rd7];
fma.rn.f32 %f4, %f2, %f1, %f3;
st.global.f32 [%rd7], %f4;

BB0_2:
ret;
}
```

Figure 2.3: PTX code corresponding to compute kernel in Figure 2.2 (compiled with CUDA 8.0).

```
.visible .entry _Z5saxpyifPfS_(
.param .u32 _Z5saxpyifPfS__param_0,
.param .f32 _Z5saxpyifPfS__param_1,
.param .u64 _Z5saxpyifPfS__param_2,
.param .u64 _Z5saxpyifPfS__param_3
)
{
.reg .pred %p<2>;
.reg .f32 %f<5>;
.reg .b32 %r<6>;
.reg .b64 %rd<8>;


ld.param.u32 %r2, [_Z5saxpyifPfS__param_0];
ld.param.f32 %f1, [_Z5saxpyifPfS__param_1];
ld.param.u64 %rd1, [_Z5saxpyifPfS__param_2];
ld.param.u64 %rd2, [_Z5saxpyifPfS__param_3];
mov.u32 %r3, %ctaid.x;
mov.u32 %r4, %ntid.x;
mov.u32 %r5, %tid.x;
mad.lo.s32 %r1, %r4, %r3, %r5;
setp.ge.s32 %p1, %r1, %r2;
@%p1 bra BB0_2;

cvta.to.global.u64 %rd3, %rd2;
cvta.to.global.u64 %rd4, %rd1;
mul.wide.s32 %rd5, %r1, 4;
add.s64 %rd6, %rd4, %rd5;
ld.global.f32 %f2, [%rd6];
add.s64 %rd7, %rd3, %rd5;
ld.global.f32 %f3, [%rd7];
fma.rn.f32 %f4, %f2, %f1, %f3;
st.global.f32 [%rd7], %f4;

BB0_2:
ret;
}
```

图 2.3：图 2.2 中计算内核对应的 PTX 代码（使用 CUDA 8.0 编译）。

The decuda project developed a sufficiently detailed understanding of the SASS instruction set that it was possible to develop an assembler. This helped in developing support for SASS up to NVIDIA's GT200 architecture in GPGPU-Sim 3.2.2 [Tor M. Aamodt et al.]. NVIDIA eventually introduced a tool, called `cuobjdump`, and started to partially document SASS. NVIDIA's SASS documentation [NVIDIA Corporation, c] currently (April 2018) provides only a list of the assembly opcode names but no details on operand formats or SASS instruction semantics. More recently, with the explosive growth in the use of GPUs for machine learning and the need for performance-optimized code, others have developed tools similar to decuda for subsequent architectures such as NVIDIA's Fermi [Yunqing] and NVIDIA's Maxwell architecture [Gray].

Figure 2.4 illustrates SASS code for our SAXPY kernel compiled for NVIDIA's Fermi architecture [NVI, 2009] and extracted with NVIDIA's `cuobjdump` (part of the CUDA Toolkit). The first column in Figure 2.4 is the address of the instruction. The second column is assembly and the third column is the encoded instruction. As noted above, NVIDIA only partially documents their hardware assembly. Comparing Figure 2.3 and Figure 2.4, one can note both similarities and differences between the virtual and hardware ISA levels. At a high level there are important similarities such as both being RISC (both used loads and stores to access memory) and both using predication [Allen et al., 1983]. More subtle differences include: (1) the PTX version has an essentially infinite set of registers available so that each definition typically uses a new register much like static single assignment [Cytron et al., 1991] whereas SASS uses a limited set of registers; and (2) the kernel parameters are passed via banked constant memory which can be accessed by non load/store instructions in SASS, whereas parameters are allocated into their own separate "parameter" address space in PTX.

Figure 2.5 illustrates SASS code for SAXPY generated by the same version of CUDA but for NVIDIA's Pascal architecture and extracted with NVIDIA's `cuobjdump`. Comparing Figure 2.5 with Figure 2.4, it is clear NVIDIA's ISA has changed significantly including in terms of instruction encodings. Figure 2.5 contains some lines for which there is no disassembled instructions (e.g., at address 0x0000 on Line 3). These are special "control instructions" introduced in the NVIDIA Kepler architecture to eliminate the need for explicit dependency checking using a scoreboard [NVIDIA Corporation, b]. Lai and Seznec [2013] explored the encoding of control instructions for the Kepler architecture. As noted by Lai and Seznec [2013], these control instructions appear to be similar to the explicit-dependence lookahead on the Tera Computer System [Alverson et al., 1990]. Gray describes extensive details of the control instruction encoding that they were able to infer for NVIDIA's Maxwell architecture. According to Gray there is one control instruction for every three regular instructions in Maxwell. This appears to also be the case for NVIDIA's Pascal architecture as illustrated in Figure 2.5. According to Gray the 64-bit control instructions on Maxwell contain three groups of 21-bits encoding the following information for each of the following three instructions: a stall count; yeild hint flag; and write, read, and wait dependency barriers. Gray also describes the use of register reuse flags on regular instructions which can also be seen in Figure 2.5 (e.g., `R0.reuse` used for the first source

decuda 项目对 SASS 指令集有了足够详细的了解，因此可以开发汇编程序。这有助于在 GPGPU-Sim 3.2.2 [Tor M. Aamodt 等] 中开发对 NVIDIA GT200 架构以下 SASS 的支持。NVIDIA 最终推出了一款名为 `cuobjdump` 的工具，并开始部分记录 SASS。NVIDIA 的 SASS 文档 [NVIDIA Corporation, c] 目前（2018 年 4 月）仅提供汇编操作码名称列表，但没有提供操作数格式或 SASS 指令语义的详细信息。最近，随着 GPU 在机器学习中的使用呈爆炸式增长以及对性能优化代码的需求，其他人为后续架构（如 NVIDIA 的 Fermi [Yunqing] 和 NVIDIA 的 Maxwell 架构 [Gray]）开发了类似于 decuda 的工具。

图 2.4 展示了针对 NVIDIA 的 Fermi 架构 [NVI, 2009] 编译并使用 NVIDIA 的 CUDA Toolkit) 的 `cuobjdump` (part 提取的 SAXPY 内核的 SASS 代码。图 2.4 中的第一列是指令的地址。第二列是汇编代码，第三列是编码指令。如上所述，NVIDIA 仅部分记录了其硬件汇编代码。比较图 2.3 和图 2.4，可以注意到虚拟和硬件 ISA 级别之间的相似之处和不同之处。在高层次上，它们具有重要的相似之处，例如两者都是 RISC（都使用加载和存储来访问内存）并且都使用谓词 [Allen et al., 1983]。更细微的差别包括：（1）PTX 版本具有一组本质上无限的可用寄存器，因此每个定义通常使用一个新寄存器，很像静态单一分配 [Cytron et al., 1991]，而 SASS 使用一组有限的寄存器；（2）内核参数通过分组常量内存传递，可通过 SASS 中的非加载/存储指令访问，而 PTX 中的参数被分配到自己单独的“参数”地址空间中。

图 2.5 展示了由同一版本的 CUDA 为 NVIDIA 的 Pascal 架构生成并使用 NVIDIA 的 `cuobjdump` 提取的 SAXPY 的 SASS 代码。将图 2.5 与图 2.4 进行比较，很明显 NVIDIA 的 ISA 发生了显著变化，包括指令编码方面。图 2.5 包含一些没有反汇编指令的行（例如，第 3 行的地址 0x0000）。这些是 NVIDIA Kepler 架构中引入的特殊“控制指令”，以消除使用记分板进行显式依赖性检查的需要 [NVIDIA Corporation，b]。Lai 和 Seznec [2013] 探索了 Kepler 架构的控制指令编码。正如 Lai 和 Seznec [2013] 所指出的，这些控制指令似乎类似于 Tera 计算机系统 [Alverson et al., 1990] 上的显式依赖前瞻。Gray 描述了他们能够推断出的 NVIDIA 的 Maxwell 架构的控制指令编码的大量细节。据 Gray 说，Maxwell 中每三条常规指令就有一条控制指令。NVIDIA 的 Pascal 架构似乎也是如此，如图 2.5 所示。据 Gray 说，Maxwell 上的 64 位控制指令包含三组 21 位，为以下三条指令中的每一条编码以下信息：停顿计数；yield 提示标志；以及写入、读取和等待依赖屏障。Gray 还描述了常规指令上寄存器重用标志的使用，这也可以在图 2.5 中看到（例如，用于第一个源的 `R0.reuse`

```
1  Address     Dissassembly                                    Encoded Instruction
2  ========    ============================================    ======================
3  /*0000*/         MOV R1, c[0x1][0x100];                     /* 0x2800440400005de4 */
4  /*0008*/         S2R R0, SR_CTAID.X;                        /* 0x2c00000094001c04 */
5  /*0010*/         S2R R2, SR_TID.X;                          /* 0x2c00000084009c04 */
6  /*0018*/         IMAD R0, R0, c[0x0][0x8], R2;              /* 0x2004400020001ca3 */
7  /*0020*/         ISETP.GE.AND P0, PT, R0, c[0x0][0x20], PT; /* 0x1b0e40008001dc23 */
8  /*0028*/     @P0 BRA.U 0x78;                                /* 0x40000001200081e7 */
9  /*0030*/    @!P0 MOV32I R5, 0x4;                            /* 0x18000000100161e2 */
10 /*0038*/    @!P0 IMAD R2.CC, R0, R5, c[0x0][0x28];          /* 0x200b8000a000a0a3 */
11 /*0040*/    @!P0 IMAD.HI.X R3, R0, R5, c[0x0][0x2c];        /* 0x208a8000b000e0e3 */
12 /*0048*/    @!P0 IMAD R4.CC, R0, R5, c[0x0][0x30];          /* 0x200b8000c00120a3 */
13 /*0050*/    @!P0 LD.E R2, [R2];                             /* 0x840000000020a085 */
14 /*0058*/    @!P0 IMAD.HI.X R5, R0, R5, c[0x0][0x34];        /* 0x208a8000d00160e3 */
15 /*0060*/    @!P0 LD.E R0, [R4];                             /* 0x8400000000402085 */
16 /*0068*/    @!P0 FFMA R0, R2, c[0x0][0x24], R0;             /* 0x3000400090202000 */
17 /*0070*/    @!P0 ST.E [R4], R0;                             /* 0x9400000000402085 */
18 /*0078*/         EXIT;                                      /* 0x8000000000001de7 */
```

Figure 2.4: Low-level SASS code corresponding to compute kernel in Figure 2.2 (compiled with CUDA 8.0 for the NVIDIA Fermi Architecture, sm_20).

operand in the Integer Short Multiply Add instruction, XMAD, on Line 7). This appears to indicate an "operand reuse cache" was added in NVIDIA GPUs starting with Maxwell (see related research in Section 3.6.1). This operand reuse cache appears to enable register values to be read multiple times for each main register file access resulting in reduced energy consumption and/or improved performance.

2.2.2 AMD GRAPHICS CORE NEXT INSTRUCTION SET ARCHITECTURE

In contrast to NVIDIA when AMD introduced their Southern Islands architecture, they released a complete hardware-level ISA specification [AMD, 2012]. Southern Islands was the first generation of AMD's Graphics Core Next (GCN) architecture. The availability of documentation for AMD's hardware ISA has helped academic researchers in developing simulators that work at a lower level [Ubal et al., 2012]. AMD's compilation flow also includes a virtual instruction set architecture, called HSAIL, as part of the Heterogeneous System Architecture (HSA).

A key difference between AMD's GCN architecture and NVIDIA GPUs (including NVIDIA's most recent Volta architecture [NVIDIA Corp., 2017]) is separate scalar and vector instructions. Figures 2.6 and 2.7 reproduce an example from AMD [2012] of high-level OpenCL (similar to CUDA) code and the equivalent machine instructions for the AMD South-

```
1  Address     Dissassembly                                    Encoded Instruction
2  ========    ============================================    ======================
3  /*0000*/         MOV R1, c[0x1][0x100];                     /* 0x2800440400005de4 */
4  /*0008*/         S2R R0, SR_CTAID.X;                        /* 0x2c00000094001c04 */
5  /*0010*/         S2R R2, SR_TID.X;                          /* 0x2c00000084009c04 */
6  /*0018*/         IMAD R0, R0, c[0x0][0x8], R2;              /* 0x2004400020001ca3 */
7  /*0020*/         ISETP.GE.AND P0, PT, R0, c[0x0][0x20], PT; /* 0x1b0e40008001dc23 */
8  /*0028*/     @P0 BRA.U 0x78;                                /* 0x40000001200081e7 */
9  /*0030*/    @!P0 MOV32I R5, 0x4;                            /* 0x18000000100161e2 */
10 /*0038*/    @!P0 IMAD R2.CC, R0, R5, c[0x0][0x28];          /* 0x200b8000a000a0a3 */
11 /*0040*/    @!P0 IMAD.HI.X R3, R0, R5, c[0x0][0x2c];        /* 0x208a8000b000e0e3 */
12 /*0048*/    @!P0 IMAD R4.CC, R0, R5, c[0x0][0x30];          /* 0x200b8000c00120a3 */
13 /*0050*/    @!P0 LD.E R2, [R2];                             /* 0x840000000020a085 */
14 /*0058*/    @!P0 IMAD.HI.X R5, R0, R5, c[0x0][0x34];        /* 0x208a8000d00160e3 */
15 /*0060*/    @!P0 LD.E R0, [R4];                             /* 0x8400000000402085 */
16 /*0068*/    @!P0 FFMA R0, R2, c[0x0][0x24], R0;             /* 0x3000400090202000 */
17 /*0070*/    @!P0 ST.E [R4], R0;                             /* 0x9400000000402085 */
18 /*0078*/         EXIT;                                      /* 0x8000000000001de7 */
```

图 2.4：与图 2.2 中的计算内核相对应的低级 SASS 代码（使用 CUDA 8.0 为 NVIDIA Fermi 架构 sm_20 编译）。

操作数在整数短乘法加法指令 XMAD 的第 7 行中）。这似乎表明从 Maxwell 开始，NVIDIA GPU 中就添加了“ 操作数重用缓存 ”（请参阅第 3.6.1 节中的相关研究）。此操作数重用缓存似乎允许在每次主寄存器文件访问时多次读取寄存器值，从而降低能耗和/或提高性能。

2.2.2 AMD 图形核心 Next 指令集架构

与 NVIDIA 不同的是，AMD 在推出其 Southern Islands 架构时发布了完整的硬件级 ISA 规范 [AMD，2012]。Southern Islands 是 AMD 的 Graphics Core Next (GCN) 架构的第一代产品。AMD 硬件 ISA 文档的可用性帮助学术研究人员开发了在较低级别工作的模拟器 [Ubal et al.，2012]。AMD 的编译流程还包括一个虚拟指令集架构，称为 HSAIL，是异构系统架构 (HSA) 的一部分。

AMD 的 GCN 架构与 NVIDIA GPU（包括 NVIDIA 最新的 Volta 架构 [NVIDIA Corp.，2017]）之间的一个关键区别是标量和矢量指令是分开的。图 2.6 和 2.7 重现了 AMD [2012] 的高级 OpenCL（类似于 CUDA）代码示例以及 AMD South-

```
1  Address     Dissassembly                                          Encoded Instruction
2  ========    ============================================          ======================
3                                                                    /* 0x001c7c00e22007f6 */
4  /*0008*/       MOV R1, c[0x0][0x20];                             /* 0x4c98078000870001 */
5  /*0010*/       S2R R0, SR_CTAID.X;                               /* 0xf0c8000002570000 */
6  /*0018*/       S2R R2, SR_TID.X;                                 /* 0xf0c8000002170002 */
7                                                                    /* 0x001fd840fec20ff1 */
8  /*0028*/       XMAD.MRG R3, R0.reuse, c[0x0] [0x8].H1, RZ;       /* 0x4f107f8000270003 */
9  /*0030*/       XMAD R2, R0.reuse, c[0x0] [0x8], R2;              /* 0x4e00010000270002 */
10 /*0038*/       XMAD.PSL.CBCC R0, R0.H1, R3.H1, R2;               /* 0x5b30011800370000 */
11                                                                   /* 0x081fc400ffa007ed */
12 /*0048*/       ISETP.GE.AND P0, PT, R0, c[0x0][0x140], PT;       /* 0x4b6d038005070007 */
13 /*0050*/    @P0 EXIT;                                            /* 0xe30000000000000f */
14 /*0058*/       SHL R2, R0.reuse, 0x2;                            /* 0x3848000000270002 */
15                                                                   /* 0x081fc440fec007f5 */
16 /*0068*/       SHR R0, R0, 0x1e;                                 /* 0x3829000001e70000 */
17 /*0070*/       IADD R4.CC, R2.reuse, c[0x0][0x148];              /* 0x4c10800005270204 */
18 /*0078*/       IADD.X R5, R0.reuse, c[0x0][0x14c];               /* 0x4c10080005370005 */
19                                                                   /* 0x0001c800fe0007f6 */
20 /*0088*/       IADD R2.CC, R2, c[0x0][0x150];                    /* 0x4c10800005470202 */
21 /*0090*/       IADD.X R3, R0, c[0x0][0x154];                     /* 0x4c10080005570003 */
22 /*0098*/       LDG.E R0, [R4];        }                          /* 0xeed4200000070400 */
23                                                                   /* 0x0007c408fc400172 */
24 /*00a8*/       LDG.E R6, [R2];                                   /* 0xeed4200000070206 */
25 /*00b0*/       FFMA R0, R0, c[0x0][0x144], R6;                   /* 0x4980030005170000 */
26 /*00b8*/       STG.E [R2], R0;                                   /* 0xeedc200000070200 */
27                                                                   /* 0x001f8000ffe007ff */
28 /*00c8*/       EXIT;                                             /* 0xe30000000007000f */
29 /*00d0*/       BRA 0xd0;                                         /* 0xe2400fffff87000f */
30 /*00d8*/       NOP;                                              /* 0x50b0000000070f00 */
31                                                                   /* 0x001f8000fc0007e0 */
32 /*00e8*/       NOP;                                              /* 0x50b0000000070f00 */
33 /*00f0*/       NOP;                                              /* 0x50b0000000070f00 */
34 /*00f8*/       NOP;                                              /* 0x50b0000000070f00 */
```

Figure 2.5: Low-level SASS code corresponding to compute kernel in Figure 2.2 (compiled with CUDA 8.0 for the NVIDIA Pascal Architecture, sm_60).

```
Address     Dissassembly                                        Encoded Instruction
========    ================================================    ======================
                                                                /* 0x001c7c00e22007f6 */
/*0008*/       MOV R1, c[0x0][0x20];                            /* 0x4c98078000870001 */
/*0010*/       S2R R0, SR_CTAID.X;                              /* 0xf0c8000002570000 */
/*0018*/       S2R R2, SR_TID.X;                                /* 0xf0c8000002170002 */
                                                                /* 0x001fd840fec20ff1 */
/*0028*/       XMAD.MRG R3, R0.reuse, c[0x0] [0x8].H1, RZ;      /* 0x4f107f8000270003 */
/*0030*/       XMAD R2, R0.reuse, c[0x0] [0x8], R2;             /* 0x4e00010000270002 */
/*0038*/       XMAD.PSL.CBCC R0, R0.H1, R3.H1, R2;              /* 0x5b30011800370000 */
                                                                /* 0x081fc400ffa007ed */
/*0048*/       ISETP.GE.AND P0, PT, R0, c[0x0][0x140], PT;      /* 0x4b6d038005070007 */
/*0050*/    @P0 EXIT;                                           /* 0xe30000000000000f */
/*0058*/       SHL R2, R0.reuse, 0x2;                           /* 0x3848000000270002 */
                                                                /* 0x081fc440fec007f5 */
/*0068*/       SHR R0, R0, 0x1e;                                /* 0x3829000001e70000 */
/*0070*/       IADD R4.CC, R2.reuse, c[0x0][0x148];             /* 0x4c10800005270204 */
/*0078*/       IADD.X R5, R0.reuse, c[0x0][0x14c];              /* 0x4c10080005370005 */
                                                                /* 0x0001c800fe0007f6 */
/*0088*/       IADD R2.CC, R2, c[0x0][0x150];                   /* 0x4c10800005470202 */
/*0090*/       IADD.X R3, R0, c[0x0][0x154];                    /* 0x4c10080005570003 */
/*0098*/       LDG.E R0, [R4];        }                         /* 0xeed4200000070400 */
                                                                /* 0x0007c408fc400172 */
/*00a8*/       LDG.E R6, [R2];                                  /* 0xeed4200000070206 */
/*00b0*/       FFMA R0, R0, c[0x0][0x144], R6;                  /* 0x4980030005170000 */
/*00b8*/       STG.E [R2], R0;                                  /* 0xeedc200000070200 */
                                                                /* 0x001f8000ffe007ff */
/*00c8*/       EXIT;                                            /* 0xe30000000007000f */
/*00d0*/       BRA 0xd0;                                        /* 0xe2400fffff87000f */
/*00d8*/       NOP;                                             /* 0x50b0000000070f00 */
                                                                /* 0x001f8000fc0007e0 */
/*00e8*/       NOP;                                             /* 0x50b0000000070f00 */
/*00f0*/       NOP;                                             /* 0x50b0000000070f00 */
/*00f8*/       NOP;                                             /* 0x50b0000000070f00 */
```

图 2.5：与图 2.2 中的计算内核相对应的低级 SASS 代码（使用 CUDA 8.0 为 NVIDIA Pascal 架构 sm_60 编译）。

ern Islands architecture. In Figure 2.7, scalar instructions are prefaced with `s_` and vector instructions are prefaced with `v_`. In the AMD GCN architecture, each compute unit (e.g., SIMT core) contains a scalar unit coupled with four vector units. Vector instructions execute on the vector units and compute different 32-bit values for each individual thread in a wavefront. In contrast, scalar instructions execute on the scalar units compute a single 32-bit value shared by all threads in a wavefront. In the example shown in Figure 2.7 the scalar instructions are related to control flow handling. In particular, `exec` is a special register used to predicate execution of individual vector lanes for SIMT execution. The use of masking for control flow handling on GPUs is described in more detail in Section 3.1.1. Another potential benefit of the scalar unit in the GCN architecture is that frequently certain portions of a computation in a SIMT program will compute the same result independent of thread ID (see Section 3.5).

```
float fn0(float a,float b)
{
    if(a>b)
        return(a * a - b);
    else
        return(b * b - a);
}
```

Figure 2.6: OpenCL code (based on Figure 2.2 in AMD [2012]).

```
// Registers r0 contains "a", r1 contains "b"
// Value is returned in r2
    v_cmp_gt_f32 r0, r1 // a>b
    s_mov_b64 s0, exec  // Save current exec mask
    s_and_b64 exec, vcc, exec // Do "if"
    s_cbranch_vccz label0 // Branch if all lanes fail
    v_mul_f32 r2, r0, r0 // result = a * a
    v_sub_f32 r2, r2, r1 // result = result - b
label0:
    s_not_b64 exec, exec // Do "else"
    s_and_b64 exec, s0, exec // Do "else"
    s_cbranch_execz label1 // Branch if all lanes fail
    v_mul_f32 r2, r1, r1 // result = b * b
    v_sub_f32 r2, r2, r0 // result = result - a
label1:
    s_mov_b64 exec, s0   // Restore exec mask
```

Figure 2.7: Southern Islands (graphics core next) microcode (based on Figure 2.2 in AMD [2012]).

ern Islands 架构。在图 2.7 中，标量指令以 s_ 开头，矢量指令以 v_ 开头。在 AMD GCN 架构中，每个计算单元（例如，SIMT 核心）包含一个标量单元和四个矢量单元。矢量指令在矢量单元上执行，并为波前中的每个单独线程计算不同的 32 位值。相反，在标量单元上执行的标量指令计算波前中所有线程共享的单个 32 位值。在图 2.7 所示的示例中，标量指令与控制流处理有关。具体而言，exec 是一个特殊寄存器，用于预测 SIMT 执行的各个矢量通道的执行。第 3.1.1 节更详细地描述了在 GPU 上使用掩码进行控制流处理的更多细节。 GCN 架构中标量单元的另一个潜在好处是，SIMT 程序中计算的某些部分通常将计算出与线程 ID 无关的相同结果（参见第 3.5 节）。

```
float fn0(float a,float b)
{
    if(a>b)
        return(a * a - b);
    else
        return(b * b - a);
}
```

图 2.6：OpenCL 代码（基于 AMD [2012] 中的图 2.2）。

```
// Registers r0 contains "a", r1 contains "b"
// Value is returned in r2
    v_cmp_gt_f32 r0, r1 // a>b
    s_mov_b64 s0, exec  // Save current exec mask
    s_and_b64 exec, vcc, exec // Do "if"
    s_cbranch_vccz label0 // Branch if all lanes fail
    v_mul_f32 r2, r0, r0 // result = a * a
    v_sub_f32 r2, r2, r1 // result = result - b
label0:
    s_not_b64 exec, exec // Do "else"
    s_and_b64 exec, s0, exec // Do "else"
    s_cbranch_execz label1 // Branch if all lanes fail
    v_mul_f32 r2, r1, r1 // result = b * b
    v_sub_f32 r2, r2, r0 // result = result - a
label1:
    s_mov_b64 exec, s0   // Restore exec mask
```

图 2.7：Southern Islands（下一个图形核心）微码（基于 AMD [2012] 中的图 2.2）。

AMD's GCN hardware instruction set manual [AMD, 2012] provides many interesting insights into AMD GPU hardware. For example, to enable data dependency resolution for long latency operations AMD's GCN architecture includes `S_WAITCNT` instructions. For each wavefront there are three counters: vector memory count, local/global data store count, and register export count. Each of these indicate the number of outstanding operations of a given type. The compiler or programmer inserts `S_WAITCNT` instructions to have the wavefront wait until the number of outstanding operations decreases below a specified threshold.

AMD 的 GCN 硬件指令集手册 [AMD，2012] 提供了许多有关 AMD GPU 硬件的有趣见解。例如，为了实现长延迟操作的数据依赖性解析，AMD 的 GCN 架构包含 `S_WAITCNT` 指令。每个波前都有三个计数器：矢量内存计数、本地/全局数据存储计数和寄存器导出计数。每个计数器都指示给定类型的未完成操作的数量。编译器或程序员插入 `S_WAITCNT` 指令，让波前等待，直到未完成操作的数量降至指定阈值以下。

CHAPTER 3

The SIMT Core: Instruction and Register Data Flow

In this and the following chapter we will examine the architecture and microarchitecture of modern GPUs. We divide our discussion of GPU architecture into two parts: (1) examining the SIMT cores that implement computation in this chapter and then (2) looking at the memory system in the next chapter.

In their traditional graphics-rendering role, GPUs access data sets such as detailed texture maps that are far too large to be fully cached on-chip. To enable high-performance programmability, which is desirable in graphics both to ease verification costs as the number of graphics modes increase and to enable games developers to more easily differentiate their products [Lindholm et al., 2001], it is necessary to employ an architecture that can sustain large off-chip bandwidths. Thus, today's GPUs execute tens of thousands of threads concurrently. While the amount of on-chip memory storage per thread is small, caches can still be effective in reducing a sizable number of off-chip memory accesses. For example, in graphics workloads, there is significant spatial locality between adjacent pixel operations that can be captured by on-chip caches.

Figure 3.1 illustrates the microarchitecture of the GPU pipeline discussed in this chapter. This figure illustrates the internal organization of a single SIMT-core shown in Figure 1.2. The pipeline can be divided into a SIMT front-end and a SIMD back-end. The pipeline consists of three scheduling "loops" acting together in a single pipeline: an instruction fetch loop, an instruction issue loop, and a register access scheduling loop. The instruction fetch loop includes the blocks labeled Fetch, I-Cache, Decode, and I-Buffer. The instruction issue loop includes the blocks labeled I-Buffer, Scoreboard, Issue, and SIMT Stack. The register access scheduling loop includes the blocks labeled Operand Collector, ALU, and Memory. In the rest of this chapter we help you build up a full understanding of the individual blocks in this picture by considering key aspects of the architecture that depend on each of these loops.

As there are many details involved in fully understanding this organization, we divide our discussion up in parts. We order these with the objective of developing an increasingly detailed view of the core microarchitecture. We start with a high-level view of the overall GPU pipeline and then fill in details. We call these increasingly accurate descriptions "approximations" to acknowledge that some details are omitted even in our most detailed descriptions. As the central organizing principle of today's GPUs is multithreading we organize these "approximations"

CHAPTER 3

SIMT 核心：指令和寄存器数据流

在本章和下一章中，我们将研究现代 GPU 的架构和微架构。我们将 GPU 架构的讨论分为两部分：(1) 本章中研究实现计算的 SIMT 核心，然后 (2) 下一章中研究内存系统。

在传统的图形渲染角色中，GPU 会访问详细纹理贴图等数据集，而这些数据集太大以致无法完全缓存在片上。为了实现高性能可编程性（这在图形处理中很有必要），有必要采用一种能够维持大量片外带宽的架构，这在图形处理中很有必要，因为随着图形模式数量的增加，这既可以降低验证成本，也可以使游戏开发人员更轻松地区分他们的产品 [Lindholm et al., 2001]。因此，当今的 GPU 会同时执行数万个线程。虽然每个线程的片上内存存储空间很小，但缓存仍可有效减少大量片外内存访问。例如，在图形工作负载中，相邻像素操作之间存在明显的空间局部性，这些局部性可以被片上缓存捕获。

图 3.1 说明了本章讨论的 GPU 流水线的微架构。该图说明了图 1.2 中所示的单个 SIMT 核心的内部组织。流水线可分为 SIMT 前端和 SIMD 后端。流水线由三个调度“循环”组成，它们在单个流水线中共同作用：指令获取循环、指令发出循环和寄存器访问调度循环。指令获取循环包括标记为 Fetch、I-Cache、Decode 和 I-Buffer 的块。指令发出循环包括标记为 I-Buffer、Scoreboard、Issue 和 SIMT Stack 的块。寄存器访问调度循环包括标记为 Operand Collector、ALU 和 Memory 的块。在本章的其余部分中，我们将通过考虑依赖于每个循环的架构的关键方面来帮助您全面了解此图中的各个块。

由于要完全理解这个组织涉及许多细节，我们将讨论分为几个部分。我们按顺序进行讨论，目的是开发越来越详细的核心微体系结构视图。我们从整个 GPU 管道的高级视图开始，然后填写详细信息。我们将这些越来越准确的描述称为“近似值”，以承认即使在我们最详细的描述中也省略了一些细节。由于当今 GPU 的中心组织原则是多线程，我们将这些“近似值”组织起来

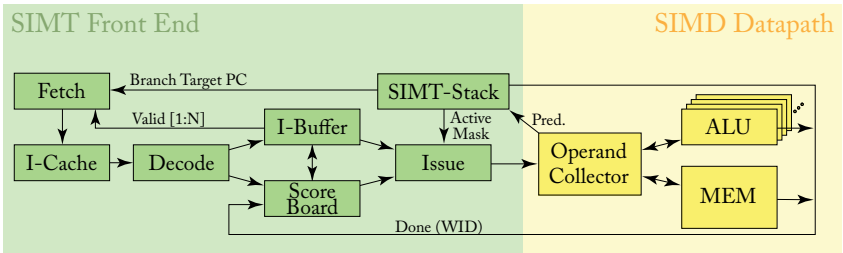


Figure 3.1: Microarchitecture of a generic GPGPU core.

around the three scheduling loops described above. We have found it convenient to organize this chapter by considering three increasingly accurate "approximation loops" that progressively take into account the details of these scheduler loops.

3.1 ONE-LOOP APPROXIMATION

We start by considering a GPU with a single scheduler. This simplified look at the hardware is not unlike what one might expect the hardware to do if they only read the description of the hardware found in the CUDA programming manual.

To increase efficiency, threads are organized into groups called "warps" by NVIDIA and "wavefronts" by AMD. Thus, the unit of scheduling is a warp. In each cycle, the hardware selects a warp for scheduling. In the one loop approximation the warp's program counter is used to access an instruction memory to find the next instruction to execute for the warp. After fetching an instruction, the instruction is decoded and source operand registers are fetched from the register file. In parallel with fetching source operands from the register file, the SIMT execution mask values are determined. The following sub-section describes how the SIMT execution mask values are determined and contrasts them with predication, which is also employed in modern GPUs.

After the execution masks and source registers are available, execution proceeds in a single-instruction, multiple-data manner. Each thread executes on the function unit associated with a lane provided the SIMT execution mask is set. As in modern CPU designs, the function units are typically heterogeneous meaning a given function unit supports only a subset of instructions. For example, NVIDIA GPUs contain a *special function unit* (SFU), *load/store unit*, *floating-point function unit*, *integer function unit*, and, as of Volta, a *Tensor Core*.

All function units nominally contain as many lanes as there are threads within a warp. However, several GPUs have used a different implementation in which a single warp or wavefront is executed over several clock cycles. This is achieved by clocking the function units at a higher frequency, which can achieve higher performance per unit area at the expense of increased energy consumption. One way to achieve higher clock frequencies for the function units is to pipeline their execution or increase their pipeline depth.

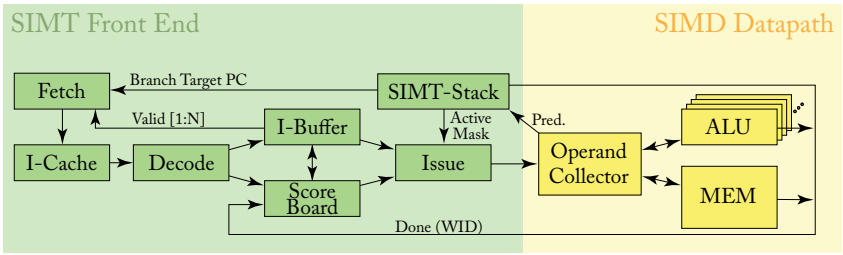


图 3.1：通用 GPGPU 核心的微架构。

围绕上述三个调度循环。我们发现，通过考虑三个越来越精确的“ 近似循环 ”来组织本章会很方便，这些循环会逐步考虑到这些调度程序循环的细节。

3.1 单圈近似

我们首先考虑具有单个调度程序的 GPU。这种对硬件的简化看法与人们仅阅读 CUDA 编程手册中的硬件描述时对硬件的预期并无不同。

为了提高效率，NVIDIA 将线程组织成“ warp ”组，AMD 将其称为“ wavefront ”。因此，调度的单位是 warp。在每个周期中，硬件选择一个 warp 进行调度。在一次循环近似中，warp 的程序计数器用于访问指令内存，以查找 warp 要执行的下一条指令。获取指令后，将解码该指令并从寄存器文件中获取源操作数寄存器。在从寄存器文件中获取源操作数的同时，确定 SIMT 执行掩码值。以下小节介绍了如何确定 SIMT 执行掩码值，并将它们与现代 GPU 中也采用的预测进行了对比。

在执行掩码和源寄存器可用后，执行以单指令、多数据的方式进行。如果设置了 SIMT 执行掩码，则每个线程都在与通道关联的功能单元上执行。与现代 CPU 设计一样，功能单元通常是异构的，这意味着给定的功能单元仅支持指令子集。例如，NVIDIA GPU 包含 *special function unit* (SFU)、*load/store unit*、*floating-point function unit*、*integer function unit*，以及 Volta 中的 *Tensor Core*。

所有功能单元名义上都包含与 warp 中的线程数一样多的通道。但是，一些 GPU 使用了不同的实现方式，即在几个时钟周期内执行单个 warp 或波前。这是通过以更高的频率对功能单元进行计时来实现的，这可以实现更高的单位面积性能，但代价是增加能耗。实现功能单元更高时钟频率的一种方法是将其执行流水线化或增加其流水线深度。

3.1.1 SIMT EXECUTION MASKING

A key feature of modern GPUs is the SIMT execution model, which from the standpoint of functionality (although not performance) presents the programmer with the abstraction that individual threads execute completely independently. This programming model can potentially be achieved via predication alone. However, in current GPUs it is achieved via a combination of traditional predication along with a stack of predicate masks that we shall refer to as the *SIMT stack*.

The SIMT stack helps efficiently handle two key issues that occur when all threads can execute independently. The first is nested control flow. In nested control flow one branch is control dependent upon another. The second issue is skipping computation entirely while all threads in a warp avoid a control flow path. For complex control flow this can represent a significant savings. Traditionally, CPUs supporting predication have handled nested control flow by using multiple predicate registers and supporting across lane predicate tests has been proposed in the literature.

The SIMT stack employed by GPUs can handle both nested control flow and skipped computation. There are several implementations described in patents and instruction set manuals. In these descriptions the SIMT stack is at least partly managed by special instructions dedicated to this purpose. Instead, we will describe a slightly simplified version introduced in an academic work that assumes the hardware is responsible for managing the SIMT stack.

To describe the SIMT stack we use an example. Figure 3.2 illustrates CUDA C code that contains two branches nested within a do-while loop and Figure 3.3 illustrates the corresponding PTX assembly. Figure 3.4, which reproduces Figure 5 in Fung et al. [Fung et al., 2007], illustrates how this code interacts with the SIMT stack assuming a GPU that has four threads per warp.

Figure 3.4a illustrates a control flow graph (CFG) corresponding to the code in Figures 3.2 and 3.3. As indicated by the label "`A/1111`" inside the top node of the CFG, initially all four threads in the warp are executing the code in Basic Block A which corresponds to the code on lines 2 through 6 in Figure 3.2 and lines 1 through 6 in Figure 3.3. These four threads follow different (divergent) control flow after executing the branch on line 6 in Figure 3.3, which corresponds to the "if" statement on line 6 in Figure 3.2. Specifically, as indicated by the label "`B/1110`" in Figure 3.4a the first three threads fall through to Basic Block B. These three threads branch to line 7 in Figure 3.3 (line 7 in Figure 3.2). As indicated by the label "`F/0001`" in Figure 3.4a, after executing the branch the fourth thread jumps to Basic Block F, which corresponds to line 14 in Figure 3.3 (line 14 in Figure 3.2).

Similarly, when the three threads executing in Basic Block B reach the branch on line 9 in Figure 3.3 the first thread diverges to Basic Block C while the second and third thread diverges to Basic Block D. Then, all three threads reach Basic Block E and execute together as indicated by the label "`E/1110`" in Figure 3.4a. At Basic Block G all four threads execute together.

How does GPU hardware enable threads within a warp to follow different paths through the code while employing a SIMD datapath that allows only one instruction to execute per cycle?

3.1.1 SIMT 执行掩码

现代 GPU 的一个关键特性是 SIMT 执行模型，从功能（而非性能）的角度来看，它为程序员提供了各个线程完全独立执行的抽象。这种编程模型可能仅通过预测就能实现。然而，在当前的 GPU 中，它是通过传统预测与我们将称为 *SIMT stack* 的谓词掩码堆栈的组合来实现的。

SIMT 堆栈有助于有效处理所有线程独立执行时发生的两个关键问题。第一个是嵌套控制流。在嵌套控制流中，一个分支的控制依赖于另一个分支。第二个问题是完全跳过计算，而 warp 中的所有线程都避开控制流路径。对于复杂的控制流，这可以节省大量成本。传统上，支持谓词的 CPU 通过使用多个谓词寄存器来处理嵌套控制流，文献中提出了支持跨通道谓词测试。

GPU 使用的 SIMT 堆栈可以处理嵌套控制流和跳过计算。专利和指令集手册中描述了几种实现。在这些描述中，SIMT 堆栈至少部分由专用于此目的的特殊指令管理。相反，我们将描述学术著作中引入的稍微简化的版本，该版本假设硬件负责管理 SIMT 堆栈。

为了描述 SIMT 堆栈，我们使用了一个示例。图 3.2 说明了包含嵌套在 do-while 循环中的两个分支的 CUDA C 代码，图 3.3 说明了相应的 PTX 程序集。图 3.4 重现了 Fung 等人的图 5。[Fung et al., 2007]，说明了此代码如何与 SIMT 堆栈交互，假设 GPU 每个 warp 有四个线程。

图 3.4a 给出了与图 3.2 和 3.3 中的代码相对应的控制流图 (CFG)。如 CFG 顶部节点内的标签“ `A/1111` ”所示，最初 warp 中的所有四个线程都在执行基本块 A 中的代码，这对应于图 3.2 中第 2 至第 6 行以及图 3.3 中第 1 至第 6 行的代码。在执行图 3.3 中第 6 行的分支之后，这四个线程遵循不同的（发散的）控制流，这对应于图 3.2 中第 6 行的“ if ”语句。具体而言，如图 3.4a 中的标签“ `B/1110` ”所示，前三个线程进入基本块 B。这三个线程分支到图 3.3 中的第 7 行（图 3.2 中的第 7 行）。如图 3.4a 中的标签“ `F/0001` ”所示，执行分支后，第四个线程跳转到基本块 F，对应图 3.3 中的第 14 行（图 3.2 中的第 14 行）。

类似地，当在基本块 B 中执行的三个线程到达图 3.3 中第 9 行的分支时，第一个线程会分叉到基本块 C，而第二个和第三个线程会分叉到基本块 D。然后，所有三个线程都会到达基本块 E 并一起执行，如图 3.4a 中的标签“ `E/1110` ”所示。在基本块 G 中，所有四个线程都会一起执行。

GPU 硬件如何使 warp 中的线程能够遵循代码中的不同路径，同时采用每个周期只允许执行一条指令的 SIMD 数据路径？

```
do {
  t1 = tid*N;       // A
  t2 = t1 + i;
  t3 = data1[t2];
  t4 = 0;
  if( t3 != t4 ) {
    t5 = data2[t2]; // B
    if( t5 != t4 ) {
      x += 1;       // C
    } else {
      y += 2;       // D
    }
  } else {
    z += 3;         // F
  }
  i++;              // G
} while( i < N );
```

Figure 3.2: Example CUDA C source code for illustrating SIMT stack operation.

```
A:    mul.lo.u32      t1, tid, N;
      add.u32         t2, t1, i;
      ld.global.u32   t3, [t2];
      mov.u32         t4, 0;
      setp.eq.u32     p1, t3, t4;
@p1   bra             F;
B:    ld.global.u32   t5, [t2];
      setp.eq.u32     p2, t5, t4;
@p2   bra             D;
C:    add.u32         x, x, 1;
      bra             E;
D:    add.u32         y, y, 2;
E:    bra             G;
F:    add.u32         z, z, 3;
G:    add.u32         i, i, 1;
      setp.le.u32     p3, i, N;
@p3   bra             A;
```

Figure 3.3: Example PTX assembly code for illustrating SIMT stack operation.

```
do {
  t1 = tid*N;       // A
  t2 = t1 + i;
  t3 = data1[t2];
  t4 = 0;
  if( t3 != t4 ) {
    t5 = data2[t2]; // B
    if( t5 != t4 ) {
      x += 1;       // C
    } else {
      y += 2;       // D
    }
  } else {
    z += 3;         // F
  }
  i++;              // G
} while( i < N );
```

图 3.2：用于说明 SIMT 堆栈操作的示例 CUDA C 源代码。

```
A:    mul.lo.u32      t1, tid, N;
      add.u32         t2, t1, i;
      ld.global.u32   t3, [t2];
      mov.u32         t4, 0;
      setp.eq.u32     p1, t3, t4;
@p1   bra             F;
B:    ld.global.u32   t5, [t2];
      setp.eq.u32     p2, t5, t4;
@p2   bra             D;
C:    add.u32         x, x, 1;
      bra             E;
D:    add.u32         y, y, 2;
E:    bra             G;
F:    add.u32         z, z, 3;
G:    add.u32         i, i, 1;
      setp.le.u32     p3, i, N;
@p3   bra             A;
```

图 3.3：用于说明 SIMT 堆栈操作的示例 PTX 汇编代码。

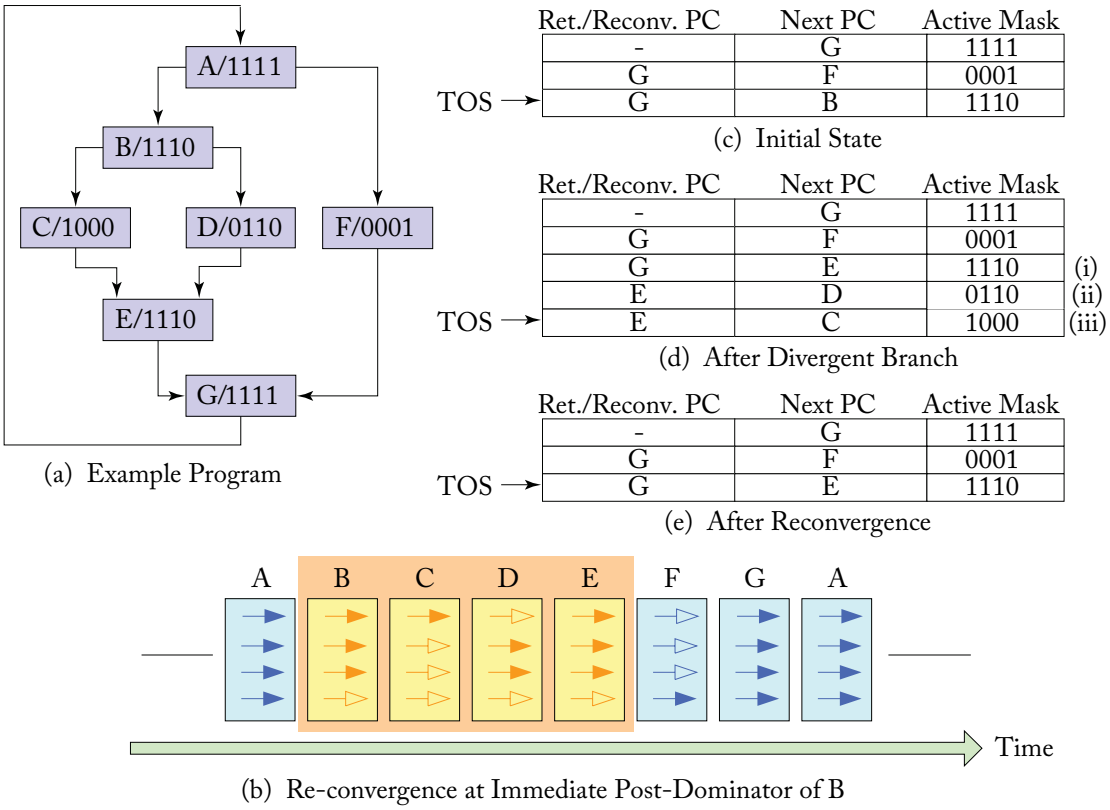


Figure 3.4: Example of SIMT stack operation (based on Figure 5 from Fung et al. [2007]).

The approach used in current GPUs is to serialize execution of threads following different paths within a given warp. This is illustrated in Figure 3.4b where arrows represent threads. A filled-in arrow indicates the thread is executing the code in the corresponding basic block (indicated by the letter on top each rectangle). An arrow with a hollow head indicates the thread is masked off. Time advances to the right in the diagram, as indicated by the arrow at the bottom. Initially, each thread is executing in Basic Block B. Then, after the branch the first three threads execute the code in Basic Block B. Note that at this point in time thread four is masked off. To maintain SIMD execution the fourth thread executes the alternate code path through Basic Block F at a different time (several cycles later in this example).

To achieve this serialization of divergent code paths one approach is to use a stack like that is illustrated in Figure 3.4c–e. Each entry on this stack contains three entries: a reconvergence program counter (RPC), the address of the next instruction to execute (Next PC), and an active mask.

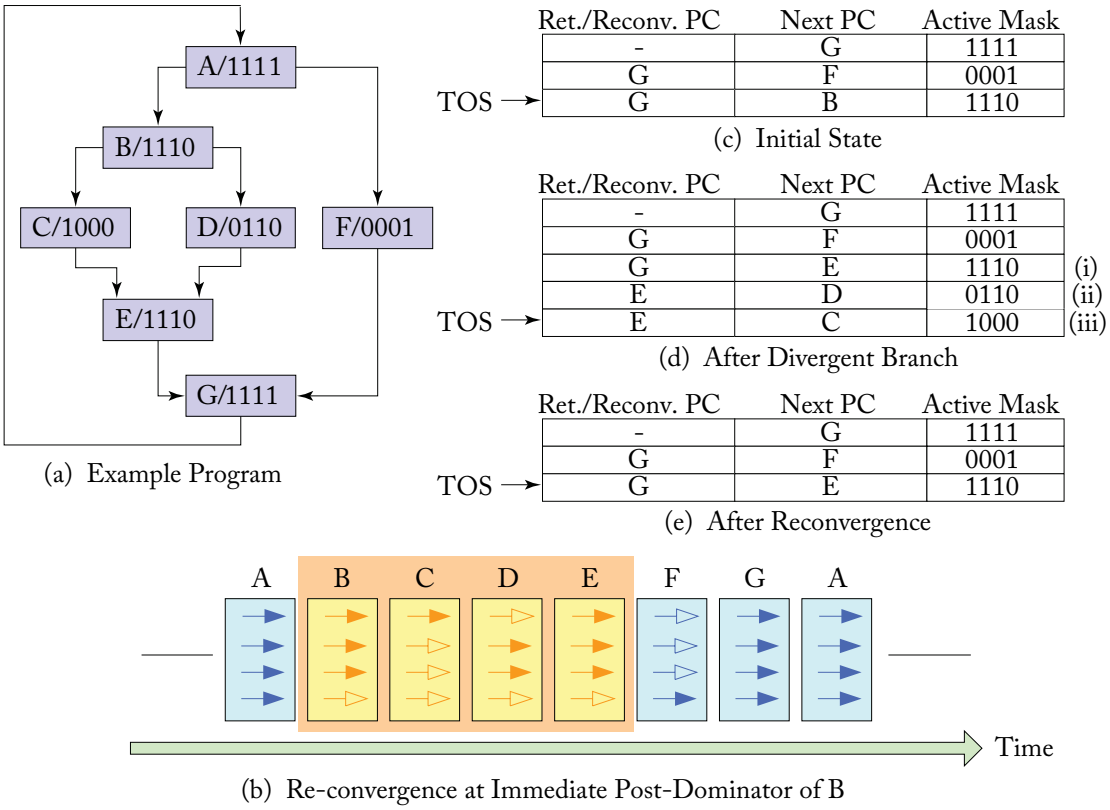


图 3.4：SIMT 堆栈操作示例（基于 Fung 等人 [2007] 的图 5）。

当前 GPU 中使用的方法是序列化在给定 warp 内遵循不同路径的线程的执行。这在图 3.4b 中进行了说明，其中箭头代表线程。实心箭头表示线程正在执行相应基本块中的代码（由每个矩形顶部的字母表示）。空心箭头表示线程已被屏蔽。时间在图中向右推进，如下方箭头所示。最初，每个线程都在基本块 B 中执行。然后，在分支之后，前三个线程执行基本块 B 中的代码。请注意，此时线程四已被屏蔽。为了保持 SIMD 执行，第四个线程在不同时间（在此示例中为几个周期后）通过基本块 F 执行备用代码路径。

为了实现这种不同代码路径的序列化，一种方法是使用如图 3.4c-e 所示的堆栈。此堆栈上的每个条目包含三个条目：一个重新收敛程序计数器 (RPC)、下一个要执行的指令的地址 (Next PC) 和一个活动掩码。

Figure 3.4c illustrates the state of the stack immediately after the warp has executed the branch on line 6 in Figure 3.3. Since three threads branch to Basic Block B and one thread branches to Basic Block F, two new entries have been added to the top of the stack (TOS). The next instruction that the warp executes is determined using the Next PC value in the top of stack (TOS) entry. In Figure 3.4c, this Next PC value is B, which represents the address for the first instruction in Basic Block B. The corresponding Active Mask entry, "1110", indicates only the first three threads in the warp should execute this instruction.

The first three threads in the warp continue executing instructions from Basic Block B until they reach the branch on line 9 in Figure 3.3. After executing this branch they diverge, as noted earlier. This branch divergence causes three changes to the stack. First, the Next PC entry of the TOS entry prior to executing the branch, labeled (i) in Figure 3.4d, is modified to the *reconvergence point* of the branch, which is the address of the first instruction in Basic Block E. Then, two entries, labeled (ii) and (iii) in Figure 3.4d, are added, one for each of the paths followed by threads in the warp after executing the branch.

A reconvergence point is a location in the program where threads that diverge can be forced to continue executing in lock-step. The nearest reconvergence point is generally preferred. The earliest point in a given program execution where it can be guaranteed at compile time that threads which diverge can again execute in lock-step is the immediate post-dominator of the branch that caused the branch divergence. At runtime it is sometimes possible to reconverge at an earlier point in the program [Coon and Lindholm, 2008, Diamos et al., 2011, Fung and Aamodt, 2011].

An interesting question is "what order should be used to add the entries to the stack following a divergent branch?" To reduce the maximum depth of the reconvergence stack to be logarithmic in the number of threads in a warp it is best to put the entry with the most active threads on the stack first and then the entry with fewer active threads [AMD, 2012]. In part (d) of Figure 3.4 we follow this order while in part (c) we used the opposite order.

3.1.2 SIMT DEADLOCK AND STACKLESS SIMT ARCHITECTURES

Recently, NVIDIA has disclosed details of their upcoming Volta GPU architecture [NVIDIA Corp., 2017]. One change they have highlighted will be to the behavior of masking under divergence and how this interacts with synchronization. The stack-based implementation of SIMT can lead to a deadlock condition called "SIMT deadlock" by ElTantawy and Aamodt [2016]. Academic work has described alternative hardware for SIMT execution [ElTantaway et al., 2014] that, with minor changes [ElTantawy and Aamodt, 2016], can avoid SIMT deadlock. NVIDIA calls their new thread divergence management approach Independent Thread Scheduling. The descriptions of independent thread scheduling suggest they achieve behavior similar to that obtained by the above academic proposals. Below, we first describe the SIMT deadlock problem, then we describe a mechanism for avoiding SIMT deadlock that is consis-

图 3.4c 显示了 Warp 执行完图 3.3 中第 6 行的分支后堆栈的状态。由于三个线程分支到基本块 B，一个线程分支到基本块 F，因此堆栈顶部 (TOS) 中添加了两个新条目。Warp 执行的下一条指令由堆栈顶部 (TOS) 条目中的下一个 PC 值确定。在图 3.4c 中，这个下一个 PC 值为 B，表示基本块 B 中第一条指令的地址。相应的活动掩码条目“ 1110 ”表示 Warp 中只有前三个线程应该执行此指令。

线程束中的前三个线程继续执行基本块 B 中的指令，直到它们到达图 3.3 中第 9 行的分支。执行此分支后，它们会发散，如前所述。此分支发散会导致堆栈发生三处变化。首先，执行分支之前的 TOS 条目的下一个 PC 条目（图 3.4d 中标记为 (i)）被修改为分支的 *reconvergence point*，即基本块 E 中第一条指令的地址。然后，添加两个条目（图 3.4d 中标记为 (ii) 和 (iii)），每个条目对应于执行分支后线程束中的每条路径。

重新收敛点是程序中可以强制分叉的线程继续以锁步方式执行的位置。通常首选最近的重新收敛点。在给定程序执行中，编译时可以保证分叉的线程可以再次以锁步方式执行的最早点是导致分支分叉的分支的直接后支配者。在运行时，有时可以在程序的较早点重新收敛 [Coon 和 Lindholm，2008 年，Diamos 等，2011 年，Fung 和 Aamodt，2011 年]。

一个有趣的问题是“ 应该使用什么顺序在发散分支之后将条目添加到堆栈中？ ”为了将重新收敛堆栈的最大深度降低到与 Warp 中的线程数成对数关系，最好先将活动线程最多的条目放在堆栈上，然后再将活动线程较少的条目放在堆栈上 [AMD，2012]。在图 3.4 的部分 (d) 中，我们遵循此顺序，而在部分 (c) 中，我们使用了相反的顺序。

3.1.2 SIMT 死锁和无堆栈 SIMT 架构

最近，NVIDIA 披露了即将推出的 Volta GPU 架构的细节 [NVIDIA Corp.，2017]。他们强调的一项变化是发散下的掩蔽行为及其与同步的交互方式。基于堆栈的 SIMT 实现可能导致死锁情况，ElTantawy 和 Aamodt [2016] 将其称为“ SIMT 死锁 ”。学术工作描述了用于 SIMT 执行的替代硬件 [ElTantaway 等，2014]，经过微小更改 [ElTantawy 和 Aamodt，2016] 可以避免 SIMT 死锁。NVIDIA 将他们的新线程发散管理方法称为独立线程调度。独立线程调度的描述表明它们实现了与上述学术提案类似的行为。下面，我们首先描述 SIMT 死锁问题，然后描述一种避免 SIMT 死锁的机制，该机制是一致的

tent with NVIDIA's description of independent thread scheduling and that was disclosed in a recent NVIDIA patent application [Diamos et al., 2015].

The left part of Figure 3.5 gives a CUDA example illustrating the SIMT deadlock problem and the middle part shows the corresponding control flow graph. Line A initializes the shared variable, mutex, to zero to indicate that a lock is free. On line B, each thread in a warp executes the `atomicCAS` operation, which performs a compare-and-swap operation on the memory location containing mutex. The `atomicCAS` operation is a compiler intrinsic that is translated to a `atom.global.cas` PTX instruction. Logically, the compare-and-swap first reads the contents of mutex, then it compares it to the second input, 0. If the current value of mutex is 0, then the compare and swap operation updates the value of mutex to the third input, 1. The value returned by `atomicCAS` is original value of mutex. Importantly, the compare-and-swap performs the above sequence of logical operations atomically for each thread. Thus, multiple accesses by `atomicCAS` to any single location, made by different threads within the same warp, are serialized. As all threads in Figure 3.5 access the same memory location, only one thread will see the value of mutex as 0, and the remaining threads will see the value as 1. Next, while keeping the SIMT-stack in mind, consider what happens with the `while` loop on line B after `atomicCAS` returns. Different threads see different loop conditions. Specifically, one thread will want to exit the loop while the remaining threads will want to stay in the loop. The thread that exits the loop will have reached the reconvergence point and thus will no longer be active on the SIMT-stack and thus unable to execute the `atomicExch` operation to release the lock on line C. The threads that remain in the loop will be active at the top of the SIMT-stack and will spin indefinitely. The resulting circular dependence between threads introduces a new form of deadlock, called SIMT-deadlock by ElTantawy and Aamodt [2016] that would not exist had the threads executed on a MIMD architecture.

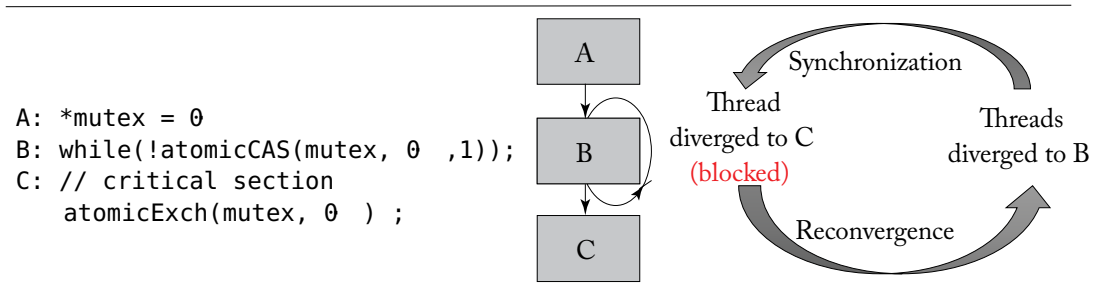


Figure 3.5: SIMT deadlock example (based on Figure 1 from ElTantawy and Aamodt [2016]).

Next, we summarize a stackless branch reconvergence mechanism like that in a recent US Patent Application from NVIDIA [Diamos et al., 2015]. This mechanism is consistent with NVIDIA's descriptions to date of Volta's reconvergence handling mechanisms [Nvidia, 2017]. The key idea is to replace the stack with per warp convergence barriers. Figure 3.6 shows various fields maintained per warp as described in NVIDIA's patent application and Figure 3.8

与 NVIDIA 对独立线程调度的描述一致，并在最近的 NVIDIA 专利申请 [Diamos et al., 2015] 中进行了披露。

图 3.5 的左侧部分给出了一个说明 SIMT 死锁问题的 CUDA 示例，中间部分显示了相应的控制流图。行 A 将共享变量 mutex 初始化为零，以指示锁是空闲的。在行 B 上，warp 中的每个线程执行 `atomicCAS` 操作，该操作对包含 mutex 的内存位置执行比较和交换操作。`atomicCAS` 操作是编译器内在函数，它被转换为 `atom.global.cas` PTX 指令。从逻辑上讲，比较和交换首先读取 mutex 的内容，然后将其与第二个输入 0 进行比较。如果 mutex 的当前值为 0，则比较和交换操作会将 mutex 的值更新为第三个输入 1。`atomicCAS` 返回的值是 mutex 的原始值。重要的是，比较和交换对每个线程都以原子方式执行上述一系列逻辑操作。因此，同一 warp 内不同线程对任何单个位置的多次访问（`atomicCAS`）都是串行的。由于图 3.5 中的所有线程都访问相同的内存位置，因此只有一个线程会看到互斥锁的值为 0，其余线程会看到该值为 1。接下来，在考虑 SIMT 堆栈的同时，考虑在`atomicCAS` 返回后，B 行上的`while` 循环会发生什么。不同的线程会看到不同的循环条件。具体来说，一个线程想要退出循环，而其余线程想要留在循环中。退出循环的线程将到达重新收敛点，因此将不再在 SIMT 堆栈上处于活动状态，因此无法执行`atomicExch` 操作来释放 C 行上的锁。留在循环中的线程将在 SIMT 堆栈的顶部处于活动状态，并将无限期地旋转。线程之间产生的循环依赖引入了一种新形式的死锁，ElTantawy 和 Aamodt [2016] 将其称为 SIMT 死锁，如果线程在 MIMD 架构上执行，这种死锁就不会出现。

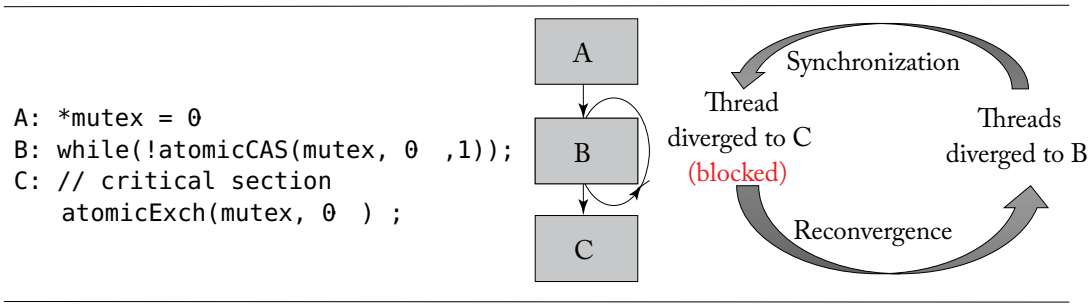


图 3.5：SIMT 死锁示例（基于 ElTantawy 和 Aamodt [2016] 中的图 1）。

接下来，我们总结了一种无堆栈分支重收敛机制，类似于 NVIDIA 最近的美国专利申请 [Diamos et al., 2015]。该机制与 NVIDIA 迄今为止对 Volta 重收敛处理机制的描述一致 [Nvidia, 2017]。关键思想是用每个 warp 收敛屏障取代堆栈。图 3.6 显示了 NVIDIA 专利申请中所述的每个 warp 维护的各种字段，图 3.8

provides a corresponding example to illustrate the operation of convergence barriers. Effectively, the proposal provides an alternative implementation of Multi-Path IPDOM [ElTantaway et al., 2014], which will be described in Section 3.4.2 along with earlier academic works. The convergence barrier mechanism has some similarities to the notion of a *warp barrier* described in Fung and Aamodt [2011]. To help explain the convergence barrier mechanism below we consider the execution of a single warp on the code in Figure 3.8, which shows the control flow graph that results from CUDA code like that shown in Figure 3.7.

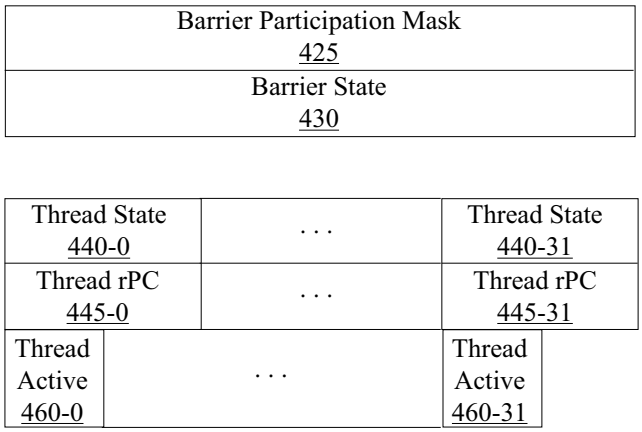


Figure 3.6: Alternate stack-less convergence barrier based branch divergence handling mechanism recently described by NVIDIA (based on Figure 4B from Diamos et al. [2015]).

```
1  // id = warp ID
2  // BB_A Basic Block "A"
3  if(id%2==0){
4      // BB_B
5  }else{
6      // BB_C
7      if(id==1){
8          // BB_D
9      }else{
10         // BB_E
11     }
12     // BB_F
13 }
14 // BB_G
```

Figure 3.7: Nested control flow example (based on Figure 6(a) from ElTantaway et al. [2014]).

提供了一个相应的示例来说明收敛屏障的操作。实际上，该提案提供了多路径 IPDOM 的替代实现 [ElTantaway 等，2014]，这将在第 3.4.2 节与早期的学术著作一起描述。收敛屏障机制与 Fung 和 Aamodt [2011] 中描述的 *warp barrier* 概念有一些相似之处。为了帮助解释下面的收敛屏障机制，我们考虑在图 3.8 中的代码上执行单个 warp，它显示了由图 3.7 中所示的 CUDA 代码产生的控制流图。

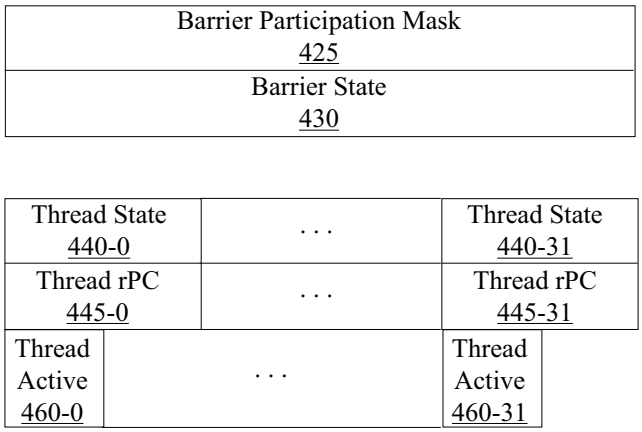


图 3.6：NVIDIA 最近描述的替代无堆栈收敛屏障的基于分支发散处理机制（基于 Diamos 等人 [2015] 的图 4B）。

```
1  // id = warp ID
2  // BB_A Basic Block "A"
3  if(id%2==0){
4      // BB_B
5  }else{
6      // BB_C
7      if(id==1){
8          // BB_D
9      }else{
10         // BB_E
11     }
12     // BB_F
13 }
14 // BB_G
```

图 3.7：嵌套控制流示例（基于 ElTantaway 等人 [2014] 的图 6（a））。

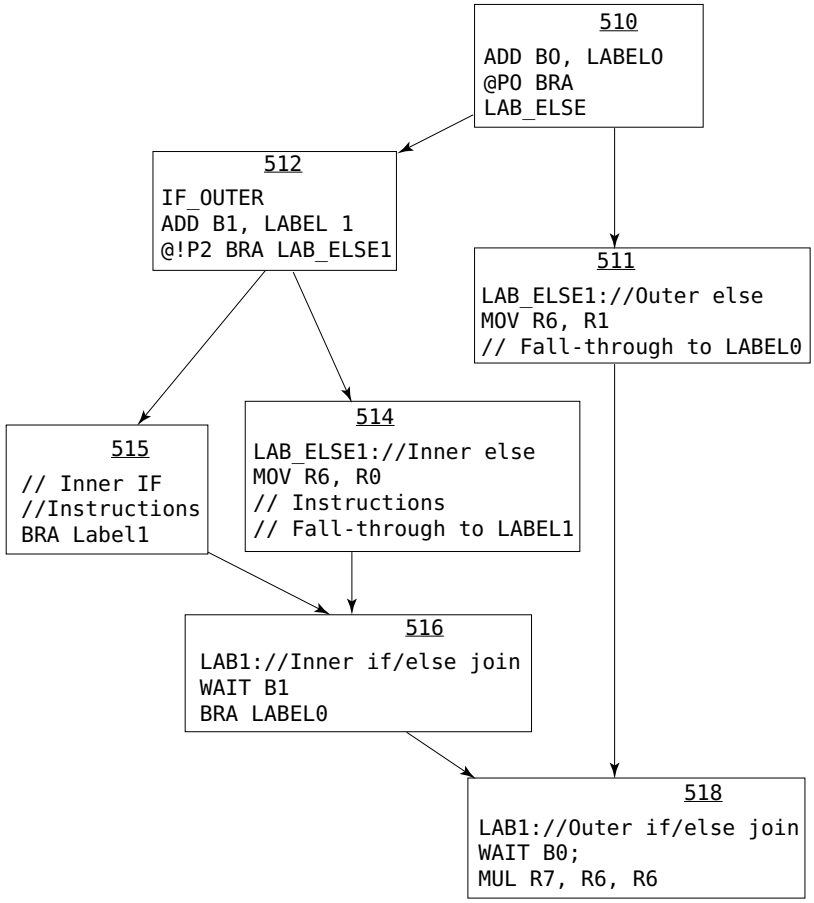


Figure 3.8: Code example for convergence barrier branch divergence handling mechanism recently described by NVIDIA (based on Figure 5B from Diamos et al. [2015]).

Next, we describe the fields in Figure 3.6. These fields are stored in registers and used by the hardware warp scheduler. Each *Barrier Participation Mask* is used to track which threads within a given warp participate in a given convergence barrier. There may be more than one barrier participation mask for a given warp. In the common case threads tracked by a given barrier participation mask will wait for each other to reach a common point in the program following a divergent branch and thereby reconverge together. To support this the *Barrier State* field is used to track which threads have arrived at a given convergence barrier. The *Thread State* tracks, for each thread in the warp whether the thread is ready to execute, blocked at a convergence barrier (and if so, which one), or has yielded. It appears the yielded state is may be used to enable other threads in the warp to make forward progress past the convergence barrier in a situation that

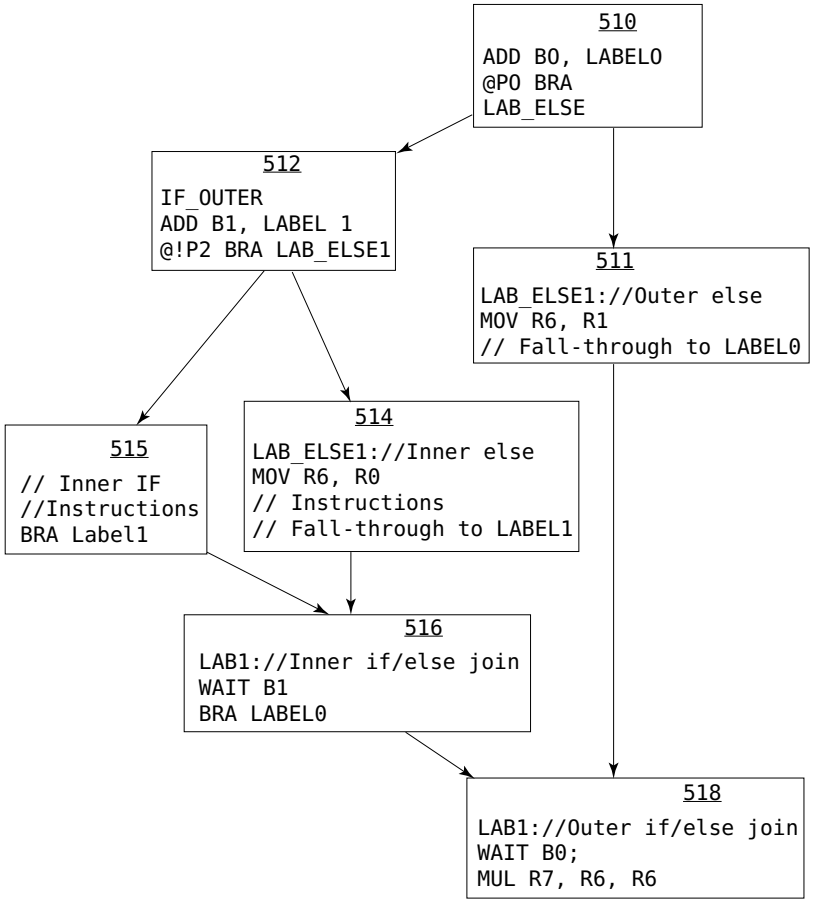


图 3.8：NVIDIA 最近描述的收敛屏障分支发散处理机制的代码示例（基于 Diamos 等人 [2015] 的图 5B）。

接下来，我们描述图 3.6 中的字段。这些字段存储在寄存器中，由硬件 Warp 调度程序使用。每个 *Barrier Participation Mask* 用于跟踪给定 Warp 中的哪些线程参与给定的收敛屏障。给定 Warp 可能有多个屏障参与掩码。在常见情况下，给定屏障参与掩码跟踪的线程将等待彼此在分叉分支后到达程序中的公共点，从而重新收敛在一起。为了支持这一点，*Barrier State* 字段用于跟踪哪些线程已到达给定的收敛屏障。*Thread State* 跟踪 Warp 中每个线程是否已准备好执行、是否在收敛屏障处受阻（如果是，则跟踪哪个屏障）或是否已让步。似乎让步状态可用于使 Warp 中的其他线程在以下情况下越过收敛屏障向前推进：

would otherwise lead to SIMT deadlock. The Thread rPC field tracks, for each thread that is not active, the address of the next instruction to execute. The Thread Active field is a bit that indicates if the corresponding thread in the warp is active.

Assuming a warp contains 32 threads, the barrier participation mask is 32-bits wide. If a bit is set, that means the corresponding thread in the warp participates in this convergence barrier. Threads diverge when they execute a branch instruction such as those at the end of basic blocks 510 and 512 in Figure 3.8. These branches correspond to the two "if" statements in Figure 3.7. The barrier participation mask is used by the warp scheduler to stop threads at a specific convergence barrier location which can be the immediate postdominator of the branch or another location. At any given time each warp may require multiple barrier participation masks to support nested control flow constructs such as the nested if statements in Figure 3.7. The registers in Figure 3.6 might be implemented using general-purpose registers or dedicated registers or some combination of the two (the patent application does not say). Given the barrier participation mask is only 32-bits wide, it would be redundant if each thread had a copy of the barrier participation mask as it might if naively using the general-purpose register file to store it. However, as the control flow can be nested to arbitrary depths, a given warp may need an arbitrary number of barrier participation masks making software management of the mask desirable.

To initialize the convergence barrier participation mask a special "ADD" instruction is employed. All threads that are active when the warp executes this ADD instruction have their bit set in the convergence barrier indicated by the ADD instruction. After executing a branch some threads may diverge, meaning the address of the next instruction (i.e., PC) to execute will differ. When this happens the scheduler will select a subset of threads with a common PC and update the Thread Active field to enable execution for these threads of the warp. Academic proposals refer to such a subset of threads as a "warp split" [ElTantaway et al., 2014, ElTantawy and Aamodt, 2016, Meng et al., 2010]. In contrast to a stack-based SIMT implementation, with a convergence barrier implementation the scheduler is free to switch between groups of diverged threads. This enables forward progress between threads in a warp when some threads have acquired a lock while others have not.

A "WAIT" instruction is used to stop a warp split when it reaches a convergence barrier. As described in NVIDIA's patent application, the WAIT instruction includes an operand to indicate the identity of the convergence barrier. The effect of the WAIT instruction is to add the threads in the warp split to the Barrier State register for the barrier and change the threads' state to blocked. Once all threads in the barrier participation mask have executed the corresponding WAIT instruction the thread scheduler can switch all the threads from the original warp split to active and SIMD efficiency is maintained. The example in Figure 3.8 has two convergence barriers, B1 and B2 with WAIT instructions in basic blocks 516 and 518. To enable switching between warp splits NVIDIA describes using a YIELD instruction along with other details such as support for indirect branches that we omit in this discussion [Diamos et al., 2015].

否则会导致 SIMT 死锁。Thread rPC 字段跟踪每个非活动线程的下一条要执行的指令的地址。Thread Active 字段是一个位，用于指示 warp 中的相应线程是否处于活动状态。

假设一个 Warp 包含 32 个线程，屏障参与掩码为 32 位宽。如果设置了某个位，则表示 Warp 中的相应线程参与此收敛屏障。线程在执行分支指令（例如图 3.8 中基本块 510 和 512 末尾的指令）时会发散。这些分支对应于图 3.7 中的两个“ if ”语句。屏障参与掩码由 Warp 调度程序用于将线程停止在特定的收敛屏障位置，该位置可以是分支的直接后支配者或其他位置。在任何给定时间，每个 Warp 可能需要多个屏障参与掩码来支持嵌套控制流构造（例如图 3.7 中的嵌套 if 语句）。图 3.6 中的寄存器可以使用通用寄存器或专用寄存器或两者的某种组合来实现（专利申请未说明）。鉴于屏障参与掩码只有 32 位宽，如果每个线程都拥有屏障参与掩码的副本，那么这将是多余的，就像单纯使用通用寄存器文件来存储它一样。但是，由于控制流可以嵌套到任意深度，给定的 warp 可能需要任意数量的屏障参与掩码，这使得掩码的软件管理成为可取之道。

为了初始化收敛屏障参与掩码，需要使用特殊的“ ADD ”指令。在 Warp 执行此 ADD 指令时，所有处于活动状态的线程都会在 ADD 指令指示的收敛屏障中设置其位。执行分支后，某些线程可能会发散，这意味着要执行的下一条指令（即 PC）的地址将有所不同。发生这种情况时，调度程序将选择具有共同 PC 的线程子集并更新线程活动字段以启用 Warp 中这些线程的执行。学术提案将这样的线程子集称为“ Warp 分裂 ”[ElTantaway 等，2014 年，ElTantawy 和 Aamodt，2016 年，Meng 等，2010 年]。与基于堆栈的 SIMT 实现相比，使用收敛屏障实现，调度程序可以自由地在发散线程组之间切换。当某些线程已获得锁而其他线程尚未获得锁时，这可以使 Warp 中的线程之间向前推进。

“ WAIT ”指令用于在 Warp Split 达到收敛屏障时停止 Warp Split。如 NVIDIA 专利申请中所述，WAIT 指令包含一个操作数，用于指示收敛屏障的身份。WAIT 指令的作用是将 Warp Split 中的线程添加到屏障的屏障状态寄存器中，并将线程的状态更改为阻塞。一旦屏障参与掩码中的所有线程都执行了相应的 WAIT 指令，线程调度程序就可以将所有线程从原始 Warp Split 切换到活动状态，同时保持 SIMD 效率。图 3.8 中的示例有两个收敛屏障，B1 和 B2，基本块 516 和 518 中有 WAIT 指令。为了实现 Warp Split 之间的切换，NVIDIA 描述了使用 YIELD 指令以及其他细节，例如对间接分支的支持，我们在此讨论中省略了这些细节 [Diamos et al., 2015]。

Figure 3.9 shows an example of the timing of stack-based reconvergence and Figure 3.10 illustrates potential timing using independent thread scheduling as described in NVIDIA's Volta whitepaper. In Figure 3.10, we can see statements A and B are interleaved with statements X and Y by Volta in contrast with the behavior in Figure 3.9. This behavior is consistent with the convergence barrier mechanism described above (as well as Multi-Path IPDOM [ElTantaway et al., 2014]). Finally, Figure 3.11 illustrates how a stackless architecture might execute the spin look code from Figure 3.5 so as to avoid SIMT deadlock.

```
if (threadIdx.x < 4) {
    A;
    B;
} else {
    X;
    Y;
}
Z;
```

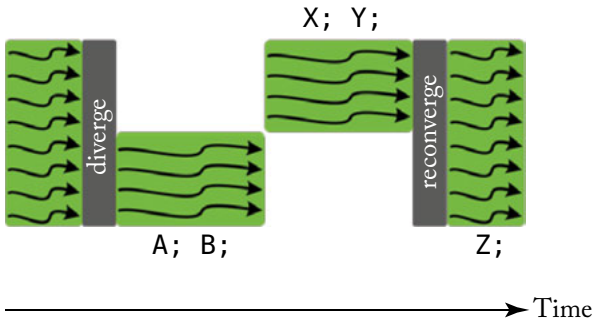


Figure 3.9: Example showing behavior of stack-based reconvergence (based on Figure 20 from Nvidia [2017]).

```
if (threadIdx.x < 4) {
    A;
    B;
} else {
    X;
    Y;
}
Z;
__syncwarp()
```

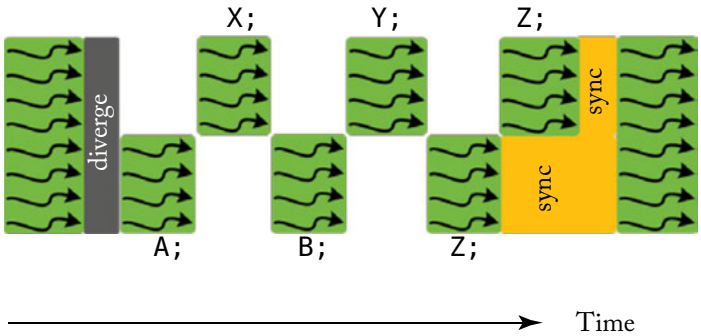


Figure 3.10: Example showing behavior of Volta reconvergence (based on Figure 23 from Nvidia [2017]).

3.1.3 WARP SCHEDULING

Each core in a GPU hosts contains many warps. A very interesting question is which order these warps should be scheduled in. To simplify the discussion we assume that each warp issues only a single instruction when it is scheduled and furthermore that the warp is not eligible

图 3.9 显示了基于堆栈的重新收敛的时序示例，图 3.10 说明了使用 NVIDIA Volta 白皮书中所述的独立线程调度的潜在时序。在图 3.10 中，我们可以看到 Volta 将语句 A 和 B 与语句 X 和 Y 交错，这与图 3.9 中的行为形成对比。此行为与上面描述的收敛屏障机制一致（以及多路径 IPDOM [ElTantaway 等，2014]）。最后，图 3.11 说明了无堆栈架构如何执行图 3.5 中的自旋外观代码以避免 SIMT 死锁。

```
if (threadIdx.x < 4) {
    A;
    B;
} else {
    X;
    Y;
}
Z;
```

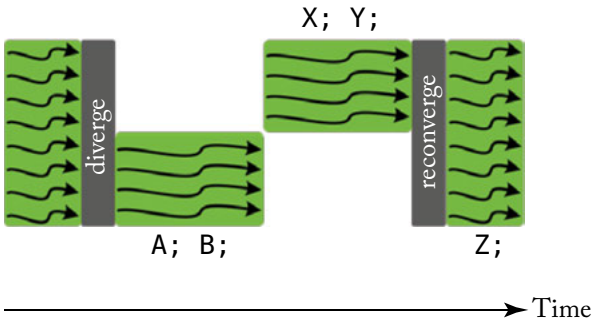


图 3.9：基于堆栈的重新收敛行为的示例（基于 Nvidia [2017] 的图 20）。

```
if (threadIdx.x < 4) {
    A;
    B;
} else {
    X;
    Y;
}
Z;
__syncwarp()
```

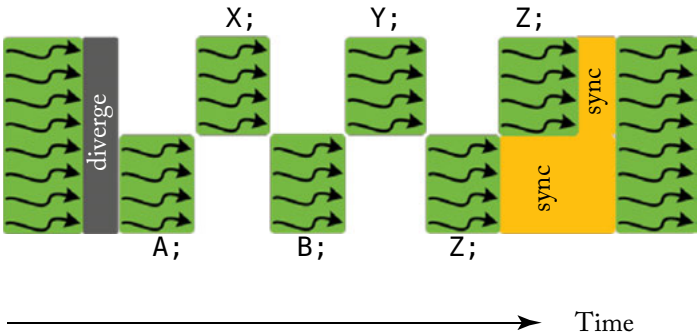


图 3.10：展示 Volta 重新收敛行为的示例（基于 Nvidia [2017] 的图 23）。

3.1.3 WARP 调度

GPU 主机中的每个核心都包含许多 Warp。一个非常有趣的问题是这些 Warp 应该按什么顺序进行调度。为了简化讨论，我们假设每个 Warp 在调度时只发出一条指令，而且 Warp 不符合条件

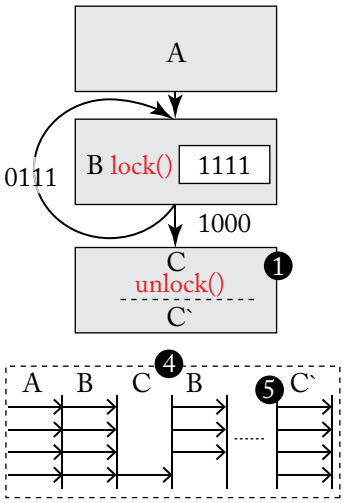


Figure 3.11: Example showing behavior of an academic mechanism similar to convergence-barrier on spin lock code from Figure 3.5 (based on Figure 6(a) from ElTantawy and Aamodt [2016]).

to issue another instruction until the first instruction completes execution. We will revisit this assumption later in this chapter.

If the memory system were "ideal" and responded to memory requests within some fixed latency it would, in theory, be possible to design the core to support enough warps to hide this latency using fine-grained multithreading. In this case it can be argued that we can reduce the area of the chip for a given throughput by scheduling warps in "round robin" order. In round robin the warps are given some fixed ordering, for example ordered by increasing thread identifiers, and warps are selected by the scheduler in this order. One property of this scheduling order is that it allows roughly equal time to each issued instruction to complete execution. If the number of warps in a core multiplied by the issue time of each warp exceeds the memory latency the execution units in the core will always remain busy. So, increasing the number of warps up to this point can in principle increase throughput per core.

However, there is an important trade off: to enable a different warp to issue an instruction each cycle it is necessary that each thread have its own registers (this avoids the need to copy and restore register state between registers and memory). Thus, increasing the number of warps per core increases the fraction of chip area devoted to register file storage relative to the fraction dedicated to execution units. For a fixed chip area increasing warps per core will decrease the total number of cores per chip.

In practice, the response latency of memory depends upon the application's locality properties and the amount of resulting contention encountered by off-chip memory accesses. What

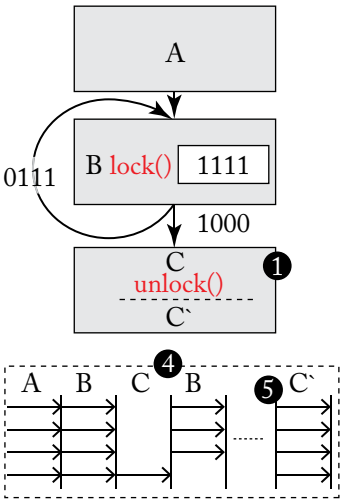


图 3.11：展示与图 3.5 中的自旋锁代码上的收敛屏障类似的学术机制的行为的示例（基于 ElTantawy 和 Aamodt [2016] 中的图 6（a））。

发出另一条指令，直到第一条指令执行完毕。我们将在本章后面重新讨论这一假设。

如果内存系统是“ 理想的 ”，能以某个固定的延迟响应内存请求，那么理论上可以设计内核来支持足够多的 warp，从而使用细粒度多线程来隐藏这种延迟。在这种情况下，可以说我们可以通过“ 循环 ”顺序调度 warp 来减少给定吞吐量的芯片面积。在循环中，warp 被赋予某种固定的顺序，例如按线程标识符递增的顺序排列，并且调度程序按照此顺序选择 warp。此调度顺序的一个特性是，它允许每个发出的指令在大致相同的时间内完成执行。如果内核中的 warp 数乘以每个 warp 的发出时间超过了内存延迟，那么内核中的执行单元将始终保持繁忙。因此，将 warp 数增加到这个程度原则上可以增加每个内核的吞吐量。

但是，有一个重要的权衡：要使不同的 Warp 能够在每个周期发出指令，每个线程必须有自己的寄存器（这样就无需在寄存器和内存之间复制和恢复寄存器状态）。因此，增加每个核心的 Warp 数量会增加用于寄存器文件存储的芯片面积相对于专用于执行单元的面积的比例。对于固定的芯片面积，增加每个核心的 Warp 数量将减少每个芯片的核心总数。

实际上，内存的响应延迟取决于应用程序的局部性属性以及片外内存访问所遇到的争用量。

impact does scheduling play when considering the memory system of the GPU? This has been the topic of considerable research in the past few years and we will return to this question after adding more detail about the memory system to our model of GPU microarchitecture. However, briefly, locality properties can either favor or discourage round-robin scheduling: when different threads share data at a similar point in their execution, such as when accessing texture maps in graphics pixel shaders, it is beneficial for threads to make equal progress as this can increase the number of memory references which "hit" in on-chip caches, which is encouraged by round-robin scheduling [Lindholm et al., 2015]. Similarly, accessing DRAM is more efficient when nearby locations in the address space are accessed nearby in time and this is also encouraged by round-robin scheduling [Narasiman et al., 2011]. On the other hand, when threads mostly access disjoint data, as tends to occur with more complex data structures, it can be beneficial for a given thread to be scheduled repeatedly so as to maximize locality [Rogers et al., 2012].

3.2 TWO-LOOP APPROXIMATION

To help reduce the number of warps that each core must support to hide long execution latencies it is helpful to be able to issue a subsequent instruction from a warp while earlier instructions have not yet completed. However, the one-loop microarchitecture described earlier prevents this because the scheduling logic in that design only has access to the thread identifier and the address of the next instruction to issue. Specifically, it does not know whether the next instruction to issue for the warp has a dependency upon an earlier instruction that has not yet completed execution. To provide such dependency information it is necessary to first fetch the instruction from memory so as to determine what data and/or structural hazards exists. For this purpose, GPUs implement an instruction buffer were instructions are placed after cache access. A separate scheduler is used to decide which of several instructions in the instruction buffer should be issued next to the rest of the pipeline.

Instruction memory is implemented as a first-level instruction cache backed by one or more levels of secondary (typically unified) caches. The instruction buffer can also help in hiding instruction cache miss latencies in combination with instruction miss-status holding registers (MSHRs) [Kroft, 1981]. After a cache hit or a fill from a cache miss, the instruction information is placed into the instruction buffer. The organization of the instruction buffer can take many forms. One particularly straightforward approach is to have storage for one or more instructions per warp.

Next, let us consider how to detect data dependencies between instructions within the same warp. Their are two traditional approaches to detecting dependencies between instructions found in traditional CPU architectures: a scoreboard and reservation stations. Reservation stations are used for eliminating name dependencies and introduce the need for associative logic that is expensive in terms of area and energy. Scoreboards can be designed to support either in-order execution or out-of-order execution. Scoreboards supporting out-of-order execution, like that used in the CDC 6600, are also fairly complex. On the other hand, the scoreboard for a

在考虑 GPU 的内存系统时，调度会产生什么影响？这在过去几年中一直是大量研究的主题，在将有关内存系统的更多细节添加到我们的 GPU 微架构模型后，我们将回到这个问题。然而，简而言之，局部性可以支持或阻碍循环调度：当不同的线程在其执行的相似点共享数据时，例如在访问图形像素着色器中的纹理贴图时，线程取得同等进展是有益的，因为这可以增加在片上缓存中“命中”的内存引用数量，这是循环调度所鼓励的 [Lindholm 等人，2015]。同样，当在时间上就近访问地址空间中的邻近位置时，访问 DRAM 会更有效率，这也是循环调度所鼓励的 [Narasiman 等人，2011]。另一方面，当线程主要访问不相交的数据时（这在更复杂的数据结构中往往会发生），重复调度给定的线程以最大化局部性是有益的 [Rogers et al., 2012]。

3.2 双环路近似

为了减少每个核心必须支持的 Warp 数量以隐藏较长的执行延迟，能够在先前的指令尚未完成时从 Warp 发出后续指令会很有帮助。但是，前面描述的单循环微架构阻止了这一点，因为该设计中的调度逻辑只能访问线程标识符和要发出的下一个指令的地址。具体来说，它不知道 Warp 发出的下一个指令是否依赖于尚未完成执行的先前指令。要提供此类依赖关系信息，必须先从内存中获取指令，以确定存在哪些数据和/或结构风险。为此，GPU 实现了指令缓冲区，指令放在缓存访问之后。使用单独的调度程序来决定指令缓冲区中的哪些指令应该在其余管道之后发出。

指令存储器实现为第一级指令缓存，由一个或多个级别的二级（通常是统一的）缓存支持。指令缓冲区还可以与指令未命中状态保持寄存器 (MSHR) 结合使用，帮助隐藏指令缓存未命中延迟 [Kroft，1981]。缓存命中或缓存未命中填充后，指令信息将放入指令缓冲区。指令缓冲区的组织形式多种多样。一种特别直接的方法是每个 warp 存储一个或多个指令。

接下来，让我们考虑如何检测同一 warp 中指令之间的数据依赖关系。在传统 CPU 架构中，检测指令之间的依赖关系有两种传统方法：记分板和保留站。保留站用于消除名称依赖关系，并引入了对关联逻辑的需求，而关联逻辑在面积和能量方面都很昂贵。记分板可以设计为支持按序执行或无序执行。支持无序执行的记分板（如 CDC 6600 中使用的记分板）也相当复杂。另一方面，

single threaded in-order CPU is very simple: each register is represented in the scoreboard with a single bit that is set whenever an instruction issues that will write to that register. Any instruction that wants to read or write to a register that has its corresponding bit set in the scoreboard is stalled until the bit is cleared by the instruction writing to the register. This prevents both read-after-write and write-after-write hazards. When combined with in-order instruction issue this simple scoreboard can prevent write-after-read hazards provided that reading of the register file is constrained to occur in-order which is typically the case in in-order CPU designs. Given it is the simplest design and therefore will consume the least amount of area and energy, GPUs implement in-order scoreboards. However, as discussed next, there are challenges to using an in-order scoreboard when supporting multiple warps.

The first concern with the simple in-order scoreboard design described above is the very large number of registers contained in modern GPUs. With up to 128 registers per warp and up to 64 warps per core a total of 8192 bits per core is required to implement the scoreboard.

Another concern with the simple in-order scoreboard design described above is that an instruction that encounters a dependency must repeatedly lookup its operands in the scoreboard until the prior instruction it depends upon writes its results to the register file. With a single-threaded design this introduces little complexity. However, in an in-order issue multithreaded processor instructions from multiple threads may be waiting for earlier instructions to complete. If all such instructions must probe the scoreboard additional read ports are required. Recent GPUs support up to 64 warps per core and with up to 4 operands allowing all warps to probe the scoreboard every cycle would require 256 read ports, which would be very expensive. One alternative would be to restrict the number of warps that can probe the scoreboard each cycle, but this restricts the number of warps that can be considered for scheduling. Also, if none of the instructions checked are free of dependencies no instruction may be issued even if other instructions that could not be checked happened to be free of dependencies.

Both issues can be addressed using a design proposed by Coon et al. [2008]. Rather than hold a single bit per register per warp, the design contains a small number (estimated to be around 3 or 4 in one recent study [Lashgar et al., 2016]) of entries per warp, where each entry is the identifier of a register that will be written by an instruction that has been issued but not yet completed execution. A regular in-order scoreboard is accessed both when instructions issue and when they write back. Instead, Coon et al.'s scoreboard is accessed when an instruction is placed into the instruction buffer and when an instruction writes its results into the register file.

When an instruction is fetched from the instruction cache and placed in the instruction buffer the scoreboard entries for the corresponding warp are compared against that instructions' source and destination registers. This results in a short bit vector, with one bit for each entry in the scoreboard for that warp (e.g., 3 or 4 bits). A bit is set if the corresponding entry in the scoreboard matched any of the operands of the instruction. This bit vector is then copied alongside the instruction in the instruction buffer. An instruction is not eligible to be considered by the instruction scheduler until all bits are cleared, which can be determined by feeding each

单线程有序 CPU 非常简单：每个寄存器在记分板中都用一个位表示，只要发出要写入该寄存器的指令，该位就会被设置。任何想要读取或写入在记分板中设置了相应位的寄存器的指令都会被阻止，直到写入寄存器的指令清除该位。这可以防止写后读和写后写危险。当与有序指令发出相结合时，这个简单的记分板可以防止读后写危险，前提是寄存器文件的读取被限制为按顺序进行，这在有序 CPU 设计中通常就是这种情况。鉴于它是最简单的设计，因此将消耗最少的面积和能量，GPU 实现了有序记分板。但是，如下所述，在支持多个 warp 时使用有序记分板存在挑战。

上述简单有序记分板设计的第一个问题是现代 GPU 中包含的寄存器数量非常多。每个 Warp 最多有 128 个寄存器，每个核心最多有 64 个 Warp，因此实现记分板需要每个核心总共 8192 位。

上述简单按序记分板设计的另一个问题是，遇到依赖关系的指令必须反复在记分板中查找其操作数，直到它所依赖的上一条指令将其结果写入寄存器文件。对于单线程设计，这几乎不会带来复杂性。但是，在按序发出中，来自多个线程的多线程处理器指令可能正在等待早期指令完成。如果所有这些指令都必须探测记分板，则需要额外的读取端口。最近的 GPU 支持每个核心最多 64 个 Warp，最多 4 个操作数允许所有 Warp 每个周期探测记分板，这将需要 256 个读取端口，这将非常昂贵。一种替代方法是限制每个周期可以探测记分板的 Warp 数量，但这会限制可以考虑进行调度的 Warp 数量。此外，如果所检查的指令都不存在依赖关系，则即使其他无法检查的指令恰好不存在依赖关系，也不会发出任何指令。

这两个问题都可以使用 Coon 等人 [2008] 提出的设计来解决。该设计不是在每个 warp 中为每个寄存器保存一个位，而是在每个 warp 中包含少量条目（在最近的一项研究中估计约为 3 或 4 个 [Lashgar 等人，2016]），其中每个条目都是将由已发出但尚未完成执行的指令写入的寄存器的标识符。常规的按顺序记分板在指令发出和写回时都会被访问。相反，Coon 等人的记分板是在将指令放入指令缓冲区时以及指令将其结果写入寄存器文件时访问的。

当从指令缓存中取出一条指令并将其放入指令缓冲区时，相应 Warp 的记分板条目将与该指令的源寄存器和目标寄存器进行比较。这会产生一个短位向量，该 Warp 的记分板中每个条目对应一个位（例如 3 位或 4 位）。如果记分板中的相应条目与该指令的任何操作数匹配，则设置一个位。然后，将此位向量与指令一起复制到指令缓冲区中。直到所有位都被清除后，指令调度程序才会考虑该指令，这可以通过将每个位输入到寄存器中来确定

bit of the vector into a NOR gate. Dependency bits in the instruction buffer are cleared as instructions write their results to the register file. If all entries are used up for a given warp then either fetch stalls for all warps or the instruction is discarded and must be fetched again. When an instruction that has executed is ready to write to the register file it clears the entry that was allocated to it in the scoreboard and also clears the corresponding dependency bit for any instructions from the same warp that are stored in the instruction buffer.

In the two-loop architecture, the first loop selects a warp that has space in the instruction buffer, looks up its program counter and performs an instruction cache access to obtain the next instruction. The second loop selects an instruction in the instruction buffer that has no outstanding dependencies and issues it to the execution units.

3.3 THREE-LOOP APPROXIMATION

As described earlier, to hide long memory latencies it is necessary to support many warps per core and to support cycle by cycle switching between warps it is necessary to have a large register file that contains separate physical registers for every warp that is executing. For example, such register contain 256 KB on recent GPU architectures from NVIDIA (e.g., Kepler, Maxwell, and Pascal architectures). Now, the area of an SRAM memory is proportional to the number of ports. A naive implementation of a register file requires one port per operand per instruction issued per cycle. One way to reduce the area of the register file is to simulate the large number of ports using multiple banks of single-ported memories. While it is possible to achieve such effects by exposing these banks to the instruction set architecture, in some GPU designs it appears a structure known as the operand collector [Coon et al., 2009, Lindholm et al., 2008b, Lui et al., 2008] is used to achieve this in a more transparent way. The operand collector effective forms a third scheduling loop as described below.

To better understand the problem solved by the operand collector, first consider Figure 3.12, which shows a naive microarchitecture for providing increased register file bandwidth. This figure shows the register read stage of a GPU instruction pipeline where the register file is composed for four single-ported logical banks of registers. In practice, as the register file is very large, each logical bank may be further decomposed into a larger number of physical banks (not shown). The logical banks are connected via a crossbar to staging registers (labeled "pipeline register") that buffer source operands before passing them to a SIMD execution unit. An arbiter controls access to the individual banks and routes results through a crossbar to the appropriate staging register.

Figure 3.13 shows a naive layout of each warp's registers to logical banks. In this figure, register `r0` from warp 0 (`w0`) is stored in the first location in Bank 0, register `r1` from warp 0 is stored in the first location in Bank 1, and so on. If the number of registers required by the computation is larger than the number of logical banks, the allocation wraps around. For example, register `r4` for warp 0 is stored in the second location in Bank 0.

将向量的位放入 NOR 门。当指令将其结果写入寄存器文件时，指令缓冲区中的依赖位将被清除。如果给定 warp 的所有条目都已用完，则所有 warp 的提取都会停止，或者指令将被丢弃并必须再次提取。当已执行的指令准备好写入寄存器文件时，它会清除记分板中分配给它的条目，并清除指令缓冲区中存储的来自同一 warp 的任何指令的相应依赖位。

在双循环架构中，第一个循环选择指令缓冲区中有空间的 Warp，查找其程序计数器并执行指令缓存访问以获取下一条指令。第二个循环选择指令缓冲区中没有未完成依赖项的指令并将其发送到执行单元。

3.3 三环路近似

如前所述，为了隐藏较长的内存延迟，需要支持每个核心的多个 Warp，并且为了支持 Warp 之间的逐周期切换，需要有一个大型寄存器文件，其中包含每个正在执行的 Warp 的单独物理寄存器。例如，在 NVIDIA 的最新 GPU 架构（例如 Kepler、Maxwell 和 Pascal 架构）上，此类寄存器包含 256 KB。现在，SRAM 内存的面积与端口数量成正比。寄存器文件的简单实现要求每个操作数每个周期发出的每条指令都有一个端口。减少寄存器文件面积的一种方法是使用多个单端口存储器组来模拟大量端口。虽然可以通过将这些库暴露给指令集架构来实现这种效果，但在某些 GPU 设计中，似乎使用一种称为操作数收集器的结构 [Coon 等人，2009 年，Lindholm 等人，2008b 年，Lui 等人，2008 年] 以更透明的方式实现这一点。操作数收集器有效形成第三个调度循环，如下所述。

为了更好地理解操作数收集器所解决的问题，首先考虑图 3.12，它展示了一种用于提供增加的寄存器文件带宽的简单微架构。该图显示了 GPU 指令流水线的寄存器读取阶段，其中寄存器文件由四个单端口逻辑寄存器组组成。实际上，由于寄存器文件非常大，每个逻辑寄存器组可以进一步分解为大量物理寄存器组（未显示）。逻辑寄存器组通过交叉开关连接到暂存寄存器（标记为“流水线寄存器”），暂存寄存器在将源操作数传递给 SIMD 执行单元之前对其进行缓冲。仲裁器控制对各个寄存器组的访问，并通过交叉开关将结果路由到适当的暂存寄存器。

图 3.13 显示了每个 Warp 的寄存器到逻辑组的简单布局。在该图中，Warp 0 (`w0`) 中的寄存器 `r0` 存储在组 0 的第一个位置，Warp 0 中的寄存器 `r1` 存储在组 1 的第一个位置，依此类推。如果计算所需的寄存器数量大于逻辑组的数量，则分配会回绕。例如，Warp 0 的寄存器 `r4` 存储在组 0 的第二个位置。

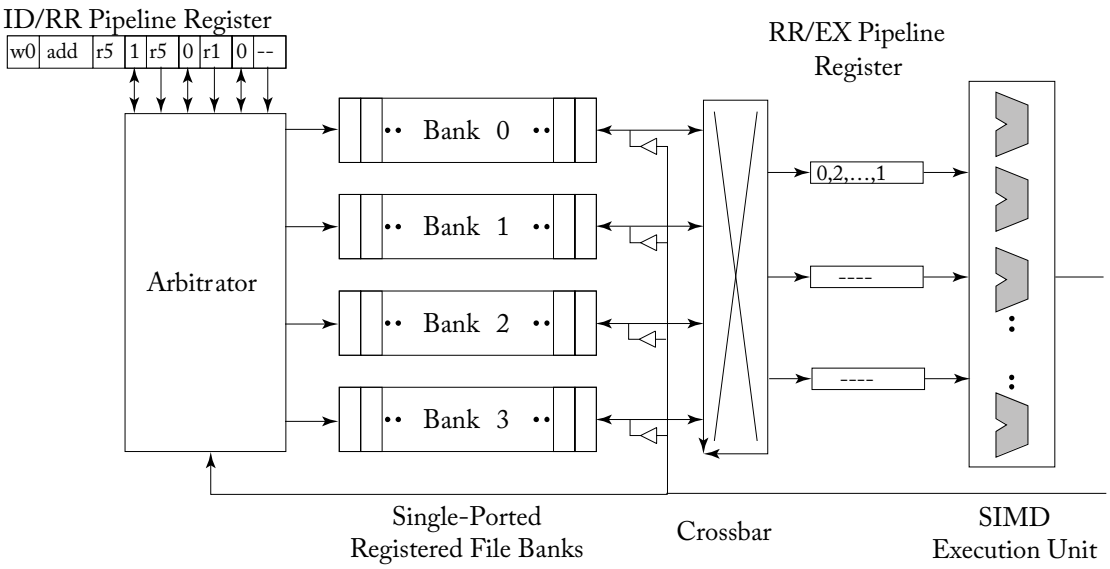


Figure 3.12: Naive banked register file microarchitecture.

Bank 0	Bank 1	Bank 2	Bank 3
...	...	...	...
w1:r4	w1:r5	w1:r6	w1:r7
w1:r0	w1:r1	w1:r2	w1:r3
w0:r4	w0:r5	w0:r6	w0:r7
w0:r0	w0:r1	w0:r2	w0:r3

Figure 3.13: Naive banked register layout.

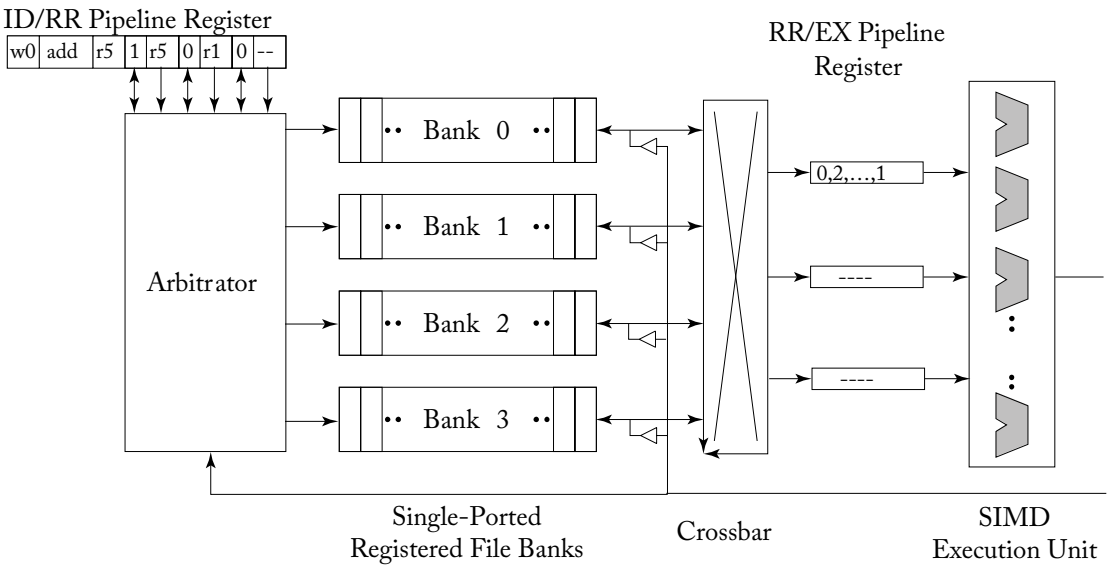


图 3.12：简单的分组寄存器文件微架构。

Bank 0	Bank 1	Bank 2	Bank 3
…	…	…	…
w1:r4	w1:r5	w1:r6	w1:r7
w1:r0	w1:r1	w1:r2	w1:r3
w0:r4	w0:r5	w0:r6	w0:r7
w0:r0	w0:r1	w0:r2	w0:r3

图 3.13：简单的分组寄存器布局。

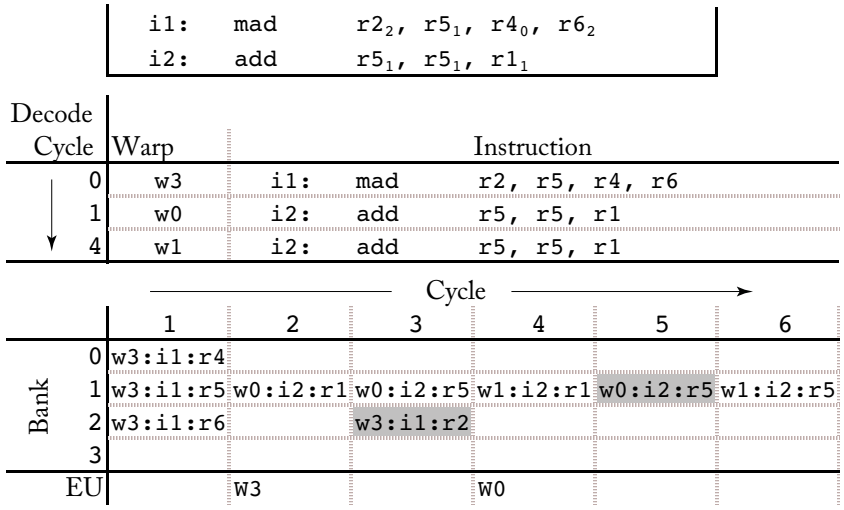


Figure 3.14: Timing of naive banked register file.

Figure 3.14 illustrates a timing example that highlights how this microarchitecture cause hurt performance. The example involves two instructions shown at the top. The first instruction, `i1`, is a multiple-add operation that reads from registers `r5`, `r4`, and `r6` which are allocated in banks 1, 0, and 2 (indicated by subscripts in the figure). The second instruction, `i2`, is an addition instruction that reads from registers `r5` and `r1` both allocated in bank 1. The middle part of the figure shows the order that instructions are issued. On cycle 0 warp 3 issues instruction `i1`, on cycle 1 warp 0 issues instruction `i2` and on cycle 4 warp 1 issues instruction `i2` after a delay due to bank conflicts as described next. The bottom portion of the figure illustrates the timing of accesses to different banks by the different instructions. On cycle 1, instruction `i1` from warp 3 is able to read all three of its source registers on cycle 1 because they map to distinct logical banks. However, on cycle 2, instruction `i2` from warp 0 is only able to read one of its two source registers because both map to bank 1. On cycle 3, the second source register for this instruction is read in parallel with the writeback of instruction `i1` from warp 3. On cycle 4, instruction `i2` from warp 1 is able to read its first-source operand but not the second as, again, both map to bank 1. On cycle 5, the second-source operand from instruction `i2` from warp 1 is prevented from being read from the register file due to the fact the bank is already being accessed by the higher-priority writeback of instruction `i2` issued earlier by warp 0. Finally, on cycle 6 the second source operand of `i2` from warp 1 is read from the register file. In summary, it takes six cycles for three instructions to finish reading their source registers and during this time many of the banks are not accessed.

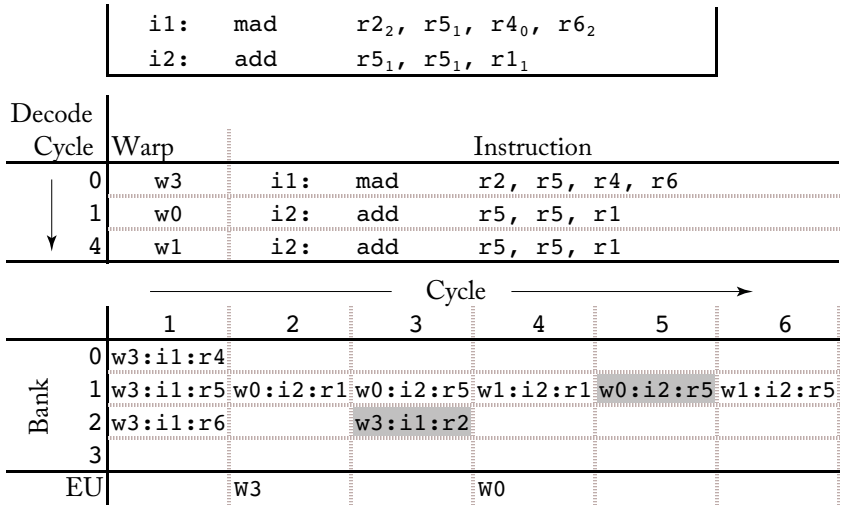


图 3.14：简单寄存器组文件的时间安排。

图 3.14 说明了一个时序示例，突出显示了此微架构如何导致性能下降。该示例涉及顶部显示的两条指令。第一条指令 `i1` 是一个多重加法运算，它从分配在存储体 1、0 和 2 中的寄存器 `r5`、`r4` 和 `r6` 中读取（在图中用下标表示）。第二条指令 `i2` 是一个加法指令，它从分配在存储体 1 中的寄存器 `r5` 和 `r1` 中读取。图的中间部分显示了指令的发出顺序。在周期 0，warp 3 发出指令 `i1`，在周期 1，warp 0 发出指令 `i2`，在周期 4，warp 1 在因存储体冲突而延迟之后发出指令 `i2`，如下所述。图的底部说明了不同指令访问不同存储体的时序。在周期 1 上，来自 warp 3 的指令 `i1` 能够在周期 1 上读取其所有三个源寄存器，因为它们映射到不同的逻辑存储体。但是，在第 2 个周期，来自 Warp 0 的指令 `i2` 只能读取两个源寄存器中的一个，因为两个源寄存器都映射到存储体 1。在第 3 个周期，此指令的第二个源寄存器与来自 Warp 3 的指令 `i1` 的写回并行读取。在第 4 个周期，来自 Warp 1 的指令 `i2` 能够读取其第一个源操作数，但不能读取第二个，因为两个源操作数也都映射到存储体 1。在第 5 个周期，来自 Warp 1 的指令 `i2` 的第二个源操作数无法从寄存器文件中读取，因为该存储体已经被 Warp 0 先前发出的高优先级指令 `i2` 的写回访问。最后，在第 6 个周期，来自 Warp 1 的第二个源操作数 `i2` 将从寄存器文件中读取。总之，三条指令需要六个周期才能完成读取其源寄存器，在此期间许多存储体都未被访问。

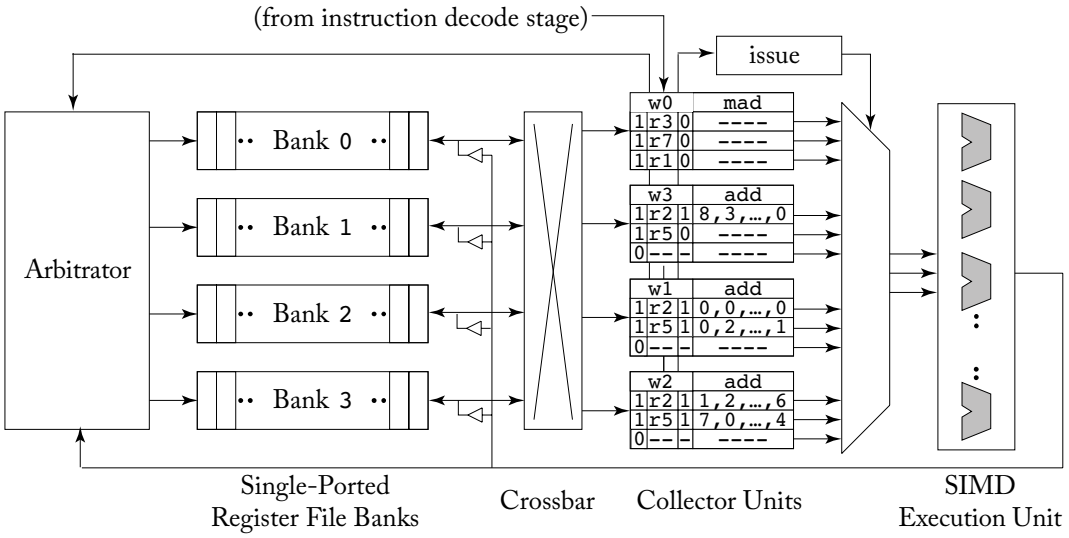


Figure 3.15: Operand collector microarchitecture (based on Figure 6 from Tor M. Aamodt et al.).

3.3.1 OPERAND COLLECTOR

The operand collector microarchitecture [Lindholm et al., 2008b] is illustrated in Figure 3.15. The key change is that the staging registers have been replaced with *collector units*. Each instruction is allocated a collector unit when it enters the register read stage. There are multiple collector units so that multiple instructions can overlap reading of source operands which can help improve throughput in the presence of bank conflicts between the source operands of individual instructions. Each collector unit contains buffering space for all source operands required to execute an instruction. Given the larger number of source operands for multiple instructions the arbiter is more likely to achieve increased bank-level parallelism to allow accessing multiple register file banks in parallel.

The operand collector uses scheduling to tolerate bank conflicts when they occur. This leaves open the question of how to reduce the number of bank conflicts. Figure 3.16 illustrates a modified register layout that Coon et al. describe for helping to reduce bank conflicts. The idea is to allocate equivalent registers from different warps in different banks. For example, in Figure 3.16 register `r0` for warp 0 is allocated to bank 0, but register `r0` for warp 1 is allocated to bank 1. This does not address bank conflicts between register operands of a single instruction. However, where it does help is in reducing bank conflicts between instructions from different warps. In particular, whenever warps are making relatively even progress (e.g., due to round-robin scheduling or two-level scheduling [Narasiman et al., 2011] in which individual warps in a fetch group are scheduled in round-robin order).

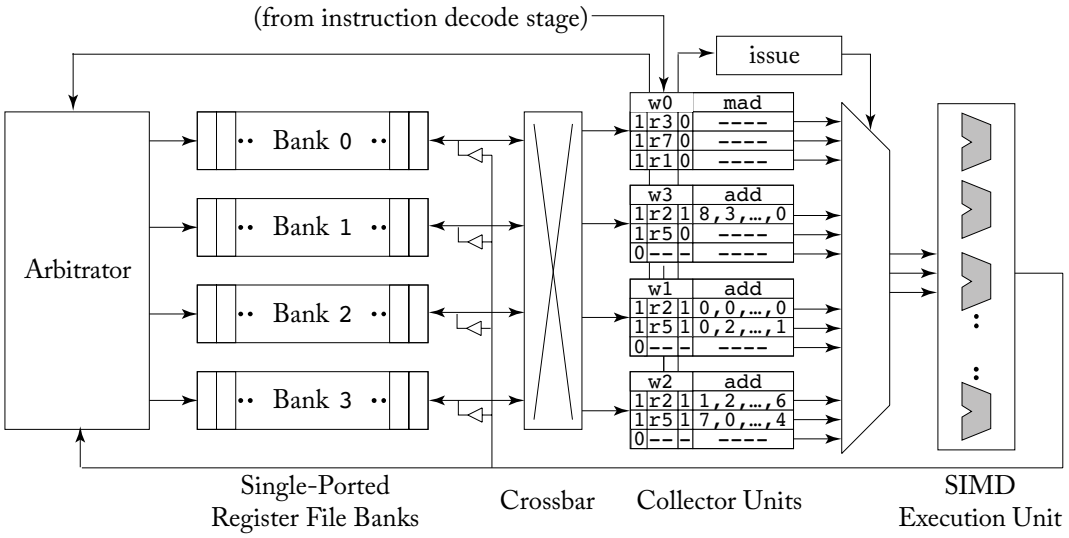


图 3.15：操作数收集器微架构（基于 Tor M. Aamodt 等人的图 6）。

3.3.1 操作数收集器

图 3.15 显示了操作数收集器微架构 [Lindholm et al., 2008b]。关键变化是暂存寄存器已被 *collector units* 取代。每条指令在进入寄存器读取阶段时都会分配一个收集器单元。有多个收集器单元，因此多条指令可以重叠读取源操作数，这有助于在各个指令的源操作数之间存在存储体冲突的情况下提高吞吐量。每个收集器单元都包含执行一条指令所需的所有源操作数的缓冲空间。鉴于多条指令的源操作数数量较多，仲裁器更有可能实现更高的存储体级并行性，以允许并行访问多个寄存器文件存储体。

操作数收集器使用调度来容忍发生的库冲突。这留下了一个如何减少库冲突的问题。图 3.16 展示了 Coon 等人为帮助减少库冲突而描述的经过修改的寄存器布局。其思想是将来自不同 warp 的等效寄存器分配到不同的库中。例如，在图 3.16 中，warp 0 的寄存器 `r0` 分配给库 0，而 warp 1 的寄存器 `r0` 分配给库 1。这并不能解决单个指令的寄存器操作数之间的库冲突。但是，它确实有助于减少来自不同 warp 的指令之间的库冲突。特别地，每当 warp 取得相对均匀的进展时（例如，由于循环调度或两级调度 [Narasiman et al., 2011]，其中获取组中的各个 warp 按循环顺序进行调度）。

Bank 0	Bank 1	Bank 2	Bank 3
...	...	...	...
w1:r7	w1:r4	w1:r5	w1:r6
w1:r3	w1:r0	w1:r1	w1:r2
w0:r4	w0:r5	w0:r6	w0:r7
w0:r0	w0:r1	w0:r2	w0:r3

Figure 3.16: Swizzled banked register layout.

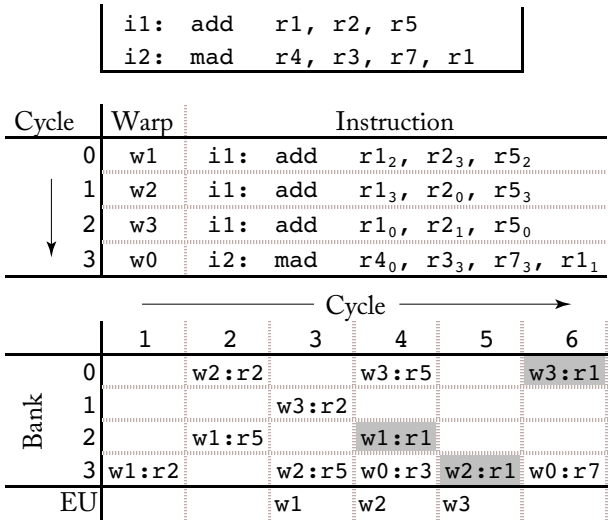


Figure 3.17: Timing of operand collector.

Figure 3.17 shows a timing example with a sequence of addition and multiply-add instructions shown at the top. In the middle the issue order is shown. Three instances of `i1` from warps 1 through 3 are issued on cycles 0 through 2. An instance of instruction `i2` from warp 0 issues on cycle 3. Notice that the add instructions write to register `r1`, which for any given warp is allocated in the same bank as source register `r5`. However, unlike the case using the register layout in Figure 3.13, here different warps access different banks which helps reduce conflicts between writeback of one warp and reading source operands in other warps. The bottom portion shows the bank level timing of accesses due to the operand collector. On cycle 1 register `r2` from warp 1 reads Bank 3. On cycle 4, notice the writeback of register `r1` from warp 1 proceeds in parallel with reading register `r5` from warp 3 and reading register `r3` from warp 0.

A subtle issue with the operand collector as described so far is that because it does not impose any order between when different instructions are ready to issue, it may allow write-after-read (WAR) hazards [Mishkin et al., 2016]. This can occur if two instructions from the

Bank 0	Bank 1	Bank 2	Bank 3
…	…	…	…
w1:r7	w1:r4	w1:r5	w1:r6
w1:r3	w1:r0	w1:r1	w1:r2
w0:r4	w0:r5	w0:r6	w0:r7
w0:r0	w0:r1	w0:r2	w0:r3

图 3.16： 混合寄存器布局。

```
i1:  add   r1, r2, r5
i2:  mad   r4, r3, r7, r1
```

Cycle	Warp	Instruction
0	w1	i1: add $r1_2$, $r2_3$, $r5_2$
1	w2	i1: add $r1_3$, $r2_0$, $r5_3$
2	w3	i1: add $r1_0$, $r2_1$, $r5_0$
3	w0	i2: mad $r4_0$, $r3_3$, $r7_3$, $r1_1$

Cycle →

Bank	1	2	3	4	5	6
0		w2:r2		w3:r5		w3:r1
1			w3:r2			
2		w1:r5		w1:r1		
3	w1:r2		w2:r5	w0:r3	w2:r1	w0:r7
EU			w1	w2	w3	

图3.17：操作数收集器时序。

图 3.17 显示了时序示例，其中顶部显示了加法和乘加指令序列。中间显示了发出顺序。来自 warp 1 至 3 的三个 `i1` 实例在第 0 至 2 个周期发出。来自 warp 0 的指令 `i2` 的一个实例在第 3 个周期发出。请注意，加法指令写入寄存器 `r1`，对于任何给定 warp，该寄存器与源寄存器 `r5` 分配在同一个存储体中。但是，与使用图 3.13 中的寄存器布局的情况不同，这里不同的 warp 访问不同的存储体，这有助于减少一个 warp 的写回和读取其他 warp 中的源操作数之间的冲突。底部显示了由于操作数收集器而导致的存储体级别访问时序。在周期 1 上，来自 Warp 1 的寄存器 `r2` 读取 Bank 3。在周期 4 上，注意到来自 Warp 1 的寄存器 `r1` 的写回与来自 Warp 3 的寄存器 `r5` 的读取以及来自 Warp 0 的寄存器 `r3` 的读取同时进行。

到目前为止描述的操作数收集器的一个微妙问题是，由于它没有在不同指令准备发出时强加任何顺序，因此它可能允许写后读 (WAR) 风险 [Mishkin et al., 2016]。如果来自

same warp are present in an operand collector with the first instruction reading a register that the second instruction will write. If the first instruction's source operand accesses encounter repeated bank conflicts the second instruction can conceivably write a new value to the register before the first register has read the (correct) older value. One way to prevent this WAR hazard is simply to require that instructions from the same warp leave the operand collector to execution units in program order. Mishkin et al. [2016] explore three potential solutions with low hardware complexity and evaluate their performance impact. The first, a *release-on-commit warpboard*, allows at most one instruction per warp to be executing. Unsurprisingly, they find this impacts performance negatively, reducing performance by almost a factor of two in some cases. Their second proposal is a *release-on-read warpboard* which allows only one instruction at a time per warp to be collecting operands in the operand collector. This scheme results in slowdowns of at most 10% on the workloads they studied. Finally, to allow for instruction level parallelism in the operand collector they propose a *bloomboard* mechanism that uses a small bloom filter to track outstanding register reads. This results in impact of less than a few percent vs. (incorrectly) allowing WAR hazards. Separately, an analysis performed by Gray suggests NVIDIA's Maxwell GPU introduced a "read dependency barrier" which is managed by special "control instructions" and which may be used to avoid WAR hazards for certain instructions (see Section 2.2.1).

3.3.2 INSTRUCTION REPLAY: HANDLING STRUCTURAL HAZARDS

There are many potential causes of structural hazards in GPU pipelines. For example, the register read stage may run out of operand collector units. Many sources of structural hazards relate to the memory system, which we will discuss in more detail the next chapter. In general, a single memory instruction executed by a warp may need to be broken down into multiple separate operations. Each of these separate operations may fully utilize a portion of the pipeline on a given cycle.

What happens when an instruction encounters a structural hazard in a GPU pipeline? In a single-threaded in-order CPU pipeline a standard solution is to stall younger instructions until the instruction encountering the stall condition can make further progress. This approach is arguable less desirable in a highly multithreaded throughput architecture for at least two reasons. First, given the large size of the register file along with the many pipeline stages required to support a full graphics pipeline distributing a stall signal may impact the critical path. Pipelining stall-cycle distribution leads to the need to introduce additional buffering increasing area. Second, stalling an instruction from one warp may cause instructions from other warps to stall behind it. If those instructions do not require the resource required by the instruction that caused the stall, throughput may suffer.

To avoid these issues GPUs implement a form of instruction replay. Instruction replay is found in some CPU designs where it is used as a recovery mechanism when speculatively scheduling a dependent instruction upon a earlier instruction that has variable latency. For example, loads may either hit or miss in a first-level cache but CPU designs that are clocked at a high

同一个 warp 中的指令同时存在于操作数收集器中，第一条指令读取第二条指令将要写入的寄存器。如果第一条指令的源操作数访问遇到重复的存储体冲突，那么第二条指令很可能在第一个寄存器读取到（正确的）旧值之前将新值写入寄存器。防止这种 WAR 风险的一种方法是要求同一个 warp 中的指令按程序顺序离开操作数收集器到执行单元。Mishkin 等人 [2016] 探索了三种具有低硬件复杂度的潜在解决方案，并评估了它们对性能的影响。第一个方案是 *release-on-commit warpboard*，它允许每个 warp 最多执行一条指令。不出所料，他们发现这会对性能产生负面影响，在某些情况下性能几乎降低了两倍。他们的第二个方案是 *release-on-read warpboard*，它只允许每个 warp 一次一条指令在操作数收集器中收集操作数。这种方案导致他们研究的工作负载最多减慢 10%。最后，为了在操作数收集器中实现指令级并行，他们提出了一种 *bloomboard* 机制，该机制使用小型布隆过滤器来跟踪未完成的寄存器读取。与（错误地）允许 WAR 风险相比，这导致的影响不到百分之几。另外，Gray 进行的分析表明，NVIDIA 的 Maxwell GPU 引入了一种“读取依赖屏障”，它由特殊的“控制指令”管理，可用于避免某些指令的 WAR 风险（参见第 2.2.1 节）。

3.3.2 指令重放：处理结构性危险

GPU 流水线中存在许多导致结构性危险的潜在原因。例如，寄存器读取阶段可能会用尽操作数收集器单元。许多结构性危险的来源与内存系统有关，我们将在下一章中更详细地讨论这一点。通常，warp 执行的单个内存指令可能需要分解为多个单独的操作。每个单独的操作都可能在给定周期内充分利用流水线的一部分。

当一条指令在 GPU 流水线中遇到结构性危险时会发生什么？在单线程有序 CPU 流水线中，标准解决方案是暂停较新的指令，直到遇到暂停条件的指令可以取得进一步进展。这种方法在高度多线程吞吐量架构中可能不太可取，至少有两个原因。首先，考虑到寄存器文件的大小以及支持完整图形流水线所需的许多流水线阶段，分发暂停信号可能会影响关键路径。流水线暂停周期分布导致需要引入额外的缓冲来增加面积。其次，暂停一个 warp 中的指令可能会导致其他 warp 中的指令在其后暂停。如果这些指令不需要导致暂停的指令所需的资源，则吞吐量可能会受到影响。

为了避免这些问题，GPU 实现了一种指令重放形式。指令重放出现在一些 CPU 设计中，在推测性地将依赖指令调度到具有可变延迟的较早指令时，它被用作恢复机制。例如，负载可能在第一级缓存中命中或未命中，但以高时钟频率运行的 CPU 设计

frequency may pipeline first-level cache access over as many as four clock cycles. Some CPUs speculative wake up instructions depending upon a load so as to improve single threaded performance. In contrast, GPUs avoid speculation as it tends to waste energy and reduce throughput. Instead, instruction replay is used in GPUs to avoid clogging the pipeline and the circuit area and/or timing overheads resulting from stalling.

To implement instruction replay a GPU can hold instructions in the instruction buffer either until it is known that they have completed or all individual portions of the instruction have executed [Lindholm et al., 2015].

3.4 RESEARCH DIRECTIONS ON BRANCH DIVERGENCE

This section is based on Wilson Fung's Ph.D. dissertation [Fung, 2015].

Ideally, threads within the same warp execute through the same control flow path, so that the GPU can execute them in lockstep on SIMD hardware. Given the autonomy of the threads, a warp may encounter a *branch divergence* when its threads diverge to different targets at a data-dependent branch. Modern GPUs contains special hardware to handle branch divergence in a warp. Section 3.1.1 describes the baseline SIMT stack, which is used by the baseline GPU architecture in this book. The baseline SIMT stack handles branch divergence in a warp by serializing the execution of the different targets. While the baseline SIMT stack handles the branch divergence correctly for most existing GPU applications, it has the following deficiencies.

Lower SIMD Efficiency In the presence of branch divergence, the baseline SIMT stack serializes the execution of each branch target. As each target is executed, the SIMT stack only activates the subset of scalar threads running the target. This causes some lanes in the SIMD hardware to be idle, lowering the overall *SIMD efficiency*.

Needless Serialization The serialized execution of each branch target by the baseline SIMT stack is not required for functional correctness. The GPU programming model does not impose any implicit data dependency between scalar threads within a warp—they have to communicate explicitly via Shared Memory and barrier. The GPU can interleave the execution of all branch targets of a diverged warp to make use of idle cycles in the SIMD hardware.

Inadequate MIMD Abstraction By forcing diverged warps to reconverge at a compiler-defined reconvergence point, the baseline SIMT stack implicitly imposes a warp-wide synchronization point at each reconvergence point. This works for many existing GPU applications. However, this implicit synchronization may interact pathologically with other user-implemented synchronization mechanisms, such as fine-grained locks, causing the warp to deadlock. The compiler-defined reconvergence point also does not account for the control-flow divergence introduced by system-level constructs such as exceptions and interrupts.

频率可以在多达四个时钟周期内流水线一级缓存访问。一些 CPU 根据负载推测唤醒指令，以提高单线程性能。相比之下，GPU 避免推测，因为它往往会浪费能源并降低吞吐量。相反，GPU 中使用指令重放来避免堵塞流水线和电路面积和/或因停顿而导致的时序开销。

为了实现指令重放，GPU 可以将指令保存在指令缓冲区中，直到知道它们已经完成或指令的所有各个部分都已执行 [Lindholm et al., 2015]。

3.4 分支分歧的研究方向

This section is based on Wilson Fung's Ph.D. dissertation [Fung, 2015].

理想情况下，同一个 Warp 中的线程通过相同的控制流路径执行，这样 GPU 就可以在 SIMD 硬件上同步执行它们。考虑到线程的自主性，当 Warp 的线程在数据相关分支处分叉到不同目标时，该 Warp 可能会遇到 *branch divergence*。现代 GPU 包含特殊硬件来处理 Warp 中的分支分叉。第 3.1.1 节介绍了本书中的基本 GPU 架构所使用的基本 SIMT 堆栈。基本 SIMT 堆栈通过序列化不同目标的执行来处理 Warp 中的分支分叉。虽然基本 SIMT 堆栈可以正确处理大多数现有 GPU 应用程序的分支分叉，但它存在以下缺陷。

较低的 SIMD 效率 在存在分支发散的情况下，基线 SIMT 堆栈会序列化每个分支目标的执行。在执行每个目标时，SIMT 堆栈仅激活运行目标的标量线程子集。这会导致 SIMD 硬件中的某些通道处于空闲状态，从而降低整体 *SIMD efficiency*。

无需序列化 基线 SIMT 堆栈对每个分支目标的序列化执行对于功能正确性而言并非必需的。GPU 编程模型不会在 warp 中的标量线程之间施加任何隐式数据依赖性 - 它们必须通过共享内存和屏障进行显式通信。GPU 可以交错执行分叉 warp 的所有分支目标，以利用 SIMD 硬件中的空闲周期。

MIMD 抽象不足 通过强制发散的 Warp 在编译器定义的重新收敛点重新收敛，基线 SIMT 堆栈在每个重新收敛点隐式地施加了一个 Warp 范围的同步点。这适用于许多现有的 GPU 应用程序。但是，这种隐式同步可能会与其他用户实现的同步机制（例如细粒度锁）发生不良交互，从而导致 Warp 死锁。编译器定义的重新收敛点也没有考虑由系统级构造（例如异常和中断）引入的控制流发散。

Area Cost While the area requirement of the baseline SIMT stack for each warp is only 32 × 64 bits (or as low as 6 × 64 bits), the area scales with the number of in-flight warps in the GPU. In typical GPU applications where branch divergence is rare, the SIMT stack occupies area that can otherwise be used to boos the application throughput in other ways (e.g., large cache, more ALU units, etc.).

Both the industry and academia has proposed alternatives to address the above deficiencies. The various proposals can be classified into the following categories: warp compaction, intra-warp divergent path management, adding MIMD capability and complexity reduction. Some proposals contain improvements that capture aspects from multiple categories, and thus are mentioned multiple times.

3.4.1 WARP COMPACTION

With GPU implementing fine-grained multi-threading to tolerate long-memory access latency, there are many warps in each SIMT core, hundreds to thousands of scalar threads in total. Since these warps are usually running the same compute kernel, they are likely to follow the same execution path, and encounter branch divergence at the same set of data-dependent branches. Consequently, each target of a divergent branch is probably executed by a large number of threads, but these threads are scattered among multiple static warps, with each warp handling the divergence individually.

In this section, we summarize a series of research that exploits this observation to improve the performance of GPU applications that suffer from branch divergence. Proposals in this series all involve novel hardware mechanisms to *compact* threads from different *static warps* into new *dynamic warps* to improve the overall SIMD efficient of these divergent GPU applications. Here, static warps refers to the warps that are formed by the GPU hardware when the scalar threads are spawned from a kernel launch. In our baseline GPU architecture, this arrangement is fixed throughout the execution of the warp. The arrangement of scalar threads into static warps is an arbitrary grouping imposed by the GPU hardware that is largely invisible to the programming model.

Dynamic Warp Formation. Dynamic warp formation (DWF) [Fung et al., 2007, Fung et al., 2009] exploits this observation by rearranging these scattered threads that execute the same instruction into new dynamic warps. At a divergent branch, DWF can boost the overall SIMD efficiency of an application by compacting threads scattered among multiple diverged static warps into fewer non-divergent dynamic warps. In this way, DWF can capture a significant fraction of the benefits of MIMD hardware on SIMD hardware. However, DWF requires warps to encounter the same divergent branch within a short time window. This timing-dependent nature of DWF makes it very sensitive to the warp scheduling policy.

The follow-up work by Fung and Aamodt [2011] identifies two major of performance pathologies for DWF: (1) a greedy scheduling policy can starve some threads, leading to a SIMD

面积成本 虽然每个 Warp 的基线 SIMT 堆栈的面积要求仅为 32 × 64 位（或低至 6 × 64 位），但面积会随着 GPU 中正在运行的 Warp 数量而变化。在分支发散很少见的典型 GPU 应用中，SIMT 堆栈占用的面积原本可用于以其他方式提高应用程序吞吐量（例如，大缓存、更多 ALU 单元等）。

业界和学术界都提出了一些替代方案来解决上述缺陷。各种提案可以分为以下几类：warp 压缩、warp 内发散路径管理、添加 MIMD 功能和降低复杂性。一些提案包含涵盖多个类别的改进，因此被多次提及。

3.4.1 经纱压缩

由于 GPU 实现了细粒度多线程来容忍较长的内存访问延迟，因此每个 SIMT 核心中都有许多 Warp，总共有数百到数千个标量线程。由于这些 Warp 通常运行相同的计算内核，因此它们很可能遵循相同的执行路径，并在同一组数据相关分支中遇到分支分歧。因此，分歧分支的每个目标可能由大量线程执行，但这些线程分散在多个静态 Warp 中，每个 Warp 单独处理分歧。

在本节中，我们总结了一系列利用这一观察结果来改善受分支发散影响的 GPU 应用程序性能的研究。本系列中的提案都涉及新颖的硬件机制，将来自不同 *static warps* 的 *compact* 线程转换为新的 *dynamic warps*，以提高这些不同 GPU 应用程序的整体 SIMD 效率。此处，静态 warp 是指从内核启动生成标量线程时由 GPU 硬件形成的 warp。在我们的基准 GPU 架构中，这种安排在整个 warp 执行过程中都是固定的。标量线程到静态 warp 的排列是 GPU 硬件强加的任意分组，对于编程模型来说基本上是不可见的。

动态 Warp 形成。动态 Warp 形成 (DWF) [Fung et al., 2007, Fung et al., 2009] 利用这一观察结果，将执行相同指令的分散线程重新排列为新的动态 Warp。在发散分支中，DWF 可以将分散在多个发散静态 Warp 中的线程压缩为更少的非发散动态 Warp，从而提高应用程序的整体 SIMD 效率。通过这种方式，DWF 可以在 SIMD 硬件上捕获 MIMD 硬件的很大一部分优势。但是，DWF 要求 Warp 在短时间窗口内遇到相同的发散分支。DWF 的这种时间依赖性使其对 Warp 调度策略非常敏感。

Fung 和 Aamodt [2011] 的后续工作确定了 DWF 的两种主要性能病症：（1）贪婪的调度策略可能会使某些线程挨饿，从而导致 SIMD

efficiency reduction; and (2) Thread regrouping in DWF increases non-coalesced memory accesses and shared memory bank conflicts. These pathologies cause DWF to slowdown many existing GPU applications. Moreover, applications that rely on implicit synchronization in a static warp execute incorrectly with DWF.

The above pathologies can be partially addressed with an improved scheduling policy that effectively separates the compute kernel into two sets of regions: divergent and non-divergent (coherent) regions. The divergent regions benefit significantly from DWF, whereas the coherent regions are free of branch divergence but are prone to the DWF pathologies. We found that the impact of the DWF pathologies can be significantly reduced by forcing DWF to rearrange scalar threads back to their static warps in the coherent regions.

Thread Block Compaction. Thread block compaction (TBC) [Fung and Aamodt, 2011] builds upon this insight with the observation that rearrangement of threads into new dynamic warps continually does not yield additional benefit. Instead, the rearrangement, or *compaction*, only needs to happen right after a divergent branch, the start of a divergent region, and before its reconvergence point, the start to a coherent region. We note the existing per-warp SIMT stack (described in Chapter 3.1.1) implicitly synchronizes threads diverged to different execution paths at the reconvergence point of the divergent branch, merging these diverged threads back into a static warp before executing a coherent region. TBC extends the SIMT stack to encompass all warps executing in the same core, forcing them to synchronize and compact at divergent branches and reconvergence points to achieve robust DWF performance benefits. However, synchronizing all the warps within a core at each divergent branch for compaction can greatly reduce the available thread-level parallelism (TLP). GPU architectures rely on the abundance of TLP to tolerate pipeline and memory latency.

TBC settles for a compromise between SIMD efficiency and TLP availability by restricting compaction to only occur within a *thread block*. GPU applications usually execute multiple thread blocks concurrently on a single core to overlap the synchronization and memory latency. TBC leverages this software optimization to overlap the compaction overhead at divergent branches—when warps in one thread block synchronize for compaction at a divergent branch, warps in other thread blocks can keep the hardware busy. It extends the per-warp SIMT stack to encompass warps in a thread block. The warp scheduling logic uses this thread-block-wide SIMT stack to determine when warps in a thread block should synchronizes and be compacted into a new sets of warps. The result is a far more robust and simple mechanism that captures much of the benefits of DWF, without the pathological behaviors.

Large Warp Microarchitecture. The large warp microarchitecture [Narasiman et al., 2011] extends the SIMT stack, similar to TBC, to manage the reconvergence of a group of warps. However, instead of restricting the compaction at branches and reconvergence points. LWM requires warps within the group to execute in complete lockstep, so that it can compact the group at every instruction. This reduces the available TLP even more so than TBC, but allows

效率降低；（2）DWF 中的线程重组增加了非合并内存访问和共享内存库冲突。这些缺陷导致 DWF 减慢了许多现有的 GPU 应用程序的速度。此外，依赖于静态 warp 中的隐式同步的应用程序在 DWF 中执行不正确。

上述问题可以通过改进的调度策略得到部分解决，该策略有效地将计算内核分为两组区域：发散区域和非发散（连贯）区域。发散区域从 DWF 中受益匪浅，而连贯区域没有分支发散，但容易出现 DWF 问题。我们发现，通过强制 DWF 将标量线程重新排列回连贯区域中的静态扭曲，可以显著减少 DWF 问题的影响。

线程块压缩。线程块压缩 (TBC) [Fung and Aamodt, 2011] 建立在此见解之上，其观察结果是不断将线程重新排列到新的动态 Warp 中不会产生额外的好处。相反，重新排列或 *compaction* 只需要在发散分支之后、发散区域的开始处以及其重新收敛点之前、连贯区域的开始处进行。我们注意到，现有的每个 Warp SIMT 堆栈（在第 3.1.1 章中描述）隐式同步发散到不同执行路径的线程，在执行连贯区域之前将这些发散的线程合并回静态 Warp。 TBC 扩展了 SIMT 堆栈以涵盖在同一核心中执行的所有 warp，迫使它们在不同的分支和重新收敛点处同步和压缩，以实现强大的 DWF 性能优势。但是，在每个不同的分支处同步核心内的所有 warp 以进行压缩会大大减少可用的线程级并行 (TLP)。GPU 架构依靠丰富的 TLP 来容忍管道和内存延迟。

TBC 通过将压缩限制在仅发生在 *thread block* 内来在 SIMD 效率和 TLP 可用性之间达成妥协。GPU 应用程序通常在单个核心上同时执行多个线程块以重叠同步和内存延迟。TBC 利用这种软件优化来重叠不同分支处的压缩开销 - 当一个线程块中的 warp 同步以在不同分支处进行压缩时，其他线程块中的 warp 可以让硬件保持繁忙。它扩展了每个 warp 的 SIMT 堆栈以包含线程块中的 warp。warp 调度逻辑使用这个线程块范围的 SIMT 堆栈来确定何时应同步线程块中的 warp 并将其压缩成新的 warp 集。结果是一种更加健壮和简单的机制，它捕获了 DWF 的许多好处，而没有病态行为。

大型 Warp 微架构。大型 Warp 微架构 [Narasiman et al., 2011] 扩展了 SIMT 堆栈，类似于 TBC，用于管理一组 Warp 的重新收敛。但是，LWM 不会限制分支和重新收敛点处的压缩。它要求组内的 Warp 完全同步执行，以便它可以在每条指令中压缩组。这比 TBC 更能减少可用的 TLP，但允许

LWM to perform compaction with predicated instructions as well as unconditional jumps. Similar to TBC, LWM splits warps running on the same core into multiple groups, and restricts compaction to occur only within a group. It also opts for a more complex scoreboard microarchitecture that tracks register dependency at thread-granularity. This allows some warps in the group to execute slightly ahead of others compensate the lost TLP due to lockstep execution.

Compaction-Adequacy Predictor. Rhu and Erez [2012] extend TBC with a compaction-adequacy predictor (CAPRI). The predictor identifies the effectiveness of compacting threads into few warps at each branch, and only synchronizes the threads at branches where the compaction is predicted to yield a benefit. This reclaims the TLP lost due to non-beneficial stall and compaction with TBC. Rhu and Erez [2012] also show that a simple history-based predictor similar to a single-level branch predictor is sufficient to achieve high accuracy.

Intra-Warp Compaction. Vaidya et al. [2013] propose a low-complexity compaction technique that benefits wide SIMD execution groups that executes multiple cycle on narrower hardware units. Their basic technique divides a single execution group into multiple subgroups that match the hardware width. SIMD execution group that suffers from divergence can run faster on the narrow hardware by skipping subgroups that are completely idle. To create more completely idle subgroups, they propose a swizzle mechanism that compacts elements into fewer subgroups at divergence.

Simultaneous Warp Interweaving. Brunie et al. [2012] propose simultaneous branch and warp interweaving (SBI and SWI). They extend the GPU SIMT front-end to support issuing two different instructions per cycle. They compensate this increased complexity by widening the warp to twice its original size. SWI co-issues an instruction from a warp suffering from divergence with instructions from another diverged warp to fill the gaps left by branch divergence.

Impact on Register File Microarchitecture

To avoid introducing extra communication traffics between SIMT cores, the hardware compaction proposals usually takes place locally within a SIMT core. Since the compacted threads all located on the same core sharing the same register file, it is possible to perform compaction without moving their architectural states with a more flexible register file design [Fung et al., 2007].

As discussed earlier in this chapter, GPU register files are implemented with large single-ported SRAM banks to maximize their area efficiency. Registers for threads in the same warp are stored in consecutive regions in the same SRAM bank, so that they can be accessed together via a single wide port. This allows high bandwidth register file access while amortizing register file access control hardware. Hardware warp compaction creates dynamic warps that may not obey this arrangement of registers. Fung et al. [2007] propose a more flexible register file design featuring SRAM banks with narrow ports. This design has more SRAM banks to maintain the same bandwidth.

LWM 使用预测指令和无条件跳转执行压缩。与 TBC 类似，LWM 将在同一核心上运行的 Warp 拆分为多个组，并限制压缩仅在一个组内进行。它还选择了一种更复杂的记分板微架构，以线程粒度跟踪寄存器依赖性。这允许组中的某些 Warp 略微领先于其他 Warp 执行，以补偿由于锁步执行而丢失的 TLP。

压缩充分性预测器。Rhu 和 Erez [2012] 使用压缩充分性预测器 (CAPRI) 扩展了 TBC。该预测器确定将线程压缩为每个分支上的几个 warp 的有效性，并且仅在预测压缩会产生效益的分支上同步线程。这可以回收由于无益的停滞和使用 TBC 压缩而丢失的 TLP。Rhu 和 Erez [2012] 还表明，类似于单级分支预测器的简单基于历史的预测器足以实现高精度。

Warp 内部压缩。Vaidya 等人 [2013] 提出了一种低复杂度压缩技术，该技术有利于在较窄的硬件单元上执行多个周期的宽 SIMD 执行组。他们的基本技术是将单个执行组划分为与硬件宽度匹配的多个子组。通过跳过完全空闲的子组，受发散影响的 SIMD 执行组可以在窄硬件上运行得更快。为了创建更多完全空闲的子组，他们提出了一种 swizzle 机制，在发散时将元素压缩成更少的子组。

同时进行 Warp 交织。Brunie 等人 [2012] 提出了同时进行分支和 Warp 交织 (SBI 和 SWI)。他们扩展了 GPU SIMT 前端，以支持每个周期发出两个不同的指令。他们通过将 Warp 加宽到其原始大小的两倍来补偿这种增加的复杂性。SWI 同时发出来自出现发散的 Warp 的指令和来自另一个发散 Warp 的指令，以填补分支发散留下的空白。

对寄存器文件微架构的影响
为了避免在 SIMT 核心之间引入额外的通信流量，硬件压缩方案通常在 SIMT 核心内本地进行。由于压缩线程都位于共享相同寄存器文件的同一核心上，因此可以使用更灵活的寄存器文件设计执行压缩而无需移动其架构状态 [Fung et al., 2007]。

如本章前面所述，GPU 寄存器文件采用大型单端口 SRAM 组来实现，以最大程度地提高其面积效率。同一 Warp 中的线程的寄存器存储在同一 SRAM 组中的连续区域中，因此可以通过单个宽端口一起访问它们。这允许高带宽寄存器文件访问，同时分摊寄存器文件访问控制硬件。硬件 Warp 压缩会创建可能不遵循这种寄存器排列的动态 Warp。Fung 等人 [2007] 提出了一种更灵活的寄存器文件设计，其特点是 SRAM 组具有窄端口。此设计具有更多 SRAM 组来维持相同的带宽。

Dynamic Micro-Kernels. Steffen and Zambreno [2010] improved SIMD efficiency of ray tracing on GPUs with *dynamic micro-kernels*. The programmer is given primitives to break iterations in a data-dependent loop into successive micro-kernel launches. This decomposition by itself does not improve parallelism, because each iteration depends on data from the previous iteration. Instead, the launch mechanism improves the load imbalance between different threads in the same core by compacting the remaining active threads into few warps. It also differs from the rest of the hardware warp compaction techniques in that the compaction migrates the threads with their architectural states, using the per-core scratchpad memory as a staging area.

Section 3.4.1 summarizes a series of research that implements warp compaction in software, which does not require the more flexible register file design. Instead, these proposals introduces extra memory traffic to relocate threads from one SIMT core to another.

Warp Compaction in Software

On existing GPUs, one way to improve SIMD efficiency of an application is through software warp compaction—using software to group threads/work items according to their control flow behavior. The regrouping involves moving the thread and its private data in memory, potentially introducing a significant memory bandwidth overhead. Below we highlight several works on software compaction techniques.

Conditional streams [Kapasi et al., 2000] apply this concept to stream computing. It splits a compute kernel for stream processors with potentially divergent control flow into multiple kernels. At a divergent branch, a kernel splits its data stream into multiple streams according to branch outcome of each data element. Each stream is then processed by a separate kernel, and merges back at the end of the control flow divergence.

Billeter et al. [2009] proposed to use a parallel prefix sum to implement SIMD *stream compaction*. The stream compaction reorganizes streams of elements with assorted tasks into compact substreams of identical tasks. This implementation leverages the access flexibility of the GPU on-chip scratchpad to achieve high efficiency. Hoberock et al. [2009] proposed a deferred shading technique for ray tracing that uses stream compaction to improve the SIMD efficiency of pixel shading in a complex scene with many material classes. Each material class requires its unique computation. A pixel shader combining the computation for every material class runs inefficiently on GPUs. Stream compaction groups the rays hitting objects with similar material classes, allowing the GPU SIMD hardware to execute the shader for these pixels efficiently.

Zhang et al. [2010] proposed a runtime system that remaps thread into different warps on the fly to improve SIMD efficiency as well as memory access spatial locality. The runtime system features a pipelined system, with the CPU performing the on-the-fly remapping and the GPU performing computations on the remapped data/threads.

Khorasani et al. [2015] proposed *Collective Context Collection* (CCC), a compiler technique that transforms a given GPU compute kernel with potential branch divergence penalty

动态微内核。Steeffen 和 Zambreno [2010] 使用 *dynamic micro-kernels* 提高了 GPU 上光线追踪的 SIMD 效率。程序员可以使用原语将数据相关循环中的迭代分解为连续的微内核启动。这种分解本身并不能提高并行性，因为每次迭代都依赖于前一次迭代的数据。相反，启动机制通过将剩余的活动线程压缩为几个 warp 来改善同一核心中不同线程之间的负载不平衡。它与其他硬件 warp 压缩技术的不同之处在于，压缩使用每个核心的暂存器内存作为暂存区，将线程与其架构状态一起迁移。

3.4.1 节总结了一系列在软件中实现 warp 压缩的研究，这些研究不需要更灵活的寄存器文件设计。相反，这些提案引入了额外的内存流量来将线程从一个 SIMT 核心重新定位到另一个 SIMT 核心。

软件中的 Warp Compaction

在现有的 GPU 上，提高应用程序 SIMD 效率的一种方法是通过软件 Warp 压缩 - 使用软件根据线程/工作项的控制流行为对其进行分组。重新分组涉及在内存中移动线程及其私有数据，这可能会带来很大的内存带宽开销。下面我们重点介绍几项关于软件压缩技术的工作。

条件流 [Kapasi et al., 2000] 将这一概念应用于流计算。它将具有潜在发散控制流的流处理器的计算内核拆分为多个内核。在发散分支处，内核根据每个数据元素的分支结果将其数据流拆分为多个流。然后，每个流由单独的内核处理，并在控制流发散结束时合并回来。

Billeter 等人 [2009] 提出使用并行前缀和来实现 SIMD *stream compaction*。流压缩将具有各种任务的元素流重新组织为相同任务的紧凑子流。此实现利用 GPU 片上暂存器的访问灵活性来实现高效率。Hoberock 等人 [2009] 提出了一种用于光线追踪的延迟着色技术，该技术使用流压缩来提高具有许多材质类的复杂场景中像素着色的 SIMD 效率。每种材质类都需要其独特的计算。结合每种材质类的计算的像素着色器在 GPU 上的运行效率低下。流压缩将击中具有相似材质类的对象的光线分组，从而允许 GPU SIMD 硬件高效地执行这些像素的着色器。

Zhang 等人 [2010] 提出了一种运行时系统，该系统可以动态地将线程重新映射到不同的 warp 中，以提高 SIMD 效率以及内存访问空间局部性。该运行时系统采用流水线系统，其中 CPU 执行动态重新映射，GPU 对重新映射的数据/线程执行计算。

Khorasani 等人 [2015] 提出了 *Collective Context Collection* (CCC)，这是一种编译器技术，它转换给定的 GPU 计算内核，并产生潜在的分支发散惩罚

to improve its SIMD efficiency on existing GPUs. CCC focuses on compute kernels in which each thread performs an irregular amount of computation at each step, such as a breath-first-search through a irregular graph. Instead of assigning one node (or task in other applications) per thread, CCC first transforms the compute kernel so that each thread processes multiple nodes, with the node to warp (note: not thread) assignment determined ahead of the kernel launch. CCC then transforms the compute kernel so that each thread in a warp can offload the context of a task to a warp-specific stack stored in the shared memory. A warp that experiences low SIMD efficiency at its current set of tasks can offload the tasks to the stack, and uses these offloaded tasks to fill up threads that went idle they process a later set of tasks. In effect, CCC performs "warp compaction" by grouping the tasks from multiple warps into a fewer set of warps, and then compacting the divergent tasks into fewer iterations within each warp via the warp-specific stack stored in the fast, on-chip, shared memory.

Impacts of Thread Assignment within a Warp

In the baseline GPU architecture studied throughout this book, threads with consecutive thread IDs are statically fused together to form warps. Little academic work has gone into the static assignment of threads to warps or lanes in a warp. This default sequential mapping works well for most workloads since adjacent threads tend to access adjacent data, improving memory coalescing. However, some research has looked at alternatives.

SIMD Lane Permutation. Rhu and Erez [2013b] make the observation that the sequential mapping of thread IDs to consecutive threads in a warp is suboptimal for warp compaction techniques described earlier in this section. A key limitation of most warp compaction and formation work is that when threads are assigned to a new warp, they cannot be assigned to a different lane, or else their register file state would have to be moved to a different lane in the vector register. Rhu and Erez observe that the structure of a program biases certain control flow paths to certain SIMD lanes. This biasing makes it more difficult to achieve compaction, since threads that take the same path tend to be in the same lane, preventing those threads from being merged together. Rhu and Erez propose several different thread mapping permutations that remove these programmatic biases and significantly improve the rate of compaction.

Intra-warp Cycle Compaction. Vaidya et al. [2013] exploit the fact that the width of the SIMD datapath does not always equal the warp width. For example, in NVI [2009], the SIMD width is 16, but the warp size is 32. This means a 32-thread warp is executed over 2 core cycles. Vaidya et al. [2013] observe that when divergence occurs, if a sequential SIMD-worth of threads are masked off for an instruction, then the instruction can be issued in only one cycle, skipping the masked off lanes. They call their technique cycle compression. However, if the masked-off threads are not contiguous, the basic technique does not yield any performance improvement. To address this, they propose a swizzled cycle compression that re-arranges which threads are in which lanes in order to create more opportunities for cycle compression.

以提高其在现有 GPU 上的 SIMD 效率。CCC 专注于计算内核，其中每个线程在每个步骤中执行不规则量的计算，例如通过不规则图进行广度优先搜索。CCC 不会为每个线程分配一个节点（或其他应用程序中的任务），而是首先转换计算内核，以便每个线程处理多个节点，并在内核启动之前确定节点到 warp（注意：不是线程）的分配。然后，CCC 转换计算内核，以便 warp 中的每个线程都可以将任务的上下文卸载到存储在共享内存中的特定于 warp 的堆栈。在当前任务集中经历低 SIMD 效率的 warp 可以将任务卸载到堆栈，并使用这些卸载的任务来填充空闲的线程，然后处理后面的一组任务。实际上，CCC 通过将多个 warp 中的任务分组为更少的一组 warp，然后通过存储在快速片上共享内存中的特定于 warp 的堆栈将分散的任务压缩为每个 warp 中更少的迭代，从而执行“ warp 压缩 ”。

Warp 中线程分配的影响

在本书研究的基准 GPU 架构中，具有连续线程 ID 的线程会静态融合在一起以形成 Warp。很少有学术研究将线程静态分配给 Warp 或 Warp 中的通道。这种默认顺序映射适用于大多数工作负载，因为相邻线程倾向于访问相邻数据，从而改善内存合并。然而，一些研究已经研究了替代方案。

SIMD 通道排列。Rhu 和 Erez [2013b] 观察到，对于本节前面描述的 Warp 压缩技术，将线程 ID 顺序映射到 Warp 中的连续线程并非最佳选择。大多数 Warp 压缩和形成工作的一个关键限制是，当线程被分配到新的 Warp 时，它们不能被分配到不同的通道，否则它们的寄存器文件状态必须移动到矢量寄存器中的不同通道。Rhu 和 Erez 观察到，程序的结构将某些控制流路径偏向某些 SIMD 通道。这种偏向使得实现压缩变得更加困难，因为采用相同路径的线程往往位于同一通道中，从而阻止这些线程合并在一起。Rhu 和 Erez 提出了几种不同的线程映射排列，以消除这些程序偏差并显著提高压缩率。

内部 Warp 周期压缩。Vaidya 等人 [2013] 利用了 SIMD 数据路径的宽度并不总是等于 Warp 宽度这一事实。例如，在 NVI [2009] 中，SIMD 宽度为 16，但 Warp 大小为 32。这意味着 32 线程 Warp 将在 2 个核心周期内执行。Vaidya 等人 [2013] 观察到，当出现分歧时，如果为一条指令屏蔽了连续的 SIMD 线程，则该指令可以在一个周期内发出，跳过屏蔽的关闭通道。他们将这种技术称为周期压缩。但是，如果屏蔽的关闭线程不连续，则基本技术不会带来任何性能改进。为了解决这个问题，他们提出了一种混合周期压缩，重新排列哪些线程位于哪些通道中，以创造更多的周期压缩机会。

Warp Scalarization. Other work, such as that by [Yang et al., 2014], argues that the SIMT programming model is inefficient when the threads with a warp operate on the same data. A number of solutions propose including a scalar unit in the pipeline for work that the compiler or programmer can identify as scalar a-priori. AMD's Graphics Core Next (GCN) architecture includes a scalar pipeline for this purpose. See Section 3.5 for more details.

3.4.2 INTRA-WARP DIVERGENT PATH MANAGEMENT

While a SIMT stack with immediate post-dominator reconvergence points can handle branch divergence with arbitrary control flow, it can be further improved in various aspects.

1. Threads diverged to different branch targets of a diverged warp can interleave their execution to make use of idle cycles in the SIMD hardware.

2. While the immediate post-dominator of a divergent branch is the definite convergence point, threads diverged to different branch targets may be able to converge before the immediate post-dominator of the divergent branch.

The following subsections highlight several works that attempt to improve the SIMT stack in these two aspects.

Multi-Path Parallelism

When a warp diverges at a branch, the threads are split into multiple groups, called *warp-splits*. Each warp-split consists of threads following the same branch target. In the baseline, *single path*, SIMT stack, warp-splits from the same warp are executed one-by-one, until the warp-split reaches its reconvergence point. This serialization lends itself to a relatively simple hardware implementation, but is not necessary for functional correctness. Threads in the warp have independent registers, and communicate among themselves explicitly via memory operations and synchronization operations such as barriers. In other words, every warp-split from the same warp may execute in *parallel*. We call this mode of execution *multi-path execution mode*.

While the different warp-splits may not execute on the same hardware at the same cycle (afterall, they runs different instructions), they may interleave their execution on the same hardware just like multiple warps interleaving their execution on the same datapath. In this way, multi-path execution mode boosts thread-level-parallelism (TLP) available in an application to tolerate memory access latency. Even though the SIMD efficiency is not improved, multi-path execution boosts the overall performance of memory-bound applications, where the SIMT cores has plenty of idle cycles to be filled with useful work.

Example 3.1 shows a short compute kernel that may benefit from multi-path execution. In this example, the code paths in both branch targets contain a load from memory. In the single-path SIMT stack, each of block B and C are executed serially until the corresponding warp-split reaches block D (the reconvergence point), even when the warp-split stalled waiting for data from memory. This stalls the entire warp, introducing idle cycles in the datapath, as shown

Warp 标量化。其他研究（例如 [Yang et al., 2014] 的研究）认为，当具有 Warp 的线程对相同数据进行操作时，SIMT 编程模型效率低下。许多解决方案建议在管道中包含一个标量单元，以便编译器或程序员可以先验地识别为标量。AMD 的 Graphics Core Next (GCN) 架构为此目的包含一个标量管道。有关更多详细信息，请参阅第 3.5 节。

3.4.2 经线内发散路径管理

虽然具有立即后支配器重新收敛点的 SIMT 堆栈可以处理具有任意控制流的分支发散，但它可以在各个方面得到进一步改进。

1. 发散到发散 warp 的不同分支目标的线程可以交错执行，以利用 SIMD 硬件中的空闲周期。
2. 虽然发散分支的直接后支配者是明确的收敛点，但是发散到不同分支目标的线程可能能够在发散分支的直接后支配者之前收敛。

以下小节重点介绍了几项尝试从这两个方面改进 SIMT 堆栈的工作。

多路径并行

当 Warp 在分支处发散时，线程会被分成多个组，称为 *warp-splits*。每个 Warp 分裂都由遵循相同分支目标的线程组成。在基线 *single path*、SIMT 堆栈中，来自同一 Warp 的 Warp 分裂会逐个执行，直到 Warp 分裂达到其重新收敛点。这种序列化适合相对简单的硬件实现，但对于功能正确性而言并非必需。Warp 中的线程具有独立的寄存器，并通过内存操作和同步操作（如屏障）明确地相互通信。换句话说，来自同一 Warp 的每个 Warp 分裂都可以在 *parallel* 中执行。我们将这种执行模式称为 *multi-path execution mode*。

虽然不同的 warp-split 可能无法在同一硬件上以同一周期执行（毕竟它们运行不同的指令），但它们可以在同一硬件上交错执行，就像多个 warp 在同一数据路径上交错执行一样。这样，多路径执行模式可以提高应用程序中可用的线程级并行性 (TLP)，以容忍内存访问延迟。即使 SIMD 效率没有提高，多路径执行也可以提高内存受限应用程序的整体性能，其中 SIMT 核心有大量空闲周期可以用于有用的工作。

示例 3.1 展示了一个可能受益于多路径执行的短计算内核。在此示例中，两个分支目标中的代码路径都包含来自内存的加载。在单路径 SIMT 堆栈中，块 B 和 C 中的每一个都按顺序执行，直到相应的 Warp 拆分到达块 D（重新收敛点），即使 Warp 拆分因等待来自内存的数据而停滞。这会停滞整个 Warp，从而在数据路径中引入空闲周期，如图所示

in Figure 3.18, that has to be filled by works from other warps. With multi-path execution, the warp-splits of block B and C can interleave their execution, eliminating these idle cycles introduced by memory accesses.

Algorithm 3.1 Example of multi-path parallelism with branch divergence.

```
X = data[i];          // block A
if( X > 3 )
  result = Y[i] * i;  // block B
else
  result = Z[i] + i;  // block C
return result;        // block D
```

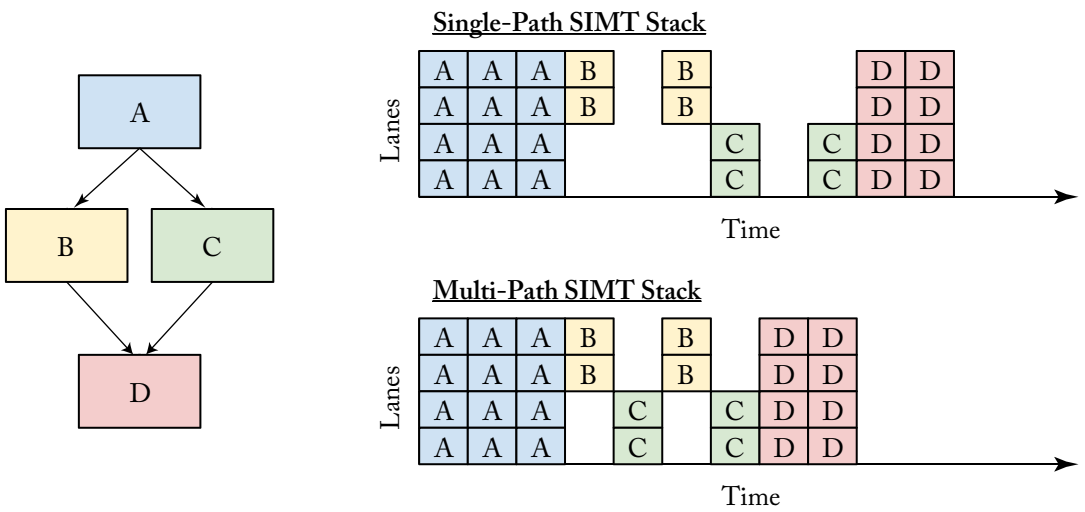


Figure 3.18: Multi-path execution at branch divergence.

Dynamic Warp Subdivision. Meng et al. [2010] propose that *dynamic warp subdivision* (DWS) is the first proposal to exploit on the TLP boost from multi-path execution. DWS extends the SIMT stack with a warp-split table to subdivide a diverged warp into concurrent warp-splits. The warp-splits, each executing a divergent branch target, can execute in parallel to reclaim hardware idleness due to memory accesses. Warp-splits are also created at memory divergences—when only a portion of the threads in a warp hit in the L1 data cache. Instead of waiting for all threads to obtain their data, DWS split the warp and allow the warp-split that hits in the cache to execute ahead, potentially prefetching data for those who have missed the cache.

在图 3.18 中，该位置必须由其他 Warp 中的工作填充。通过多路径执行，块 B 和 C 的 Warp 拆分可以交错执行，从而消除由内存访问引入的这些空闲周期。

Algorithm 3.1 Example of multi-path parallelism with branch divergence.

```
X = data[i];          // block A
if( X > 3 )
  result = Y[i] * i;  // block B
else
  result = Z[i] + i;  // block C
return result;        // block D
```

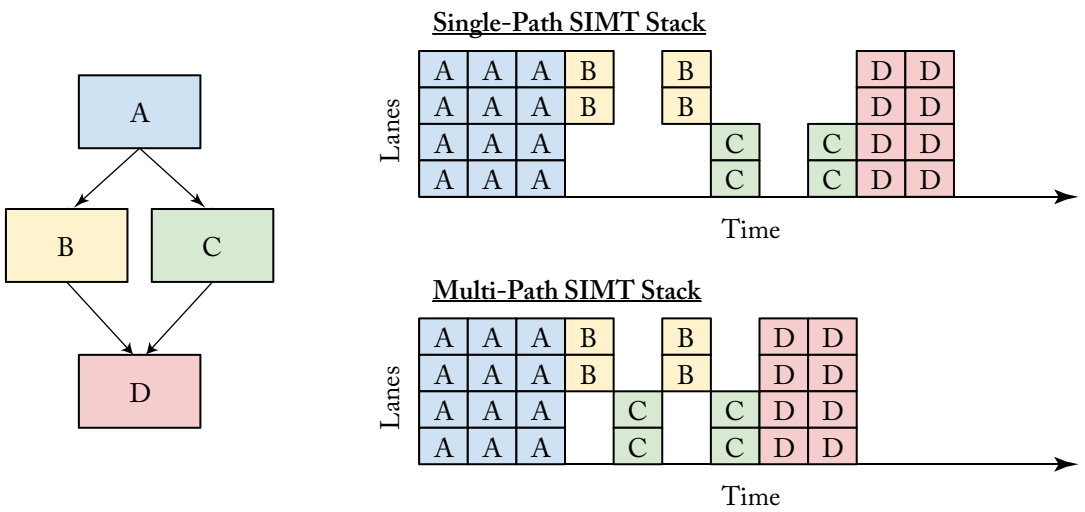


图 3.18：分支分歧处的多路径执行。

动态 Warp 细分。Meng 等人 [2010] 提出 *dynamic warp subdivision* (DWS) 是第一个利用多路径执行带来的 TLP 提升的方案。DWS 使用 Warp 拆分表扩展了 SIMT 堆栈，将发散的 Warp 细分为并发 Warp 拆分。每个 Warp 拆分都执行发散的分支目标，可以并行执行以回收由于内存访问而导致的硬件空闲。当 Warp 中只有一部分线程命中 L1 数据缓存时，内存发散时也会创建 Warp 拆分。DWS 不会等待所有线程获取其数据，而是拆分 Warp 并允许命中缓存的 Warp 拆分先执行，从而可能为未命中缓存的线程预取数据。

Dual-Path Execution. Rhu and Erez [2013a] propose the dual-path SIMT stack (DPS), which addresses some of the implementation shortcomings of DWS by restricting each warp to execute only two concurrent warp-splits. While this restriction enables DPS to capture a good portion of the full DWS advantage, it leads to a far simpler hardware design. DPS only requires extending the baseline SIMT stack with an extra set of PC and active mask to encode the extra warp-split. Only the two warp-splits at the top-of-stack entry of a warp are executed in parallel; every other warp-split in the same warp is paused until its entry reaches top-of-stack. DPS is also accompanied with an extension of the scoreboard to track the register dependency of each warp-split independently. This allows a dual-path execution model to achieve greater TLP than DWS with the baseline scoreboard.

Multi-Path Execution. ElTantaway et al. [2014] remove the dual-path limitation with a multi-path execution model (MPM). MPM replaces the SIMT stack with two tables: a warp-split table maintaining the set of warp-splits from the diverged warp, and a reconvergence table synchronizing all the warp-splits with the same reconvergence point.

At a divergent branch, one new entry is created in the reconvergence table, with the reconvergence point of the divergent branch (its immediate post-dominator). Multiple (usually two) entries is created in the warp-split table, one for each warp-split. Each warp-split entry maintains the current PC of the warp-split, its active mask, the reconvergence PC (RPC), and the R-index pointing to the corresponding entry in the reconvergent table. Every warp-split in the warp-split table is available for execution, until its PC == RPC. At this point, the corresponding reconvergence table entry is updated to reflect that threads from this warp-split has arrived at the reconvergence. When all pending threads have arrived at the reconvergence point, the reconvergence table entry is deallocated, and a new warp-split entry is created with the reconverged threads active, starting at the RPC.

MPM also extended the scoreboard to track the register dependency for each thread, without fully duplicating the scoreboard for each thread (which would render MPM impractical due to the significant area overhead in doing so). This is a crucial extension that allows the warp-splits to execute in a truly independent manner—without the extension, register dependency for one warp-split may be mistaken for the dependency of another warp-split from the same warp.

MPM is further extended with opportunistic early reconvergence, boosting the SIMD efficiency for unstructured control flow (see Section 3.4.2).

DWS, as well as other techniques discussed in this section, are orthogonal the warp compaction techniques discussed in Section 3.4.1. For example, the block-wide SIMT stack in TBC can be extended with DWS to boost the available TLP.

Better Convergence

The post-dominator (PDOM) stack-based reconvergence mechanism [Fung et al., 2007, Fung et al., 2009] uses reconvergence points identified using a unified algorithm rather than by translating control flow idioms in the source code into instructions [AMD, 2009, Coon and

双路径执行。Rhu 和 Erez [2013a] 提出了双路径 SIMT 堆栈 (DPS)，它通过限制每个 warp 仅执行两个并发的 warp-split 来解决 DWS 的一些实现缺陷。虽然这种限制使 DPS 能够充分利用 DWS 的大部分优势，但它导致硬件设计变得简单得多。DPS 只需要扩展基线 SIMT 堆栈，添加一组额外的 PC 和活动掩码来编码额外的 warp-split。只有 warp 堆栈顶部条目处的两个 warp-split 是并行执行的；同一 warp 中的每个其他 warp-split 都会暂停，直到其条目到达堆栈顶部。DPS 还附带记分板的扩展，以独立跟踪每个 warp-split 的寄存器依赖性。这使得双路径执行模型能够实现比使用基线记分板的 DWS 更大的 TLP。

多路径执行。ElTantaway 等人 [2014] 使用多路径执行模型 (MPM) 消除了双路径限制。MPM 用两个表替换了 SIMT 堆栈：一个 Warp-split 表，用于维护来自发散 Warp 的 Warp-split 集；一个重收敛表，用于同步所有 Warp-split 到相同的重收敛点。

在发散分支处，在重收敛表中创建一个新条目，其中包含发散分支的重收敛点（其直接后支配者）。在 Warp-split 表中创建多个（通常为两个）条目，每个 Warp-split 一个。每个 Warp-split 条目维护 Warp-split 的当前 PC、其活动掩码、重收敛 PC (RPC) 和指向重收敛表中相应条目的 R 索引。Warp-split 表中的每个 Warp-split 都可供执行，直到其 PC == RPC。此时，相应的重收敛表条目会更新，以反映来自此 Warp-split 的线程已到达重收敛。当所有待处理线程都已到达重新汇聚点时，重新汇聚表条目将被释放，并从 RPC 开始创建一个新的 warp-split 条目，其中重新汇聚的线程处于活动状态。

MPM 还扩展了记分板，以跟踪每个线程的寄存器依赖性，而无需完全复制每个线程的记分板（这会导致 MPM 不切实际，因为这样做会产生很大的面积开销）。这是一个至关重要的扩展，它允许 warp-split 以真正独立的方式执行 - 如果没有此扩展，一个 warp-split 的寄存器依赖性可能会被误认为是来自同一 warp 的另一个 warp-split 的依赖性。MPM 通过机会性早期重新收敛得到进一步扩展，从而提高了非结构化控制流的 SIMD 效率（参见第 3.4.2 节）。

DWS 以及本节讨论的其他技术与 3.4.1 节讨论的 warp 压缩技术是正交的。例如，TBC 中的块宽 SIMT 堆栈可以使用 DWS 进行扩展，以提升可用的 TLP。

更好的融合

基于后支配器（PDOM）堆栈的重新收敛机制[Fung et al.，2007，Fung et al.，2009]使用统一算法确定的重新收敛点，而不是将源代码中的控制流习语转换为指令[AMD，2009，Coon and

Lindholm, 2008, Levinthal and Porter, 1984]. The immediate post-dominator of a divergent branch selected as the reconvergence point is the earliest point in a program where the divergent threads are *guaranteed* to reconverge. In certain situations, threads can reconverge at an *earlier point*, and if hardware can exploit this, it would improve SIMD efficiency. We believe this observation motivates the inclusion of the `break` instruction in recent NVIDIA GPUs [Coon and Lindholm, 2008].

The code in Example 3.2 (from [Fung and Aamodt, 2011]) exhibits this earlier reconvergence. It results in the control flow graph in Figure 3.19 where edges are marked with the probability with which individual scalar threads follow that path. Block `F` is the immediate post-dominator of `A` and `C` since `F` is the first location where *all* paths starting at `A` (or `C`) coincide. In the baseline mechanism, when a warp diverges at `A`, the reconvergence point is set to `F`. However, the path from `C` to `D` is rarely followed and hence in *most* cases threads can reconverge earlier at `E`.

Algorithm 3.2 Example for branch reconvergence earlier than immediate post-dominator.

```
while (i < K) {
  X = data[i];       // block A
  if( X == 0 )
    result[i] = Y;   // block B
  else if ( X == 1 ) // block C
    break;           // block D
  i++;               // block E
}
return result[i];    // block F
```

Likely-Convergence Points. Fung and Aamodt [2011] propose extending the SIMT stack with *likely convergence points*. This extension adds two new fields to each SIMT stack entry: one for the PC of the likely convergence point (LPC) and the other (LPos), a pointer that records the stack position of a special likely convergence entry created when a branch has a likely convergence point that differs from the immediate post-dominator. The likely convergence point of each branch can be identified with either control flow analysis or profile information (potentially collected at runtime). The proposal by Fung and Aamodt [2011] restrictes likely convergence points to the closest enclosing backward-taken branch to capture the impact of "`break`" statements within loops [Coon and Lindholm, 2008].

When a warp diverges at a branch with a likely-convergence point, three entries are pushed onto the SIMT stack. The first entry, an LPC entry, is created for the likely convergence point of the branch. Two other entries for the taken and fall through of the branch are created as in the baseline mechanism. The LPC field in each of these other entries is populated with the likely convergence point of the divergent branch, the LPos field populated with the stack position of the LPC entry. The LPC entry has its RPC set to the immediate post-dominator, i.e., the definite

Lindholm，2008，Levinthal and Porter，1984]。选择作为重新收敛点的发散分支的直接后支配者是程序中发散线程重新收敛的最早点。在某些情况下，线程可以在 *earlier* 和 *point* 处重新收敛，如果硬件可以利用这一点，它将提高 SIMD 效率。我们相信这一观察促使在最近的 NVIDIA GPU 中加入 `break` 指令 [Coon and Lindholm，2008]。

示例 3.2（来自 [Fung and Aamodt, 2011]）中的代码展示了这种早期的重新收敛。它产生了图 3.19 中的控制流图，其中的边缘标有各个标量线程遵循该路径的概率。块 `F` 是 `A` 和 `C` 的直接后支配者，因为 `F` 是从 `A`（或 `C`）开始的 *all* 路径重合的第一个位置。在基线机制中，当 warp 在 `A` 处发散时，重新收敛点设置为 `F`。但是，从 `C` 到 `D` 的路径很少被遵循，因此在 *most* 情况下，线程可以在 `E` 处更早地重新收敛。

Algorithm 3.2 Example for branch reconvergence earlier than immediate post-dominator.

```
while (i < K) {
  X = data[i];       // block A
  if( X == 0 )
    result[i] = Y;   // block B
  else if ( X == 1 ) // block C
    break;           // block D
  i++;               // block E
}
return result[i];    // block F
```

可能收敛点。Fung 和 Aamodt [2011] 建议使用 *likely convergence points* 扩展 SIMT 堆栈。此扩展为每个 SIMT 堆栈条目添加了两个新字段：一个用于可能收敛点 (LPC) 的 PC，另一个用于 (LPos)，这是一个指针，用于记录当分支具有与直接后支配者不同的可能收敛点时创建的特殊可能收敛条目的堆栈位置。每个分支的可能收敛点可以通过控制流分析或配置文件信息（可能在运行时收集）来识别。Fung 和 Aamodt [2011] 的提议将可能收敛点限制为最近的封闭后向分支，以捕获循环内“`break`”语句的影响 [Coon and Lindholm，2008]。

当 Warp 在具有可能收敛点的分支处发散时，三个条目会被推送到 SIMT 堆栈上。第一个条目是 LPC 条目，是为分支的可能收敛点创建的。与基线机制一样，还会为分支的执行和失败创建另外两个条目。这些其他条目中的每个条目中的 LPC 字段都填充了发散分支的可能收敛点，LPos 字段填充了 LPC 条目的堆栈位置。LPC 条目的 RPC 设置为直接后支配者，即确定的

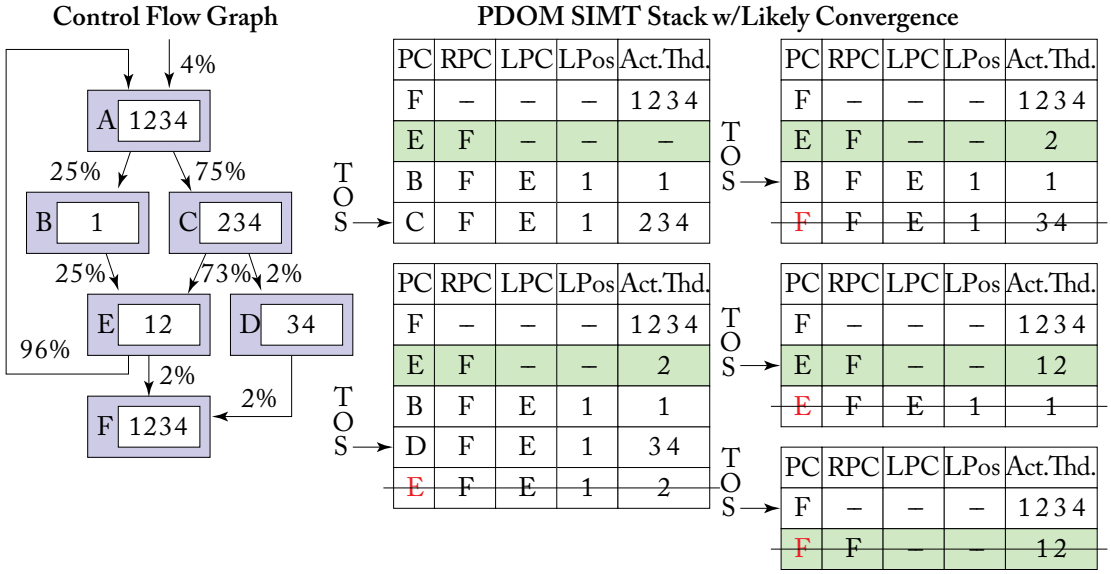


Figure 3.19: Earlier reconvergence points *before* the immediate post-dominator. Likely convergence point capture this earlier reconvergence at E.

reconvergence point, of the divergent branch, so that threads in this entry will reconverge to the definite reconvergence point.

As a warp executes with the top entry in the SIMT stack, it compares its PC against both the RPC field (just as it does with baseline SIMT stack), as well as the LPC field. If PC == LPC, the SIMT stack is popped, and threads in this popped entry is merged into the LPC entry. Otherwise, if PC == RPC, the SIMT stack is simply popped—the RPC entry already records these threads in its active mask. When the LPC entry reaches the top of the SIMT stack, it is executed just like any other SIMT stack entries, or popped directly if its active mask is empty.

Thread Frontiers. Diamos et al. [2011] depart from the SIMT stack altogether and instead propose to reconverged threads after divergence via *thread frontiers*. A compiler supporting thread frontiers sorts the basic blocks in a kernel according to their topological order. In this way, threads executing at an instruction at a higher PC can never jump to an instruction at a lower PC. Loops are handled by placing the loop exit at the end of the loop body. With this sorted code layout, a diverged warp will eventually reconverge by prioritizing threads with lower PCs (allowing them to catch up).

Compared to SIMT stacks with immediate post-dominator reconvergence, reconvergence via thread frontiers yields higher SIMD efficiency for applications with unstructured control flow. The evaluation semantics of multi-expression conditional statements and the use of exceptions can both generate code with unstructured control flow. SIMT stacks extended with likely

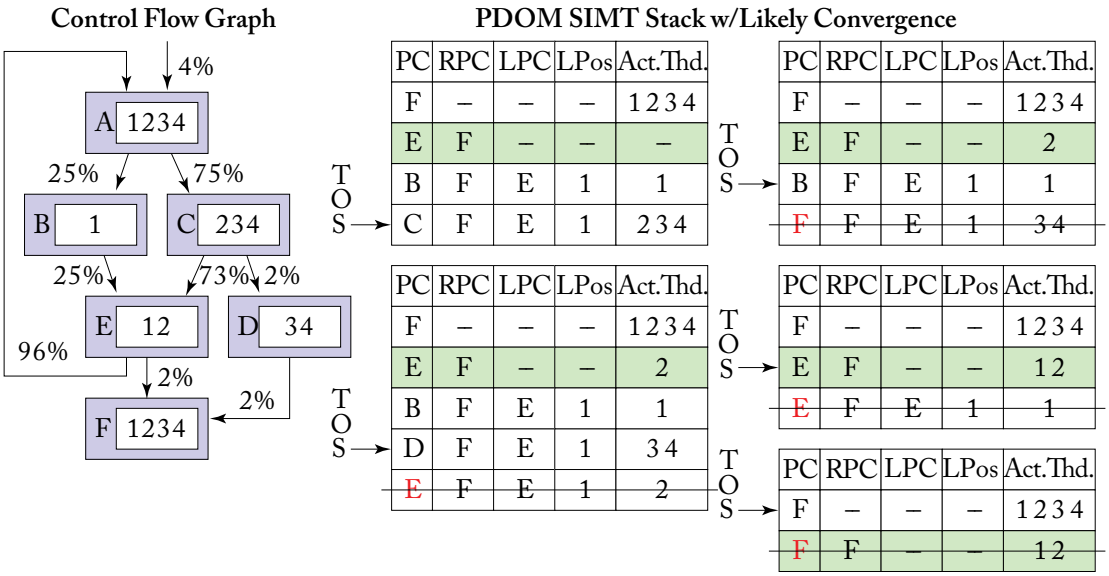


图 3.19：早期的再收敛点 *before* 紧接着后支配点。可能的收敛点捕捉到 E 处的早期再收敛。

重新收敛点，即发散分支的重新收敛点，以便此条目中的线程将重新收敛到确定的重新收敛点。

当一个 warp 使用 SIMT 堆栈中的顶部条目执行时，它会将其 PC 与 RPC 字段（就像它与基线 SIMT 堆栈所做的一样）以及 LPC 字段进行比较。如果 PC == LPC，则弹出 SIMT 堆栈，并且此弹出条目中的线程将合并到 LPC 条目中。否则，如果 PC == RPC，则简单地弹出 SIMT 堆栈 — RPC 条目已将这些线程记录在其活动掩码中。当 LPC 条目到达 SIMT 堆栈的顶部时，它会像任何其他 SIMT 堆栈条目一样执行，或者如果其活动掩码为空，则直接弹出。

线程边界。Diamos 等人 [2011] 完全脱离了 SIMT 堆栈，而是提出通过 *thread frontiers* 在发散后重新收敛线程。支持线程边界的编译器根据内核中基本块的拓扑顺序对其进行排序。这样，在较高 PC 的指令处执行的线程永远不会跳转到较低 PC 的指令。通过将循环出口放在循环体的末尾来处理循环。通过这种排序的代码布局，发散的 warp 最终将通过优先考虑具有较低 PC 的线程（允许它们赶上）来重新收敛。

与具有立即后支配器重新收敛的 SIMT 堆栈相比，通过线程边界重新收敛可为具有非结构化控制流的应用程序带来更高的 SIMD 效率。多表达式条件语句的求值语义和异常的使用都可以生成具有非结构化控制流的代码。扩展的 SIMT 堆栈可能

convergence points can yield similar SIMD efficiency improvement on applications with unstructured control flow; however, each entry in the SIMT stack may only have a finite number of likely convergence points, whereas the thread frontier approach has no such restriction.

Oppportunistic Early Reconvergence. ElTantaway et al. [2014] proposes *opportunistic early reconvergence* (OREC), boosting the SIMD efficiency for GPU applications with unstructured control flow without any additional compiler analysis. OREC builds upon the Multi-Path (MP) SIMT Stack introduced in the same paper (see Section 3.4.2). The MP SIMT Stack uses a separate warp-split table holding the current set of warp-splits available for execution. At a divergent branch, new warp-splits are created with the branch target PCs and the reconvengence PC of the divergent branch. With OREC, instead of simply inserting these new warp-splits into the warp-split table, the hardware searches through the warp-split table for an existing warp-split with the same starting PC and RPC. If such a warp-split exists, the hardware create an early reconvergence point in the Reconvergence Table to converge the two warp-splits before the original RPC. The early reconvergence point synchronizes the two warp-split at a particular PC, so that they can be merged even when the existing warp-split have advanced through the diverged path. In ElTantaway et al. [2014] the early reconvergence point is the next PC of the existing warp-split.

3.4.3 ADDING MIMD CAPABILITY

The following proposals improve GPUs' compatibility with divergent control flow by incorporating some limited amount of MIMD capability. All of these proposals offer two modes of operation:

- a SIMD mode, where the front-end issues one instruction to be executed across all threads in a warp; or
- a MIMD mode, where the front-end issues different instructions for each thread in a diverged warp.

When a warp is not diverged, it executes in SIMD mode to capture the control-flow locality exhibited by the threads in the warp, with energy efficiency comparable to traditional SIMD architectures. The warp switches to MIMD mode when it diverges. The warp runs less efficiently in this mode, but the performance penalty is lower than the penalty on a traditional SIMD architecture.

Vector-Thread Architecture. Vector-thread (VT) architecture [Krashinsky et al., 2004] combines aspects of both SIMD and MIMD architectures, with the goal of capturing the best of both approaches. A VT architecture features a set of lanes that are connected to a common L1 instruction cache. In SIMD mode, all lanes receive instructions directly from the L1 instruction cache for lockstep execution, but each lane may switch to a MIMD model, running at its own

收敛点可以在具有非结构化控制流的应用程序上产生类似的 SIMD 效率改进；然而，SIMT 堆栈中的每个条目可能只有有限数量的可能收敛点，而线程边界方法没有这样的限制。

机会性早期重收敛。ElTantaway 等人 [2014] 提出了 *opportunistic early reconvergence* (OREC)，无需任何额外的编译器分析，即可提高具有非结构化控制流的 GPU 应用程序的 SIMD 效率。OREC 以同一篇论文中介绍的多路径 (MP) SIMT 堆栈为基础（参见第 3.4.2 节）。MP SIMT 堆栈使用单独的 warp-split 表来保存当前可供执行的 warp-split 集。在发散分支处，将使用分支目标 PC 和发散分支的重收敛 PC 创建新的 warp-split。使用 OREC，硬件不会简单地将这些新的 warp-split 插入 warp-split 表，而是在 warp-split 表中搜索具有相同起始 PC 和 RPC 的现有 warp-split。如果存在这样的 Warp-split，硬件会在重聚表中创建一个早期重聚点，以便在原始 RPC 之前将两个 Warp-split 收敛。早期重聚点将两个 Warp-split 同步到特定的 PC，这样即使现有 Warp-split 已经通过分叉路径前进，它们也可以合并。在 ElTantaway 等人 [2014] 中，早期重聚点是现有 Warp-split 的下一个 PC。

3.4.3 添加 MIMD 功能

以下提案通过整合一定数量的 MIMD 功能来提高 GPU 与发散控制流的兼容性。所有这些提案都提供了两种操作模式：

- SIMD 模式，其中前端发出一条指令，在 warp 中的所有线程中执行；或者
- MIMD 模式，其中前端为发散 warp 中的每个线程发出不同的指令。

当 Warp 未发散时，它会在 SIMD 模式下执行，以捕获 Warp 中线程所表现出的控制流局部性，其能效可与传统 SIMD 架构相媲美。当 Warp 发散时，它会切换到 MIMD 模式。Warp 在此模式下运行效率较低，但性能损失低于传统 SIMD 架构上的损失。

矢量线程架构。矢量线程 (VT) 架构 [Krashinsky 等，2004] 结合了 SIMD 和 MIMD 架构的特点，目的是兼顾两种方法的优点。VT 架构具有一组连接到公共 L1 指令缓存的通道。在 SIMD 模式下，所有通道都直接从 L1 指令缓存接收指令以进行锁步执行，但每个通道都可以切换到 MIMD 模型，以自己的方式运行

pace with instructions from its L0 cache. A recent comparison with traditional SIMT architectures (e.g., GPUs) by Lee et al. [2011] shows that VT architectures have comparable efficiency with regular parallel applications, while performing much more efficiently with irregular parallel applications.

Temporal SIMT. Temporal SIMT [Keckler et al., 2011, Krashinsky, 2011] permits each lane to execute in MIMD fashion, similar to VT architecture. However, instead of running a warp across all lanes in lockstep, it time-multiplexes the execution of a warp through a single lane, and each lane runs a separate set warp. Temporal SIMT achieves the efficiency of SIMD hardware by fetching each instruction only once for the whole warp. This amortizes the control flow overhead across time, while the traditional SIMD architecture amortizes the same overhead across multiple lanes in space.

Variable Warp Size Architecture. Variable Warp Size (VWS) Architecture [Rogers et al., 2015] contains multiple (e.g., 8) slices, each of which contains a fetch and decode unit, so that each slice may execute different instructions simultaneously, similar to VT and Temporal SIMT. Instead of time-multiplexing large warps via narrow data path, each slice in VWS consists of narrow (4-wide) warps. These narrow warps are then grouped into larger execution entities called *gangs*. Each gang contains a warp from each slice.

In applications with no branch divergence, warps in a gang executes in lock-steps, obtaining instructions from a shared fetch unit and a shared L1 instruction cache. Upon encountering a branch divergence (or a memory divergence), the gang splits into multiple gangs. The new gangs may further splits until the point where every warp is in its own gang. At that point, these single-warp gangs are executed on their own slices individually via the slices' fetch unit and private L0 instruction cache. These split gangs are merged back together into the original gang opportunistically via hardware that compares the PC of the individual gangs. If they all match, the original gang is recreated. Rogers et al. [2015] also proposed inserting a gang-level synchronization barrier at the immediate post-dominator of the first divergent branch.

This book also evaluated performance impact of the capacity of the L0 instruction cache in each slice, in relation to the shared L1 instruction cache bandwidth. In non-ganged mode, the L0 caches in the slices may all requests instructions from the L1 cache simultaneously, creating a bandwidth bottleneck. Their evaluation shows that even for divergent applications, a 256-Byte L0 cache can filter out most of the requests to the shared L1 cache, As a result, the L1 cache can cover most of the bandwidth deficiency with just 2× bandwidth of the baseline SIMT architecture.

Simultaneous Branch Interweaving. Brunie et al. [2012] propose simultaneous branch and warp interweaving (SBI and SWI) after the publication of thread block compaction. They extend the GPU SIMT front-end to support issuing two different instructions per cycle. SBI co-issues instructions from the same warp when it encounters a branch divergence. Executing both targets of a divergence branch at the same time eliminates its performance penalty significantly.

与其 L0 缓存中的指令同步。Lee 等人 [2011] 最近与传统 SIMT 架构（例如 GPU）进行了比较，结果表明 VT 架构与常规并行应用程序具有相当的效率，而与不规则并行应用程序相比，其执行效率要高得多。

时间 SIMT。时间 SIMT [Keckler 等，2011，Krashinsky，2011] 允许每个通道以 MIMD 方式执行，类似于 VT 架构。但是，它不是同步在所有通道上运行 warp，而是通过单个通道对 warp 的执行进行时间复用，并且每个通道运行单独的 warp 集。时间 SIMT 通过为整个 warp 仅获取一次每条指令来实现 SIMD 硬件的效率。这分摊了时间上的控制流开销，而传统的 SIMD 架构则在空间上的多个通道上分摊相同的开销。

可变 Warp 大小架构。可变 Warp 大小 (VWS) 架构 [Rogers et al., 2015] 包含多个（例如 8 个）切片，每个切片包含一个提取和解码单元，因此每个切片可以同时执行不同的指令，类似于 VT 和时间 SIMT。VWS 中的每个切片由窄（4 宽）Warp 组成，而不是通过窄数据路径对大型 Warp 进行时间复用。然后，这些窄 Warp 被分组到称为 *gangs* 的较大执行实体中。每个组包含来自每个切片的 Warp。

在没有分支发散的应用程序中，一个组中的 warp 以锁步执行，从共享提取单元和共享 L1 指令缓存中获取指令。一旦遇到分支发散（或内存发散），该组就会分裂成多个组。新的组可能会进一步分裂，直到每个 warp 都在自己的组中。此时，这些单 warp 组将通过切片的提取单元和私有 L0 指令缓存在自己的切片上单独执行。这些分裂的组会通过硬件比较各个组的 PC 而适时地合并回原始组。如果它们都匹配，则重新创建原始组。Rogers 等人 [2015] 还建议在第一个发散分支的紧邻后支配者处插入组级同步屏障。

本书还评估了每个切片中的 L0 指令缓存容量与共享 L1 指令缓存带宽之间的性能影响。在非联动模式下，切片中的 L0 缓存可能同时从 L1 缓存请求指令，从而造成带宽瓶颈。他们的评估表明，即使对于不同的应用程序，256 字节的 L0 缓存也可以过滤掉对共享 L1 缓存的大部分请求，因此，L1 缓存仅使用基线 SIMT 架构的 2× 带宽就可以弥补大部分带宽不足。

同时分支交织。Brunie 等人 [2012] 在线程块压缩发布后提出了同时分支和 Warp 交织 (SBI 和 SWI)。他们扩展了 GPU SIMT 前端，以支持每个周期发出两个不同的指令。当 SBI 遇到分支分歧时，它会从同一个 Warp 中同时发出指令。同时执行分歧分支的两个目标可显著消除其性能损失。

3.4.4 COMPLEXITY-EFFECTIVE DIVERGENCE MANAGEMENT

The area requirement of the baseline SIMT stack for each warp is only 32 × 64 bits (or as low as 6 × 64 bits with the optimization used in AMD GCN). While this area is small in comparison to the register file in a SIMT core, this area scales directly with the number of in-flight warps in the GPU, as well as the number of threads per warp. Moreover, in typical GPU applications where branch divergence is rare, the SIMT stack occupies area that can otherwise be used to boost the application throughput in other ways. A number of proposals replace the SIMT stack with alternative mechanisms, which share resources that can be used in other ways when a warp does not encounter any branch divergence.

SIMT Stack in Scalar Register File. AMD GCN [AMD, 2012] features a scalar register file that is shared by all threads in a warp. Its registers can be used as predication registers to control the activity of each thread in a warp. The compiler uses this scalar register file to emulate a SIMT stack in software when it detects potentially divergent branch in the compute kernel. The GCN architecture features special instructions to accelerate the SIMT stack emulation.

One optimization that minimizes the number of scalar registers required to support the worst-case divergence is to prioritize the execution of the target with the fewer number of active threads. This allows the worst-case divergence to be supported with $\log_2(\#threads\ per\ warp)$ scalar registers, which is far fewer than the entries required by the baseline SIMT stack. Furthermore, when the compute kernel has no potentially divergent branch, the compiler can use the scalar registers reserved for the SIMT stack for other scalar computation.

Thread Frontiers. As mentioned in Section 3.4.2, Diamos et al. [2011] replaces the SIMT stack with *thread frontiers*. With thread frontiers, each thread maintains its own PC in the register file, and the code is topologically sorted so that reconvergence point of a branch always have a higher PC. When a warp diverges, it always prioritize the threads with the lowest PC among all of its threads. This set of threads is known as the thread frontier of the warp. Prioritizing the execution of the threads in the frontier implicitly forces all threads further ahead in the program to wait at the reconvergence point of the branch to be merged.

Since the per-thread PCs are only needed when the compute kernel contains potentially divergent branches, the compiler only needs to allocate a PC register in these compute kernels. In other compute kernels, the extra register storage works to improve the warp occupancy, increasing the number of warps that each SIMT core can sustain to better tolerate memory latency.

Stackless SIMT. Asanovic et al. [2013] proposes extending the temporal SIMT architecture with a `syncwarp` instruction. In this proposal, threads in a warp run in lock-step when they execute in the *convergent regions* of the compute kernel, where the compiler guarantees that the warp can never diverge. At a divergent branch, each thread in a warp follows its own control-flow path with its private PC, leveraging the MIMD capability in the temporal SIMT architecture.

每个 Warp 的基准 SIMT 堆栈的面积要求仅为 32×64 位（或使用 AMD GCN 中使用的优化低至 6×64 位）。虽然与 SIMT 核心中的寄存器文件相比，这个面积很小，但这个面积与 GPU 中正在运行的 Warp 数量以及每个 Warp 的线程数量成正比。此外，在分支发散很少见的典型 GPU 应用中，SIMT 堆栈占用的区域原本可用于以其他方式提高应用程序吞吐量。许多提案用替代机制取代 SIMT 堆栈，这些机制共享资源，当 Warp 未遇到任何分支发散时，这些资源可用于其他方式。

标量寄存器文件中的 SIMT 堆栈。AMD GCN [AMD，2012] 具有一个标量寄存器文件，该文件由 warp 中的所有线程共享。其寄存器可用作预测寄存器，以控制 warp 中每个线程的活动。当编译器检测到计算内核中可能存在分歧的分支时，它会使用此标量寄存器文件在软件中模拟 SIMT 堆栈。GCN 架构具有特殊指令来加速 SIMT 堆栈模拟。

最小化支持最坏情况发散所需的标量寄存器数量的一种优化是优先执行活动线程数较少的目标。这样就可以使用 $\log_2(\#threads\ per\ warp)$ 个标量寄存器来支持最坏情况发散，这比基线 SIMT 堆栈所需的条目要少得多。此外，当计算内核没有潜在发散分支时，编译器可以使用为 SIMT 堆栈保留的标量寄存器进行其他标量计算。

线程边界。如第 3.4.2 节所述，Diamos 等人 [2011] 用 *thread frontiers* 替换了 SIMT 堆栈。使用线程边界，每个线程在寄存器文件中维护自己的 PC，并且代码按拓扑排序，以便分支的重新收敛点始终具有更高的 PC。当 Warp 发散时，它始终优先考虑其所有线程中 PC 最低的线程。这组线程称为 Warp 的线程边界。优先执行边界中的线程会隐式强制程序中所有更靠前的线程在要合并的分支的重新收敛点处等待。

由于只有当计算内核包含潜在发散分支时才需要每个线程的 PC，因此编译器只需在这些计算内核中分配一个 PC 寄存器即可。在其他计算内核中，额外的寄存器存储可以提高 warp 占用率，从而增加每个 SIMT 内核可以维持的 warp 数量，以更好地容忍内存延迟。

无堆栈 SIMT。Asanovic 等人 [2013] 提议使用 `syncwarp` 指令扩展时间 SIMT 架构。在此提议中，warp 中的线程在计算内核的 *convergent regions* 中执行时以锁步方式运行，其中编译器保证 warp 永远不会发散。在发散分支处，warp 中的每个线程都遵循其私有 PC 的控制流路径，从而利用时间 SIMT 架构中的 MIMD 功能。

The compiler places a `syncwarp` instruction at the reconvergence point of the divergent branch. This forces all threads in the diverged warp to sync up at the reconvergence point before entering another convergent region of the compute kernel.

While this mechanism does not capture reconvergence possible with nested divergent branches, it is nevertheless a cheaper alternative that can provide comparable performance to the baseline SIMT stack for GPU applications that rarely exhibit branch divergence. The paper introduces a combined convergent and variant analysis that allows the compiler to determine operations in an arbitrary compute kernel that is eligible for *scalarization* and/or *affine transformation*. In the context of stackless SIMT, the same analysis allows the compiler to determine the convergent and divergent regions in an arbitrary compute kernel.

1. The compiler first assumes all basic-blocks to be *thread-invariant*.

2. It marks all instructions dependent on the thread ID, atomic instructions, and memory instructions on volatile memory as *thread-variant*.

3. It then iteratively marks all instruction dependent on thread-variant instructions as thread-variant as well.

4. All instructions in basic-blocks that are *control dependent* on a thread-variant branch instruction are also thread-variant. Essentially, instructions beyond the immediate post-dominator of a thread-variant branch may remain thread-invariant, as long as they are not marked as thread-variant for other conditions.

This analysis allows the compiler to detect branches that are taken uniformly by all threads in each warp. Since these branches do not cause warp to diverge, the compiler does not need to insert code to detect dynamic divergence for these branches, nor does it need to insert `syncwarp` instructions at their immediate post-dominators to enforce reconvergence.

Predication. Before incorporating a full SIMT stack into the architecture, GPUs with programmable shaders has been supporting limited control-flow constructs in the shader programs via *predications*, just like traditional vector processors. Prediction remains in modern GPUs as a low-overhead way to handle simple if-branches, avoiding the overhead of pushing and popping the SIMT stack. In NVIDIA's implementation, each instruction is extended with an extra operand field to specify its predication registers. Predication registers are essentially scalar registers dedicated to control flow.

Lee et al. [2014b] proposes a *thread-aware prediction algorithm* that extends the application of predication to arbitrary control-flow, with performance comparable to the NVIDIA's SIMT stack. The thread-aware predication algorithm extends the standard Control-Flow Dependency Graph (CDG) with predication nodes at each branch. The prediction required by each basic block can then be computed based on their control-flow dependency, and further optimized rigorously without breaking the functional behavior. The paper then describes two

编译器将 `syncwarp` 指令放置在发散分支的重新收敛点。这会强制发散 warp 中的所有线程在进入计算内核的另一个收敛区域之前在重新收敛点同步。

虽然这种机制无法捕捉嵌套发散分支可能出现的重新收敛，但它仍然是一种更便宜的替代方案，可以为很少出现分支发散的 GPU 应用程序提供与基线 SIMT 堆栈相当的性能。本文介绍了一种组合收敛和变体分析，允许编译器确定适用于 *scalarization* 和/或 *affine transfor- mation* 的任意计算内核中的操作。在无堆栈 SIMT 的上下文中，相同的分析允许编译器确定任意计算内核中的收敛和发散区域。

1. 编译器首先假设所有基本块都是*thread-invariant*。

2.它将所有依赖于线程ID的指令，原子指令以及易失性存储器上的内存指令标记为*thread-variant*。

3. 然后，它迭代地将所有依赖于线程变体指令的指令也标记为线程变体。

4. 基本块中所有在线程变量分支指令上为 *control dependent* 的指令也是线程变量。本质上，线程变量分支的直接后支配者之外的指令可以保持线程不变，只要它们没有因其他条件而被标记为线程变量。

此分析允许编译器检测每个 Warp 中所有线程一致执行的分支。由于这些分支不会导致 Warp 发散，因此编译器无需插入代码来检测这些分支的动态发散，也无需在其紧邻的后支配者处插入 `syncwarp` 指令来强制重新收敛。

预测。在将完整的 SIMT 堆栈纳入架构之前，具有可编程着色器的 GPU 一直通过 *predications* 在着色器程序中支持有限的控制流构造，就像传统的矢量处理器一样。预测在现代 GPU 中仍然是一种处理简单 if 分支的低开销方式，可避免推送和弹出 SIMT 堆栈的开销。在 NVIDIA 的实现中，每条指令都扩展了一个额外的操作数字段来指定其预测寄存器。预测寄存器本质上是专用于控制流的标量寄存器。

Lee 等人 [2014b] 提出了一种 *thread-aware prediction algorithm*，将预测的应用扩展到任意控制流，其性能可与 NVIDIA 的 SIMT 堆栈相媲美。线程感知预测算法扩展了标准控制流依赖图 (CDG)，每个分支都有预测节点。然后可以根据每个基本块的控制流依赖关系计算每个基本块所需的预测，并在不破坏功能行为的情况下进一步严格优化。然后，本文描述了两种

optimizations based on this thread-aware CDG, and the convergence and variance analysis in their prior work [Asanovic et al., 2013].

- *Static branch-uniformity optimization* is applied when the compiler can guarantee that the branch can be taken uniformly across the entire warp, as deduced by the convergence analysis. In this case, the compiler can replace the predication generation with a uniform branch instruction.
- *Runtime branch-uniformity optimization* is applied in other cases. The compiler issue consensual branches (`cbranch.ifnone`) that are only taken when given a null predicate (i.e., all threads disabled). This allows the warp to skip through code with null predicates—a key advantage offered by SIMT stacks. This approach differs from prior efforts for vector processors, such as BOSCC, in that it relies on structure analysis to determine the candidates for this optimization.

While both predication and SIMT stacks are fundamentally providing the same functionality at a similar energy and area cost, Lee et al. [2014b] highlighted the following trade-offs between the two approaches.

- Since different branch targets are guarded by different predication registers, the compiler can schedule instructions from different branch targets, interleaving the execution of different branch targets to exploit thread-level parallelism (TLP) that would otherwise require more advanced hardware branch divergence management.
- Predication tends to increase register pressure, which in turn reduces warp occupancy and imposes an overall performance penalty. This happens because a conservative register allocation cannot reuse registers for both sides of a branch. It cannot robustly prove that instructions from different branch targets are operating no exclusive sets of lanes in the registers. The uniform and consensual branch instructions inserted by the two proposed optimizations alleviate this problem.
- Predication may impact the dynamic instruction count in multiple ways. In some cases, the overhead of checking uniform branches increases the dynamic instruction count significantly. Alternatively, not performing the check means that some paths are executed with a null predication mask. In other cases, it removes push/pop instructions required to maintain the SIMT stack.

In the end, the paper proposes new instructions to reduce the overhead of predication.

- For function calls and indirect branches, they propose a new `find_unique` instruction to serially execute each branch target/function via a loop.
- `cbranch.ifany` (in addition to existing consensual branch instructions `cbranch.ifnone` and `cbranch.ifall`) will help reduce the instruction count overhead introduced by dynamic uniform branch detection.

基于该线程感知 CDG 的优化，以及他们先前工作中的收敛和方差分析 [Asanovic et al., 2013]。

- 当编译器可以保证分支可以在整个 Warp 中均匀执行（由收敛分析得出）时，将应用 *Static branch-uniformity optimization*。在这种情况下，编译器可以用统一分支指令取代谓词生成。 • 在其他情况下，将应用 *Runtime branch-uniformity optimization*。编译器发出一致分支 (`cbranch.ifnone`)，只有在给定空谓词（即禁用所有线程）时才会执行这些分支。这允许 Warp 跳过带有空谓词的代码 — 这是 SIMT 堆栈提供的一个关键优势。这种方法不同于之前针对矢量处理器（如 BOSCC）的努力，因为它依靠结构分析来确定此优化的候选对象。

虽然预测和 SIMT 堆栈从根本上以相似的能源和面积成本提供相同的功能，但 Le e 等人 [2014b] 强调了两种方法之间的以下权衡。

- 由于不同的分支目标由不同的预测寄存器保护，因此编译器可以调度来自不同分支目标的指令，交错执行不同的分支目标以利用线程级并行 (TLP)，否则将需要更高级的硬件分支发散管理。 • 预测往往会增加寄存器压力，从而降低 warp 占用率并造成整体性能损失。发生这种情况的原因是保守的寄存器分配无法重用分支两侧的寄存器。它无法稳健地证明来自不同分支目标的指令没有在寄存器中操作独占的通道集。两个提议的优化插入的统一和一致分支指令缓解了这个问题。 • 预测可能会以多种方式影响动态指令数。在某些情况下，检查统一分支的开销会显著增加动态指令数。或者，不执行检查意味着某些路径使用空预测掩码执行。在其他情况下，它会删除维护 SIMT 堆栈所需的推送/弹出指令。

最后，论文提出了新的指令来减少预测的开销。

- 对于函数调用和间接分支，他们提出了一种新的 `find_unique` 指令，通过循环按顺序执行每个分支目标/函数。 • 除了现有的一致分支指令 `cbranch.ifnone` 和 `cbranch.ifall`) 之外，`cbranch.ifany` (将有助于减少动态统一分支检测引入的指令数开销。

3.5 RESEARCH DIRECTIONS ON SCALARIZATION AND AFFINE EXECUTION

As described in Chapter 2, GPU computing APIs, such as CUDA and OpenCL, feature a MIMD-like programming model that allows the programmer to launch a large array of scalar threads onto the GPU. While each of these scalar threads can follow its unique execution path and may access arbitrary memory locations, in the common case, they all follow a small set of execution paths and perform similar operations. The convergent control-flow among GPU threads is exploited on most, if not all, modern GPUs via the SIMT execution model, where scalar threads are grouped into warps that runs on SIMD hardware (see Section 3.1.1).

This section summarizes a series of research that further exploit the similarity of these scalar threads via *scalarization* and *affine execution*. The key insight of these research lies in the observation of *value structure* [Kim et al., 2013] across threads executing the same compute kernel. The two types of value structure, *uniform* and *affine*, are illustrated the compute kernel in Example 3.3.

Uniform Variable A variable that has the same constant value for every thread in the compute kernel. In Algorithm 3.3, the variable `a`, as well as the literals `THRESHOLD` and `Y_MAX_VALUE`, all have uniform value across all threads in the compute kernel. A uniform variable can be stored in a single scalar register, and reused by all threads in the compute kernel.

Affine Variable A variable with values that is a linear function of thread ID for every thread in the compute kernel. In Algorithm 3.3, the memory address of the variable `y[idx]` can be represented as an *affine* transform of the thread ID `threadIdx.x`:

```
&(y[idx]) = &(y[0]) + size(int) * threadIdx.x;
```

This affine representation can be stored as a pair of scalar values, a *base* and a *stride*, which is far more compact than the fully expanded vector.

There are multiple research proposal on how to *detect* and *exploit* uniform or affine variables in GPUs. The rest of this section summarizes these proposals in these two aspects.

3.5.1 DETECTION OF UNIFORM OR AFFINE VARIABLES

There are two main approaches to detect the existence of uniform or affine variables in a GPU compute kernel: Compiler-Driven Detection and Detection via Hardware.

Compiler-Driven Detection

One way to detect the existence of uniform or affine variables in a GPU compute kernel is to do so via a special compiler analysis. This is possible because the existing GPU programming models, CUDA and OpenCL, already provides means for the programmer to declare a variable as

3.5. 标量化和仿射执行的研究方向 57 3.5 标量化和仿射执行的研究方向

如第 2 章所述，GPU 计算 API（例如 CUDA 和 OpenCL）具有类似 MIMD 的编程模型，允许程序员在 GPU 上启动大量标量线程。虽然这些标量线程中的每一个都可以遵循其独特的执行路径并可以访问任意内存位置，但在常见情况下，它们都遵循一小组执行路径并执行类似的操作。大多数（如果不是全部）现代 GPU 都通过 SIMT 执行模型利用 GPU 线程之间的收敛控制流，其中标量线程被分组为在 SIMD 硬件上运行的 warp（参见第 3.1.1 节）。

本节总结了一系列研究，这些研究通过 *scalarization* 和 *affine execution* 进一步利用了这些标量线程的相似性。这些研究的关键见解在于对执行相同计算内核的线程之间的 *value structure* [Kim et al., 2013] 的观察。示例 3.3 中的计算内核说明了两种类型的值结构，即 *uniform* 和 *affine*。

统一变量 计算内核中每个线程都具有相同常量值的变量。在算法 3.3 中，变量 `a` 以及文字 `THRESHOLD` 和 `Y_MAX_VALUE` 在计算内核的所有线程中都具有统一的值。统一变量可以存储在单个标量寄存器中，并由计算内核中的所有线程重复使用。

线性变量 一个变量，其值是计算内核中每个线程的线程 ID 的线性函数。在算法 3.3 中，变量 `y[idx]` 的内存地址可以表示为线程 ID `threadIdx.x` 的 *affine* 变换：

```
&(y[idx]) = &(y[0]) + size(int) * threadIdx.x;
```

这种精细表示可以存储为一对标量值，*base* 和 *stride*，这比完全展开的向量紧凑得多。

关于如何在 GPU 中实现 *detect* 和 *exploit* 均匀或仿射变量，有多个研究提案。本节的其余部分从这两个方面总结了这些提案。

3.5.1 均匀或仿射变量的检测

检测 GPU 计算内核中统一或仿射变量的存在主要有两种方法：编译器驱动检测和通过硬件检测。

编译器驱动检测

检测 GPU 计算内核中是否存在统一或仿射变量的一种方法是通过特殊的编译器分析来实现。这是可能的，因为现有的 GPU 编程模型 CUDA 和 OpenCL 已经为程序员提供了将变量声明为

Algorithm 3.3 Example of scalar and affine operations in a compute kernel (from [Kim et al., 2013]).

```
__global__ void vsadd( int y[], int a )
{
  int idx = threadIdx.x;
  y[idx] = y[idx] + a;
  if ( y[idx] > THRESHOLD )
    y[idx] = Y_MAX_VALUE;
}
```

constant through out the compute kernel, as well as providing special variable for the thread ID. The compiler can perform a control-dependency analysis to detect variables that are dependent purely on constants and thread IDs, and mark them as uniform/affine. Operations that work solely on uniform/affine variables are then candidates for *scalarization*.

AMD GCN [AMD, 2012] relies on the compiler to detect uniform variables and scalar operations that can be stored and processed by a dedicated scalar processor.

Asanovic et al. [2013] introduce a combined convergent and variant analysis that allows the compiler to determine operations in an arbitrary compute kernel that is eligible for *scalarization* and/or *affine transformation*. Instructions within the convergent regions of a compute kernel can be converted into scalar/affine instructions. At any transition from divergent to convergent regions of a compute kernel, the compiler inserts a `syncwarp` instruction to handle control-flow induced register dependencies between the two regions. Asanovic et al. [2013] adopted this analysis to generate scalar operations for the Temporal-SIMT architecture [Keckler et al., 2011, Krashinsky, 2011].

Decoupled Affine Computation (DAC) [Wang and Lin, 2017] relies on a similar compiler analysis to extract scalar and affine candidates to be decoupled into a separate warp. Wang and Lin [2017] augments the process with a divergent affine analysis, with the goal to extract strands of instructions that has been affine from the start of the compute kernel. These strands of affine instructions are decoupled from the main kernel into an affine kernel that feeds data into the main kernel via a hardware queue.

Hardware Detection

Detecting uniform/affine variables in hardware offers two potential advantage over compiler-driven detection.

1. This allows scalarization and affine execution to be applied with the original GPU instruction set architecture. It saves the effort to co-develop a special scalarization compiler along with the hardware.

算法 3.3 计算内核中的标量和仿射运算示例（来自 [Kim et al., 2013]）。

```
__global__ void vsadd( int y[], int a )
{
  int idx = threadIdx.x;
  y[idx] = y[idx] + a;
  if ( y[idx] > THRESHOLD )
    y[idx] = Y_MAX_VALUE;
}
```

在整个计算内核中，常量以及为线程 ID 提供特殊变量。编译器可以执行控制依赖性分析，以检测仅依赖于常量和线程 ID 的变量，并将它们标记为统一/仿射。仅对统一/仿射变量起作用的操作将成为 *scalarization* 的候选。

AMD GCN [AMD，2012] 依靠编译器来检测可以由专用标量处理器存储和处理的统一变量和标量运算。

Asanovic 等人 [2013] 引入了一种综合收敛和变体分析，允许编译器确定任意计算内核中适用于 *scalariza-*、*tion* 和/或 *affine transformation* 的操作。计算内核收敛区域内的指令可以转换为标量/仿射指令。在计算内核从发散区域到收敛区域的任何转换中，编译器都会插入一条 `syncwarp` 指令来处理两个区域之间控制流引起的寄存器依赖关系。Asanovic 等人 [2013] 采用此分析为 Temporal-SIMT 架构生成标量操作 [Keckler 等人，2011，Krashinsky，2011]。

解耦仿射计算 (DAC) [Wang and Lin, 2017] 依靠类似的编译器分析来提取标量和仿射候选，并将其解耦为单独的 warp。Wang and Lin [2017] 通过发散仿射分析增强了该过程，目标是提取从计算内核开始就已仿射的指令链。这些仿射指令链从主内核解耦为仿射内核，该内核通过硬件队列将数据输入主内核。

硬件检测
在硬件中检测均匀/仿射变量比编译器驱动的检测有两个潜在优势。

1. 这使得标量化和仿射执行可以与原始的 GPU 指令集架构一起应用。它节省了与硬件共同开发专用标量化编译器的努力。

2. The hardware detection happens during the compute kernel execution. As a result, it is capable of detecting uniform/affine variables that occurs dynamically, but are missed by the static analysis.

Tag-Based Detection. Collange et al. [2010] introduce a tag-based detection system. In this system, each GPU register is extended with a tag, indicating if the register contains uniform, affine, or generic vector values. At the launch of a compute kernel, the tag of the register that contains the thread ID is set to affine state. Instructions that broadcast values from a single location in constant or shared memory set the tag of the destination register to the uniform state. During kernel execution, states of the registers are propagated across arithmetic instruction from source to destination operands according to simple rules in Table 3.1. While this tag-based detection has little hardware overhead, it tends to be conservative—for example, it conservatively rules the output of multiplication between uniform and affine variables as vector variable.

Table 3.1: Examples of rules of uniform and affine state propagation across instructions from Collange et al. [2010]. For each operation, the first row and column shows states of the input operands, and the remaining entries show the state of the output operand for every permutation of input operand states (U = uniform, A = affine, V = vector).

+	U	A	V
U	U	A	V
A	A	V	V
V	V	V	V

×	U	A	V
U	U	V	V
A	V	V	V
V	V	V	V

<<	U	A	V
U	U	A	V
A	V	V	V
V	V	V	V

FG-SIMT architecture [Kim et al., 2013] extends the tag-based detection mechanism from Collange et al. [2010] with better support for branches. Affine branches, or branches that compares between affine operands, is resolved via the scalar datapath if one of the operand is a uniform. Kim et al. [2013] also introduce a *lazy expansion* scheme, where affine registers are lazily expanded into full vector registers after a divergent branch or a predicated instruction. This expansion is required to allow a subset of threads in a divergent warp to update their slots in the destination register, while leaving other slots unchanged—this maintains the SIMT execution semantics. In contrast to a more naive, eager, expansion scheme that expands every affine register after the first divergent branch, the lazy expansion scheme eliminates many unnecessary expansion.

Comparison at Write-Back. Gilani et al. [2013] introduce a more aggressive mechanism to detect uniform variables by comparing the register values from all threads in a warp at each write-back of a vector instruction. At the detection of an uniform variable, the detection logic reroutes the write-back to a scalar register file, and updates an internal table to remember the state of the register. Subsequent use of the register is then redirected to the scalar register file. Instructions with all operands from the scalar register file are executed on a separate scalar pipeline.

2. 硬件检测发生在计算内核执行期间。因此，它能够检测动态发生的均匀/仿射变量，但静态分析会遗漏这些变量。

基于标签的检测。Collange 等人 [2010] 介绍了一种基于标签的检测系统。在这个系统中，每个 GPU 寄存器都用一个标签进行扩展，指示寄存器是否包含统一、仿射或通用向量值。在启动计算内核时，包含线程 ID 的寄存器的标签设置为仿射状态。从常量或共享内存中的单个位置广播值的指令将目标寄存器的标签设置为统一状态。在内核执行期间，寄存器的状态根据表 3.1 中的简单规则在算术指令中从源操作数传播到目标操作数。虽然这种基于标签的检测几乎没有硬件开销，但它往往比较保守——例如，它保守地将统一和仿射变量之间的乘法输出规定为向量变量。

表 3.1：来自 Collange 等人的指令间均匀和仿射状态传播规则示例 [2010]。对于每个操作，第一行和第一列显示输入操作数的状态，其余条目显示输入操作数状态每次变换的输出操作数状态（U = 均匀，A = 仿射，= 矢量）。

+	U	A	V	×	U	A	V	<<	U	A	V
U	U	A	V	U	U	V	V	U	U	A	V
A	A	V	V	A	V	V	V	A	V	V	V
V	V	V	V	V	V	V	V	V	V	V	V

FG-SIMT 架构 [Kim et al., 2013] 扩展了 Collange et al. [2010] 的基于标签的检测机制，更好地支持分支。如果其中一个操作数是统一的，则通过标量数据路径解析仿射分支或比较仿射操作数的分支。Kim et al. [2013] 还引入了一种 *lazy expansion* 方案，其中仿射寄存器在发散分支或谓词指令之后被惰性扩展为全矢量寄存器。需要进行此扩展以允许发散 warp 中的线程子集更新其在目标寄存器中的槽，同时保持其他槽不变 - 这保持了 SIMT 执行语义。与在第一个发散分支之后扩展每个仿射寄存器的更简单、更急切的扩展方案相比，惰性扩展方案消除了许多不必要的扩展。

写回时的比较。Gilani 等人 [2013] 引入了一种更积极的机制来检测统一变量，方法是在每次写回矢量指令时比较来自 warp 中所有线程的寄存器值。在检测到统一变量时，检测逻辑将写回重新路由到标量寄存器文件，并更新内部表以记住寄存器的状态。随后将寄存器的使用重定向到标量寄存器文件。所有操作数都来自标量寄存器文件的指令在单独的标量流水线上执行。

Lee et al. [2015] use a similar detection scheme. Instead of a simple uniform detector, they augment the register write-back stage with a register value compressor that transforms an incoming vector of values into a tuple of $< base, delta, immediate >$ (BDI) using the algorithm introduced by Pekhimenko et al. [2012].

Wong et al. [2016] introduce *Warp Approximation*, a framework to exploit approximate computing within a warp which also features a detection at register write-back. The detector computes the smallest *d-similarity*, two given values sharing *d*-MSBs, among all values in the vector written back to the register file. The register with a higher than threshold d-similarity is marked as *similar*, which is then used to determine the eligibility of approximation execution in subsequent dependent instructions.

Like the proposal from Lee et al. [2015], G-Scalar [Liu et al., 2017] features a register value compressor at the register write-back stage, but the compressor employs a simpler algorithm that only extracts common bytes used by all values across all lanes in the warp. If all bytes are common, the register contains a uniform variable. Any instruction that operates only on uniform variables can be scalarized.

G-Scalar also extended the register value compressor to detect operation that is eligible for scalar execution under branch divergence. All prior proposals revert back to vector execution as soon as the warp diverge. Liu et al. [2017] observe that in many instruction under branch divergence, the operand values of the active lanes are uniform. Instructions using these partially uniform registers are actually eligible for scalar execution. They then extend the register value compressor to check only values from the active lanes using a special logic. This greatly increases the number of scalar instructions across various GPU compute workloads. Note that under divergence, the written register is not compressed.

3.5.2 EXPLOITING UNIFORM OR AFFINE VARIABLES IN GPU

The design of a GPU may exploit the existence of value structure in compute kernels in multiple ways.

Compressed Register Storage

The compact representation of uniform and affine variables allows them to be stored in the Register File with far fewer bits. The salvaged storage can be used to sustain more inflight warps, increasing a GPU's tolerance to memory latency with the same register file resource.

Scalar Register File. Many proposals/designs exploit uniform or affine variables in GPU features a dedicated register file for scalar/affine values.

- AMD GCN architecture features a scalar register file that is accessible by both scalar and vector pipelines.
- FG-SIMT architecture [Kim et al., 2013] stores uniform/affine values in a separate Affine SIMT Register File (ASRF). The ASRF records the state (affine/uniform/vector) of each

Lee 等人 [2015] 使用了类似的检测方案。他们没有使用简单的统一检测器，而是使用寄存器值压缩器增强了寄存器写回阶段，该压缩器使用 Pekhimenko 等人 [2012] 引入的算法将传入的值向量转换为 $< base, delta, immediate >$ (BDI) 的元组。

Wong 等人 [2016] 引入了 *Warp Approximation*，这是一个利用 warp 中的近似计算的框架，它还具有寄存器写回检测功能。检测器计算写回到寄存器文件的向量中的所有值中最小的 *d-similarity*，即两个共享 *d*-MSB 的给定值。d 相似度高于阈值的寄存器被标记为 *similar*，然后用于确定后续相关指令中近似执行的资格。

与 Lee et al. [2015] 的提议一样，G-Scalar [Liu et al., 2017] 在寄存器写回阶段采用了寄存器值压缩器，但该压缩器采用了一种更简单的算法，该算法仅提取所有通道中所有值使用的公共字节。如果所有字节都是公共的，则寄存器包含一个统一变量。任何仅对统一变量进行操作的指令都可以标量化。

G-Scalar 还扩展了寄存器值压缩器，以检测在分支发散下适合标量执行的操作。所有先前的提案都会在 warp 发散后立即恢复为矢量执行。Liu 等人 [2017] 观察到，在分支发散下的许多指令中，活动通道的操作数值是统一的。使用这些部分统一寄存器的指令实际上适合标量执行。然后，他们使用特殊逻辑扩展了寄存器值压缩器，以仅检查来自活动通道的值。这大大增加了跨各种 GPU 计算工作负载的标量指令数量。请注意，在发散下，写入的寄存器不会被压缩。

3.5.2 在 GPU 中利用均匀或仿射变量

GPU 的设计可以通过多种方式利用计算内核中价值结构的存在。

压缩寄存器存储

均匀和仿射变量的紧凑表示允许它们以更少的位存储在寄存器文件中。回收的存储空间可用于维持更多的飞行中扭曲，从而提高 GPU 对相同寄存器文件资源的内存延迟的容忍度。

标量寄存器文件。许多提案/设计利用 GPU 中的均匀或仿射变量，为标量/仿射值提供专用的寄存器文件。

- AMD GCN 架构具有标量寄存器文件，可由标量和矢量管道访问。
- FG-SIMT 架构 [Kim et al., 2013] 将均匀/仿射值存储在单独的仿射 SIMT 寄存器文件 (ASRF) 中。 ASRF 记录每个寄存器的状态（仿射/均匀/矢量）

register, allowing the control logic to detect operations eligible for direction execution on the Control Processor.

- The dynamic uniform detection proposal from Gilani et al. [2013] stores the dynamically detected uniform values into a dedicated scalar register file.

Partial Register File Access. Lee et al. [2015] apply base, delta, immediate (BDI) compression to registers written back to the register file. The compressed registers are decompressed back to normal vectors as it is read back as a source operand. In this scheme, each compressed register still occupies the same storage slot as an uncompressed register, but only a subset of the register banks, thus it takes less energy to read out the compressed representation of the register.

Warp Approximate architecture [Wong et al., 2016] reduces the register read/write energy use by only accessing the lane corresponding to the representative thread selected via similarity detection.

Similarly, G-Scalar [Liu et al., 2017] features compressed registers that occupy only a subset of banks allocated or the uncompressed register to reduce the energy for register read.

Dedicated Affine Warp. Decoupled Affine Computation (DAC) [Wang and Lin, 2017] buffers all compiler-extracted affine variables in the registers of a dedicated affine warp. This affine warp shares the same vector register file storage just as the rest of the non-affine warps, but the affine warp uses individual lanes of each individual register entry to store the base, and the deltas for different non-affine warps.

Scalarize Operations

Aside from efficient storage, the operations with uniform or affine variables can be *scalarized*. Instead of repeating the same operation across all threads in a warp via a SIMD datapath, a scalar operation can be done once in single scalar datapath, consuming far less energy in the process. In general, an arithmetic operation can be scalarized if its input operands consist only of uniform or affine variables.

Dedicated Scalar Pipeline. AMD's GCN architecture features a dedicated scalar pipeline that executes scalar instructions generated by the compiler. FG-SIMT architecture [Kim et al., 2013] features a control processor that is capable of executing dynamically detected affine operations directly without evoking the SIMD datapath.

In both implementations, the scalar pipeline also handles the control flow and predication of the SIMD pipeline. The decoupling means that many system-related features (for example, communication with host processor) can also be offloaded to the scalar pipeline, freeing the SIMD datapath from the burden of implementing the full instruction set.

Clock-Gated SIMD Datapath. Warp Approximate architecture [Wong et al., 2016] and G-Scalar [Liu et al., 2017] both executes dynamically detected scalar instructions on one of the

寄存器，允许控制逻辑检测控制处理器上有资格直接执行的操作。

- Gilani 等人 [2013] 提出的动态统一检测提案将动态检测到的统一值存储到专用的标量寄存器文件中。

部分寄存器文件访问。Lee 等人 [2015] 将基数、增量、立即数 (BDI) 压缩应用于写回寄存器文件的寄存器。压缩寄存器在作为源操作数读回时被解压缩回法向量。在此方案中，每个压缩寄存器仍占用与未压缩寄存器相同的存储槽，但仅占用寄存器组的子集，因此读出寄存器的压缩表示所需的能量较少。

Warp Approximate 架构 [Wong et al., 2016] 通过仅访问与相似性检测选择的代表线程相对应的通道来减少寄存器读/写能耗。

类似地，G-Scalar [Liu et al.， 2017] 具有压缩寄存器，其仅占用分配的寄存器的子集或未压缩的寄存器，以减少寄存器读取的能量。

专用仿射扭曲。解耦仿射计算 (DAC) [Wang and Lin，2017] 在专用仿射扭曲的寄存器中缓冲所有编译器提取的仿射变量。此仿射扭曲与其他非仿射扭曲共享相同的矢量寄存器文件存储，但仿射扭曲使用每个单独寄存器条目的单独通道来存储基数以及不同非仿射扭曲的增量。

标量化操作

除了高效的存储之外，具有均匀或仿射变量的运算可以是 *scalarized*。标量运算可以在单个标量数据路径中执行一次，而不是通过 SIMD 数据路径在 warp 中的所有线程上重复相同的运算，从而在此过程中消耗的能量要少得多。一般来说，如果算术运算的输入操作数仅由均匀或仿射变量组成，则可以将其标量化。

专用标量流水线。AMD 的 GCN 架构具有专用标量流水线，可执行编译器生成的标量指令。FG-SIMT 架构 [Kim et al., 2013] 具有控制处理器，能够直接执行动态检测到的仿射运算，而无需调用 SIMD 数据路径。

在这两种实现中，标量流水线还处理 SIMD 流水线的控制流和预测。解耦意味着许多与系统相关的功能（例如，与主机处理器的通信）也可以卸载到标量流水线，从而使 SIMD 数据路径摆脱实现完整指令集的负担。

时钟门控 SIMD 数据路径。Warp Approximate 架构 [Wong et al., 2016] 和 G-Scalar [Liu et al., 2017] 均在其中一个上执行动态检测的标量指令

lanes in the SIMD datapath. When this happens, other lanes are clock-gated to reduce dynamic power consumption.

This approach eliminates the duplicating effort of supporting the full instruction set on the dedicated scalar datapath, or having to triage the subset to be implemented on the scalar datapath. For example, G-Scalar [Liu et al., 2017] can scalarize instructions supported by the special function units with relatively low overhead.

Aggregate to Affine Warp. Decoupled Affine Computation (DAC) [Wang and Lin, 2017] aggregates the affine operations from multiple warps into a single affine warp per SIMT core. This affine warp executes on the SIMD datapath just as other warps, but each instruction executed operates simultaneously on the affine representation of multiple warps.

Memory Access Acceleration

When a uniform or affine variable is used to represent the address of a memory operation (load/store), the memory locations touched by the memory operation is highly predictable—each successive location is separate by a known stride. This allows for various optimizations. For example, the memory coalescing of memory locations with a known stride is far simpler than coalescing of arbitrary random locations. The affine variable can also be used to represent a bulk transfer with a single instruction rather than via loops of load/store instructions.

FG-SIMT architecture [Kim et al., 2013] features a special address generation unit in the control process to expand memory accesses with affine addresses into the actual addresses. Since affine addresses have a fixed stride between threads, coalescing these affine memory accesses into cache lines can be done with simpler hardware.

Decoupled Affine Computation (DAC) [Wang and Lin, 2017] also features similar optimizations to exploit fixed strides in affine memory accesses. In addition, it uses an affine warp to execute ahead of the rest of the non-affine warps, prefetching data for these warps. The prefetched data is stored in the L1 cache, retrieved later by the corresponding non-affine warp via a special dequeue instruction.

3.6 RESEARCH DIRECTIONS ON REGISTER FILE ARCHITECTURE

Modern GPUs employ a large number of hardware threads (warps), multiplexing their execution on a far fewer (still large) number of ALUs, to tolerate both pipeline and memory access latencies. To allow fast and efficient switching between warps, GPUs use hardware warp schedulers and store the registers of all hardware threads in on-chip register files. On many GPU architectures, the capacity of these register files is substantial, and sometimes exceeding the capacity of the last-level cache, due to the wide-SIMD datapaths used in GPU, as well as the shear number of warps required to tolerate hundreds of cycles of memory access latency. For example,

SIMD 数据路径中的通道。发生这种情况时，其他通道将进行时钟门控以降低动态功耗。

这种方法消除了在专用标量数据路径上支持完整指令集的重复工作，也消除了必须对要在标量数据路径上实现的子集进行分类的重复工作。例如，G-Scalar [Liu et al., 2017] 可以以相对较低的开销标量化特殊功能单元支持的指令。

聚合到 Affine Warp。解耦 Affine 计算 (DAC) [Wang and Lin, 2017] 将来自多个 Warp 的 Affine 操作聚合到每个 SIMT 核心的单个 Affine Warp 中。此 Affine Warp 与其他 Warp 一样在 SIMD 数据路径上执行，但执行的每条指令同时在多个 Warp 的 Affine 表示上运行。

内存访问加速
当使用统一或仿射变量来表示内存操作（加载/存储）的地址时，内存操作涉及的内存位置是高度可预测的——每个连续位置都由已知步长分隔。这允许进行各种优化。例如，具有已知步长的内存位置的内存合并比任意随机位置的合并简单得多。仿射变量还可用于表示使用单个指令而不是通过加载/存储指令循环的批量传输。

FG-SIMT 架构 [Kim et al., 2013] 在控制过程中采用特殊的地址生成单元，将具有近似地址的内存访问扩展为实际地址。由于近似地址在线程之间具有固定的步长，因此可以使用更简单的硬件将这些近似内存访问合并到缓存行中。

解耦仿射计算 (DAC) [Wang and Lin, 2017] 也具有类似的优化，以利用仿射内存访问中的固定步长。此外，它使用仿射 Warp 在其他非仿射 Warp 之前执行，预取这些 Warp 的数据。预取的数据存储在 L1 缓存中，稍后由相应的非仿射 Warp 通过特殊的出队指令检索。

3.6 寄存器文件架构的研究方向

现代 GPU 使用大量硬件线程（warp），在数量少得多（但数量仍然很大）的 ALU 上多路复用执行，以容忍流水线和内存访问延迟。为了在 warp 之间快速高效地切换，GPU 使用硬件 warp 调度程序并将所有硬件线程的寄存器存储在片上寄存器文件中。在许多 GPU 架构中，这些寄存器文件的容量很大，有时会超过最后一级缓存的容量，这是由于 GPU 中使用的宽 SIMD 数据路径以及容忍数百个内存访问延迟周期所需的 warp 数量。例如，

NVIDIA's Fermi GPU can sustain over 20,000 in-flight threads, and has an aggregate register capacity of 2 MB.

To minimize the area consumed by register file storage, register files on GPUs are generally implemented via low-port count SRAM banks. The SRAM banks are accessed in parallel to supply the operand bandwidth required to sustain instructions running on the wide-SIMD pipeline at peak throughput. As described earlier in this chapter, some GPUs use an operand collector to orchestrate operand accesses from multiple instructions to minimize bank-conflict penalties.

Accessing these large register files consumes a high amount of dynamic energy at each access, and their large size also leads to high static power consumption. On a NVIDIA GTX280 GPU, nearly 10% of the total GPU power is consumed by the register file. This provides clear incentives to innovate on GPU register file architectures to reduce their energy consumption. As a result, there have been a large number of research papers on this topic in recent years. The rest of this section summarizes several research proposals aimed to achieve this goal.

3.6.1 HIERARCHICAL REGISTER FILE

Gebhart et al. [2011b] observe that among a set of real-world graphics and compute workloads, up to 70% of the values produced by an instruction are only read once, and only 10% are only read more than twice. To capture this short lifetime among a majority of the register values, they propose extending the main register file on GPU with a *register file cache* (RFC). This forms a hierarchy of the register file, and dramatically reduces the access frequency to the main register file.

In this work, the RFC allocates a new entry, via a FIFO replacement policy, for the destination operand of every instruction. Source operands that miss the RFC are not loaded onto the RFC to reduce pollution of the already small RFC. By default, every value evicted from RFC is written back to the main register file. However, since many of these values are never read again, Gebhart et al. [2011b] extend the hardware-only RFC with compile time-generated static liveness information. An extra bit is added to the instruction encoding to indicate the last instruction consuming a register value. A register that has been read for the last time is marked dead in the RFC. At eviction, it will not be written back to the main register file.

To further reduce the size of the RFC, Gebhart et al. [2011b] combine it with a two-level warp scheduler. This two-level warp scheduler restricts execution to a pool of *active* warps, which only consists of a small subset of the warps in each SIMT core. This work considers an active warp pool of 4–8 warps, out of 32 warps in total for each SIMT core. The RFC only holds values from the active warps, and is therefore smaller. A warp is removed from the active pool at long-latency operations, such as global memory loads or texture fetches. When this occurs, the RFC entries of the warp are flushed, freeing up space for a different warp made active by the second-level scheduler.

NVIDIA 的 Fermi GPU 可以支持超过 20,000 个运行线程，并且总寄存器容量为 2 MB。

为了最大限度地减少寄存器文件存储所占用的区域，GPU 上的寄存器文件通常通过低端口数 SRAM 组实现。SRAM 组被并行访问，以提供维持在宽 SIMD 流水线上以峰值吞吐量运行的指令所需的操作数带宽。如本章前面所述，一些 GPU 使用操作数收集器来协调来自多个指令的操作数访问，以最大限度地减少组冲突惩罚。

访问这些大型寄存器文件每次访问都会消耗大量动态能量，而且它们的大尺寸也会导致高静态功耗。在 NVIDIA GTX280 GPU 上，寄存器文件消耗了近 10% 的 GPU 总功率。这为创新 GPU 寄存器文件架构以降低其能耗提供了明确的激励。因此，近年来出现了大量关于这一主题的研究论文。本节的其余部分总结了旨在实现这一目标的几项研究提案。

3.6.1 分层寄存器文件

Gebhart 等人 [2011b] 观察到，在一组现实世界的图形和计算工作负载中，一条指令产生的值中多达 70% 只被读取一次，只有 10% 只被读取两次以上。为了捕捉大多数寄存器值的这种短暂生命周期，他们建议使用 *register file cache* (RFC) 扩展 GPU 上的主寄存器文件。这形成了寄存器文件的层次结构，并大大降低了对主寄存器文件的访问频率。

在本研究中，RFC 通过 FIFO 替换策略为每条指令的目标操作数分配一个新条目。未命中 RFC 的源操作数不会加载到 RFC 中，以减少对已经很小的 RFC 的污染。默认情况下，从 RFC 中逐出的每个值都会写回主寄存器文件。然而，由于其中许多值再也不会被读取，Gebhart 等人 [2011b] 使用编译时生成的静态活跃信息扩展了纯硬件 RFC。在指令编码中添加了一个额外的位，以指示最后一条使用寄存器值的指令。最后一次读取的寄存器在 RFC 中被标记为死寄存器。在逐出时，它不会写回主寄存器文件。

为了进一步减小 RFC 的大小，Gebhart 等人 [2011b] 将其与两级 Warp 调度程序相结合。这个两级 Warp 调度程序将执行限制在一个 *active* Warp 池中，该池仅由每个 SIMT 核心中的一小部分 Warp 组成。这项工作考虑了 4-8 个 Warp 的活动 Warp 池，每个 SIMT 核心总共有 32 个 Warp。RFC 只保存来自活动 Warp 的值，因此较小。在执行长延迟操作（例如全局内存加载或纹理提取）时，Warp 会从活动池中移除。发生这种情况时，Warp 的 RFC 条目将被刷新，从而为由第二级调度程序激活的另一个 Warp 释放空间。

Compile-Time Managed Register File Hierarchy. Gebhart et al. [2011a] further extend this register file hierarchy with a Last Result File (LRF), which just buffers the register value produced by the last instruction of each active warp. This work also replaces a hardware-managed RFC with a compile-time managed operand register file (ORF). Movements of values in and out of the ORF is managed explicitly by the compiler. This removes the tag-lookup required by the RFC. The compiler also has a more holistic view of the register usage pattern in most GPU workloads, allowing it to make more optimal decisions. This work also extends the two-level warp scheduler so that the compiler indicates when a warp can be switched out of the active pool. This is required to coordinate the content of the ORF with the activeness of the warp, moving all live data from the ORF back to the main register file before the warp is switched out.

3.6.2 DROWSY STATE REGISTER FILE

Abdel-Majeed and Annavaram [2013] proposes a tri-modal register file design that reduces the leakage power of the large GPU register file. Each entry in the tri-modal register file can switch between ON, OFF, and Drowsy mode. ON mode is the normal mode of operation; OFF mode does not retain the value of the register; Drowsy mode retains the value of the register, but the entry needs to be awaken to ON mode before access. In this work, all non-allocated registers are in OFF mode, and all allocated registers are placed into drowsy state immediately after each access. This policy takes advantage of the long delays between consecutive access to the same register on GPU, due to the fine-grained multithreading on GPU, to allow registers in the register file to spend most time in drowsy mode. The long pipeline in GPU also means that additional latency of waking up a register from drowsy state does not introduce significant performance penalty.

3.6.3 REGISTER FILE VIRTUALIZATION

Tarjan and Skadron [2011] observe that while waiting for memory operations the number of live registers in a GPU thread tends to be small. For some GPU applications they claim up to 60% of registers go unused. They propose reducing the size of the physical register file by up to 50% or double the number of concurrently executing threads by using register renaming to virtualize physical registers. In the proposed mechanism, a thread starts executing with no registers allocated and physical registers are allocated to destination registers as instructions are decoded. Tarjan and Skadron further suggest that deallocation of physical registers can be enhanced by employing compiler analysis to determine the last read of a register. They propose "final read annotations" and suggest adding "a bit for each operand to indicate whether it is a last read" and point out this may require additional bits in the instruction encoding.

Jeon et al. [2015] quantify the impact of reducing the GPU register file size by spilling registers to memory. They find that reducing the size of the register file by 50% by employing spilling increased execution time by an average of 73%. They review older proposals for reclaiming physical registers early when employing register renaming on CPUs employing out-of-order

编译时管理的寄存器文件层次结构。Gebhart 等人 [2011a] 进一步扩展了此寄存器文件层次结构，添加了最后结果文件 (LRF)，该文件仅缓冲每个活动 Warp 的最后一条指令产生的寄存器值。这项工作还用编译时管理的操作数寄存器文件 (ORF) 取代了硬件管理的 RFC。ORF 中值的进出移动由编译器明确管理。这消除了 RFC 所需的标签查找。编译器还可以更全面地了解大多数 GPU 工作负载中的寄存器使用模式，从而做出更优化的决策。这项工作还扩展了两级 Warp 调度程序，以便编译器指示何时可以将 Warp 切换出活动池。这需要协调 ORF 的内容和 Warp 的活跃度，在 Warp 切换出之前将所有实时数据从 ORF 移回主寄存器文件。

3.6.2 瞌睡状态寄存器文件

Abdel-Majeed 和 Annavaram [2013] 提出了一种三模式寄存器文件设计，以降低大型 GPU 寄存器文件的泄漏功率。三模式寄存器文件中的每个条目都可以在 ON、OFF 和 Drowsy 模式之间切换。ON 模式是正常操作模式；OFF 模式不保留寄存器的值；Drowsy 模式保留寄存器的值，但条目需要在访问前唤醒到 ON 模式。在这项工作中，所有未分配的寄存器都处于 OFF 模式，并且所有分配的寄存器在每次访问后立即进入休眠状态。由于 GPU 上的细粒度多线程，此策略利用了 GPU 上对同一寄存器的连续访问之间的长延迟，以允许寄存器文件中的寄存器大部分时间处于休眠模式。GPU 中的长管道还意味着从休眠状态唤醒寄存器的额外延迟不会带来显着的性能损失。

3.6.3 寄存器文件虚拟化

Tarjan 和 Skadron [2011] 观察到，在等待内存操作时，GPU 线程中的活动寄存器数量往往很少。对于某些 GPU 应用程序，他们声称多达 60% 的寄存器未被使用。他们建议通过使用寄存器重命名来虚拟化物理寄存器，将物理寄存器文件的大小减少多达 50% 或将并发执行的线程数量增加一倍。在提出的机制中，线程开始执行时没有分配任何寄存器，并且在解码指令时将物理寄存器分配给目标寄存器。Tarjan 和 Skadron 进一步建议，可以通过使用编译器分析来确定寄存器的最后一次读取，从而增强物理寄存器的释放。他们提出了“ 最终读取注释 ”，并建议为每个操作数添加“ 一个位来指示它是否是最后一次读取 ”，并指出这可能需要在指令编码中添加额外的位。

Jeon 等人 [2015] 量化了通过将寄存器溢出到内存来减少 GPU 寄存器文件大小的影响。他们发现，通过使用溢出将寄存器文件的大小减少 50% 会使执行时间平均增加 73%。他们回顾了在使用无序的 CPU 上使用寄存器重命名时尽早回收物理寄存器的旧提案

execution. The propose addressing the issue of the additional bits required to add "final read annotations" by adding "metadata instructions" that efficiently encode when physical registers can be reclaimed and generate these using register lifetime liveness analysis. An important observation they make is that branch divergence must be taken into account when determining where it is safe to reclaim physical registers (further elaborated upon by Kloosterman et al. [2017]). For a 128 KB register file a straightforward implementation of Jeon et al.'s renaming technique requires 3.8 KB of renaming hardware. They show this overhead can be reduced to 1 KB by not renaming registers with long lifetimes. To exploit this opportunity they propose using renaming only for registers with logical register numbers larger than a compiler determined threshold. Jeon et al. further propose using renaming to enable power gating of register file subarrays. They evaluate the effectiveness of their detailed proposal for supporting register file virtualization via register renaming showing that a reduction in the size of the register file by 50% with no performance loss is indeed obtainable.

3.6.4 PARTITIONED REGISTER FILE

Abdel-Majeed et al. [2017] introduces the *Pilot Register File*, which partitions the GPU register file into a fast and a slow register file (FRF and SRF). The FRF is implemented using regular SRAMs, whereas the SRF is implemented using near-threshold voltage (NTV) SRAMs. Compared to regular SRAMs, NTV SRAMs feature far lower access energy as well as a much lower leakage power. In exchange, access latency to the NTV SRAMs are far slower, often consists of several cycles (instead of one cycle in regular SRAMs). In this work, the SRF is significantly larger than the FRF. Every warp has 4 entries in the FRF. The key is to use FRF to service most of the access to compensate for the slowness of the SRF. The additional latency for accessing the SRF is handled by the operand collector. The FRF is further enhanced with a low power mode using FinFET's back gate control. This allow the FRF of an inactive warp to switch to low power mode. This allows FRF to reap the benefit of two-level scheduler without explicit scheduling warps in and out of an active pool.

This work is distinct from the hierarchical register file in that the different partitions hold an exclusive set of registers, and the partition remains constant throughout the lifetime of the warp. Instead of using a compiler to determine the set of registers to be placed in the FRF, Abdel-Majeed et al. [2017] employs a pilot CTA at each kernel launch to profile the most-frequently used registers. This set of high-use registers are recorded in a lookup table that is accessible to every subsequent warp from the kernel launch.

3.6.5 REGLESS

Kloosterman et al. [2017] introduces *RegLess*, which aims at eliminating the register file and replace it with an operand staging buffer. The paper observes that over relatively short spans of time the number of registers accessed is a small fraction of the total register file capacity. For example, over a period of 100 cycles many of the applications they evaluated accessed less than

执行。他们建议通过添加“元数据指令”来解决添加“最终读取注释”所需的额外位的问题，这些指令可以有效地编码何时可以回收物理寄存器，并使用寄存器生存期活跃度分析生成这些指令。他们提出的一个重要观察是，在确定在哪里可以安全回收物理寄存器时，必须考虑分支发散（Kloosterman 等人 [2017] 进一步阐述）。对于 128 KB 的寄存器文件，Jeon 等人的重命名技术的直接实现需要 3.8 KB 的重命名硬件。他们表明，通过不重命名具有长生存期的寄存器可以将这个开销减少到 1 KB。为了利用这个机会，他们建议仅对逻辑寄存器编号大于编译器确定的阈值的寄存器使用重命名。Jeon 等人进一步建议使用重命名来启用寄存器文件子阵列的电源门控。他们评估了通过寄存器重命名支持寄存器文件虚拟化的详细提案的有效性，结果表明确实可以将寄存器文件的大小减少 50%，而性能不会有任何损失。

3.6.4 分区寄存器文件

Abdel-Majeed 等人 [2017] 引入了 *Pilot Register File*，它将 GPU 寄存器文件划分为快速和慢速寄存器文件（FRF 和 SRF）。FRF 使用常规 SRAM 实现，而 SRF 使用近阈值电压 (NTV) SRAM 实现。与常规 SRAM 相比，NTV SRAM 的访问能耗低得多，泄漏功率也低得多。作为交换，NTV SRAM 的访问延迟要慢得多，通常包含几个周期（而不是常规 SRAM 中的一个周期）。在这项工作中，SRF 明显大于 FRF。每个 warp 在 FRF 中都有 4 个条目。关键是使用 FRF 来服务大部分访问以弥补 SRF 的缓慢。访问 SRF 的额外延迟由操作数收集器处理。通过使用 FinFET 的背栅控制，FRF 进一步增强了低功耗模式。这允许非活动 Warp 的 FRF 切换到低功耗模式。这样，FRF 就可以享受两级调度程序的好处，而无需明确调度 Warp 进出活动池。

这项工作与分层寄存器文件不同，因为不同的分区保存一组独有的寄存器，并且分区在 Warp 的整个生命周期中保持不变。Abdel-Majeed 等人 [2017] 不使用编译器来确定要放置在 FRF 中的寄存器集，而是在每次内核启动时使用引导 CTA 来分析最常用的寄存器。这组高使用率的寄存器记录在查找表中，内核启动后的每个后续 Warp 都可以访问该查找表。

3.6.5 调整

Kloosterman 等人 [2017] 引入了 *RegLess*，旨在消除寄存器文件并将其替换为操作数暂存缓冲区。本文观察到，在相对较短的时间内，访问的寄存器数量只是总寄存器文件容量的一小部分。例如，在 100 个周期内，他们评估的许多应用程序访问的次数少于

10% of a 2048 KB register file when using a GTO or two-level warp scheduler. To take advantage of this observation RegLess uses a compiler algorithm to divide up kernel execution into regions. Regions are contiguous instructions within a single basic block. The boundary between regions is selected so as to limit the number of live registers. Using the region annotations, a Capacity Manager (CM) determines which warps are eligible for scheduling. When a warp begins executing instructions from a new region the registers used in that region are brought into an Operand Staging Unit (OSU) from a backing storage area allocated in global memory and potentially cached in the L1 data cache. The OSU is essentially a cache consisting of eight banks which provides enough bandwidth to service two instructions per cycle. To avoid stalling while accessing data in the OSU, the CM preloads registers before issuing the first instruction in a region. To manage the preloading process the CM maintains a state machine for each warp indicating whether the registers needed for the next region are present in the OSU. To reduce the amount of memory traffic generated between the OSU and memory hierarchy, RegLess employs register compression techniques that exploit affine values (see Section 3.5) are employed.

Kloosterman et al. performed a detailed evaluation of their proposal including Verilog synthesis and extraction of parasitic capacitance and resistance values of the hardware units introduced by RegLess. Their evaluation shows a 512 entry OSU can achieves slightly better performance versus a 2048 KB register file while occupying only 25% of the space and reducing overall GPU energy consumption by 11%.

使用 GTO 或两级 Warp 调度程序时，约占 2048 KB 寄存器文件的 10%。为了利用这一观察结果，RegLess 使用编译器算法将内核执行划分为多个区域。区域是单个基本块内的连续指令。选择区域之间的边界是为了限制活动寄存器的数量。容量管理器 (CM) 使用区域注释确定哪些 Warp 有资格进行调度。当 Warp 开始执行新区域的指令时，该区域中使用的寄存器将从全局内存中分配的备用存储区域带入操作数暂存单元 (OSU)，并可能缓存在 L1 数据缓存中。OSU 本质上是一个由 8 个存储体组成的缓存，它提供的带宽足以每周期处理两条指令。为了避免在访问 OSU 中的数据时停顿，CM 会在发出区域中的第一条指令之前预加载寄存器。为了管理预加载过程，CM 为每个 warp 维护一个状态机，指示下一个区域所需的寄存器是否存在于 OSU 中。为了减少 OSU 和内存层次结构之间产生的内存流量，RegLess 采用了利用近似值的寄存器压缩技术（参见第 3.5 节）。

Kloosterman 等人对其提案进行了详细评估，包括 Verilog 综合和提取 RegLess 引入的硬件单元的寄生电容和电阻值。他们的评估表明，512 个条目的 OSU 可以实现比 2048 KB 寄存器文件略好的性能，同时仅占用 25% 的空间并将整体 GPU 能耗降低 11%。

CHAPTER 4

Memory System

This chapter explores the memory system of GPUs. GPU computing kernels interact with the memory system through load and store instructions. Traditional graphics applications interact with several memory spaces such as texture, constant, and render surfaces. While access to these memory spaces is available in GPGPU programming APIs like CUDA, we will focus on memory spaces employed in GPGPU programming in this chapter and in particular the microarchitecture support required to implement them.

CPUs typically include two separate memory spaces: The register file and memory. Modern GPUs logically subdivide memory further into local and global memory spaces. The local memory space is private per thread and typically used for register spilling while global memory is used for data structures that are shared among multiple threads. In addition, modern GPUs typically implement a programmer managed scratchpad memory with shared access among threads that execute together in a cooperative thread array. One motivation for including a shared address space is that in many applications a programmer knows which data needs to be accessed at a given step in a computation. By loading all of this data into shared memory at once they can overlap long latency off-chip memory accesses and avoid long latency accesses to memory while performing computation on this data. More importantly, the number of bytes that can be transferred between the GPU and off-chip memory in a given amount of time (DRAM bandwidth) is small relative to the number of instructions that can be executed in that same amount of time. Moreover, the energy consumed to transfer data between off-chip memory and the GPU is orders of magnitude higher than the energy consumed accessing data from on-chip memory. Thus, accessing data from on-chip memory yields higher performance and saves energy.

We divide our discussion of the memory system into two parts reflecting the division of memory into portions that reside within the GPU cores and within memory partitions that connect to off-chip DRAM chips.

4.1 FIRST-LEVEL MEMORY STRUCTURES

This section describes the first-level cache structures found on GPUs with a focus on the unified L1 data cache and scratch pad "shared memory" and how these interact with the core pipeline. We also include a brief discussion of a typical microarchitecture for an L1 texture cache. We include discussion of the texture cache, which has found limited use in GPU computing applications, as it provides some insights and intuition as to how GPUs differ from CPUs. A recent patent describes how one might unify the texture cache and L1 data (e.g., as found in NVIDIA's

CHAPTER 4

记忆系统

本章探讨了 GPU 的内存系统。GPU 计算内核通过加载和存储指令与内存系统交互。传统图形应用程序与多个内存空间（如纹理、常量和渲染表面）交互。虽然 CUDA 等 GP GPU 编程 API 可以访问这些内存空间，但本章我们将重点介绍 GPGPU 编程中使用的内存空间，特别是实现它们所需的微架构支持。

CPU 通常包括两个独立的内存空间：寄存器文件和内存。现代 GPU 在逻辑上将内存进一步细分为本地和全局内存空间。本地内存空间是每个线程私有的，通常用于寄存器溢出，而全局内存用于多个线程共享的数据结构。此外，现代 GPU 通常实现程序员管理的暂存器内存，在协作线程阵列中一起执行的线程之间共享访问权限。包含共享地址空间的一个动机是，在许多应用程序中，程序员知道在计算的给定步骤中需要访问哪些数据。通过一次将所有这些数据加载到共享内存中，它们可以重叠长延迟的片外内存访问，并在对这些数据执行计算时避免对内存的长延迟访问。更重要的是，在给定时间内（DRAM 带宽）在 GPU 和片外内存之间传输的字节数相对于在相同时间内可以执行的指令数较小。此外，在片外存储器和 GPU 之间传输数据所消耗的能量比从片上存储器访问数据所消耗的能量高出几个数量级。因此，从片上存储器访问数据可以获得更高的性能并节省能源。

我们将对内存系统的讨论分为两部分，分别反映内存分为驻留在 GPU 核心内的部分和连接到片外 DRAM 芯片的内存分区内的部分。

4.1 一级存储器结构

本节介绍 GPU 上的一级缓存结构，重点介绍统一的 L1 数据缓存和暂存器“ 共享内存 ”，以及它们如何与核心管道交互。我们还简要讨论了 L1 纹理缓存的典型微架构。我们讨论了纹理缓存，它在 GPU 计算应用中的使用有限，因为它提供了一些关于 GPU 与 CPU 不同之处的见解和直觉。最近的一项专利描述了如何统一纹理缓存和 L1 数据（例如，在 NVIDIA 的

Maxwell and Pascal GPUs) [Heinrich et al., 2017]. We defer discussion of this design until after first considering how texture caches are organized. An interesting aspect of the first-level memory structures in GPUs is how they interact with the core pipeline when hazards are encountered. As noted in Chapter 3, pipeline hazards can be handled by replaying instructions. We expand on our earlier discussion of replay in this chapter with a focus on hazards in the memory system.

4.1.1 SCRATCHPAD MEMORY AND L1 DATA CACHE

In the CUDA programming model, "shared memory" refers to a relatively small memory space that is expected to have low latency but which is accessible to all threads within a given CTA. In other architectures, such a memory space is sometimes referred to as a scratchpad memory [Hofstee, 2005]. The latency to access this memory space is typically comparable to register file access latency. Indeed, early NVIDIA patents refer to CUDA "shared memory" as a Global Register File [Acocella and Goudy, 2010]. In OpenCL this memory space is referred to as "local memory." From a programmer perspective a key aspect to consider when using shared memory, beyond its limited capacity, is the potential for *bank conflicts*. The shared memory is implemented as a static random access memory (SRAM) and is described in some patents [Minkin et al., 2012] as being implemented with one bank per lane with each bank having one read port and one write port. Each thread has access to all of the banks. A *bank conflict* arises when more than one thread accesses the same bank on a given cycle and the threads wish to access distinct locations in that bank. Before considering in detail how the shared memory is implemented we first look at the L1 data cache.

The L1 data cache maintains a subset of the global memory address space in the cache. In some architectures the L1 cache contains only locations that are not modified by kernels, which helps avoid complications due to the lack of cache coherence on GPUs. From a programmer perspective a key consideration when accessing global memory is the relationship, with respect to each other, of memory locations accessed by different threads within a given warp. If all threads in a warp access locations that fall within a single L1 data cache block and that block is not present in the cache, then only a single request needs to be sent to lower level caches. Such accesses are said to be "coalesced." If the threads within a warp access different cache blocks then multiple memory accesses need to be generated. Such accesses are said to be uncoalesced. Programmers try to avoid both bank conflicts and uncoalesced accesses, but to ease programming the hardware allows both.

Figure 4.1 illustrates a GPU cache organization like that described by Minkin et al. [2012]. The design pictured implements a unified shared memory and L1 data cache, which is a feature introduced in NVIDIA's Fermi architecture that is also present in the Kepler architecture. At the center of the diagram is an SRAM data array ❺ which can be configured [Minkin et al., 2013] partly for direct mapped access for shared memory and partly as a set associative cache. The design supports a non-stalling interface with the instruction pipeline by using a replay mechanism when handling bank conflicts and L1 data cache misses. To help explain the operation of this

Maxwell 和 Pascal GPU）[Heinrich 等人，2017]。我们将这种设计的讨论推迟到首先考虑纹理缓存的组织方式之后。GPU 中第一级内存结构的一个有趣方面是它们在遇到风险时如何与核心管道交互。如第 3 章所述，管道风险可以通过重放指令来处理。我们在本章中扩展了之前关于重放的讨论，重点关注内存系统中的风险。

4.1.1 暂存器和 L1 数据缓存

在 CUDA 编程模型中，“ 共享内存 ” 是指相对较小的内存空间，预期延迟较低，但给定 CTA 内的所有线程都可以访问。在其他架构中，这种内存空间有时被称为暂存器内存 [Hofstee，2005]。访问此内存空间的延迟通常与寄存器文件访问延迟相当。事实上，早期的 NVIDIA 专利将 CUDA “ 共享内存 ” 称为全局寄存器文件 [Acocella 和 Goudy，2010]。在 OpenCL 中，此内存空间称为 “ 本地内存 ”。从程序员的角度来看，使用共享内存时要考虑的一个关键方面是其有限的容量之外的 *bank conflicts* 的潜力。共享内存实现为静态随机存取存储器 (SRAM)，在一些专利 [Minkin 等，2012] 中描述为每通道一个存储体，每个存储体有一个读取端口和一个写入端口。每个线程都可以访问所有存储体。当多个线程在给定周期内访问同一个存储体，并且这些线程希望访问该存储体中的不同位置时，就会出现 *bank conflict*。在详细考虑共享内存的实现方式之前，我们首先来看一下 L1 数据缓存。

L1 数据缓存维护缓存中全局内存地址空间的子集。在某些架构中，L1 缓存仅包含未被内核修改的位置，这有助于避免由于 GPU 上缺乏缓存一致性而导致的复杂性。从程序员的角度来看，访问全局内存时的一个关键考虑因素是给定 warp 内不同线程访问的内存位置之间的关系。如果 warp 中的所有线程都访问单个 L1 数据缓存块内的位置，并且该块不在缓存中，则只需向较低级别的缓存发送单个请求。这种访问被称为 “ 合并 ”。如果 warp 中的线程访问不同的缓存块，则需要生成多个内存访问。这种访问被称为未合并。程序员试图避免库冲突和未合并访问，但为了简化编程，硬件允许两者。

图 4.1 展示了 Minkin 等人 [2012] 描述的 GPU 缓存组织。图示中的设计实现了统一的共享内存和 L1 数据缓存，这是 NVIDIA Fermi 架构中引入的一项功能，Kepler 架构中也有此功能。图表的中心是 SRAM 数据阵列 5，可以配置 [Minkin et al., 2013] 部分用于共享内存的直接映射访问，部分配置为组相联缓存。该设计在处理存储体冲突和 L1 数据缓存未命中时使用重放机制，支持与指令流水线的非停顿接口。为了帮助解释此操作，

cache architecture we first consider how shared memory accesses are processed, then consider coalesced cache hits, and finally consider cache misses and uncoalesced accesses. For all cases, a memory access request is first sent from the load/store unit inside the instruction pipeline to the L1 cache ❶. A memory access request consists of a set of memory addresses, one for each thread in a warp along with the operation type.

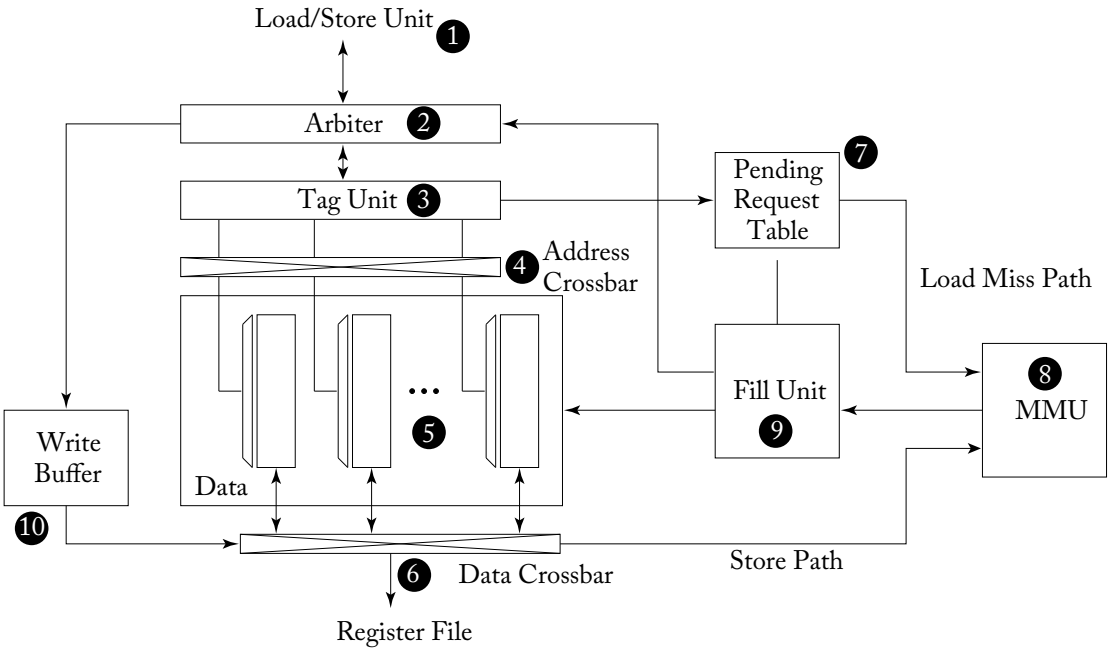


Figure 4.1: Unified L1 data cache and shared memory [Minkin et al., 2012].

Shared Memory Access Operations

For a shared memory accesses the arbiter determines whether the requested addresses within the warp will cause bank conflicts. If the requested addresses would cause one or more bank conflicts, the arbiter splits the request into two parts. The first part includes addresses for a subset of threads in the warp which do not have bank conflicts. This part of the original request is accepted by the arbiter for further processing by the cache. The second part contains those addresses that cause bank conflicts with addresses in the first part. This part of the original request is returned to the instruction pipeline and must be executed again. This subsequent execution is known as a "replay." There is a tradeoff in where the replay part of the original shared memory request is stored. While area can be saved by replaying the memory access instruction from the instruction buffer this consumes energy in accessing the large register file. A better alternative for energy efficiency may be to provide limited buffering for replaying memory access instructions in the

缓存架构我们首先考虑如何处理共享内存访问，然后考虑合并的缓存命中，最后考虑缓存未命中和未合并的访问。对于所有情况，内存访问请求首先从指令管道内的加载/存储单元发送到 L1 缓存 1 。内存访问请求由一组内存地址组成，每个线程都有一个内存地址以及操作类型。

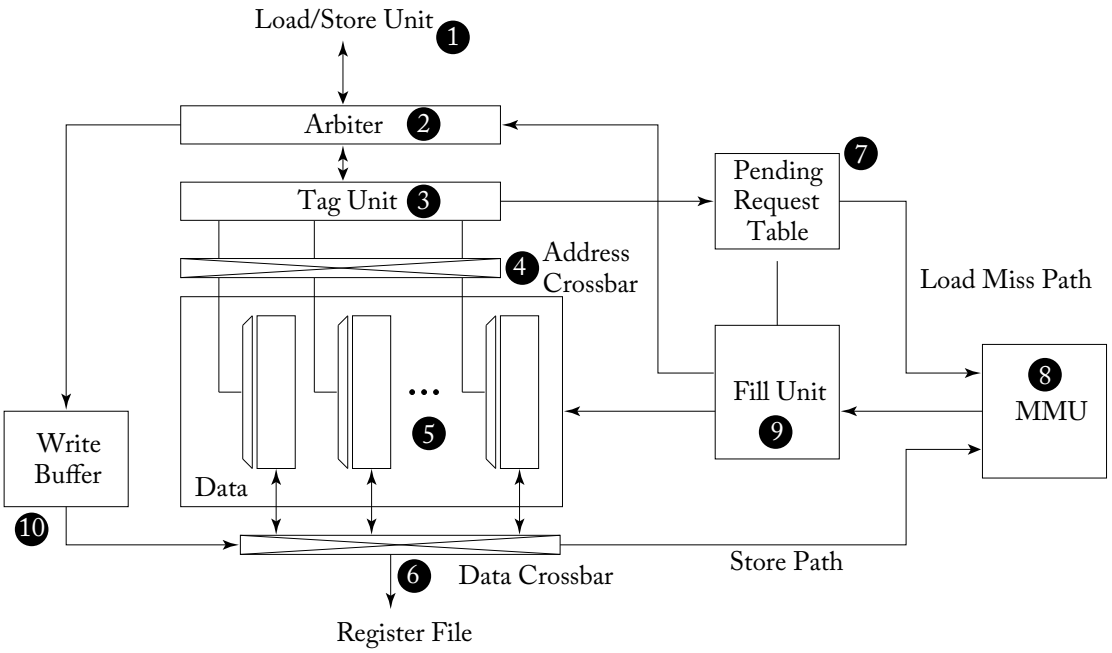


图 4.1：统一的 L1 数据缓存和共享内存 [Minkin et al., 2012]。

共享内存访问操作

对于共享内存访问，仲裁器会确定 Warp 中请求的地址是否会导致存储体冲突。如果请求的地址会导致一个或多个存储体冲突，仲裁器会将请求拆分为两部分。第一部分包括 Warp 中没有存储体冲突的线程子集的地址。仲裁器会接受原始请求的这一部分，以供缓存进一步处理。第二部分包含与第一部分中的地址导致存储体冲突的地址。原始请求的这一部分将返回到指令管道，必须再次执行。此后续执行称为“ 重放 ”。原始共享内存请求的重放部分的存储位置存在权衡。虽然可以通过从指令缓冲区重放内存访问指令来节省面积，但这会消耗访问大型寄存器文件的能量。更好的节能替代方案可能是在 Warp 中提供有限的缓冲来重放内存访问指令

load/store unit and avoiding scheduling memory access operations from the instruction buffer when free space in this buffer beings to run out. Before considering what happens to the replay request, let us consider how the accepted portion of the memory request is processed.

The accepted portion of a shared memory request bypasses tag lookup inside the tag unit ❸ as shared memory is direct mapped. When accepting a shared memory load request the arbiter schedules a writeback event to the register file inside the instruction pipeline as the latency of the direct mapped memory lookup is constant in the absence of bank conflicts. The tag unit determines which bank each thread's request maps to so as to control the address crossbar ❹ which distributes addresses to the individual banks within the data array. Each bank inside the data array ❺ is 32-bits wide and has its own decoder allowing for independent access to different rows in each bank. The data is returned to the appropriate thread's lane for storage in the register file via the data crossbar ❻. Only lanes corresponding to active threads in the warp write a value to the register file.

Assuming a single-cycle latency for shared memory lookup, the replayed portion of a shared memory request can access the L1 cache arbiter the cycle after the previous accepted portion. If this replayed portion encounters bank conflicts it is further subdivided into an accepted and replayed portion.

Cache Read Operations

Next, let us consider how a load to the global memory space is processed. As only a subset of the global memory space is cached in the L1 the tag unit will need to check whether the data is present in the cache or not. While the data array is highly banked to enable flexible access to shared memory by individual warps, access to global memory is restricted to a single cache block per cycle. This restriction helps to reduce tag storage overhead relative to the amount of cached data and is also a consequence of the standard interface to standard DRAM chips. The L1 cache block size is 128 bytes in Fermi and Kepler and is further divided into four 32-byte sectors [Liptay, 1968] in Maxwell and Pascal [NVIDIA Corp.]. The 32-byte sector size corresponds to the minimum size of data that can be read from a recent graphics DRAM chip in a single access (e.g., GDDR5). Each 128-byte cache block is composed of 32-bit entries at the same row in each of the 32 banks.

The load/store unit ❶ computes memory addresses and applies the coalescing rules to break a warp's memory access into individual coalesced accesses which are then fed into the arbiter ❷. The arbiter may reject a request if enough resources are not available. For example, if all ways in the cache set that the access maps to are busy or there are no free entries in the pending request table ❼, which is described below. Assuming enough resources are available to handle a miss, the arbiter requests the instruction pipeline schedule a writeback to the register file a fixed number of cycles in the future corresponding to a cache hit. In parallel the arbiter also requests the Tag Unit ❸ check whether the access in fact leads to a cache hit or a miss. In the event of a cache hit, the appropriate row of the data array ❺ is accessed in all banks and

加载/存储单元，并避免在指令缓冲区的可用空间即将用完时从指令缓冲区调度内存访问操作。在考虑重放请求会发生什么之前，让我们先考虑一下如何处理内存请求的已接受部分。

由于共享内存是直接映射的，因此共享内存请求的已接受部分将绕过标签单元 3 内的标签查找。当接受共享内存加载请求时，仲裁器将安排一个写回事件到指令管道内的寄存器文件，因为在没有存储体冲突的情况下，直接映射内存查找的延迟是恒定的。标签单元确定每个线程的请求映射到哪个存储体，以控制地址交叉开关 4，地址交叉开关 4 将地址分配给数据阵列中的各个存储体。数据阵列 5 内的每个存储体都是 32 位宽，并且具有自己的解码器，允许独立访问每个存储体中的不同行。数据通过数据交叉开关 6 返回到相应线程的通道以存储在寄存器文件中。只有与 warp 中的活动线程相对应的通道才会将值写入寄存器文件。

假设共享内存查找的延迟为单周期，则共享内存请求的重放部分可以在前一个接受部分之后的周期访问 L1 缓存仲裁器。如果此重放部分遇到库冲突，则进一步细分为接受和重放部分。

缓存读取操作

接下来，让我们考虑如何处理全局内存空间的加载。由于 L1 中仅缓存了全局内存空间的子集，因此标记单元需要检查数据是否存在于缓存中。虽然数据阵列高度存储，以便各个 warp 能够灵活访问共享内存，但对全局内存的访问仅限于每个周期一个缓存块。此限制有助于减少相对于缓存数据量的标记存储开销，也是标准 DRAM 芯片标准接口的结果。Fermi 和 Kepler 中的 L1 缓存块大小为 128 字节，Maxwell 和 Pascal [NVIDIA Corp.] 中进一步分为四个 32 字节扇区 [Liptay，1968]。32 字节扇区大小对应于单次访问中可从最新图形 DRAM 芯片（例如 GDDR5）读取的最小数据大小。每个 128 字节缓存块由 32 个存储体中同一行的 32 位条目组成。

加载/存储单元 1 计算内存地址并应用合并规则将 warp 的内存访问分解为单独的合并访问，然后将其输入到仲裁器 2 。如果资源不足，仲裁器可能会拒绝请求。例如，如果访问映射到的缓存集中的所有路径都处于繁忙状态，或者待处理请求表 7 中没有空闲条目（如下所述）。假设有足够的资源来处理未命中，仲裁器请求指令流水线在未来固定数量的周期内安排对寄存器文件的写回，以对应缓存命中。同时，仲裁器还请求标记单元 3 检查访问实际上是否导致缓存命中或未命中。如果发生缓存命中，则在所有存储体中访问数据阵列 5 的相应行，并且

the data is returned ❻ to the register file in the instruction pipeline. As in the case of shared memory accesses, only register lanes corresponding to active threads are updated.

When accessing the Tag Unit, if it is determined that a request triggers a cache miss, the arbiter informs the load/store unit it must replay the request and in parallel it sends the request information to the pending request table (PRT) ❼. The pending request table provides functionality not unlike that supported by traditional miss-status holding registers [Kroft, 1981] in CPU cache memory systems. There are at least two versions of the pending request table described in NVIDIA patents [Minkin et al., 2012, Nyland et al., 2011]. The version associated with the L1 cache architecture shown in Figure 4.1 appears somewhat similar to a traditional MSHR. Traditional MSHRs for data caches contain the block address of a cache miss along with information on the block offset and associated register that needs to be written when the block is filled into the cache. Multiple misses to the same block are supported by recording multiple block offsets and registers. The PRT in Figure 4.1 supports merging two requests to the same block and records information needed to inform the instruction pipeline which deferred memory access to replay.

The L1 data cache shown in Figure 4.1 is virtually indexed and virtually tagged. This may be surprising when contrasted with modern CPU microarchitectures which mostly employ virtual indexed/physically tagged L1 data caches. CPUs use this organization to avoid the overheads of flushing the L1 data cache on context switches [Hennessy and Patterson, 2011]. While GPUs effectively perform a context switch every cycle that a warp issues, the warps are part of the same application. Page-based virtual memory is still advantageous within a GPUs even when it is limited to running a single OS application at a time, because it helps simplify memory allocation and reduces memory fragmentation. After an entry is allocated in the PRT a memory request is forwarded to the memory management unit (MMU) ❽ for virtual to physical address translation and from there over a crossbar interconnect to the appropriate memory partition unit. As will be expanded upon in Section 4.3, the memory partition units contain a bank of L2 cache along with a memory access scheduler. Along with information about which physical memory address to access and how many bytes to read, the memory request contains a "subid" that can be used to lookup the entry in the PRT containing information about the request when the memory request returns to the core.

Once a memory request response for the load is returned to the core it is passed by the MMU to the fill unit ❾. The fill unit in turn uses the subid field in the memory request to lookup information about the request in the PRT. This includes information that can be passed by the fill unit to the load/store unit via the arbiter ❷ to reschedule the load which is then guaranteed to hit in the cache by locking the line in the cache after it has been placed into the data array ❺.

Cache Write Operations

The L1 data cache in Figure 4.1 can support both write through and write back policies. Thus, store instructions (writes) to global memory can be handled in several ways. The specific memory

数据返回到指令流水线中的寄存器文件。与共享内存访问的情况一样，只有与活动线程相对应的寄存器通道才会更新。

访问标记单元时，如果确定请求触发了缓存未命中，仲裁器将通知加载/存储单元它必须重放该请求，同时将请求信息发送到待处理请求表 (PRT) 7。待处理请求表提供的功能与 CPU 缓存系统中传统未命中状态保持寄存器 [Kroft, 1981] 支持的功能类似。NVIDIA 专利 [Minkin et al., 2012, Nyland et al., 2011] 中描述了至少两个版本的待处理请求表。图 4.1 中显示的与 L1 缓存架构相关的版本看起来与传统的 MSHR 有点相似。数据缓存的传统 MSHR 包含缓存未命中的块地址以及块偏移和相关寄存器的信息，这些信息需要在块填充到缓存中时写入。通过记录多个块偏移和寄存器，可以支持对同一块的多次未命中。图 4.1 中的 PRT 支持将两个请求合并到同一个块，并记录通知指令管道哪些延迟的内存访问需要重放所需的信息。

图 4.1 中显示的 L1 数据缓存是虚拟索引和虚拟标记的。与大多数采用虚拟索引/物理标记 L1 数据缓存的现代 CPU 微架构相比，这可能令人惊讶。CPU 使用这种组织方式来避免在上下文切换时刷新 L1 数据缓存的开销 [Hennessy and Patterson，2011]。虽然 GPU 会在 warp 发出的每个周期有效地执行上下文切换，但 warp 是同一应用程序的一部分。即使 GPU 一次只能运行一个 OS 应用程序，基于页面的虚拟内存仍然具有优势，因为它有助于简化内存分配并减少内存碎片。在 PRT 中分配条目后，内存请求将转发到内存管理单元 (MMU) 8 进行虚拟到物理地址转换，然后通过交叉互连转发到适当的内存分区单元。如第 4.3 节中所述，内存分区单元包含一组 L2 缓存以及一个内存访问调度程序。除了有关访问哪个物理内存地址以及读取多少字节的信息之外，内存请求还包含一个“ subid ”，当内存请求返回核心时，可以使用该“ subid ”来查找 PRT 中包含有关该请求的信息的条目。

一旦将加载的内存请求响应返回到核心，MMU 就会将其传递给填充单元 9 。填充单元依次使用内存请求中的 subid 字段在 PRT 中查找有关该请求的信息。这包括填充单元可以通过仲裁器 2 传递给加载/存储单元的信息，以重新安排加载，然后通过在将加载放入数据阵列 5 后锁定缓存中的行来保证加载命中缓存。

缓存写入操作

图 4.1 中的 L1 数据缓存可以支持直写和回写策略。因此，可以以多种方式处理全局内存的存储指令（写入）。特定内存

space written to determines whether the write is treated as write through or write back. Accesses to global memory in many GPGPU applications can be expected to have very poor temporal locality as commonly kernels are written in such a way that threads write out data to a large array right before exiting. For such accesses a write through with no write allocate [Hennessy and Patterson, 2011] policy might make sense. In contrast, local memory writes for spilling registers to the stack may show good temporal locality with subsequent loads justifying a write back with write allocate policy [Hennessy and Patterson, 2011].

The data to be written either to shared memory or global memory is first placed write data buffer (WDB) ⑩. For uncoalesced accesses or when some threads are masked off, only a portion of a cache block is written to. If the block is present in the cache the data can be written to the data array via the data crossbar ⑥. If the data is not present in the cache the block must first be read from the L2 cache or DRAM memory. Coalesced writes which completely fill a cache block may bypass the cache if they invalidate tags for any stale data in the cache.

Note that the cache organization described in Figure 4.1 does *not* support cache coherence. For example, suppose a thread executing on SM 1 reads memory location A and the value is stored in SM 1's L1 data cache and then another thread executing on SM 2 writes memory location A. If any thread on SM 1 subsequently reads memory location A before it is evicted from SM 1's L1 data cache it will obtain the old value instead of the new value. To avoid this issue, NVIDIA GPUs starting with Kepler only permitted local memory accesses for register spills and stack data or read-only global memory data to be placed in the L1 data cache. Recent research has explored how to enable coherent L1 data caches on GPUs [Ren and Lis, 2017, Singh et al., 2013] and the need for clearly defined GPU memory consistency models [Alglave et al., 2015].

4.1.2 L1 TEXTURE CACHE

Recent GPU architectures from NVIDIA combine the L1 data cache and texture cache to save area. To better understand how such a cache may work it is first necessary to understand a bit about the design of a stand alone texture cache. The details covered here should help provide additional intuition into how to develop microarchitectures for throughput processors. Much of the discussion here is based upon a paper by Igehy et al. [1998] that aimed to fill in for a lack of literature on how industrial texture cache designs tolerate long off-chip latencies for cache misses. Recent industry GPU patents [Minken et al., 2010, Minken and Rubinstein, 2003] describe closely related designs. As the focus of this book is not on graphics we provide only a brief summary of the texture operations that motivate the inclusion of texture caches.

In 3D graphics it is desirable to make scenes look as realistic as possible. To achieve this realism with the high frame rates required for real-time rendering, graphics APIs employ a technique called texture mapping [Catmull, 1974]. In texture mapping an image, called a texture, is applied to a surface in a 3D model to make the surface look more realistic. For example, a texture might be used to give the appearance of natural wood to a table in a scene. To implement

写入的空间决定了写入是被视为直写还是回写。许多 GPGPU 应用程序中对全局内存的访问可能具有非常差的时间局部性，因为内核的编写方式通常是线程在退出之前将数据写入大型数组。对于此类访问，直写而不使用写入分配 [Hennessy and Patterson, 2011] 策略可能有意义。相比之下，将寄存器溢出到堆栈的本地内存写入可能表现出良好的时间局部性，后续加载证明使用写入分配策略回写是合理的 [Hennessy and Patterson, 2011] 。

要写入共享内存或全局内存的数据首先放置在写入数据缓冲区 (WDB) 10 中。对于未合并的访问或[redacted]某些线程被屏蔽时，只会写入缓存块的一部分。如果缓存中存在该块，则可以通过数据交叉开关 6 将数据写入数据阵列。如果缓存中不存在该数据，则必须首先从 L2 缓存或 DRAM 内存中[redacted]取该块。如果合并写入完全填充缓存块，并且它们使缓存中任何陈旧数据的标签无效，则它们可能会绕过缓存。

请注意，图 4.1 中描述的缓存组织确实支持缓存一致性。例如，假设在 SM 1 上执行的线程读取内存位置 A，并且该值存储在 SM 1 的 L1 数据缓存中，然后在 SM 2 上执行的另一个线程写入内存位置 A。如果 SM 1 上的任何线程在从 SM 1 的 L1 数据缓存中逐出之前随后读取内存位置 A，它将获得旧值而不是新值。为了避免这个问题，从 Kepler 开始的 NVIDIA GPU 仅允许本地内存访问寄存器溢出和堆栈数据或只读全局内存数据放置在 L1 数据缓存中。最近的研究探索了如何在 GPU 上启用一致的 L1 数据缓存 [Ren and Lis，2017，Singh et al.，2013] 以及对明确定义的 GPU 内存一致性模型的需求 [Alglave et al.，2015]。

4.1.2 L1 纹理缓存

NVIDIA 的最新 GPU 架构将 L1 数据缓存和纹理缓存相结合以节省空间。为了更好地理解这种缓存的工作原理，首先需要了解一些独立纹理缓存的设计。这里介绍的细节应该有助于提供有关如何开发吞吐量处理器微架构的更多直觉。这里的大部分讨论都基于 Igehy 等人的一篇论文 [1998]，该论文旨在填补有关工业纹理缓存设计如何容忍缓存未命中的长片外延迟的文献不足。最近的行业 GPU 专利 [Minken et al., 2010, Minken and Rubinstein, 2003] 描述了密切相关的设计。由于本书的重点不是图形，我们仅简要概述了促使包含纹理缓存的纹理操作。

在 3D 图形中，希望使场景看起来尽可能逼真。为了在实时渲染所需的高帧速率下实现这种真实感，图形 API 采用了一种称为纹理映射的技术 [Catmull，1974]。在纹理映射中，将称为纹理的图像应用于 3D 模型中的表面，以使表面看起来更逼真。例如，纹理可用于使场景中的桌子呈现天然木材的外观。为了实现

texture mapping the rendering pipeline first determines the coordinates of one or more samples within the texture image. These samples are called texels. The coordinates are then used to find the address of the memory locations containing the texels. As adjacent pixels on the screen map to adjacent texels, and as it is common to average the values of nearby texels, there is significant locality in texture memory accesses that can be exploited by caches [Hakura and Gupta, 1997].

Figure 4.2 illustrates the microarchitecture of an L1 texture cache as described by Igehy et al. [1998]. In contrast to the L1 data cache described in Section 4.1.1, the tag array ❷ and data array ❺ are separated by a FIFO buffer ❸. The motivation for this FIFO is to hide the latency of miss requests that may need to be serviced from DRAM. In essence, the texture cache is designed assuming that cache misses will be frequent and that the working set size is relatively small. To keep the size of the tag and data arrays small, the tag array essentially runs ahead of the data array. The contents of the tag array reflect what the data array in the future after an amount of time roughly equal to the round trip time of a miss request to memory and back. While throughput is improved relative to regular CPU design with limited capacity and miss handling resources, both cache hits and misses experience roughly the same latency.

In detail, the texture cache illustrated in Figure 4.2 operates as follows. The load/store unit ❶ sends the computed addresses for texels to perform a lookup in the tag array ❷. If the access hits, a pointer to the location of the data in the data array is placed in an entry at the tail of the fragment FIFO ❸ along with any other information required to complete the texture operation. When the entry reaches the head of the fragment FIFO a controller ❹ uses the pointer to lookup the texel data from the data array ❺ and return it to the texture filter unit ❻. While not shown in detail, for operations such as bilinear and trilinear filtering (mipmap filtering) there are actually four or eight texel lookups per fragment (i.e., screen pixel). The texture filter unit combines the texels to produce a single color value which is returned to the instruction pipeline via the register file.

In the event of a cache miss during tag lookup, the tag array sends a memory request via the miss request FIFO ❽. The miss request FIFO sends requests to lower levels of the memory system ❾. DRAM bandwidth utilization in GPU memory systems can be improved by the use of memory access scheduling techniques [Eckert, 2008, 2015]. that may service memory requests out-of-order to reduce row switch penalties. To ensure the contents of the data array ❺ reflect the time-delayed state of the tag array ❷, data must be returned from the memory system in order. This is accomplished using a reorder buffer ❿.

4.1.3 UNIFIED TEXTURE AND DATA CACHE

In recent GPU architectures from NVIDIA and AMD caching of data and texture values is performed using a unified L1 cache structure. To accomplish this in this most straightforward way, only data values that can be guaranteed to read-only are cached in the L1. For data that follows this restriction the texture cache hardware can be used almost unmodified except for changes to the addressing logic. Such a design is described in a recent patent [Heinrich et al.,

纹理映射渲染管道首先确定纹理图像中一个或多个样本的坐标。这些样本称为纹素。然后使用坐标查找包含纹素的内存位置的地址。由于屏幕上的相邻像素映射到相邻的纹素，并且通常会对附近纹素的值取平均值，因此纹理内存访问中存在明显的局部性，缓存可以利用这一点 [Hakura and Gupta，1997]。

图 4.2 说明了 Igehy 等人 [1998] 描述的 L1 纹理缓存的微架构。与第 4.1.1 节中描述的 L1 数据缓存不同，标记数组 2 和数据数组 5 由 FIFO 缓冲区 3 分隔。此 FIFO 的动机是隐藏可能需要从 DRAM 提供服务的未命中请求的延迟。本质上，纹理缓存的设计假设缓存未命中会频繁发生，并且工作集大小相对较小。为了保持标记和数据数组的大小较小，标记数组基本上在数据数组之前运行。标记数组的内容反映了在经过大约等于未命中请求到内存并返回的往返时间的一段时间后数据数组的内容。虽然吞吐量相对于容量和未命中处理资源有限的常规 CPU 设计有所提高，但缓存命中和未命中的延迟大致相同。

具体来说，图 4.2 中所示的纹理缓存的操作如下。加载/存储单元 1 发送计算出的纹素地址以在标签数组 2 中执行查找。如果访问命中，则指向数据数组中数据位置的指针将与完成纹理操作所需的任何其他信息一起放置在片段 FIFO 3 尾部的条目中。当条目到达片段 FIFO 的头部时，控制器 4 使用指针从数据数组 5 中查找纹素数据并将其返回到纹理过滤单元 6 。虽然没有详细显示，但对于双线性和三线性过滤（mipmap 过滤）等操作，每个片段（即屏幕像素）实际上有四次或八次纹素查找。纹理过滤单元将纹素组合起来以产生单个颜色值，该颜色值通过寄存器文件返回到指令管道。

如果在标签查找期间发生缓存未命中，标签阵列将通过未命中请求 FIFO 8 发送内存请求。未命中请求 FIFO 将请求发送到内存系统 9 的较低级别。通过使用内存访问调度技术 [Eckert，2008，2015]，可以提高 GPU 内存系统中的 DRAM 带宽利用率。这些技术可以无序处理内存请求，以减少行切换惩罚。为了确保数据阵列 5 的内容反映标签阵列 2 的时间延迟状态，必须按顺序从内存系统返回数据。这是使用重新排序缓冲区 10 来实现的。

4.1.3 统一纹理和数据缓存

在 NVIDIA 和 AMD 的最新 GPU 架构中，数据和纹理值的缓存是使用统一的 L1 缓存结构进行的。为了以最直接的方式实现这一点，只有可以保证只读的数据值才会缓存在 L1 中。对于遵循此限制的数据，除了对寻址逻辑进行更改外，几乎无需修改即可使用纹理缓存硬件。最近的一项专利 [Heinrich et al.,

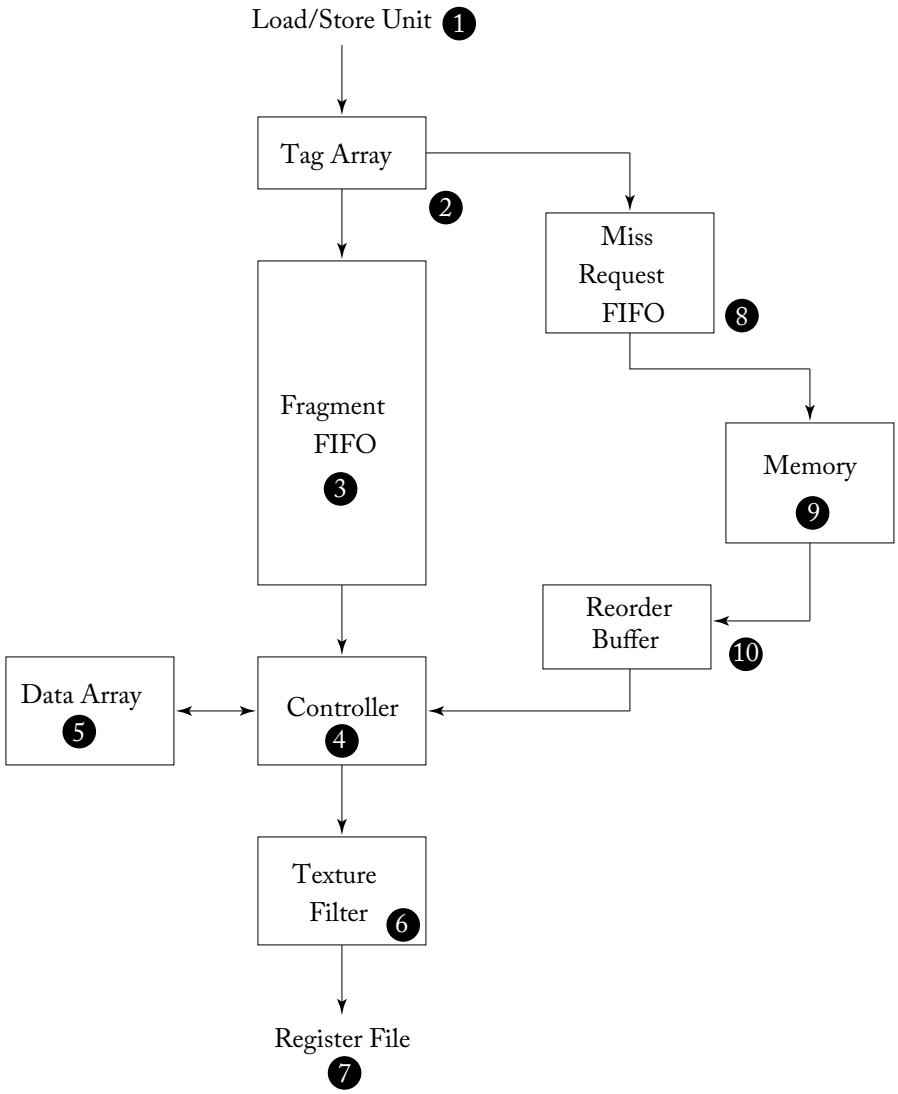


Figure 4.2: L1 texture cache (based in part on Figure 2 in [Igehy et al., 1998]).

2017]. In AMD's GCN GPU architecture all vector memory operations are processed through the texture cache [AMD, 2012].

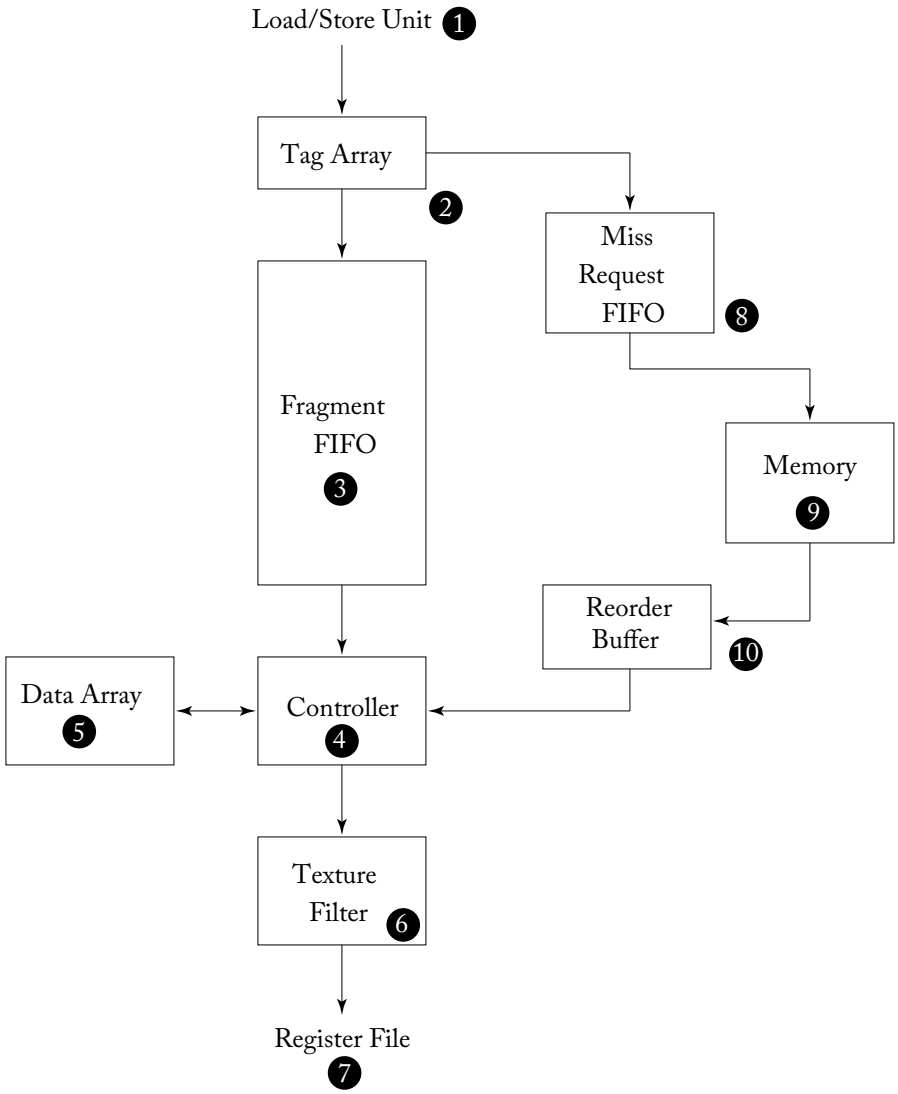


图 4.2：L1 纹理缓存（部分基于 [Igehy et al., 1998] 中的图 2）。

2017]. In AMD's GCN GPU architecture all vector memory operations are processed through the texture cache [AMD, 2012].

4.2 ON-CHIP INTERCONNECTION NETWORK

To supply the large amount of memory bandwidth required to supply the SIMT cores, high-performance GPUs connect to multiple DRAM chips in parallel via memory partition units (described in Section 4.3). Memory traffic is distributed across the memory partition units using address interleaving. An NVIDIA patent describes address interleaving schemes for balancing traffic among up to 6 memory partitions at granularities of 256 bytes or 1,024 bytes [Edmondson and Van Dyke, 2011].

The SIMT cores connect to the memory partition units via an on-chip interconnection network. The on-chip interconnection networks described in recent patents for NVIDIA are crossbars [Glasco et al., 2013, Treichler et al., 2015]. GPUs from AMD have sometimes been described as using ring networks [Shrout, 2007].

4.3 MEMORY PARTITION UNIT

Below, we describe the microarchitecture of a memory partition unit corresponding to several recent NVIDIA patents. In terms of historical context, these patents were filed about a year prior to the release of NVIDIA's Fermi GPU architecture. As shown in Figure 4.3, each memory partition unit contains a portion of the second-level (L2) cache along with a one or more memory access schedulers also called a "frame buffer," or FB, and a raster operation (ROP) unit. The L2 cache contains both graphics and compute data. The memory access scheduler reorders memory read and write operations to reduce overheads of accessing DRAM. The ROP unit is primarily used in graphics operation such as alpha blending and supports compression of graphics surfaces. The ROP unit also supports atomic operations like those found in the CUDA programming model. All three units are tightly coupled and will be described below in some detail.

4.3.1 L2 CACHE

The L2 cache design includes several optimizations to improve overall throughput per unit area for the GPU. The L2 cache portion inside each memory partition is composed of two slices [Edmondson et al., 2013]. Each slice contains separate tag and data arrays and processes incoming requests in order [Roberts et al., 2012]. To match the DRAM atom size of 32 bytes in GDDR5, each cache line inside the slice has four 32-byte sectors. Cache lines are allocated for use either by store instructions or load instructions. To optimize throughput in the common case of coalesced writes that completely overwrite each sector on a write miss no data is first read from memory. This is quite different from how CPU caches are commonly described in standard computer architecture textbooks. How uncoalesced writes, which do not completely cover a sector, are handled is not described in the patents we examined, but two solutions are storing byte-level valid bits and bypassing the L2 entirely. To reduce area of the memory access scheduler, data that is being written to memory is buffered in cache lines in the L2 while writes awaiting scheduling.

为了提供 SIMT 核心所需的大量内存带宽，高性能 GPU 通过内存分区单元并行连接到多个 DRAM 芯片（如第 4.3 节所述）。内存流量使用地址交错分布在内存分区单元中。NVIDIA 的一项专利描述了地址交错方案，用于在粒度为 256 字节或 1,024 字节的最多 6 个内存分区之间平衡流量 [Edmondson and Van Dyke，2011]。

SIMT 核心通过片上互连网络连接到内存分区单元。NVIDIA 最近的专利中描述的片上互连网络是交叉开关 [Glasco 等人，2013 年，Treichler 等人，2015 年]。AMD 的 GPU 有时被描述为使用环形网络 [Shrout，2007 年]。

4.3 内存分区单元

下面，我们描述了与最近的几项 NVIDIA 专利相对应的内存分区单元的微架构。从历史背景来看，这些专利是在 NVIDIA Fermi GPU 架构发布前一年左右提交的。如图 4.3 所示，每个内存分区单元包含一部分二级 (L2) 缓存以及一个或多个内存访问调度程序（也称为“ 帧缓冲区 ”或 FB）和一个光栅操作 (ROP) 单元。L2 缓存包含图形和计算数据。内存访问调度程序重新排序内存读写操作以减少访问 DRAM 的开销。ROP 单元主要用于图形操作（例如 alpha 混合），并支持图形表面的压缩。ROP 单元还支持原子操作，如 CUDA 编程模型中的原子操作。这三个单元紧密耦合，下面将详细介绍。

4.3.1 二级缓存

L2 缓存设计包括多项优化，以提高 GPU 单位面积的总吞吐量。每个内存分区内的 L2 缓存部分由两个切片组成 [Edmondson 等人，2013]。每个切片包含单独的标签和数据阵列，并按顺序处理传入请求 [Roberts 等人，2012]。为了匹配 GDDR5 中 32 字节的 DRAM 原子大小，切片内的每个缓存行都有四个 32 字节扇区。缓存行分配给存储指令或加载指令使用。为了在写入未命中时完全覆盖每个扇区的合并写入的常见情况下优化吞吐量，首先不会从内存中读取任何数据。这与标准计算机架构教科书中通常描述的 CPU 缓存完全不同。我们研究的专利中没有描述如何处理未完全覆盖扇区的未合并写入，但有两种解决方案是存储字节级有效位和完全绕过 L2。为了减少内存访问调度程序的面积，正在写入内存的数据被缓冲在 L2 中的缓存行中，同时写入等待调度。

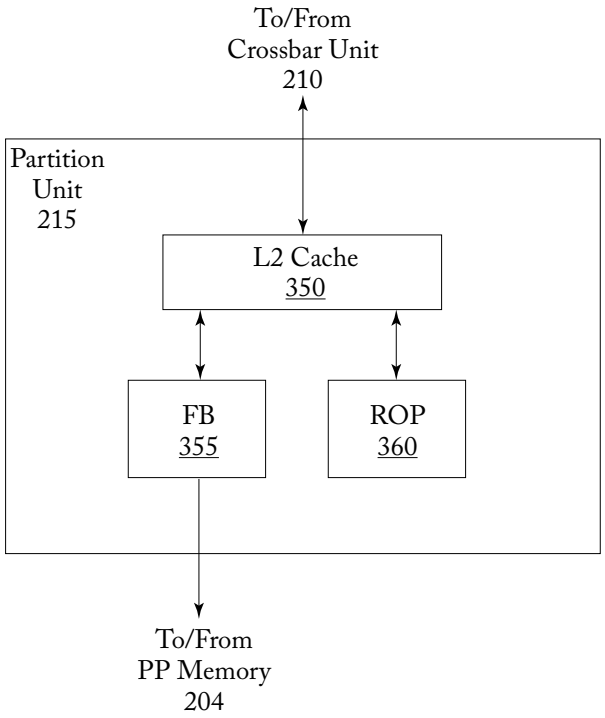


Figure 4.3: Memory partition unit overview (based on Figure 3B in Edmondson et al. [2013]).

4.3.2 ATOMIC OPERATIONS

As described in Glasco et al. [2012] the ROP unit includes function units for executing atomic and reduction operations. A sequence of atomic operations accessing the same memory location can be pipelined as the ROP unit includes a local ROP cache. Atomic operations can be used for implementing synchronization across threads running in different thread blocks.

4.3.3 MEMORY ACCESS SCHEDULER

To store large amounts of data GPUs employ special dynamic random access memory (DRAM) such as GDDR5 gdd. DRAM stores individual bits in small capacitors. To, for example, read values from these capacitors a row of bits, called a page, is first read into a small memory structure called a row buffer. To accomplish this operation the bitlines connecting the individual storage capacitors to the row buffer, and which have capacitance themselves, must first be precharged to a voltage half way between 0 and the supply voltage. When the capacitor is connected to the bit line through an access transistor during an activate operation the voltage of the bit line is pulled either up or down slightly as charge flow in or out of the storage cell from the bitline. A sense

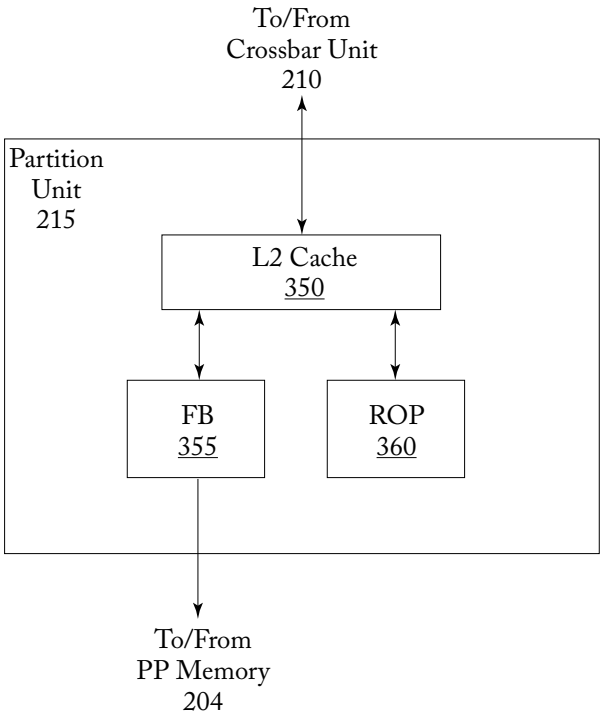


图 4.3：内存分区单元概览（基于 Edmondson 等人 [2013] 中的图 3B）。

4.3.2 原子操作

如 Glasco 等人 [2012] 所述，ROP 单元包括用于执行原子和归约操作的功能单元。由于 ROP 单元包含本地 ROP 缓存，因此可以对访问同一内存位置的一系列原子操作进行流水线处理。原子操作可用于实现在不同线程块中运行的线程之间的同步。

4.3.3 内存访问调度程序

为了存储大量数据，GPU 采用了特殊的动态随机存取存储器 (DRAM)，例如 GDDR5 gdd。DRAM 将各个位存储在小型电容器中。例如，为了从这些电容器读取值，首先将一行位（称为页面）读入称为行缓冲器的小型存储器结构中。为了完成此操作，必须首先将连接各个存储电容器和行缓冲器的位线（本身具有电容）预充电到介于 0 和电源电压之间的一半电压。当在激活操作期间通过访问晶体管将电容器连接到位线时，随着电荷从位线流入或流出存储单元，位线的电压会被略微上拉或下拉。感测

amplifier then amplifies this small change until a clean logic 0 or 1 is read. The process of reading the values into the row buffer refreshes the values stored in the capacitors. The precharge and activate operations introduce delays during which no data can be read or written to the DRAM array. To mitigate these overheads DRAMs contain multiple banks, each with their own row buffer. However, even with multiple DRAM banks it is often not possible to completely hide the latency of switching between rows when accessing data. This has led to the use of memory access schedulers [Rixner et al., 2000, Zuravleff and Robinson, 1997] that reorder DRAM memory access requests so as to reduce the number of times data must be moved between the row buffers and DRAM cells.

To enable access to DRAM, each memory partition in the GPU may contain multiple memory access schedulers [Keil and Edmondson, 2012] connecting the portion of L2 cache it contains to off-chip DRAM. The simplest approach for doing this would be for each slice of the L2 cache to have its own memory access scheduler. Each memory access scheduler contains separate logic for sorting read requests and write requests ("dirty data notifications") sent from the L2 cache [Keil et al., 2012]. To group together reads to the same row in a DRAM bank, two separate tables are employed. The first, called the read request sorter, is a set associative structure accessed by memory address and maps all read requests to the same row in a given bank to single pointer. The pointer is used to lookup a list of individual read requests in a second table called the read request store.

4.4 RESEARCH DIRECTIONS FOR GPU MEMORY SYSTEMS

4.4.1 MEMORY ACCESS SCHEDULING AND INTERCONNECTION NETWORK DESIGN

Yuan et al. [2009] explore memory access scheduler design for GPUs running GPU computing applications written in CUDA. They observe that requests generated by a single streaming multiprocessor (SM) have row-buffer locality. A sequence of memory requests to a given memory partition is said to have row buffer locality if requests that appear nearby in the sequence access the same DRAM row in the same DRAM bank. However, when memory requests from one SM are sent to the memory partitions they are intermixed with requests from other SM sending requests to the same memory partition. The result is that the row buffer locality of the sequence of requests entering the memory partition is lower. Yuan et al. [2009] propose reducing the complexity of memory access scheduling by modifying the interconnection network to maintain row buffer locality. They do this by introducing arbitration policies that prioritize granting packets containing memory requests from the same SM or with similar row-bank addresses.

Bakhoda et al. [2010, 2013] explore the design of on-chip interconnection networks for GPUs. This interconnect connects the streaming multiprocessors to the memory partitions. They argue that as the number of SMs increase it will become necessary to employ scalable topologies such as meshes. They explore how network-on-chip design impacts system throughput and find

然后，放大器放大这个微小的变化，直到读取到干净的逻辑 0 或 1。将值读入行缓冲区的过程会刷新存储在电容器中的值。预充电和激活操作会引入延迟，在此期间无法读取或写入 DRAM 阵列的任何数据。为了减轻这些开销，DRAM 包含多个存储体，每个存储体都有自己的行缓冲区。然而，即使有多个 DRAM 存储体，也常常不可能完全隐藏访问数据时在行之间切换的延迟。这导致使用内存访问调度程序[Rixner et al.，2000；Zuravleffi and Robinson，1997]来重新排序 DRAM 内存访问请求，以减少必须在行缓冲区和 DRAM 单元之间移动数据的次数。

为了能够访问 DRAM，GPU 中的每个内存分区可能包含多个内存访问调度程序 [Keil and Edmondson，2012]，将其包含的 L2 缓存部分连接到片外 DRAM。最简单的方法是让 L2 缓存的每个片段都有自己的内存访问调度程序。每个内存访问调度程序都包含单独的逻辑，用于对从 L2 缓存发送的读取请求和写入请求（“脏数据通知”）进行排序 [Keil et al.，2012]。为了将读取分组到 DRAM 组中的同一行，使用了两个单独的表。第一个表称为读取请求分类器，是一个通过内存地址访问的组关联结构，它将给定组中同一行的所有读取请求映射到单个指针。该指针用于在第二个表（称为读取请求存储库）中查找单个读取请求的列表。

4.4 GPU内存系统的研究方向

4.4.1 内存访问调度和互连网络设计

Yuan 等人 [2009] 探索了运行用 CUDA 编写的 GPU 计算应用程序的 GPU 的内存访问调度程序设计。他们观察到单个流式多处理器 (SM) 生成的请求具有行缓冲区局部性。如果对给定内存分区的一系列内存请求在序列中相邻的请求访问同一 DRAM 组中的同一 DRAM 行，则称该序列具有行缓冲区局部性。但是，当来自一个 SM 的内存请求发送到内存分区时，它们会与来自其他 SM 的请求混合在一起，这些请求向同一内存分区发送请求。结果是进入内存分区的请求序列的行缓冲区局部性较低。Yuan 等人 [2009] 建议通过修改互连网络以保持行缓冲区局部性来降低内存访问调度的复杂性。他们通过引入仲裁策略来实现这一点，这些策略优先授予包含来自同一 SM 或具有相似行缓冲区地址的内存请求的数据包。

Bakhoda 等人 [2010, 2013] 探索了 GPU 的片上互连网络的设计。该互连将流式多处理器连接到内存分区。他们认为，随着 SM 数量的增加，将有必要采用可扩展拓扑，例如网格。他们探索了片上网络设计如何影响系统吞吐量，并发现

that throughput of many CUDA applications is relatively insensitive to interconnect latency. They analyze the interconnect traffic and find it has a many-to-few-to-many pattern. They propose a more restricted scalable topology consisting of "half-routers" that reduces the area cost of routers by leveraging this traffic pattern.

4.4.2 CACHING EFFECTIVENESS

Bakhoda et al. [2009] studied the impact of adding L1 and/or L2 caches for global memory accesses to a CUDA-enabled GPU simulated using their GPGPU-Sim simulator and showed that while some applications benefited others did not.

Subsequent work by Jia et al. [2012] characterized the effectiveness of caching by enabling or disabling caches on NVIDIA Fermi GPU hardware and found similar results. It is observed that applications that read data into the scratchpad shared memory via the L1 cache do not benefit from enabling the L1 cache. Even excluding such applications Jia et al. [2012] observe that cache hit rates alone are insufficient to predict whether caching will improve performance. They find that instead it is necessary to consider the impact of caching on request traffic to the L2 caches (e.g., memory partitions). On the Fermi GPU they study, the L1 cache is not sectored and as a consequence enabling caching can incur larger 128-byte off-chip memory accesses on a miss. On memory bandwidth-limited applications this additional off-chip memory traffic can result in a reduction in performance. Jia et al. [2012] introduce a taxonomy of three forms of locality: within-warp, within-block, and cross-instruction. Within warp locality occurs when memory read accesses from a single load executed by different threads within a single warp access the same cache block. Within block locality occurs when memory read accesses from a single load executed by threads in different warps from the same thread block access the same cache block. Cross-instruction locality occurs when memory read access from different load instructions execute by threads in the same thread block access the same cache block. Jia et al. [2012] introduce a compile time algorithm using this taxonomy to help infer when enabling caching is helpful for individual load instructions.

4.4.3 MEMORY REQUEST PRIORITIZATION AND CACHE BYPASSING

Following up on the above characterization study [Jia et al., 2012] and work by Rogers et al. [2012] which demonstrated warp scheduling can improve cache effectiveness (described in Section 5.1.2), Jia et al. [2014] proposed memory request prioritization and cache bypassing techniques for GPUs. Caches with low associativity relative to number of threads can suffer from significant conflict misses [Chen and Aamodt, 2009]. Jia et al. [2014] noted that several GPGPU applications written in CUDA include array indexing that causes an individual memory request from a single warp to result in misses that map to the same cache set when using a standard modulo cache set indexing function [Hennessy and Patterson, 2011, Patterson and Hennessy, 2013]. Jia et al. [2014] call this intra-warp contention. Assuming space in the cache is allocated when a

许多 CUDA 应用程序的吞吐量对互连延迟相对不敏感。他们分析了互连流量，发现它具有多对少对多的模式。他们提出了一种由“ 半路由器 ”组成的更受限制的可扩展拓扑，通过利用这种流量模式来降低路由器的面积成本。

4.4.2 缓存有效性

Bakhoda 等人 [2009] 研究了使用他们的 GPGPU-Sim 模拟器模拟的对支持 CUDA 的 GPU 添加 L1 和/或 L2 缓存以进行全局内存访问的影响，并表明虽然一些应用程序受益，但其他应用程序却没有。

Jia 等人 [2012] 随后的研究通过启用或禁用 NVIDIA Fermi GPU 硬件上的缓存来描述缓存的有效性，并发现了类似的结果。据观察，通过 L1 缓存将数据读入暂存器共享内存的应用程序不会从启用 L1 缓存中受益。即使排除此类应用程序，Jia 等人 [2012] 也观察到单靠缓存命中率不足以预测缓存是否会提高性能。他们发现，有必要考虑缓存对 L2 缓存（例如，内存分区）请求流量的影响。在他们研究的 Fermi GPU 上，L1 缓存没有分区，因此启用缓存会在未命中时引发更大的 128 字节片外内存访问。在内存带宽有限的应用程序中，这种额外的片外内存流量可能会导致性能下降。Jia 等人[2012] 引入了三种局部性的分类法：warp 内、块内和跨指令。当由单个 warp 内的不同线程执行的单个加载的内存读取访问访问同一缓存块时，就会发生 warp 内局部性。当由同一线程块的不同 warp 中的线程执行的单个加载的内存读取访问访问同一缓存块时，就会发生块内局部性。当由同一线程块中的线程执行的不同加载指令的内存读取访问访问同一缓存块时，就会发生跨指令局部性。Jia 等人 [2012] 引入了一种使用此分类法的编译时算法，以帮助推断何时启用缓存对各个加载指令有帮助。

4.4.3 内存请求优先级和缓存绕过

继上述特性研究 [Jia et al., 2012] 和 Rogers et al. [2012] 的研究（证明了 warp 调度可以提高缓存效率（见第 5.1.2 节））之后，Jia et al. [2014] 提出了用于 GPU 的内存请求优先级和缓存绕过技术。相对于线程数而言关联性较低的缓存可能会出现严重的冲突未命中 [Chen and Aamodt, 2009]。Jia et al. [2014] 指出，用 CUDA 编写的几个 GPGPU 应用程序包含数组索引，当使用标准模缓存集索引函数时，这会导致来自单个 warp 的单个内存请求导致映射到同一缓存集的未命中 [Hennessy and Patterson, 2011, Patterson and Hennessy, 2013]。Jia et al. [2014] 将此称为内部 Warp 争用。假设在

miss is detected[1] and a limited number of miss-status holding registers[2] intra-warp contention can lead to memory pipeline stalls.[3] To address intra-warp contention Jia et al. [2014], propose bypassing the L1 data cache when a miss occurs and a cache block cannot be allocated due to an associativity stall. An associativity stall occurs when all blocks in a cache set are reserved to provide space for data that will be supplied by outstanding cache misses.

Jia et al. [2014] also examine what they call cross-warp contention. This form of cache contention results when one warp evicts data brought in by another warp. To address this form of contention, Jia et al. [2014] suggest employing a structure they call a "memory request prioritization buffer" (MRPB). Like CCWS [Rogers et al., 2012] the MRPB reduces capacity misses by modifying the order of accesses to the cache so as to increase locality. However, unlike CCWS which achieves this indirectly via thread scheduling, the MRPB attempts to increase locality by changing the order of individual memory accesses after threads have been scheduled.

The MRPB implements memory request reordering right before the first-level data cache. The input of the MRPB is feed memory requests generated in the instruction issue pipeline stage after memory request coalescing has been performed. The output of the MRPB feeds memory requests into the first-level cache. Internally, the MRPB contains several parallel first-in, first-out (FIFO) queues. A cache request is distributed to these FIFOs using a "signature." Among several options they evaluated they found the most effective signature was to use "warp ID" (a number between 0 to the maximum number of warps that can run on a streaming multiprocessor). The MRPB employs a "drain policy" to determine which FIFO to select a memory request from to use to access the cache next. Among several options explored the best version was a simple fixed-priority scheme in which each queue is assigned a static priority and the queue with highest priority that contains requests is served first.

A detailed evaluation shows the combined mechanism of bypassing and reordering using the MRPB achieves geometric mean speedup of 4% over a 64-way 16 KB. Jia et al. [2014] also perform some comparisons with CCWS showing larger improvements. We note in passing that the evaluation in Rogers et al. [2012] employed a baseline architecture with a more sophisticated set index hashing function[4] to reduce the impact of associativity stalls. Also, subsequent, work by Nugteren et al. [2014] worked to reverse engineer the details of the actual set index hash function employed in the NVIDIA Fermi architecture and found it uses XOR-ing (which also tends to reduce such conflicts).

Similar to Rogers et al. [2013], Jia et al. [2014] show that their programmer transparent approach to improving performance can narrow the gap between simple code that uses caches and more highly optimized code that uses the scratchpad shared memory.

[1] The default in GPGPU-Sim where it is used to avoid protocol deadlock.
[2] Consistent with a limited set of pending request table entries—see Section 4.1.1.
[3] GPGPU-Sim version 3.2.0, used by Jia et al. [2014], does not model instruction replay described in Sections 3.3.2 and 4.1.
[4] See `cache_config::set_index_hashed` in https://github.com/tgrogers/ccws-2012/blob/master/simulator/ccws_gpgpu-sim/distribution/src/gpgpu-sim/gpu-cache.cc

检测到未命中[1]，且有限数量的未命中状态保持寄存器[2]，warp 内争用可能导致内存流水线停顿。[3] 为了解决 warp 内争用问题，Jia 等人 [2014] 建议在发生未命中且由于关联性停顿而无法分配缓存块时绕过 L1 数据缓存。当缓存集中的所有块都被保留以提供空间用于由未解决的缓存未命中提供的数据时，就会发生关联性停顿。

Jia 等人 [2014] 还研究了他们所谓的跨 Warp 争用问题。当一个 Warp 逐出另一个 Warp 带来的数据时，就会发生这种形式的缓存争用。为了解决这种形式的争用问题，Jia 等人 [2014] 建议采用一种他们称之为“ 内存请求优先级缓冲区 ”（MRPB）的结构。与 CCWS [Rogers et al., 2012] 一样，MRPB 通过修改访问缓存的顺序来增加局部性，从而减少容量未命中。然而，与通过线程调度间接实现这一点的 CCWS 不同，MRPB 试图通过在线程调度后更改各个内存访问的顺序来增加局部性。

MRPB 在第一级数据缓存之前实现内存请求重新排序。MRPB 的输入是在执行内存请求合并之后在指令发出管道阶段生成的内存请求。MRPB 的输出将内存请求输入到第一级缓存中。在内部，MRPB 包含几个并行的先进先出 (FIFO) 队列。缓存请求使用“ 签名 ”分发到这些 FIFO。在评估的几个选项中，他们发现最有效的签名是使用“ warp ID ”（一个介于 0 到流式多处理器上可以运行的最大 warp 数之间的数字）。MRPB 采用“ 排出策略 ”来确定从哪个 FIFO 中选择内存请求以用于下一次访问缓存。在探索的几个选项中，最好的版本是一个简单的固定优先级方案，其中每个队列都分配有一个静态优先级，并且包含请求的最高优先级队列首先得到服务。

详细评估表明，使用 MRPB 的绕过和重新排序组合机制在 64 路 16 KB 上实现了 4% 的几何平均加速。Jia 等人 [2014] 还与 CCWS 进行了一些比较，显示出更大的改进。我们顺便指出，Rogers 等人 [2012] 的评估采用了基线架构和更复杂的集合索引哈希函数 [4] 来减少关联性停顿的影响。此外，随后，Nugteren 等人 [2014] 对 NVIDIA Fermi 架构中使用的实际集合索引哈希函数的细节进行了逆向工程，发现它使用了 XOR-ing（这也倾向于减少此类冲突）。

与 Rogers 等人 [2013] 类似，Jia 等人 [2014] 表明，他们采用的对程序员透明的性能提升方法可以缩小使用缓存的简单代码与使用暂存器共享内存的高度优化代码之间的差距。

[1] The default in GPGPU-Sim where it is used to avoid protocol deadlock.
[2] Consistent with a limited set of pending request table entries—see Section 4.1.1.
[3] GPGPU-Sim version 3.2.0, used by Jia et al. [2014], does not model instruction replay described in Sections 3.3.2 and 4.1.
[4] See `cache_config::set_index_hashed` in `https://github.com/tgrogers/ccws-2012/blob/master/simulator/ccws_gpgpu-sim/distribution/src/gpgpu-sim/gpu-cache.cc`

Arunkumar et al. [2016] explore the effect of bypassing and varying the cache line size, based on the level of memory divergence present in a static instruction. They use observed reuse distance patterns and memory divergence degree to predict bypassing and optimal cache line size.

Lee and Wu [2016] propose a control-loop based cache bypassing method that attempts to predict reuse behavior on an instruction-by-instruction basis at runtime. The reuse behavior of cache lines are monitored. If cache lines loaded by a particular program counter do not experience enough reuse, accesses for that instruction are bypassed.

4.4.4 EXPLOITING INTER-WARP HETEROGENEITY

Ausavarungnirun et al. [2015] propose a series of improvements at the GPU's shared L2 and memory controller that mitigate memory latency divergence in irregular GPU applications. The techniques, collectively named Memory Divergence Correction (MeDiC), exploit the observation that there is heterogeneity in the level of memory latency divergence across warps in the same kernel. Based on how they interact with the shared L2 cache, each warp in a kernel can be characterized as all/mostly hit, all/mostly miss, or balanced. The authors demonstrate that there is little benefit in having warps that are not all hit, since warps the mostly hit must wait for the slowest access to return before they are able to proceed. They also demonstrate that queueing latency at the L2 cache can have a non-trivial performance impact and that this effect can be mitigated by bypassing the L2 cache for all requests (even those that may hit) for all warps that are not all-hit. This decreases the access latency for all-hit warps by reducing queueing delay. In addition to the adaptive bypassing technique, they propose modifications to the cache replacement policy and the memory controller scheduler in an attempt to minimize latency for warps detected to be all-hit warps. They also demonstrate that even for warps that are all-hit, the difference in queueing delay among L2 cache banks can cause additional potentially avoidable queuing delay since there is an unbalance in the queuing delay between L2 banks.

The microarchitectural mechanism proposed by the authors consists of four components: (1) a warp-type detection block—which classifies warps in the GPU as being one of the five potential types: All-miss, mostly-miss, balanced, mostly-hit, or all-hit; (2) a warp-type-aware bypass logic block which decides if requests should bypass the L2 cache; (3) a warp-type-aware insertion policy, which determines where insertions in the L2 will be placed in the LRU stack; and (4) a warp-type-aware memory scheduler that orders how L2 misses/bypasses are sent to DRAM.

The detection mechanism operates by sampling the hit ratio of each warp (total hits/accesses) on an interval basis. Based on this ratio, the warp takes on one of the five classifications listed above. The exact hit ratios that determine the boundaries for these classifications are tuned dynamically for each workload. During the classification interval, no request bypasses the cache in order to react to phase changes in each warp's L2 characteristic.

Arunkumar 等人 [2016] 探索了基于静态指令中内存发散程度的绕过和改变缓存行大小的影响。他们使用观察到的重用距离模式和内存发散程度来预测绕过和最佳缓存行大小。

Lee 和 Wu [2016] 提出了一种基于控制环路的缓存绕过方法，该方法试图在运行时逐条预测重用行为。缓存行的重用行为受到监控。如果特定程序计数器加载的缓存行没有经历足够的重用，则绕过该指令的访问。

4.4.4 利用经线间的异质性

Ausavarungnirun 等人 [2015] 提出了一系列改进 GPU 共享 L2 和内存控制器的方法，以缓解不规则 GPU 应用中的内存延迟差异。这些技术统称为内存差异校正 (MeDiC)，利用了以下观察结果：同一内核中各个 Warp 的内存延迟差异水平存在差异。根据它们与共享 L2 缓存的交互方式，内核中的每个 Warp 可分为全部/大部分命中、全部/大部分未命中或平衡。作者表明，拥有并非全部命中的 Warp 几乎没有好处，因为大部分命中的 Warp 必须等待最慢的访问返回后才能继续。他们还表明，L2 缓存的排队延迟会对性能产生不小的影响，可以通过对所有非全部命中的 Warp 的所有请求（即使是可能命中的请求）绕过 L2 缓存来减轻这种影响。通过减少排队延迟，这可以降低全命中 Warp 的访问延迟。除了自适应绕过技术之外，他们还建议修改缓存替换策略和内存控制器调度程序，以尽量减少检测到的全命中 Warp 的延迟。他们还证明，即使对于全命中 Warp，L2 缓存库之间的排队延迟差异也可能导致额外的潜在可避免的排队延迟，因为 L2 库之间的排队延迟不平衡。

作者提出的微架构机制由四个部分组成：（1）一个扭曲类型检测块 - 将 GPU 中的扭曲分类为五种潜在类型之一：全部未命中、大部分未命中、平衡、大部分命中或全部命中；（2）一个扭曲类型感知旁路逻辑块，用于决定请求是否应绕过 L2 缓存；（3）一个扭曲类型感知插入策略，用于确定 L2 中的插入在 LRU 堆栈中的什么位置；（4）一个扭曲类型感知内存调度程序，用于排序如何将 L2 未命中/旁路发送到 DRAM。

检测机制通过按间隔对每个 Warp 的命中率（总命中数/访问数）进行采样来运行。根据此比率，Warp 会采用上面列出的五种分类之一。确定这些分类边界的准确命中率会针对每个工作负载进行动态调整。在分类间隔期间，没有请求会绕过缓存来对每个 Warp 的 L2 特性中的相位变化做出反应。

The bypassing mechanism sits in front of the L2 cache and receives memory requests tagged with the warp-type that generated them. This mechanism attempts to eliminate accesses from all-miss warps and transforms mostly-miss warps into all-miss warps. The block simply sends all requests marked as coming from all-miss and mostly-miss warps directly to the memory scheduler.

The cache management policy of MeDiC operates by changing where requests returned from DRAM are placed in the L2's LRU stack. Cache lines requested by a mostly-miss warp are inserted into the LRU position, while all other requests are inserted into the traditional MRU position.

Finally, MeDic modifies the baseline memory request scheduler to contain two memory access queues: a high-priority queue for all-hit and mostly-hit warps and a low-priority queue for balanced, mostly-miss, and all-miss warps. The memory scheduler simply prioritizes all requests in the high-priority queue of any of the requests in the low priority queue.

4.4.5 COORDINATED CACHE BYPASSING

Xie et al. [2015] explore the potential to selectively enable cache bypassing for improving cache hit rates. They employ profiling to determine for each static load instruction in the GPGPU application whether it has good locality, poor locality, or moderate locality. They mark the instructions accordingly. Load operations marked as having good locality are permitted to use the L1 data cache. Load operations marked as having poor locality are always bypassed. Load instructions marked with moderate locality employ an adaptive mechanism that works as follows. The adaptive mechanism operates at thread block granularity. For a given thread block, all moderate locality loads executed are treated uniformly. They either use the L1 or bypass. The behavior is determined at the time the thread blocks are launched based upon a threshold that is adapted online using a performance metric that takes account of L1 cache hits and pipeline resource conflicts. Their evaluation shows this approach improves cache hit rates significantly more than static warp limiting.

4.4.6 ADAPTIVE CACHE MANAGEMENT

Chen et al. [2014b] propose coordinated cache bypassing and warp throttling that takes advantage of both warp throttling and cache bypassing to improve performance on highly cache-sensitive applications. The proposed mechanism detects cache contention and memory resource contention at runtime, and coordinates throttling and bypassing policy accordingly. The mechanism implements cache bypassing via an existing CPU cache bypassing technique of protection distance, which prevents a cache line from being evicted for a number of accesses. Upon insertion into the cache, the line is assigned a protection distance and counters track the remaining protection distance for lines. Once the remaining protection distance reaches 0, the line is no longer protected and can be evicted. When a new memory request attempts to insert a new line into a set with no unprotected lines, the memory request bypasses the cache.

旁路机制位于 L2 缓存的前面，接收标记为生成它们的 warp 类型的内存请求。此机制尝试消除来自全未命中 warp 的访问，并将大部分未命中 warp 转换为全未命中 warp。该块只是将所有标记为来自全未命中和大部分未命中 warp 的请求直接发送到内存调度程序。

MeDiC 的缓存管理策略通过改变从 DRAM 返回的请求在 L2 的 LRU 堆栈中的位置来运行。大部分未命中的 warp 请求的缓存行被插入到 LRU 位置，而所有其他请求则被插入到传统的 MRU 位置。

最后，MeDic 修改了基线内存请求调度程序，使其包含两个内存访问队列：一个用于全命中和大部分命中的 Warp 的高优先级队列，以及一个用于平衡、大部分未命中和全未命中 Warp 的低优先级队列。内存调度程序只是将高优先级队列中的所有请求优先于低优先级队列中的任何请求。

4.4.5 协调缓存绕过

Xie 等人 [2015] 探索了选择性启用缓存旁路以提高缓存命中率的潜力。他们使用分析来确定 GPGPU 应用程序中的每个静态加载指令是否具有良好局部性、较差局部性或中等局部性。他们相应地标记指令。标记为具有良好局部性的加载操作可以使用 L1 数据缓存。标记为具有较差局部性的加载操作始终被绕过。标记为中等局部性的加载指令采用自适应机制，其工作方式如下。自适应机制以线程块粒度运行。对于给定的线程块，所有执行的中等局部性加载都得到统一处理。它们要么使用 L1，要么旁路。行为是在启动线程块时根据阈值确定的，该阈值使用考虑了 L1 缓存命中和管道资源冲突的性能指标在线调整。他们的评估表明，这种方法比静态扭曲限制更能提高缓存命中率。

4.4.6 自适应缓存管理

Chen 等人 [2014b] 提出了协调的缓存绕过和 warp 节流，利用 warp 节流和缓存绕过的优点来提高高度缓存敏感应用程序的性能。所提出的机制在运行时检测缓存争用和内存资源争用，并相应地协调节流和绕过策略。该机制通过现有的 CPU 缓存绕过保护距离技术实现缓存绕过，这可防止缓存行因多次访问而被逐出。在插入缓存后，该行会被分配一个保护距离，并且计数器会跟踪行的剩余保护距离。一旦剩余保护距离达到 0，该行将不再受保护并可被逐出。当新的内存请求尝试将新行插入到没有未受保护行的集合中时，内存请求会绕过缓存。

The protection distance is set globally and the optimal value differs between workloads. In this work, Chen et al. [2014b] sweep the static protection distance and demonstrate that GPU workloads are relatively insensitive to the protection distance value.

4.4.7 CACHE PRIORITIZATION

Li et al. [2015] observe that warp throttling optimize L1 cache hit rate while potentially leaving other resources such as off-chip bandwidth and L2 cache significantly underutilized. They propose a mechanism of assigning tokens to warps to determine which warps can allocate lines into the L1 cache. Additional "non-polluting warps" are not given a token so that while they can execute they are not permitted to evict data from the L1. This leads to an optimization space where both the number of warps that can be scheduled (W) and the number that have tokens (T) can be set to less than the maximum number of warps that can execute. They show that statically selecting the optimal value of W and T enables a 17% improvement over CCWS with static warp limiting.

Based on this observation, Li et al. [2015] explore two mechanisms to learn the best values for W and T. The first approach is based upon the idea of maintaining high thread level parallelism while increasing cache hit rates. In this approach, called dynPCALMTLP, a sampling period runs a kernel with W set to the maximum number of warps and then varies T across different SIMT cores. The value of T that achieves the maximum performance is then selected. This leads to comparable performance to CCWS with significantly less area overhead. The second approach, called dynPCALCCWS, initially uses CCWS to set W then uses dynPCALMTLP to determine T. Then it monitors resource usage of shared structures to dynamically increase or decrease W. This leads to an 11% performance gain versus CCWS.

4.4.8 VIRTUAL MEMORY PAGE PLACEMENT

Agarwal et al. [2015] consider the implications of supporting cache coherence across multiple physical memory types in a heterogeneous system including both capacity-optimized and bandwidth-optimized memory. Since DRAM optimized for bandwidth is more expensive in cost and energy than DRAM optimized for capacity, future systems are likely to include both. Agarwal et al. [2015] observe current OS page placement policies such as those deployed in Linux do not account for the non-uniformity of memory bandwidth. They study a future system in which a GPU can access low bandwidth/high capacity CPU memory at low latency—a penalty of 100 core cycles. Their experiments use a modified version of GPGPU-Sim 3.2.2 configured with additional MSHR resources to model more recent GPUs.

With this setup, they first find that for memory bandwidth limited applications there is significant opportunity to gain performance by using both CPU and GPU memory to increase aggregate memory bandwidth. They find less memory latency limited GPGPU applications for which this is not the case. Under the assumption that pages are accessed uniformly and when memory capacity of bandwidth-optimized memory is not a limitation, they show that allocating

保护距离是全局设置的，最佳值因工作负载而异。在本文中，Chen 等人 [2014b] 扫描了静态保护距离，并证明 GPU 工作负载对保护距离值相对不敏感。

4.4.7 缓存优先级

Li 等人 [2015] 观察到，warp 节流优化了 L1 缓存命中率，但可能会导致其他资源（如片外带宽和 L2 缓存）的利用率大大降低。他们提出了一种向 warp 分配令牌的机制，以确定哪些 warp 可以将行分配到 L1 缓存中。额外的“ 无污染 warp ”没有被赋予令牌，因此虽然它们可以执行，但不允许从 L1 中逐出数据。这导致了一个优化空间，其中可调度的 warp 数量 (W) 和具有令牌的数量 (T) 都可以设置为小于可执行的最大 warp 数量。他们表明，静态选择 W 和 T 的最优值可以使性能比使用静态 warp 限制的 CCWS 提高 17%。

基于这一观察，Li 等人 [2015] 探索了两种机制来学习 W 和 T 的最佳值。第一种方法基于在提高缓存命中率的同时保持高线程级并行性的理念。在这种方法中，称为 dynPCALMTLP，采样周期运行一个内核，其中 W 设置为最大 warp 数，然后在不同的 SIMT 核心之间改变 T。然后选择实现最大性能的 T 值。这可以实现与 CCWS 相当的性能，但面积开销明显更少。第二种方法称为 dynPCALCCWS，最初使用 CCWS 设置 W，然后使用 dynPCALMTLP 确定 T。然后它监视共享结构的资源使用情况以动态增加或减少 W。与 CCWS 相比，这可以使性能提高 11%。

4.4.8 虚拟内存页面布局

Agarwal 等人 [2015] 考虑在包括容量优化和带宽优化内存的异构系统中支持跨多种物理内存类型的缓存一致性的影响。由于针对带宽优化的 DRAM 比针对容量优化的 DRAM 成本更高、能耗更高，因此未来的系统可能会同时包含这两种内存。Agarwal 等人 [2015] 观察到当前的操作系统页面放置策略（例如在 Linux 中部署的策略）并未考虑内存带宽的不均匀性。他们研究了一种未来的系统，其中 GPU 可以以低延迟访问低带宽/高容量 CPU 内存——损失 100 个核心周期。他们的实验使用配置了额外 MSHR 资源的修改版 GPGPU-Sim 3.2.2 来模拟更新的 GPU。

通过这种设置，他们首先发现，对于内存带宽受限的应用程序，通过使用 CPU 和 GPU 内存来增加总内存带宽，有很大的机会获得性能提升。他们发现内存延迟较少的 GPGPU 应用程序并非如此。在假设页面访问均匀且带宽优化内存的内存容量不受限制的情况下，他们表明分配

pages to memory regions in proportion to the regions' available memory bandwidth is optimal. Assuming capacity of bandwidth limited memory is not an issue, they find a simple policy of randomly allocating pages to bandwidth- or capacity-optimized memory with probability in proportion to memory bandwidth works in practice with real GPGPU programs. However, when bandwidth-optimized memory capacity is insufficient to meet application demands, they find it is necessary to refine the page placement to consider frequency of access.

To refine the page placement, they propose a system involving a profiling pass implemented using a modified version of the NVIDIA developer tools nvcc and ptxas along with an extension of the existing CUDA API to include page placement hints. Using profile-guided page placement hints obtains about 90% of the benefits of an oracle page placement algorithm. They leave page migration strategies to future work.

4.4.9 DATA PLACEMENT

Chen et al. [2014a] propose PORPLE, a portable data placement strategy that consists of specification language, a source-to-source compiler and an adaptive runtime data placer. They capitalize on the observation that with all the various flavors of memory available on the GPU, choosing what data should be placed where is difficult for the programmer to determine and is often not portable from one GPU architecture to the next. The goal of PORPLE is to be extensible, input-adaptive, and generally applicable to both regular and irregular data accesses. Their approach relies on three solutions.

The first solution is a memory specification language to help with extensibility and portability. The memory specification language describes all the various forms of memory on the GPU based on the conditions under which accesses to these spaces are serialized. For example, accesses to adjacent global data are coalesced, hence accessed concurrently, but accesses to the same bank of shared memory must be serialized.

The second solution is a source-to-source compiler named PORPLE-C which transforms the original GPU program into a placement agnostic version. The compiler inserts guards around accesses to memory, selecting the access that corresponds to the predicted best placement of the data.

Finally, to predict which data placement would be most optimal, they use PORPLE-C to find static access patterns through code analysis. When the static analysis cannot make a determination on the access pattern, the compiler generates a function that traces the runtime access patterns and attempts to make a prediction. This function is run on the CPU for a short period of time and helps determine the best GPU-based data placement prior to launching the kernel. In the scope of this work, the system only handles the placement of arrays, as they are the most prevalent data structure used in GPU kernels.

The lightweight model used to make the placement prediction in PORPLE generates an estimate of the number of transactions generated based on the serialization conditions of the memory. For memories that have a cache hierarchy, it uses a reuse distance estimation of cache

按照区域可用内存带宽的比例将页面分配到内存区域是最佳的。假设带宽受限内存的容量不是问题，他们发现一个简单的策略，即按照内存带宽的比例将页面随机分配到带宽或容量优化的内存中，这种策略在实际的 GPGPU 程序中是可行的。然而，当带宽优化的内存容量不足以满足应用程序需求时，他们发现有必要优化页面布局以考虑访问频率。

为了优化页面布局，他们提出了一个系统，该系统涉及使用 NVIDIA 开发工具 nvcc 和 ptxas 的修改版本以及现有 CUDA API 的扩展来实现的分析过程，以包含页面布局提示。使用配置文件引导的页面布局提示可获得 Oracle 页面布局算法的约 90% 的好处。他们将页面迁移策略留待将来研究。

4.4.9 数据放置

Chen 等人 [2014a] 提出了一种可移植数据放置策略 PORPLE，它由规范语言、源到源编译器和自适应运行时数据放置器组成。他们利用了以下观察：由于 GPU 上有各种类型的内存，程序员很难决定将哪些数据放置在何处，而且通常无法从一个 GPU 架构移植到另一个 GPU 架构。PORPLE 的目标是可扩展、输入自适应，并且通常适用于常规和不规则数据访问。他们的方法依赖于三种解决方案。

第一个解决方案是内存规范语言，以帮助实现可扩展性和可移植性。内存规范语言根据对这些空间的访问进行序列化的条件来描述 GPU 上各种形式的内存。例如，对相邻全局数据的访问是合并的，因此可以同时访问，但对同一共享内存组的访问必须进行序列化。

第二种解决方案是名为 PORPLE-C 的源到源编译器，它将原始 GPU 程序转换为与位置无关的版本。编译器在内存访问周围插入保护，选择与预测的最佳数据位置相对应的访问。

最后，为了预测哪种数据放置方式最为理想，他们使用 PORPLE-C 通过代码分析找到静态访问模式。当静态分析无法确定访问模式时，编译器会生成一个函数来跟踪运行时访问模式并尝试进行预测。此函数在 CPU 上运行一小段时间，有助于在启动内核之前确定最佳的基于 GPU 的数据放置方式。在本工作范围内，系统仅处理数组的放置，因为它们是 GPU 内核中最常用的数据结构。

PORPLE 中用于进行放置预测的轻量级模型根据内存的序列化条件生成事务数量的估计值。对于具有缓存层次结构的内存，它使用缓存的重用距离估计

hit rate. When multiple arrays share a cache, the estimate of how much cache is devoted to each array is based on a linear partitioning of the cache based on the size of the array.

4.4.10 MULTI-CHIP-MODULE GPUS

Arunkumar et al. [2017] note that the slowing of Moore's Law will result in slowing increases in GPU performance. They propose to extend performance scaling by building large GPUs out of smaller GPU modules on a multichip module (see Figure 4.4). They demonstrate it is possible to attain with 10% of the performance of a single large (and unimplementable) monolithic GPU by combining local caching of remote data, CTA scheduling to modules that considers locality and first-touch page allocation. According to their analysis this is 45% better performance than possible using the largest implementable monolithich GPU in the same process technology.

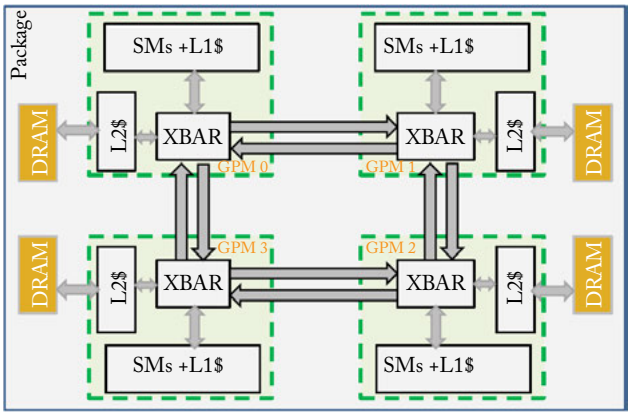


Figure 4.4: A multi-chip-module GPU (based on Figure 3 from Arunkumar et al. [2017]).

命中率。当多个阵列共享一个缓存时，对每个阵列分配多少缓存的估计是基于基于阵列大小的缓存的线性分区。

4.4.10 多芯片模块 GPU

Arunkumar 等人 [2017] 指出，摩尔定律的放缓将导致 GPU 性能提升放缓。他们建议通过在多芯片模块上用较小的 GPU 模块构建大型 GPU 来扩展性能扩展（见图 4.4）。他们证明，通过结合远程数据的本地缓存、考虑局部性的模块 CTA 调度和首次接触页面分配，可以实现单个大型（且不可实现）单片 GPU 的 10% 性能。根据他们的分析，这比使用相同工艺技术的最大可实现单片 GPU 的性能高出 45%。

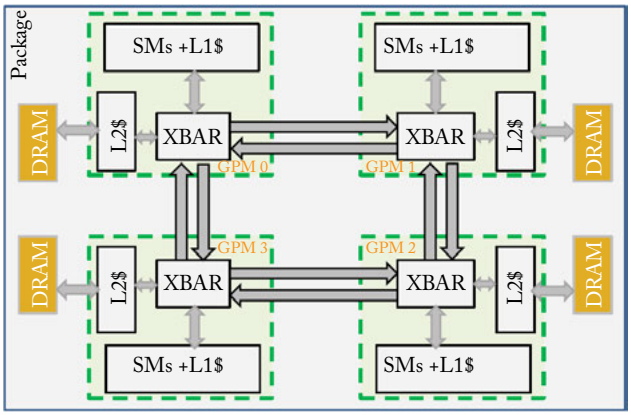


图 4.4：多芯片模块 GPU（基于 Arunkumar 等人 [2017] 的图 3）。

CHAPTER 5

Crosscutting Research on GPU Computing Architectures

This chapter details several research directions in GPGPU architecture that do not fit neatly into earlier chapters which focused on specific portions of the GPU architecture. Section 5.1 explores work on how threads are scheduled in GPUs. Section 5.2 looks at alternative programming methodologies and Section 5.4 examines work on heterogeneous CPU/GPU computing.

5.1 THREAD SCHEDULING

Contemporary GPUs are fundamentally different from CPUs in that they rely on massive parallelism. Independent of how a program is specified (e.g., using OpenCL, CUDA, OpenACC, etc.), workloads without extensive software-defined parallelism are not suitable for GPU acceleration. GPUs employ several mechanisms to aggregate and schedule all these threads. There are three primary ways threads on the GPU are organized and scheduled.

Assignment of Threads to Warps Since GPUs use SIMD units to execute threads defined by a MIMD programming model, the threads must be fused together for lock-step execution in the form of warps. In the baseline GPU architecture studied throughout this book, threads with consecutive thread IDs are statically fused together to form warps. Section 3.4.1 summarizes the research proposals on alternative thread arrangements within warps for better warp compaction.

Dynamic Assignment of Threadblocks to Cores Unlike in CPUs where threads can be assigned to hardware threads one at a time, in GPUs, work is assigned to GPU cores in bulk. This work-unit consists of multiple warps in the form of a threadblock. In our baseline GPU, threadblocks are assigned to cores in round-robin order. The core's resources (like warp-slots, register file, and shared memory space) are subscribed at the threadblock granularity. Due to the large amount of states associated with each threadblock, contemporary GPUs do not preempt their execution. Threads in the threadblock run to completion before their resources can be assigned to another threadblock.

Cycle-by-cycle Scheduling Decisions After a threadblock have been assigned to a GPU core, a collection of fine-grained hardware schedulers decides, at each cycle, which set of warps to

CHAPTER 5

GPU 计算架构的交叉研究

本章详细介绍了 GPGPU 架构中的几个研究方向，这些方向与之前专注于 GPU 架构特定部分的章节不太契合。第 5.1 节探讨了如何在 GPU 中调度线程。第 5.2 节介绍了替代编程方法，第 5.4 节研究了异构 CPU/GPU 计算方面的工作。

5.1 线程调度

当代 GPU 与 CPU 的根本区别在于它们依赖于大规模并行性。无论程序是如何指定的（例如，使用 OpenCL、CUDA、OpenACC 等），没有大量软件定义并行性的工作负载都不适合 GPU 加速。GPU 采用多种机制来聚合和调度所有这些线程。GPU 上的线程组织和调度主要有三种方式。

将线程分配到 Warp 由于 GPU 使用 SIMD 单元来执行由 MIMD 编程模型定义的线程，因此必须将线程融合在一起，以 Warp 的形式进行锁步执行。在本书研究的基线 GPU 架构中，具有连续线程 ID 的线程会静态融合在一起以形成 Warp。第 3.4.1 节总结了有关 Warp 内替代线程排列的研究提案，以实现更好的 Warp 压缩。

将线程块动态分配给核心 与 CPU 不同，在 CPU 中，线程可以一次分配给一个硬件线程，而在 GPU 中，工作是批量分配给 GPU 核心的。此工作单元由多个线程块形式的 warp 组成。在我们的基准 GPU 中，线程块按循环顺序分配给核心。核心的资源（如 warp 槽、寄存器文件和共享内存空间）以线程块粒度订阅。由于每个线程块都与大量状态相关，当代 GPU 不会抢占它们的执行。线程块中的线程运行完毕后，才能将其资源分配给另一个线程块。

逐周期调度决策将线程块分配给 GPU 核心后，一组细粒度的硬件调度程序将在每个周期决定要使用哪组 Warp

fetch instructions, which warps to issue instructions for execution, and when to read/write operands for each issued instruction.

Scheduling Multiple Kernels Threadblock-level and cycle-by-cycle scheduling decisions can take place both within a kernel and across different kernels running concurrently on the same GPU. Legacy kernel scheduling limited just one kernel to be active on a GPU at a time. However, the introduction of NVIDIA's Streams and HyperQ scheduling mechanisms make the running of concurrent kernels possible. This situation is similar in some ways to multiprogramming on CPUs.

5.1.1 RESEARCH ON ASSIGNMENT OF THREADBLOCKS TO CORES

When a kernel is launched, threads within each kernel launch are grouped into threadblocks. A GPU-wide threadblock scheduling mechanism assigns each threadblock to one of the SIMT cores, based on resource availability. Each core has a fixed amount of scratchpad memory (called shared memory in CUDA or local memory in OpenCL), number of registers, slots for warps, and slots for threadblocks. At kernel launch time, all of these parameters are known for each threadblock. The most obvious threadblock scheduling algorithm is to assign threadblocks to cores in a round-robin fashion to maximize the number of cores involved. Threadblocks are continuously scheduled until at least one resource in each core is exhausted. Note that a kernel may be composed of more threadblocks than can be run on the GPU at once. As a result, some threadblocks in a kernel may not even be running on the GPU while others execute. Several research techniques have looked at trade-offs in the threadblock scheduling space.

Throttling at the Threadblock Level. Kayiran et al. [2013] propose throttling the number of threadblocks allocated to each core in order to reduce contention in the memory system caused by thread over-subscription. They develop an algorithm that monitors core idle cycles and memory delay cycles. The algorithm starts by assigning each core only half of its maximum threadblocks. Then the idle and memory delay cycles are monitored. If a core is predominantly waiting for memory, then no more threadblocks are assigned and existing threadblocks are potentially paused to stop them from issuing instructions. The technique achieves a course-grained parallelism throttling mechanism, which limits memory system interference and improves overall application performance, even though less CTAs are concurrently active.

Dynamically Tuning GPU Resources. Sethia and Mahlke [2014] propose Equalizer, a hardware runtime system that dynamically monitors resource contention and scales the number of threads, core frequency, and memory frequency to improve both energy consumption and performance. The system bases its decision on four parameters: (1) the number of active warps in an SM; (2) the number of warps waiting for data from memory; (3) number of warps ready to execute an arithmetic instruction; and (4) number of warps ready to execute a memory instruction. With these parameters, it first decides the number of warps to keep active on an SM, then based on this value and the values of the other three counters (which act as proxies for memory

获取指令，它会扭曲以发出需要执行的指令，以及何时读取/写入每个发出的指令的操作数。

调度多个内核 线程块级和逐周期调度决策既可以在内核中进行，也可以在同一 GPU 上同时运行的不同内核之间进行。传统内核调度限制一次只能在 GPU 上激活一个内核。但是，NVIDIA 的 Streams 和 HyperQ 调度机制的引入使得并发内核的运行成为可能。这种情况在某些方面类似于 CPU 上的多道编程。

5.1.1 线程块与核心的分配研究

启动内核时，每个内核启动中的线程被分组为线程块。GPU 范围的线程块调度机制根据资源可用性将每个线程块分配给 SIMT 内核之一。每个内核都有固定数量的暂存器内存（在 CUDA 中称为共享内存，在 OpenCL 中称为本地内存）、寄存器数量、用于 warp 的插槽和用于线程块的插槽。在内核启动时，每个线程块的所有这些参数都是已知的。最明显的线程块调度算法是以循环方式将线程块分配给内核，以最大化所涉及的内核数量。线程块被连续调度，直到每个内核中的至少一个资源耗尽。请注意，内核可能由比 GPU 上一次可以运行的线程块更多的线程块组成。因此，内核中的一些线程块甚至可能没有在 GPU 上运行，而其他线程块则在执行。有几种研究技术研究了线程块调度空间中的权衡。

线程块级别的节流。Kayiran 等人 [2013] 建议限制分配给每个内核的线程块数量，以减少线程超额订阅导致的内存系统争用。他们开发了一种监控内核空闲周期和内存延迟周期的算法。该算法首先为每个内核分配其最大线程块的一半。然后监控空闲和内存延迟周期。如果内核主要在等待内存，则不会再分配线程块，并且现有线程块可能会暂停以阻止它们发出指令。该技术实现了一种粗粒度并行节流机制，即使同时活动的 CTA 较少，也可以限制内存系统干扰并提高整体应用程序性能。

动态调整 GPU 资源。Sethia 和 Mahlke [2014] 提出了 Equalizer，这是一种硬件运行时系统，可动态监控资源争用情况并调整线程数、核心频率和内存频率，以改善能耗和性能。该系统根据四个参数做出决策：(1) SM 中的活动 Warp 数；(2) 等待内存数据的 Warp 数；(3) 准备执行算术指令的 Warp 数；(4) 准备执行内存指令的 Warp 数。利用这些参数，它首先决定 SM 上要保持活动的 Warp 数，然后根据此值和其他三个计数器（充当内存的代理）的值

contention, compute intensity, and memory intensity) it decides how best to scale the frequency of both the core and the memory system.

Equalizer has two modes of operation: energy-saving mode and performance-enhancing mode. In energy-savings mode it saves energy by scaling back the under-utilized resource to minimize energy consumption while mitigating its impact on performance. In performance-enhancing mode, Equalizer boosts the bottleneck resource increasing performance in an energy-efficient manner.

They characterize a set of workloads from Rodinia and Parboil as being either compute intensive, memory intensive, cache sensitive, or unsaturated by examining the performance and energy tradeoffs associated with changing the memory frequency, compute frequency, and the number of the threads concurrently running. If minimizing energy (without sacrificing performance) is the goal, then compute intensive kernels should operate with a lower memory frequency and memory kernels should operate with a lower SIMT core frequency. This help reduce energy unnecessarily spent in the system that is not being fully utilized at the baseline rate.

Equalizer makes decisions about frequency and concurrency on an interval basis. The technique adds monitor hardware to each SIMT core that makes a local decision based on the four counters listed previously. It decides locally in each SIMT core what the three output parameters (number of CTAs, memory frequency, and compute frequency) should be for this epoch. It informs a global work distribution engine about the number of CTAs this SM should use, issuing new blocks if the SIMT core wants more work. If the SM should run with fewer CTAs, it pauses some of the CTAs on the core. After deciding the number of CTAs to run, each SIMT core submits a memory/compute voltage target to a global frequency manager which sets the chip-wide frequencies based on a majority function.

The local decision is made by observing the number of warps that are waiting to execute a memory instruction and the number of warps that are waiting to execute an ALU instruction. If the number of warps trying to wait for memory is greater than the number of warps in a CTA, then the number of CTAs running on this SIMT core is reduced, potentially helping the performance of cache-sensitive workloads. The SIMT core is considered memory (or compute) intensive if the number of warps ready to issue memory (or ALU) is more than the number of warps in a CTA. If there are fewer warps waiting on memory (or compute) than there are warps in a CTA, the SIMT core can still be considered ALU or memory bound if more than half of the active warps are waiting and there are not more than two warps waiting on memory. If this is the case, then the number of active CTAs on the core is incremented by one and the SIMT core is determined to be compute or memory bound based on if there are more compute waiting warps or more memory waiting warps.

Once the SIMT cores have made their local decisions, the frequencies of the memory and the cores are scaled by $\pm 15\%$ based on the mode of operation Equalizer is operating in.

通过权衡竞争、计算强度和内存强度，它决定了如何最好地扩展核心和内存系统的频率。

均衡器有两种运行模式：节能模式和性能增强模式。在节能模式下，它通过缩减未充分利用的资源来节省能源，以最大限度地降低能耗，同时减轻其对性能的影响。在性能增强模式下，均衡器以节能的方式增强瓶颈资源，从而提高性能。

他们通过检查与更改内存频率、计算频率和同时运行的线程数相关的性能和能源权衡，将 Rodinia 和 Parboil 的一组工作负载描述为计算密集型、内存密集型、缓存敏感型或不饱和型。如果目标是最小化能源（而不牺牲性能），那么计算密集型内核应该以较低的内存频率运行，内存内核应该以较低的 SIMT 核心频率运行。这有助于减少系统中不必要消耗的能源，这些能源在基准速率下没有得到充分利用。

均衡器会根据间隔做出有关频率和并发性的决策。该技术为每个 SIMT 核心添加监控硬件，该硬件会根据前面列出的四个计数器做出本地决策。它在每个 SIMT 核心中本地决定此时期的三个输出参数（CTA 数量、内存频率和计算频率）应该是什么。它会通知全局工作分配引擎此 SM 应使用的 CTA 数量，如果 SIMT 核心需要更多工作，则会发出新块。如果 SM 应该使用更少的 CTA 运行，它会暂停核心上的某些 CTA。在决定要运行的 CTA 数量后，每个 SIMT 核心都会向全局频率管理器提交内存/计算电压目标，该管理器根据多数函数设置芯片范围的频率。

通过观察等待执行内存指令的 Warp 数量和等待执行 ALU 指令的 Warp 数量来做出本地决策。如果尝试等待内存的 Warp 数量大于 CTA 中的 Warp 数量，则在此 SIMT 核心上运行的 CTA 数量会减少，从而可能有助于提高缓存敏感型工作负载的性能。如果准备发出内存（或 ALU）的 Warp 数量大于 CTA 中的 Warp 数量，则 SIMT 核心被视为内存（或计算）密集型。如果等待内存（或计算）的 Warp 数量少于 CTA 中的 Warp 数量，则如果超过一半的活动 Warp 正在等待并且等待内存的 Warp 不超过两个，则 SIMT 核心仍可被视为 ALU 或内存受限。如果是这种情况，则核心上活动 CTA 的数量将增加一，并且根据是否存在更多计算等待 warp 或更多内存等待 warp 来确定 SIMT 核心是计算限制还是内存限制。

一旦 SIMT 核心做出了本地决策，内存和核心的频率就会根据均衡器所运行的操作模式按 ±15% 进行缩放。

5.1.2 RESEARCH ON CYCLE-BY-CYCLE SCHEDULING DECISIONS

Early Characterizations of Cycle-by-Cycle Scheduling. Lakshminarayana and Kim [2010] explore numerous warp-scheduling policies in the context of an early GPU without hardware managed caches and show that, for applications that execute symmetric (balanced) dynamic instruction counts per warp, a fairness-based warp and DRAM access scheduling policy improves performance. This policy works well on the regular GPU workloads used in they study because regular memory requests between warps are both merged within the core and better exploit DRAM row-buffer locality. The paper also characterizes several other warp scheduling policies, including ICOUNT, which was first proposed by Tullsen et al. [1996] for simultaneously multithreaded CPUs. ICOUNT is designed to improve system throughput by prioritizing the fastest progressing warp (or thread). Lakshminarayana and Kim [2010] show that prioritizing only a few warps in their early cache-less GPU on early, regular workloads generally does not improve performance.

Two-Level Scheduling. Gebhart et al. [2011c] introduce the use of a two-level scheduler to improve energy efficiency. Their two-level scheduler divides warps in a core into two pools: an active pool of warps that are considered for scheduling in the next cycle and an inactive pool of warps that are not. A warp transition out of the active pool whenever it encounters a compiler identified global or texture memory dependency and back into the active pool in round-robin fashion from the inactive pool. Selecting from a smaller pool of warps every cycle reduces the size and energy consumption of the warp selection logic.

The two-level scheduler proposed by Narasiman et al. [2011] focuses on improving performance by allowing groups of threads to reach the same long latency operation at different times. This helps ensure cache and row-buffer locality within a fetch group is maintained. The system can then hide long latency operations by switching between fetch groups. In contrast, Cache-Conscious Warp Scheduling (see below) focuses on improving performance by adaptively limiting the amount of multithreading the system can maintain based on how much intra-warp locality is being lost.

Cache-Conscious Warp Scheduling. Rogers et al. [2012] categorize the memory locality that exists in GPU kernels as being *intra-warp*, where a warp load then references its own data, or *inter-warp*, where a warp shares data with other warps. They demonstrate that intra-warp locality is most common form of locality that occurs in cache-sensitive workloads. Based on this observation, they proposed a cache-conscious wavefront scheduling (CCWS) mechanism to exploit this locality by throttling the number of warps actively scheduled based on memory system feedback.

Actively scheduling between fewer warps enables each individual warp to consume more cache space and reduces L1 data cache contention. In particular, throttling occurs when workloads with locality are thrashing the cache. To detect this thrashing, CCWS introduces a lost-locality detection mechanism which is based on replacement victim tags from the L1 data cache.

逐周期调度的早期特征。Lakshminarayana 和 Kim [2010] 在没有硬件管理缓存的早期 GPU 背景下探索了多种 warp 调度策略，并表明对于每个 warp 执行对称（平衡）动态指令数的应用程序，基于公平性的 warp 和 DRAM 访问调度策略可以提高性能。该策略在他们研究中使用的常规 GPU 工作负载上效果很好，因为 warp 之间的常规内存请求都在核心内合并，并且更好地利用了 DRAM 行缓冲区局部性。本文还描述了其他几种 warp 调度策略，包括 ICOUNT，它是由 Tullsen 等人 [1996] 首次为同时多线程 CPU 提出的。ICOUNT 旨在通过优先考虑速度最快的 warp（或线程）来提高系统吞吐量。 Lakshminarayana 和 Kim [2010] 表明，在早期的无缓存 GPU 中，在早期的常规工作负载下仅优先考虑少数 Warp 通常不会提高性能。

两级调度。Gebhart 等人 [2011c] 引入了两级调度程序来提高能源效率。他们的两级调度程序将核心中的 Warp 分为两个池：一个是活动 Warp 池，其中的 Warp 将在下一个周期进行调度；另一个是非活动 Warp 池，其中的 Warp 不会在下一个周期进行调度。每当 Warp 遇到编译器识别的全局或纹理内存依赖关系时，它就会从活动池中转移出来，并以循环方式从非活动池中转移回活动池。每个周期从较小的 Warp 池中进行选择可减少 Warp 选择逻辑的大小和能耗。

Narasiman 等人 [2011] 提出的两级调度程序着重于通过允许线程组在不同时间达到相同的长延迟操作来提高性能。这有助于确保保持提取组内的缓存和行缓冲区局部性。然后，系统可以通过在提取组之间切换来隐藏长延迟操作。相比之下，缓存意识 Warp 调度（见下文）着重于通过根据丢失的 Warp 内局部性程度自适应地限制系统可以维持的多线程数量来提高性能。

缓存意识 Warp 调度。Rogers 等人 [2012] 将 GPU 内核中存在的内存局部性分类为 *intra-warp*，其中 Warp 加载会引用其自己的数据，或 *inter-warp*，其中 Warp 与其他 Warp 共享数据。他们证明，Warp 内局部性是缓存敏感型工作负载中最常见的局部性形式。基于这一观察，他们提出了一种缓存意识波前调度 (CCWS) 机制，通过根据内存系统反馈限制主动调度的 Warp 数量来利用这种局部性。

在较少的 Warp 之间进行主动调度可使每个 Warp 占用更多的缓存空间，并减少 L1 数据缓存争用。具体而言，当具有局部性的工作负载使缓存发生抖动时，就会发生节流。为了检测这种抖动，CCWS 引入了一种基于 L1 数据缓存中的替换受害者标签的丢失局部性检测机制。

Figure 5.1 plots the high-level microarchitecture of CCWS. On every eviction from the cache, the victim's tag is written to warp-private victim tag array. Each warp has its own victim tag array, because CCWS is only concerned with detecting intra-warp locality. On every subsequent cache miss, the victim tag array for the missing warp is probed. If the tag is found in the victim tags, then some intra-warp locality has been lost. CCWS makes the assumption that this warp might have been able to hit on this line if the warp had more exclusive access to the L1 data cache, and therefore could benefit potentially benefit from throttling.

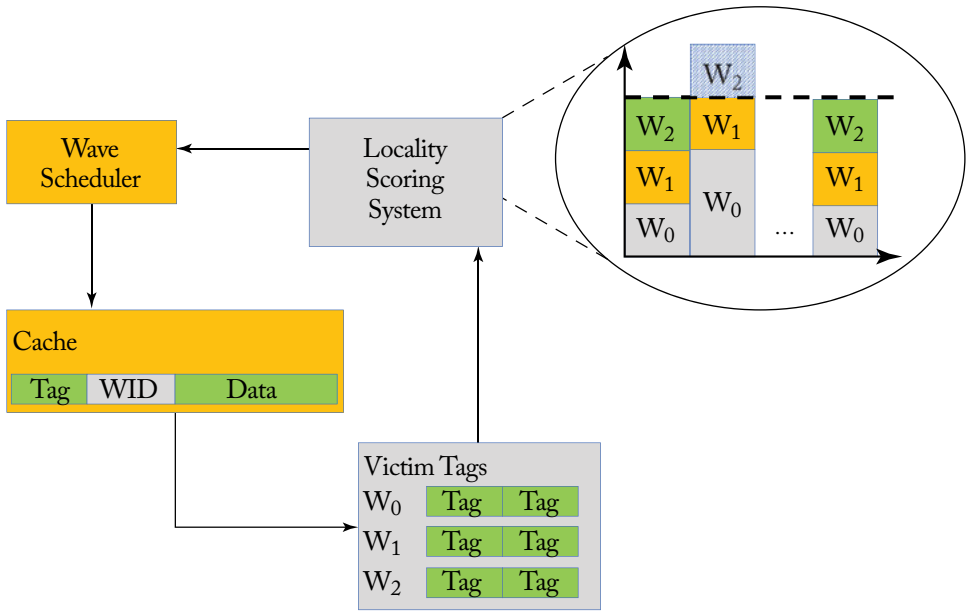


Figure 5.1: Cache-conscious wavefront scheduling microarchitecture.

To reflect this loss in locality, a signal is sent to the scheduling system. The issue scheduler uses a locality scoring system to approximate how much locality each warp in the system has lost, which is an approximation of how much additional cache capacity each warp requires. All warps in the locality scoring system are assigned an initial score, with the assumption that all warps require the same cache capacity and that no throttling occurs (the stacked bars in Figure 5.1). As time passes and lost locality is detected, the scores for individual warps are increased. In the example in Figure 5.1, warp 0 has experienced a loss in locality and its score has been increased. The increase in its score has pushed warp 3 over a threshold, which will prevent it from issuing L1 data cache requests, effectively throttling the number of warps actively scheduled on the core. Over time, if no locality is lost then the score for warp 0 is decreased until warp 2 is able to fall below the threshold and is able to issue memory requests again.

图 5.1 绘制了 CCWS 的高级微架构。每次从缓存中驱逐时，受害者的标签都会写入 warp 私有受害者标签数组。每个 warp 都有自己的受害者标签数组，因为 CCWS 只关心检测 warp 内部局部性。每次后续缓存未命中时，都会探测丢失的 warp 的受害者标签数组。如果在受害者标签中找到标签，则部分 warp 内部局部性已丢失。CCWS 假设，如果 warp 对 L1 数据缓存拥有更多的独占访问权限，则该 warp 可能能够命中此行，因此可以从节流中受益。

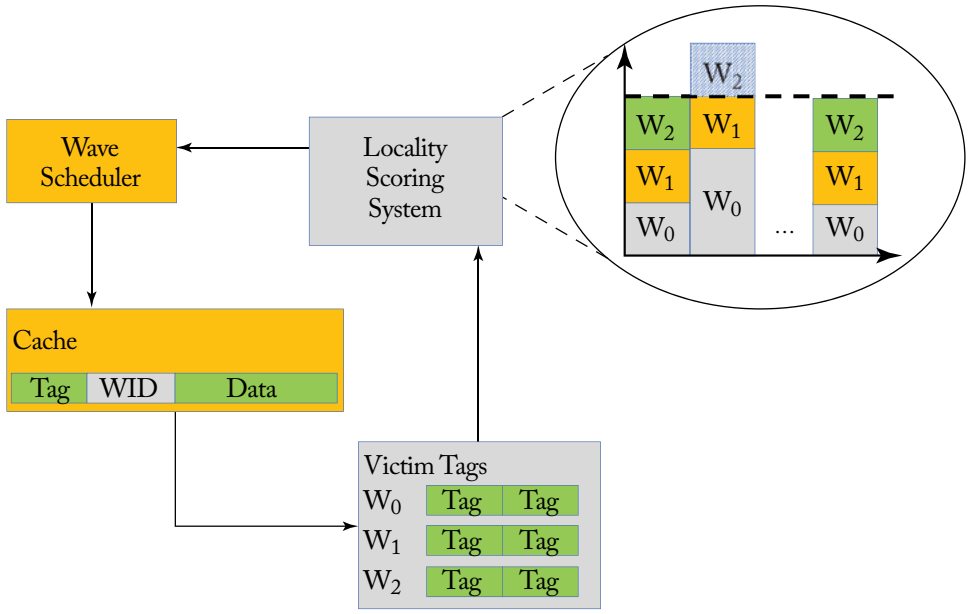


图 5.1：缓存意识的波前调度微架构。

为了反映局部性的损失，会向调度系统发送一个信号。发布调度程序使用局部性评分系统来估计系统中的每个 Warp 丢失了多少局部性，这是每个 Warp 需要多少额外缓存容量的近似值。局部性评分系统中的所有 Warp 都被分配了一个初始分数，假设所有 Warp 都需要相同的缓存容量，并且不会发生限制（图 5.1 中的堆叠条）。随着时间的推移和检测到丢失的局部性，各个 Warp 的分数会增加。在图 5.1 的示例中，Warp 0 经历了局部性的损失，其分数增加了。其分数的增加使 Warp 3 超过了阈值，这将阻止它发出 L1 数据缓存请求，从而有效地限制了核心上主动调度的 Warp 数量。随着时间的推移，如果没有丢失局部性，那么 warp 0 的分数就会减少，直到 warp 2 能够降至阈值以下并能够再次发出内存请求。

CCWS goes on to demonstrate the sensitivity cache hit rate has to scheduling decisions by comparing various scheduling mechanisms to cache replacement policies. The paper demonstrates that the decision space available to the warp scheduler is much greater than the relatively constrained decision space made by a replacement policy. The paper goes on to further demonstrate that the CCWS scheme, using an LRU replacement policy can improve cache hit rate more than prior scheduling mechanisms, even if they use a Belady optimal cache-replacement policy.

Rogers et al. [2013] propose Divergence-Aware Warp Scheduling (DAWS), which extends CCWS with a more accurate estimate of each warp's cache footprint. DAWS capitalizes on the fact that most intra-warp locality on GPU workloads occurs in loops. DAWS creates a per-warp cache footprint estimate for warps in loops. DAWS pre-emptively throttles the number of warps in loops based on the per-warp predicted loop footprint. DAWS also adapts its cache-footprint estimate with the level of control flow divergence experienced by each warp. Threads in a warp that have left the loop no longer contribute to the footprint estimate. DAWS goes on to further explore the programmability aspect of GPUs, demonstrating that with a more intelligent warp scheduler, a benchmark with no optimization on memory transfer (for example, using shared memory instead of cache) can come very close to matching a GPU-optimized version of the same benchmark.

Prefetching-Aware Warp Scheduling. Jog et al. [2013b] explore a prefetching-aware warp scheduler on the GPU. They base the scheduler on the two-level scheduling mechanism, but form fetch groups from non-consecutive warps. This policy increases the amount of bank-level parallelism in the DRAM since one DRAM bank will not be queried for contiguous accesses by the prefetcher. They further extend this idea to manipulate the prefetcher based on the warp-group assignment. By prefetching data for warps in other groups they can improve row-buffer locality and provide spacing between the prefetch request and the demand for data.

CTA-Aware Scheduling. Jog et al. [2013a] propose a CTA-aware warp scheduler that also builds on the two-level scheduler to form fetch groups based on the selectively combining CTAs. They exploit several CTA-based properties to improve performance. They employ a throttling prioritization technique that throttles the number of active warps in the core, similar to the other throttling schedulers. Combined with throttling, they exploit inter-CTA page locality among CTAs on different cores. Under a locality-aware only CTA scheduler, consecutive CTAs will often access the same DRAM bank at the same time, decreasing bank-level parallelism. They combine this with a prefetching mechanism to also improve DRAM row locality.

Impact of Scheduling on Branch Divergence Mitigation Techniques. Meng et al. [2010] introduce Dynamic Warp Subdivision (DWS) which splits warps when some lanes hit in cache and some lanes do not. This scheme allows individual scalar threads that hit in cache to make progress even if some of their warp peers miss. DWS improves performance by allowing run-

CCWS 通过将各种调度机制与缓存替换策略进行比较，进一步证明了缓存命中率对调度决策的敏感性。本文证明了 Warp 调度程序可用的决策空间比替换策略做出的相对受限的决策空间大得多。本文进一步证明了使用 LRU 替换策略的 CCWS 方案可以比以前的调度机制更好地提高缓存命中率，即使它们使用 Belady 最优缓存替换策略。

Rogers 等人 [2013] 提出了发散感知 Warp Scheduling (DAWS)，它扩展了 CCWS，可以更准确地估算每个 Warp 的缓存占用空间。DAWS 利用了 GPU 工作负载中大多数 Warp 内局部性发生在循环中的事实。DAWS 为循环中的 Warp 创建每个 Warp 的缓存占用空间估算。DAWS 根据每个 Warp 预测的循环占用空间，预先限制循环中的 Warp 数量。DAWS 还会根据每个 Warp 经历的控制流发散程度调整其缓存占用空间估算。离开循环的 Warp 中的线程不再对占用空间估算产生影响。DAWS 接着进一步探索 GPU 的可编程性方面，证明使用更智能的 Warp 调度程序，未对内存传输进行优化（例如，使用共享内存而不是缓存）的基准测试可以非常接近 GPU 优化版本的同一基准测试。

预取感知 Warp 调度。Jog 等人 [2013b] 探索了 GPU 上的预取感知 Warp 调度程序。他们基于两级调度机制设计了调度程序，但从非连续 Warp 中形成提取组。此策略增加了 DRAM 中的库级并行量，因为预取器不会查询一个 DRAM 库以进行连续访问。他们进一步扩展了这个想法，以根据 Warp 组分配来操纵预取器。通过预取其他组中 Warp 的数据，他们可以改善行缓冲区局部性，并在预取请求和数据需求之间提供间隔。

CTA 感知调度。Jog 等人 [2013a] 提出了一种 CTA 感知的 Warp 调度程序，该调度程序也基于两级调度程序，根据选择性组合的 CTA 形成提取组。他们利用了多种基于 CTA 的属性来提高性能。他们采用了一种节流优先级技术，该技术可以限制核心中的活动 Warp 数量，类似于其他节流调度程序。结合节流，他们利用不同核心上的 CTA 之间的 CTA 间页面局部性。在仅感知局部性的 CTA 调度程序下，连续的 CTA 通常会同时访问同一个 DRAM 库，从而降低库级并行性。他们将其与预取机制相结合，以改善 DRAM 行局部性。

调度对分支发散缓解技术的影响。Meng 等人 [2010] 引入了动态 Warp 细分 (DWS)，当某些通道命中缓存而某些通道未命中时，它会拆分 Warp。此方案允许命中缓存的单个标量线程即使其某些 Warp 同伴未命中也能取得进展。DWS 通过允许运行来提高性能

ahead threads to initiate their misses earlier and creates a pre-fetching effect for those left behind. DWS attempts to improve intra-warp locality by increasing the rate data is loaded into the cache.

Fung et al. [2007] explore the impact of warp scheduling policy on the effectiveness of their Dynamic Warp Formation (DWF) technique. DWF attempts to mitigate control flow divergence by dynamically creating new warps when scalar threads in the same warp take different paths on a branch instruction. They propose five schedulers and evaluate their effect on DWF.

Fung and Aamodt [2011] also propose three thread block prioritization mechanisms to compliment their Thread Block Compaction (TBC) technique. The prioritization mechanisms attempt to schedule threads within the same CTA together. Their approach is similar to concurrent work on two-level scheduling proposed by Narasiman et al. [2011], except thread blocks are scheduled together instead of fetch groups.

Section 3.4 contains more detailed summaries for DWS, DWF, and TBC.

Scheduling and Cache Re-Execution. Sethia et al. [2015] introduce Mascar which attempts to better overlap computation with memory accesses in memory intensive workloads. Mascar consists of two intertwined mechanisms.

- A memory aware warp scheduler (MAS) that prioritizes the execution of a single warp when MSHR and L1 miss queue entries in the core are oversubscribed. This prioritization helps improve performance even when workloads do not contain data locality by enabling warps executing on the in-order core to reach their computation operations faster, enabling overlap of the prioritized warp's computation with other warp's memory accesses.
- A cache access re-execution (CAR) mechanism that helps avoiding L1 data cache thrashing by enabling L1 data cache hits-under-misses when warps with data in the cache are blocked from issuing because low-locality accesses are stalling the memory pipeline.

MAS has two modes of operation: equal priority (EP) and memory access priority (MAP) mode. The system switches between EP and MP based on how full the L1 MSHRs and memory miss queue is. Once these structures are almost full, the system switches to MP mode. MAS contains two queues, one for memory warps (warps attempting to issue a memory instruction) and one for compute warps (warps attempting to issue other types of instructions). Within each queue, warps are scheduled in greedy-then-oldest order. The tracking of memory dependent instructions is done by augmenting the scoreboard to indicate when an output register is filled based on a load. The scheduler operates in EP mode when it is observed that the workload is balanced and the memory system is not oversubscribed. In EP mode, the scheduling mechanism prioritizes memory warps first. Since the memory system is not oversubscribed, it is predicted that initiating memory accesses early will improve performance. When operating in MAP mode, the scheduler prioritizes compute warps to better overlap available compute with the bottlenecked memory system. Only one of the memory warps, the "owner warp" is allowed to issue memory instructions until it reaches an operation that is dependent on a pending memory request.

提前线程启动未命中，并为落后线程创建预取效果。DWS 尝试通过提高数据加载到缓存中的速率来改善 warp 内部局部性。

Fung 等人 [2007] 探索了 Warp 调度策略对其动态 Warp 形成 (DWF) 技术效果的影响。当同一 Warp 中的标量线程在分支指令上采用不同路径时，DWF 尝试通过动态创建新 Warp 来缓解控制流发散。他们提出了五种调度程序并评估了它们对 DWF 的影响

。

Fung 和 Aamodt [2011] 还提出了三种线程块优先级机制来补充他们的线程块压缩 (TBC) 技术。优先级机制尝试将同一 CTA 内的线程一起调度。他们的方法类似于 Narasiman 等人 [2011] 提出的两级调度的并发工作，只是线程块是一起调度的，而不是按提取组进行调度。

第 3.4 节包含 DWS、DWF 和 TBC 的更详细摘要。

调度和缓存重新执行。Sethia 等人 [2015] 提出了 Mascar，旨在更好地将内存密集型工作负载中的计算与内存访问重叠。Mascar 由两个相互交织的机制组成。

- 内存感知 Warp 调度程序 (MAS)，当核心中的 MSHR 和 L1 未命中队列条目超额订阅时，它会优先执行单个 Warp。这种优先级有助于提高性能，即使工作负载不包含数据局部性，它也能使按顺序核心执行的 Warp 更快地完成计算操作，从而使优先 Warp 的计算与其他 Warp 的内存访问重叠。

- 缓存访问重新执行 (CAR) 机制。当由于低局部性访问阻碍了内存管道，导致缓存中带有数据的 warp 无法发出时，该机制可通过启用 L1 数据缓存命中 - 未命中来避免 L1 数据缓存抖动。

MAS 有两种操作模式：同等优先级 (EP) 和内存访问优先级 (MAP) 模式。系统根据 L1 MSHR 和内存未命中队列的满载程度在 EP 和 MP 之间切换。一旦这些结构几乎已满，系统就会切换到 MP 模式。MAS 包含两个队列，一个用于内存 Warp（试图发出内存指令的 Warp），另一个用于计算 Warp（试图发出其他类型指令的 Warp）。在每个队列中，Warp 都按照先贪婪后最旧的顺序进行调度。通过增强记分板来指示输出寄存器何时根据负载填满，可以跟踪内存相关指令。当观察到工作负载平衡且内存系统未超额认购时，调度程序将在 EP 模式下运行。在 EP 模式下，调度机制首先优先考虑内存 Warp。由于内存系统未超额认购，因此可以预测尽早启动内存访问将提高性能。在 MAP 模式下运行时，调度程序会优先考虑计算 Warp，以便更好地将可用计算与瓶颈内存系统重叠。只有一个内存 Warp，即 “ 所有者 Warp ”，被允许发出内存指令，直到它到达依赖于待处理内存请求的操作。

In addition to the scheduling mechanism, Sethia et al. [2015] show that memory intensive kernels perform at a much lower fraction of their peak IPC than compute intensive kernels. They illustrate that in the memory intensive applications, a large fraction of cycles are spent with the SIMT core's load store unit stalled due to memory back-pressure caused from excessive memory accesses. While the LSU is stalled, there is a significant fraction of time where data for ready warps is in the L1 data cache, but the warps are unable to issue because the LSU is backed up with memory requests other warps. The cache access re-execution (CAR) mechanism seeks to remedy this behavior by providing a buffer on the side of the LSU pipeline which stores stalled memory instructions and allows others issue into the LSU. Requests are processed from the re-execution queue only if the LSU is not stalled and has no new requests to issue unless the re-execution queue is full, in which case accesses in the re-execution queue are prioritized until space frees up in the queue.

When the re-execution queue is combined with the memory aware scheduler, special care needs to be taken since requests in the re-execution queue can come from warps other than the prioritized owner warp. When operating in MAP mode, non-owner warp's requests sent from the re-execution queue to the L1 are further delayed when they do not hit in the L1. In particular, when a request from a non-owner warp misses in the L1, the request is not relayed to the L2 cache, but rather is reinserted into the tail of the re-execution queue.

5.1.3 RESEARCH ON SCHEDULING MULTIPLE KERNELS

Supporting Pre-emption on GPUs. Park et al. [2015] tackles the challenge of supporting preemptive multitasking on GPUs. It employs a more relaxed definition of idempotence to enable flushing of computation within a threadblock. The more relaxed definition of impotence involves detecting whether execution has been idempotent from the start of thread execution. Their proposal, Chimera, dynamically selects among three approaches to achieving a context switch for each threadblock:

- a full context save/store;
- waiting until the threadblock finishes; and
- simply stopping the threadblock without saving any context if, due to idempotence, the thread block can be restarted from the beginning safely.

Each context-switching technique provides a different tradeoff between the latency to switch and the impact on system throughput. To implement Chimera an algorithm estimates the subset of threadblocks that are currently running which can be stopped with minimum impact on system throughput while meeting a user specified context switch latency target.

除了调度机制之外，Sethia 等人 [2015] 还表明，内存密集型内核的峰值 IPC 执行率远低于计算密集型内核。他们指出，在内存密集型应用中，由于过多的内存访问导致内存背压，SIMT 核心的加载存储单元会停顿，从而占用大量周期。当 LSU 停顿时，有相当一部分时间，就绪 warp 的数据位于 L1 数据缓存中，但 warp 无法发出，因为 LSU 备份了其他 warp 的内存请求。缓存访问重新执行 (CAR) 机制试图通过在 LSU 管道侧提供缓冲区来纠正此行为，该缓冲区存储停顿的内存指令并允许其他指令进入 LSU。仅当 LSU 没有停滞并且没有新的请求要发出时，才会处理来自重新执行队列的请求，除非重新执行队列已满，在这种情况下，重新执行队列中的访问将被优先处理，直到队列中的空间释放为止。

当重新执行队列与内存感知调度程序结合使用时，需要特别小心，因为重新执行队列中的请求可能来自优先级所有者 warp 以外的 warp。在 MAP 模式下操作时，从重新执行队列发送到 L1 的非所有者 warp 的请求如果未在 L1 中命中，则会进一步延迟。具体而言，当来自非所有者 warp 的请求在 L1 中未命中时，该请求不会中继到 L2 缓存，而是重新插入到重新执行队列的尾部。

5.1.3 多核调度研究

支持 GPU 上的抢占。Park 等人 [2015] 解决了支持 GPU 上的抢占式多任务处理的挑战。它采用了更宽松的幂等性定义，以支持线程块内的计算刷新。更宽松的幂等性定义涉及从线程执行开始时检测执行是否是幂等的。他们的提案 Chimera 在三种方法中动态选择以实现每个线程块的上下文切换：

- 完整的上下文保存/存储；
- 等待线程块完成；并且
- 如果由于幂等性，线程块可以安全地从头开始重新启动，则只需停止线程块而不保存任何上下文。

每种上下文切换技术在切换延迟和对系统吞吐量的影响之间提供了不同的权衡。为了实现 Chimera，一种算法会估算当前正在运行的线程块子集，这些线程块可以在对系统吞吐量影响最小的情况下停止，同时满足用户指定的上下文切换延迟目标。

5.1.4 FINE-GRAIN SYNCHRONIZATION AWARE SCHEDULING

ElTantawy and Aamodt [2018] explore the impact of warp scheduling when running codes that involve fine-grained synchronization. Using real GPU hardware they demonstrate that a significant overheads occur when threads spin waiting for locks. They point out that naively backing off execution of warps containing threads that have failed to acquire a lock can prevent or slow progress of other threads in the same warp already holding a lock. They propose a hardware structure for dynamically identifying which loops are involved in spin locks, which is made more challenging by the use of stack-based reconvergence [ElTantawy and Aamodt, 2016]. This structure uses a path history containing least significant bits of the program counter and a separate history of predicate register updates to accurately detect loops spinning on a lock. To reduce contention and improve performance they propose reducing the priority of warps identified as executing a spin loop when they execute the backwards branch of a spin loop after any threads in the warp that held locks have released those locks. They find this improves performance and reduces energy by 1.5× and 1.6×, respectively, vs. Lee and Wu [2014].

5.2 ALTERNATIVE WAYS OF EXPRESSING PARALLELISM

Fine-Grained Work Queues. Kim and Batten [2014] propose the addition of a fine-grain hardware worklist to each SIMT core in the GPU. They exploit the observation that irregular GPGPU programs often perform best when implemented in software using a data-driven approach, where work is dynamically generated and balanced between threads, as opposed to a topological approach, where a fixed number of threads are launched—many of these threads do no useful work. The data-driven approach has the potential to increase work efficiency and load balancing but may suffer from poor performance without extensive software optimizations. This paper proposes an on-chip hardware worklist with support for load balancing both within the core and among cores. They use a thread waiting mechanism and rebalance tasks generated by threads on an interval basis. They evaluate their hardware mechanism on various implementations of the irregular applications in the lonestar GPU benchmark suite that make use of both topological and data-driven work distribution.

The on-core hardware worklist addresses two major problems with data-driven software worklists: (1) contention in the memory system when threads push out generated work and (2) poor load balancing as a result of statically partitioning work based on thread ID. Software implementations that do not rely on static partitioning suffer from memory contention on both pushes and pulls. Statically partitioning the work solves the pulling contention. The hardware worklist is distributed across multiple structures, reducing contention. It improves load balancing by dynamically redistributing generated work to threads before they become idle. The authors add special instructions to the ISA for pushing and pulling from the hardware queues. Each lane in the core is assigned a small, single ported SRAM that is used as the storage for work IDs used and generated by a given lane.

ElTantawy 和 Aamodt [2018] 探讨了在运行涉及细粒度同步的代码时，warp 调度的影响。他们使用真实的 GPU 硬件证明了当线程旋转等待锁时会产生大量开销。他们指出，如果单纯地退出包含未能获取锁的线程的 warp 的执行，可能会阻止或减慢同一 warp 中已持有锁的其他线程的进度。他们提出了一种硬件结构，用于动态识别哪些循环涉及自旋锁，而使用基于堆栈的重新收敛使这一过程更具挑战性 [ElTantawy 和 Aamodt，2016]。该结构使用包含程序计数器最低有效位的路径历史记录和单独的谓词寄存器更新历史记录来准确检测在锁上旋转的循环。为了减少争用并提高性能，他们建议在执行自旋循环的反向分支时降低被识别为执行自旋循环的线程束的优先级，因为在线程束中持有锁的任何线程都释放了这些锁之后，线程束才会执行自旋循环。他们发现，与 Lee 和 Wu [2014] 相比，这分别提高了性能并降低了 1.5 个 v0 和 1.6 个 v1 的能耗。

5.2 表达并行性的替代方法

细粒度工作队列。Kim 和 Batten [2014] 建议在 GPU 中的每个 SIMT 核心中添加一个细粒度硬件工作列表。他们利用了以下观察结果：不规则的 GPGPU 程序通常在使用数据驱动方法在软件中实现时表现最佳，其中工作是动态生成并在线程之间平衡的，而不是使用拓扑方法，其中启动固定数量的线程 — 其中许多线程不做有用的工作。数据驱动方法有可能提高工作效率和负载平衡，但如果没有广泛的软件优化，可能会出现性能不佳的情况。本文提出了一种片上硬件工作列表，支持核心内和核心之间的负载平衡。他们使用线程等待机制并按间隔重新平衡线程生成的任务。他们在 lonestar GPU 基准测试套件中利用拓扑和数据驱动的工作分布的各种不规则应用程序实现上评估了他们的硬件机制。

内核硬件工作列表解决了数据驱动软件工作列表的两个主要问题：(1) 线程推送生成的工作时内存系统中的争用和 (2) 基于线程 ID 静态划分工作导致的负载平衡不佳。不依赖静态划分的软件实现在推送和拉取时都会遭遇内存争用。静态划分工作解决了拉取争用问题。硬件工作列表分布在多个结构中，从而减少了争用。它通过在线程空闲之前动态地将生成的工作重新分配给线程来改善负载平衡。作者向 ISA 添加了特殊指令，用于从硬件队列推送和拉取。内核中的每个通道都分配有一个小型单端口 SRAM，用作给定通道使用和生成的工作 ID 的存储。

The paper proposes both an interval based and demand-driven (only redistributed on push/pull requests) work redistribution method and evaluates the former in depth. The interval based method redistributes the work on either a simple threshold basis or a more complex sorting basis. The threshold method classifies lanes with more work than the threshold as greedy (having too much work) and those with less work than the threshold as needy (not having enough work). A sorting pass then redistributes work from the greedy banks to the needy banks. The sorting-based technique is more complex, but achieves better load balancing since all needy banks can also donate work to other needy banks. Their technique also includes a global sorting mechanism that can be used to distribute work between cores. Additionally, the architecture provides support for virtualizing the hardware worklist, making it extendible to workloads that generate more dynamic work than the capacity available in the hardware structure.

Nested Parallel-pattern Based Programming. Lee et al. [2014a] propose a Locality-Aware Mapping of Nested Parallel Patterns on GPUs which capitalized on the observation that there is not a universally optimal mapping of nested parallel computation to GPU threads. Algorithms with nested parallelism (such as a map/reduce operation) can have their parallelism exposed to the GPU at different levels, depending on how the GPU program is written. The authors exploit three generalizations of nested parallel mapping on GPUs:

- a 1D mapping, which parallelizes the outer loop of a sequential program;
- a thread block/thread mapping, which assigns each iteration of the sequential program's outer loop to a thread block and parallelizes the inner pattern across a thread block; and
- a warp-based mapping, which assigns each iteration of the outer loop to a warp and parallelizes the inner pattern across the warp.

This work proposes an automatic compilation framework which generates predicted performance scores based on locality and the degree of parallelism exposed in nested patterns to choose which mapping is best suited for a set of common nested parallel patterns. These patterns consist of collection operations such as map, reduce, foreach, filter, etc. The framework attempts to map threads to operations on each element of the collection. The framework handles the nesting of patterns by first assigning each nest level in the application to a dimension (x, y, z, etc.). A doubly nested pattern (i.e., a map containing a reduce) has two dimensions. The mapping then determines the number of threads in a given dimension in a CUDA thread block. After setting the dimensions and size of the thread block, the framework further controls the degree of parallelism in the kernel by assigning multiple elements to each thread using the notion of thread spanning and splitting. In a two-dimensional kernel (i.e., two levels of pattern nesting), if each dimension is assigned span(1), then each thread launched in the kernel is responsible for operating on only one element of the collection. This mapping exposes the greatest degree of parallelism. In contrast, span(all) indicates that each thread operates on all the elements in the collection. The span can be any number between (1) and (all). Span(all) is used

本文提出了一种基于间隔和需求驱动（仅根据推送/拉取请求进行重新分配）的工作重新分配方法，并对前者进行了深入评估。基于间隔的方法根据简单的阈值或更复杂的排序重新分配工作。阈值方法将工作量超过阈值的通道归类为贪婪（工作量过多），将工作量少于阈值的通道归类为贫困（工作量不足）。然后，排序过程将工作从贪婪的银行重新分配给贫困的银行。基于排序的技术更复杂，但可以实现更好的负载平衡，因为所有贫困的银行也可以将工作捐赠给其他贫困的银行。他们的技术还包括一个全局排序机制，可用于在核心之间分配工作。此外，该架构还支持虚拟化硬件工作列表，使其可扩展到生成比硬件结构中可用容量更多的动态工作的工作负载。

基于嵌套并行模式的编程。Lee 等人 [2014a] 提出了一种 GPU 上嵌套并行模式的局部感知映射，该映射利用了以下观察结果：嵌套并行计算到 GPU 线程没有普遍最优的映射。具有嵌套并行性的算法（例如 map/reduce 操作）可以在不同级别将其并行性暴露给 GPU，具体取决于 GPU 程序的编写方式。作者利用了 GPU 上嵌套并行映射的三种概括：

- 一维映射，将顺序程序的外循环并行化；
- 线程块/线程映射，将顺序程序外循环的每次迭代分配给一个线程块，并将内层模式在线程块中并行化；
- 基于 warp 的映射，将外循环的每次迭代分配给一个 warp，并在整个 warp 中并行化内层模式。

本研究提出了一个自动编译框架，该框架根据嵌套模式中公开的局部性和并行度生成预测性能分数，以选择最适合一组常见嵌套并行模式的映射。这些模式由集合操作组成，例如 map、reduce、foreach、filter 等。该框架尝试将线程映射到集合中每个元素上的操作。该框架通过首先将应用程序中的每个嵌套级别分配给一个维度（x、y、z 等）来处理模式的嵌套。双重嵌套模式（即包含 Reduce 的 Map）有两个维度。然后，映射确定 CUDA 线程块中给定维度的线程数。在设置线程块的维度和大小之后，框架通过使用线程跨越和拆分的概念为每个线程分配多个元素，进一步控制内核中的并行度。在二维内核（即两层模式嵌套）中，如果每个维度都分配有 span(1)，则内核中启动的每个线程只负责操作集合中的一个元素。此映射可实现最大程度的并行性。相反，span(all) 表示每个线程都操作集合中的所有元素。span 可以是 (1) 和 (all) 之间的任何数字。Span(all) 用于

in two special circumstances: when the size of a dimension is not known until after the kernel is launched (for example when the number of elements operated on in an inner pattern is determined dynamically) and when the pattern requires synchronization (for example, the reduce operation).

Since span(all) can severely limit the exposed parallelism and cause the GPU to be underutilized, the framework also provides the notion of a *split*. The split(2) indicates that each thread operates on half of the elements in a given dimension (think of it as span(all)/2). When split is used, the framework launches a second kernel (called a combiner kernel) to aggregate the results across the splits, producing the same result as if the kernel had been partitioned using span(all).

To select both the block size in each dimension and the split/span of each dimension, the framework uses a scoring algorithm based on hard and soft constraints. The algorithm sweeps the entire entire search space all possible dimensions, block sizes and spans. The search space is exponential to the level of loop nests. However, the base of the exponential is less than 100 and a typical kernel contains less than 3 levels. Thus, the space is completely searchable in just a few seconds. The search prunes configurations that violate hard constraints—i.e., those that cause incorrect execution, such as the maximum number of threads in a block being too high. It assigns weighted scores to soft constraints, such as ensuring that patterns sequential memory accesses are assigned to the x dimension to improve memory coalescing.

The framework also performs two common GPU optimizations: pre-allocating memory instead of dynamically allocating global memory in nested kernels and making use of shared memory when it determines that prefetching data into shared memory is beneficial for a nested pattern. Results demonstrate that the automatically generated code has competitive performance with expertly tuned code.

Dynamic Parallelism. Wang and Yalamanchili [2014] characterize the overheads of using CUDA Dynamic Parallelism on Kepler GPU hardware and find these overheads can be substantial. Specifically, they identify several key issues limiting efficiency in the workloads they studied. First, the applications employed a very large number of device launched kernels. Second, each kernel typically had only 40 threads (little more than one warp). Third, while the code executed in each dynamic kernel is similar, the launch configuration differs resulting in a large amount of storage overhead for kernel configuration information. Finally, fourth, to enable concurrency the device launched kernels are placed in separate streams to exploit the 32 parallel hardware queues (Hyper-Q) supported on Kepler. They find these factors combine to yield very poor utilization.

Wang et al. [2016a] subsequently propose dynamic thread block launch (DTBL), which modifies the CUDA programming model to enable device launched kernels to share hardware queue resources enabling greater parallelism and better utilization of the GPU hardware. A key to their proposal is enabling dynamically launched kernels to be aggregated together with an existing kernel running the same code. This is supported by maintaining a linked list of aggregated thread blocks that are used by the modified hardware when launching kernels. They

在两种特殊情况下：当维度的大小直到内核启动后才知道（例如，当内部模式中操作的元素数量是动态确定的）以及当模式需要同步时（例如，reduce 操作）。

由于 span(all) 会严重限制暴露的并行性并导致 GPU 利用率不足，因此框架还提供了 *split* 的概念。split(2) 表示每个线程对给定维度中的一半元素进行操作（将其视为 span(all)/2）。使用 split 时，框架会启动第二个内核（称为组合器内核）来聚合拆分的结果，产生的结果与使用 span(all) 对内核进行分区时的结果相同。

为了选择每个维度的块大小和每个维度的拆分/跨度，该框架使用基于硬约束和软约束的评分算法。该算法会扫描整个搜索空间的所有可能维度、块大小和跨度。搜索空间随循环嵌套的级别呈指数增长。但是，指数的底数小于 100，典型的内核包含的级别少于 3。因此，只需几秒钟即可完全搜索该空间。搜索会修剪违反硬约束的配置 - 即导致执行不正确的配置，例如块中的最大线程数过高。它为软约束分配加权分数，例如确保将模式顺序内存访问分配给 x 维度以改善内存合并。

该框架还执行两项常见的 GPU 优化：在嵌套内核中预分配内存而不是动态分配全局内存，以及在确定将数据预取到共享内存对嵌套模式有益时使用共享内存。结果表明，自动生成的代码具有与专家调整的代码相媲美的性能。

动态并行。Wang 和 Yalamanchili [2014] 描述了在 Kepler GPU 硬件上使用 CUDA 动态并行的开销，并发现这些开销可能非常大。具体而言，他们确定了限制他们所研究的工作负载效率的几个关键问题。首先，应用程序使用了非常大量的设备启动内核。其次，每个内核通常只有 40 个线程（略多于一个 warp）。第三，虽然每个动态内核中执行的代码相似，但启动配置不同，导致内核配置信息的存储开销很大。最后，第四，为了实现并发，设备启动的内核被放置在单独的流中，以利用 Kepler 上支持的 32 个并行硬件队列 (Hyper-Q)。他们发现这些因素结合在一起导致利用率非常低。

Wang 等人 [2016a] 随后提出了动态线程块启动 (DTBL)，该模型修改了 CUDA 编程模型，使设备启动的内核能够共享硬件队列资源，从而实现更高的并行性和更好地利用 GPU 硬件。他们的提案的关键是使动态启动的内核能够与运行相同代码的现有内核聚合在一起。这通过维护修改后的硬件在启动内核时使用的聚合线程块的链接列表来支持。他们

evaluate DTBL by modifying GPGPU-Sim and find DTBL improves performance by 1.4× versus CDP and 1.2× over highly tuned CUDA versions that do not employ CDP.

Wang et al. [2016b] then explore the impact of which SM dynamically launched thread blocks are scheduled to. They find that by encouraging child thread blocks to be scheduled on the same SM as parent SMs while considering workload distribution across SMs they were able to improve performance by 27% vs. a naive round-robin distribution mechanism.

5.3 SUPPORT FOR TRANSACTIONAL MEMORY

This section summarizes the various proposals to support a transactional memory (TM) [Harris et al., 2010, Herlihy and Moss, 1993] programming model on GPU architectures.

These proposals were motivated by the potential of a TM programming model to ease the challenge of managing irregular, fine-grained communications between threads in GPU applications with ample irregular parallelism. On modern GPUs, the application developers may either coarsen the synchronization between threads via barriers, or they may attempt to use single-word atomic operations, available on many modern GPUs, to implement fine-grained locks for these communications. The former approach may involve significant changes to the underlying algorithms, while the latter approach involves the uncertainty in development effort with fine-grained locking, too risky for practical, market-driven software development (with several exceptions). Enabling TM on GPUs simplifies synchronization, and provides a powerful programming model that promotes fine-grained communication and strong scaling of parallel workloads. This promise from TM hopes to encourage software developers to explore GPU acceleration with these irregular applications.

Distinct Challenges for Supporting TM on GPUs. The heavily multithreaded nature of GPU introduces a new set of challenges to TM system designs. Instead of running tens of concurrent transactions with relatively large footprint—the focus of much recent research on TM for multicore processors—TM system on a GPU aims to scale to tens of thousands of small concurrent transactions. This reflects the heavily multithreaded nature of GPU, with tens of thousands of threads working in collaboration, each performing a small task towards a common goal. These small transactions are tracked at word-level granularity, enabling finer resolution of conflict detection than cache blocks. Moreover, each per-core private cache in a GPU is shared by hundreds of GPU threads. This drastically reduces the benefit of leveraging a cache coherence protocol to detect conflicts, a technique employed on most hardware transactional memory designed for traditional CMPs with large CPU cores.

5.3.1 KILO TM

Kilo TM [Fung et al., 2011] is the first published hardware TM proposal for GPU architectures.

Kilo TM employs value-based conflict detection [Dalessandro et al., 2010, Olszewski et al., 2007] to eliminate the need for global metadata for conflict detection. Each transaction

通过修改 GPGPU-Sim 来评估 DTBL，发现 DTBL 的性能相对于 CDP 提高了 1.4×，相对于不使用 CDP 的高度调整的 CUDA 版本提高了 1.2×。

Wang 等人 [2016b] 随后探索了动态启动线程块被调度到哪个 SM 的影响。他们发现，通过鼓励将子线程块调度到与父 SM 相同的 SM 上，同时考虑跨 SM 的工作负载分配，他们能够将性能提高 27%，而使用简单的循环分配机制则不行。

5.3 对事务内存的支持

本节总结了在 GPU 架构上支持事务内存 (TM) [Harris et al., 2010, Herlihy and Moss, 1993] 编程模型的各种提案。

这些提议的动机是 TM 编程模型的潜力，它可以减轻管理具有充足不规则并行性的 GPU 应用程序中线程之间不规则、细粒度通信的挑战。在现代 GPU 上，应用程序开发人员可以通过屏障来粗化线程之间的同步，或者他们可以尝试使用许多现代 GPU 上提供的单字原子操作来实现这些通信的细粒度锁定。前一种方法可能涉及对底层算法的重大更改，而后一种方法涉及细粒度锁定的开发工作的不确定性，对于实际的、市场驱动的软件开发来说风险太大（除了几个例外）。在 GPU 上启用 TM 简化了同步，并提供了一个强大的编程模型，该模型促进了细粒度通信和并行工作负载的强大扩展。TM 的这一承诺希望鼓励软件开发人员探索这些不规则应用程序的 GPU 加速。

在 GPU 上支持 TM 的独特挑战。GPU 的高度多线程特性为 TM 系统设计带来了一系列新挑战。GPU 上的 TM 系统旨在扩展到数万个小型并发事务，而不是运行数十个占用空间相对较大的并发事务（这是最近多核处理器 TM 研究的重点）。这反映了 GPU 的高度多线程特性，数万个线程协同工作，每个线程执行一项小任务以实现共同目标。这些小事务以字级粒度进行跟踪，从而实现比缓存块更精细的冲突检测分辨率。此外，GPU 中的每个每核私有缓存都由数百个 GPU 线程共享。这大大降低了利用缓存一致性协议检测冲突的好处，这种技术在大多数为具有大 CPU 核心的传统 CMP 设计的硬件事务内存中都采用。

5.3.1 千克

Kilo TM [Fung et al.，2011] 是第一个发布的针对 GPU 架构的硬件 TM 提案。

Kilo TM 采用基于价值的冲突检测 [Dalessandro et al., 2010, Olszewski et al., 2007]，从而消除了冲突检测对全局元数据的需求。每笔交易

simply reads the existing data in global memory for validation—to determine if it has a conflict with another committed transaction. This form of validation leverages the highly parallel nature of the GPU memory subsystem, avoids any direct interaction between conflicting transactions, and detects conflicts at the finest granularity.

However, a native implementation of value-based conflict detection requires transactions to commit serially. To boost commit parallelism, Kilo TM incorporates ideas from existing TM systems [Chafi et al., 2007, Spear et al., 2008] and extended them with innovative solutions. In particular, Fung et al. [2011] introduced the *recency bloom filter*, a novel data structure that uses the notion of time and order to compress a large number of small item sets. Kilo TM uses this structure to compress the write-sets of all committing transactions. Each committing transaction queries the recency bloom filter for an approximate set of conflicting transactions—some transactions in this conflicting set are false positives. Kilo TM uses this approximate information to schedule hundreds of non-conflicting transactions for validation and commit in parallel. This approximate nature of recency bloom filter allows it to remain small, in the order of several kBs, and thus it can reside on-chip for fast access. Using the recency bloom filter to boost transaction commit parallelism is an integral part of Kilo TM.

Branch Divergence and Transactional Memory. The transactional memory programming model introduces a new type of branch divergence. When a warp finishes a transaction, each of its active threads will try to commit. Some of the threads may abort and need to reexecute their transactions, while other threads may pass the validation and commit their transactions. Since this outcome may not be unanimous across the entire warp, a warp may diverge after validation. Fung et al. [2011] proposes a simple extension to the SIMT hardware to handle this specific kind of branch divergence introduced by transaction aborts. This extension is independent of other design aspects of Kilo TM, but it is a necessary piece for supporting TM on GPUs.

Figure 5.2 shows how the SIMT stack can be extended to handle control flow divergence due to transaction aborts. When a warp enters the transaction (at line B, `tx_begin`), it pushes two special entries onto the SIMT stack ❶. The first entry of type R stores information to restart the transaction. Its active mask is initially empty, and its PC field points to the instruction after `tx_begin`. The second entry of type T tracks the current transaction attempt. At `tx_commit` (line F), any thread that fails validation sets its mask bit in the R entry. The T entry is popped when the warp finishes the commit process (i.e., its active threads have either committed or aborted) ❷. A new T entry will then be pushed onto the stack using the active mask and PC from the R entry to restart the threads that have been aborted. Then, the active mask in the R entry is cleared ❸. If the active mask in the R entry is empty, both T and R entries are popped, revealing the original N entry ❺. Its PC is then modified to point to the instruction right after `tx_commit`, and the warp resumes normal execution. Branch divergence of a warp within a transaction is handled in the same way as non-transactional divergence ❹.

只需读取全局内存中的现有数据进行验证，以确定它是否与另一个已提交的事务发生冲突。这种验证形式利用了 GPU 内存子系统的高度并行特性，避免了冲突事务之间的任何直接交互，并以最精细的粒度检测冲突。

但是，基于价值的冲突检测的本机实现要求事务串行提交。为了提高提交并行性，Kilo TM 吸收了现有 TM 系统 [Chafi 等，2007；Spear 等，2008] 中的思想，并通过创新解决方案对其进行了扩展。特别是，Fung 等人 [2011] 引入了 *recency bloom filter*，这是一种新颖的数据结构，它使用时间和顺序的概念来压缩大量小项目集。Kilo TM 使用这种结构来压缩所有提交事务的写入集。每个提交事务都会向新近布隆过滤器查询一组近似的冲突事务——该冲突集中的一些事务是误报。Kilo TM 使用这些近似信息来安排数百个非冲突事务进行验证并并行提交。新近布隆过滤器的这种近似性质使其能够保持较小，大约为几 kB，因此它可以驻留在片上以便快速访问。使用新近布隆过滤器来提高事务提交并行性是 Kilo TM 的一个组成部分。

分支发散和事务内存。事务内存编程模型引入了一种新型的分支发散。当一个 warp 完成事务时，其每个活动线程都将尝试提交。一些线程可能会中止并需要重新执行其事务，而其他线程可能会通过验证并提交其事务。由于这个结果在整个 warp 中可能不一致，因此 warp 可能会在验证后发散。Fung 等人 [2011] 提出了一种对 SIMT 硬件的简单扩展，以处理由事务中止引入的这种特定类型的分支发散。此扩展与 Kilo TM 的其他设计方面无关，但它是支持 GPU 上的 TM 的必要部分。

图 5.2 显示了如何扩展 SIMT 堆栈以处理由于事务中止而导致的控制流分歧。当 warp 进入事务时（在行 B，`tx_begin`），它会将两个特殊条目推送到 SIMT 堆栈 1 。类型 R 的第一个条目存储重新启动事务的信息。其活动掩码最初为空，其 PC 字段指向 `tx_begin` 之后的指令。类型 T 的第二个条目跟踪当前事务尝试。在 `tx_commit` (行 F)，任何未通过验证的线程都会在 R 条目中设置其掩码位。当 warp 完成提交过程（即，其活动线程已提交或中止）2 时，会弹出 T 条目。然后，将使用来自 R 条目的活动掩码和 PC 将新的 T 条目推送到堆栈上，以重新启动已中止的线程。然后，清除 R 条目中的活动掩码 3。如果 R 条目中的活动掩码为空，则 T 和 R 条目都会弹出，显示原始的 N 条目 5 ，然后修改其 PC 以指向紧接在 `tx_commit` 之后的指令，并且 warp 恢复正常执行。事务内 warp 的分支发散的处理方式与非事务发散 4 相同。

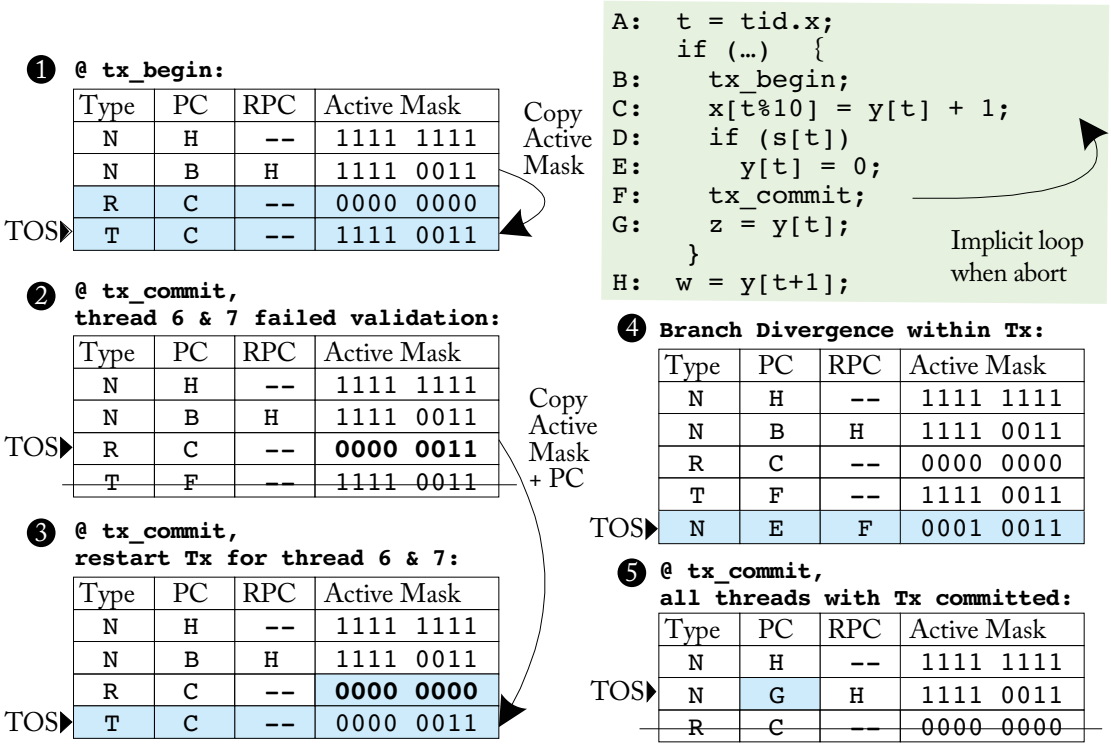


Figure 5.2: SIMT stack extension to handle divergence due to transaction aborts (validation fail). Thread 6 and 7 have failed validation and are restarted. Stack entry type: Normal (N), Transaction Retry (R), Transaction Top (T). For each scenario, added entries or modified fields are shaded.

5.3.2 WARP TM AND TEMPORAL CONFLICT DETECTION

In their follow-up paper, Fung and Aamodt [2013] propose two distinct enhancements that improves the performance and energy-efficiency of Kilo TM: warp-level transaction management (WarpTM) and temporal conflict detection (TCD).

Warp-level transaction management leverages the thread hierarchy in GPU programming models—the spatial locality among threads within a warp—to improve the efficiency of Kilo TM. In particular, WarpTM amortizes the control overhead of Kilo TM and boosts the utility of the GPU memory subsystem. These optimizations are only possible if conflicts within a warp can be resolved efficiently, and thus a low overhead intra-warp conflict resolution mechanism is crucial in maintaining the benefit from WarpTM. To this end, Fung and Aamodt [2013] propose a two-phase parallel intra-warp conflict resolution that resolves conflicts within a warp efficiently in parallel. With all intra-warp conflicts resolved, Kilo TM can merge the scalar mem-

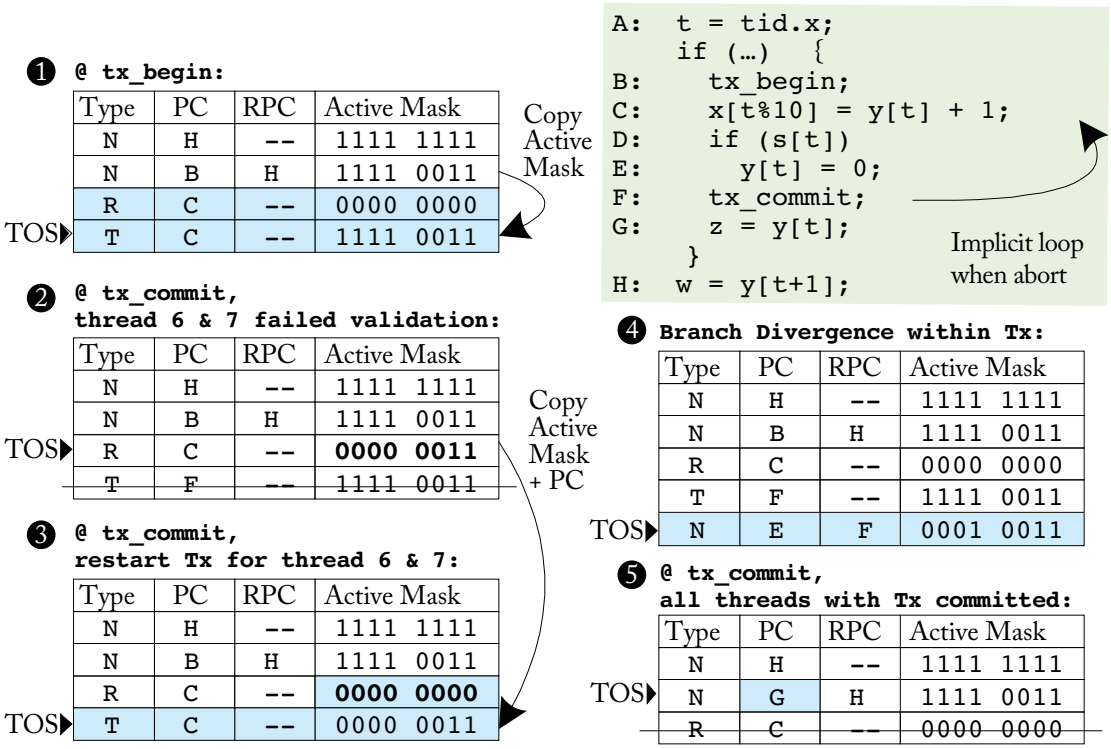


图 5.2：SIMT 堆栈扩展用于处理由于事务中止（验证失败）而导致的分歧。线程 6 和 7 验证失败并重新启动。堆栈条目类型：正常 (N)、事务重试 (R)、事务顶部 (T)。对于每个场景，添加的条目或修改的字段都用阴影表示。

5.3.2 WARP TM 和时间冲突检测

在他们的后续论文中，Fung 和 Aamodt [2013] 提出了两项不同的增强技术来提高 Kilo TM 的性能和能源效率：warp 级事务管理 (WarpTM) 和时间冲突检测 (TCD)。

Warp 级事务管理利用 GPU 编程模型中的线程层次结构（即 Warp 内线程之间的空间局部性）来提高 Kilo TM 的效率。具体而言，WarpTM 可以摊销 Kilo TM 的控制开销，并提高 GPU 内存子系统的效用。这些优化只有在可以有效解决 Warp 内的冲突时才有可能实现，因此低开销的 Warp 内冲突解决机制对于保持 WarpTM 的优势至关重要。为此，Fung 和 Aamodt [2013] 提出了一种两阶段并行的 Warp 内冲突解决机制，可以高效地并行解决 Warp 内的冲突。在解决所有 Warp 内冲突后，Kilo TM 可以合并标量内存

ory accesses for the validation and commit of multiple transactions in the same warp into wider accesses. This optimization, called *validation and commit coalescing*, is key to enable Kilo TM to take full advantage of the wide GPU memory system, which is optimized for vector-wide accesses.

Temporal conflict detection is a low overhead mechanism that uses a set of globally synchronized on-chip timers to detect conflicts for read-only transactions. Once initialized, each of these on-chip timers runs locally in its microarchitecture module and does not communicate with other timers. This implicit synchronization without communication distinguishes TCD from existing timestamp-based conflict detections used in various software TM systems [Dalessandro et al., 2010, Spear et al., 2006, Xu et al., 2014]. TCD uses timestamps captured from these timers to infer the order of the memory reads of a transaction with respect to updates from other transactions. Kilo TM incorporates TCD to detect conflict-free read-only transactions that can commit directly without value-based conflict detection. In doing so, it significantly reduces the memory bandwidth overhead for these transactions, which can occur frequently for GPU-TM applications that use transactions for data structure traversals.

5.4 HETEROGENEOUS SYSTEMS

Concurrency Management in Heterogeneous Systems. Kayiran et al. [2014] propose a concurrency-limiting scheme to throttle GPU multithreading, reducing memory and network contention in multiprogrammed CPU/GPU systems. On heterogenous system, interference from the GPU can result in a significant performance drop for concurrently executing CPU applications. Their proposed thread-level-paralleism (TLP) limiting scheme observes congestion metrics in the shared CPU/GPU memory controller and interconnection network to estimate the number of GPU warps that should be actively scheduled on each GPU core. They propose two schemes, one that focuses on increasing CPU performance only and another that seeks to optimize overall system throughput (both CPU and GPU) by balancing GPU performance degradation due to constrained multithreading with CPU interference. The authors evaluate the performance impact of warp scheduling on a tiled heterogeneous architecture with a GPU core to CPU core ratio of 2:1, justified by an NVIDIA GPU SM being roughly half the area of a modern out-of-order Intel chip using the same process technology. The baseline configuration completely shares both the network bandwidth and memory controllers between the CPU and GPU, in the interest of maximizing resource utilization. Using this scheme, the authors make the observation that limiting GPU TLP can have both a positive and negative effect on GPU performance, but never hurts CPU performance.

To boost CPU performance, the authors introduce a CPU-centric concurrency management technique that monitors stalling in the global memory controllers. This technique separately counts the number of memory requests stalled due to the memory controller input queue being full, and the number of memory requests stalled because the reply network from the MC to the cores is full. These metrics are monitored locally at each memory controller and are ag-

验证和提交同一 warp 中的多个事务的复杂度访问被简化为更宽的访问。这种优化称为 *validation and commit coalescing*，是 Kilo TM 充分利用宽 GPU 内存系统的关键，该系统针对矢量范围访问进行了优化。

时间冲突检测是一种低开销机制，它使用一组全局同步的片上计时器来检测只读事务的冲突。初始化后，每个片上计时器都在其微架构模块中本地运行，不与其他计时器通信。这种无需通信的隐式同步将 TCD 与各种软件 TM 系统中使用的现有基于时间戳的冲突检测区分开来 [Dalessandro 等，2010；Spear 等，2006；Xu 等，2014]。TCD 使用从这些计时器捕获的时间戳来推断事务相对于其他事务更新的内存读取顺序。Kilo TM 结合 TCD 来检测无冲突的只读事务，这些事务无需基于值的冲突检测即可直接提交。这样做可以显著减少这些事务的内存带宽开销，这对于使用事务进行数据结构遍历的 GPU-TM 应用程序来说可能经常发生。

5.4 异构系统

异构系统中的并发管理。Kayiran 等人 [2014] 提出了一种并发限制方案来限制 GPU 多线程，从而减少多程序 CPU/GPU 系统中的内存和网络争用。在异构系统中，来自 GPU 的干扰可能会导致并发执行的 CPU 应用程序的性能显著下降。他们提出的线程级并行 (TLP) 限制方案观察共享 CPU/GPU 内存控制器和互连网络中的拥塞指标，以估计应在每个 GPU 核心上主动调度的 GPU 扭曲数量。他们提出了两种方案，一种只注重提高 CPU 性能，另一种试图通过平衡受约束的多线程和 CPU 干扰导致的 GPU 性能下降来优化整体系统吞吐量（CPU 和 GPU）。作者评估了 Warp 调度对平铺异构架构的性能影响，该架构的 GPU 核心与 CPU 核心比率为 2:1，这是因为 NVIDIA GPU SM 的面积大约是使用相同工艺技术的现代无序 Intel 芯片的一半。基线配置完全共享 CPU 和 GPU 之间的网络带宽和内存控制器，以最大限度地提高资源利用率。使用这种方案，作者观察到限制 GPU TLP 会对 GPU 性能产生积极和消极影响，但绝不会损害 CPU 性能。

为了提高 CPU 性能，作者引入了一种以 CPU 为中心的并发管理技术，用于监控全局内存控制器中的停顿情况。该技术分别计算由于内存控制器输入队列已满而停顿的内存请求数量，以及由于从 MC 到内核的回复网络已满而停顿的内存请求数量。这些指标在每个内存控制器上进行本地监控，并由 ag-

gregated in a centralized unit that sends the information to the GPU cores. Heuristically driven scheme set both high and low thresholds for these values. If the sum of both request stall counts is low (based on the threshold), then the number of warps actively scheduled on the GPU is increased. If the sum of both counts is high, then the number of active warps is decreased in hopes that CPU performance will increase as a result of less GPU memory traffic.

A more balanced technique that attempts to maximize overall system throughput augments the CPU-centric approach to account for the impact on GPU performance from warp throttling. This balanced technique monitors the number of cycles the GPU cannot issue for during a concurrency rebalancing interval (1,024 cycles in their work). The moving average of GPU stalling for the current multithreading limit level is stored on each GPU core and used to determine if the level of multithreading should be increased or decreased. The balanced technique modulates the GPU's TLP in two phases. In the first phase its operation is identical to the CPU-centric solution, where GPU stalling is not accounted for and GPU TLP is limited based only on memory contention. In the second phase (which begins once GPU TLP throttling starts to cause a GPU performance degradation because the GPU's tolerance of latency has been reduced), the system stops throttling GPU concurrency if it predicts that doing so will harm GPU performance. This prediction is made by looking up the moving average number of GPU stalls at the target multithreading level, which has been recorded from earlier executions at this level. If the difference between observed GPU stalls at the target multithreading level and the current multithreading level exceeds a threshold k value, then the TLP level is not decreased. This k value can be set by the user and is a proxy for specifying the priority of GPU performance.

Heterogeneous System Coherence. Power et al. [2013a] propose a hardware mechanism to efficiently support cache coherence between the CPU and the GPU on an integrated system. They identify that directory bandwidth is significant bottleneck with the increased memory traffic generated by the GPU. They employ coarse-grained region coherence [Cantin et al., 2005] to cut down on excessive directory traffic caused in a traditional cache block-based coherence directory. Once permissions for the coarse-grained region have been acquired, most requests will not have to access the directory and the coherence traffic can be offloaded to an incoherent direct-access bus instead of a lower-bandwidth coherent interconnection network.

Heterogeneous TLP-Aware Cache Management for CPU-GPU Architectures. Lee and Kim [2012] evaluate the effect managing a shared last-level cache (LLC) between CPU cores and GPU cores in a heterogeneous environment. They demonstrate that while cache hit rates are a key performance metric for CPU workloads, many GPU workloads are insensitive to cache hit rate, since memory latency can be hidden by thread-level parallelism. To determine if GPU apps are cache-sensitive, they develop a per-core performance sampling technique where some cores bypass the shared LLC and some insert at the most-recently used position. Based on the relative performance of these cores, they can set the bypassing policy for the rest of the GPU cores, inserting in the LLC if performance is improved and bypassing if they are insensitive.

这些信息被聚集在一个将信息发送到 GPU 核心的集中单元中。启发式驱动方案为这些值设置了高阈值和低阈值。如果两个请求停顿计数的总和较低（基于阈值），则增加在 GPU 上主动调度的 Warp 数量。如果两个计数的总和较高，则减少活动 Warp 数量，以期通过减少 GPU 内存流量来提高 CPU 性能。

一种更平衡的技术试图最大化整个系统的吞吐量，它增强了以 CPU 为中心的方法，以解决 Warp 节流对 GPU 性能的影响。这种平衡技术监控 GPU 在并发重新平衡间隔内无法发出的周期数（其工作中为 1,024 个周期）。当前多线程限制级别的 GPU 停滞的移动平均值存储在每个 GPU 核心上，并用于确定是否应增加或减少多线程级别。平衡技术分两个阶段调节 GPU 的 TLP。在第一阶段，其操作与以 CPU 为中心的解决方案相同，其中不考虑 GPU 停滞，并且仅根据内存争用限制 GPU TLP。在第二阶段（一旦 GPU TLP 节流开始导致 GPU 性能下降，因为 GPU 对延迟的容忍度已经降低，就会开始），如果系统预测这样做会损害 GPU 性能，则停止限制 GPU 并发。此预测是通过查找目标多线程级别上 GPU 停顿的移动平均数来进行的，该平均数是从此级别的早期执行中记录下来的。如果在目标多线程级别和当前多线程级别上观察到的 GPU 停顿之间的差异超过阈值 k 值，则 TLP 级别不会降低。此 k 值可由用户设置，是指定 GPU 性能优先级的代理。

异构系统一致性。Power 等人 [2013a] 提出了一种硬件机制，以有效地支持集成系统上 CPU 和 GPU 之间的缓存一致性。他们发现，随着 GPU 产生的内存流量增加，目录带宽成为显著的瓶颈。他们采用粗粒度区域一致性 [Cantin 等人，2005] 来减少传统基于缓存块的一致性目录中造成的过多目录流量。一旦获得了粗粒度区域的权限，大多数请求就不必访问目录，并且可以将一致性流量卸载到非一致性直接访问总线，而不是带宽较低的一致性互连网络。

面向 CPU-GPU 架构的异构 TLP 感知缓存管理。Lee 和 Kim [2012] 评估了在异构环境中管理 CPU 核心和 GPU 核心之间共享的末级缓存 (LLC) 的效果。他们表明，虽然缓存命中率是 CPU 工作负载的关键性能指标，但许多 GPU 工作负载对缓存命中率不敏感，因为内存延迟可能被线程级并行性所隐藏。为了确定 GPU 应用程序是否对缓存敏感，他们开发了一种每核性能采样技术，其中一些核心绕过共享 LLC，一些核心插入最近使用的位置。根据这些核心的相对性能，他们可以为其余 GPU 核心设置绕过策略，如果性能得到改善，则插入 LLC，如果性能不敏感，则绕过。

Second, they observe that previous CPU-centric cache management favors cores with more frequent accesses. GPU cores are shown to generate five to ten times more traffic at the LLC. This increases bias cache capacity toward the GPU, decreasing the performance of CPU apps. They propose extending previously proposed work on utility-based cache partitioning [Qureshi and Patt, 2006] to account for the relative ratio of LLC accesses. When GPU cores are cache-sensitive, the CPU core's accesses cache way-allocation is increased beyond what utility-based cache partitioning provides to account for the difference in access magnitude and latency sensitivity between CPUs and GPUs.

其次，他们观察到，以前以 CPU 为中心的缓存管理更倾向于访问频率更高的内核。结果表明，GPU 内核在 LLC 上产生的流量是后者的五到十倍。这增加了 GPU 的缓存容量偏向性，从而降低了 CPU 应用程序的性能。他们建议扩展以前提出的基于实用程序的缓存分区 [Qureshi and Patt，2006] 工作，以考虑 LLC 访问的相对比率。当 GPU 内核对缓存敏感时，CPU 内核的访问缓存路数分配会超过基于实用程序的缓存分区所提供的分配，以考虑 CPU 和 GPU 之间访问量级和延迟敏感性的差异。

Bibliography

Die photo analysis. http://vlsiarch.eecs.harvard.edu/accelerators/die-photo-analysis 1

https://en.wikipedia.org/wiki/GDDR5_SDRAM 76

Top500.org 2

Tor M. Aamodt, Wilson W. L. Fung, Inderpreet Singh, Ahmed El-Shafiey, Jimmy Kwa, Tayler Hetherington, Ayub Gubran, Andrew Boktor, Tim Rogers, Ali Bakhoda, and Hadi Jooybar. *GPGPU-Sim 3.x Manual*. 16, 38

M. Abdel-Majeed and M. Annavaram. Warped register file: A power efficient register file for GPGPUs. In *Proc. of the IEEE International Symposium on High-Performance Computer Architecture (HPCA)*, February 2013. DOI: 10.1109/hpca.2013.6522337. 64

M. Abdel-Majeed, A. Shafaei, H. Jeon, M. Pedram, and M. Annavaram. Pilot register file: Energy efficient partitioned register file for GPUs. In *Proc. of the IEEE International Symposium on High-Performance Computer Architecture (HPCA)*, February 2017. DOI: 10.1109/hpca.2017.47. 65

Dominic Acocella and Mark R. Goudy. United States Patent #7,750,915: Concurrent Access of Data Elements Stored Across Multiple Banks in a Shared Memory Resource (Assignee: NVIDIA Corp.), July 2010. 68

Neha Agarwal, David Nellans, Mark Stephenson, Mike O'Connor, and Stephen W. Keckler. Page placement strategies for GPUs within heterogeneous memory systems. In *Proc. of the ACM Architectural Support for Programming Languages and Operating Systems (ASPLOS)*, 2015. DOI: 10.1145/2694344.2694381. 82

Jade Alglave, Mark Batty, Alastair F. Donaldson, Ganesh Gopalakrishnan, Jeroen Ketema, Daniel Poetzl, Tyler Sorensen, and John Wickerson. GPU concurrency: Weak behaviours and programming assumptions. In *Proc. of the ACM Architectural Support for Programming Languages and Operating Systems (ASPLOS)*, pages 577–591, 2015. DOI: 10.1145/2694344.2694391. 72

John R. Allen, Ken Kennedy, Carrie Porterfield, and Joe Warren. Conversion of control dependence to data dependence. In *Proc. of the ACM Symposium on Principles and Practices of Parallel Programming (PPoPP)*, pages 177–189, 1983. DOI: 10.1145/567067.567085. 16

参考书目

模具照片分析。http://vlsiarch.eecs.harvard.edu/accelerators/die-photo-analysis 1

https://en.wikipedia.org/wiki/GDDR5_SDRAM 76

Top500.org 2

Tor M. Aamodt、Wilson W. L. Fung、Inderpreet Singh、Ahmed El-Shafiey、Jimmy Kwa 、Tayler Hetherington、Ayub Gubran、Andrew Boktor、Tim Rogers、Ali Bakhoda 和 Had i Jooybar。 *GPGPU-Sim 3.x Manual*。 16, 38

M. Abdel-Majeed 和 M. Annavaram。扭曲寄存器文件：GPGPU 的节能寄存器文件。在 *Proc. of the IEEE International Symposium on High-Performance Computer Architecture (HPCA)* ，2013 年 2 月。DOI：10.1109/hpca.2013.6522337。64

M. Abdel-Majeed、A. Shafaei、H. Jeon、M. Pedram 和 M. Annavaram。试点寄存器文件 ：用于 GPU 的节能分区寄存器文件。在 *Proc. of the IEEE International Sympo- sium on High-Performance Computer Architecture (HPCA)*，2017 年 2 月。DOI：10.1109/h- p ca.2017.47。65

Dominic Acocella 和 Mark R. Goudy。美国专利号 7,750,915：共享内存资源中跨多个存储体存储的数据元素的并发访问（受让人：NVIDIA Corp.），2010 年 7 月。68

Neha Agarwal、David Nellans、Mark Stephenson、Mike O’Connor 和 Stephen W. Keckle r。异构内存系统中 GPU 的页面放置策略。在 *Proc. of the ACM Architectural Support for Programming Languages and Operating Systems (ASPLOS)*，2015 年。DOI：10.1145/2694344.2694381。82

Jade Alglave、Mark Batty、Alastair F. Donaldson、Ganesh Gopalakrishnan、Jeroen Ketem a、Daniel Poetzl、Tyler Sorensen 和 John Wickerson。GPU 并发性：弱行为和编程假设 。在 *Proc. of the ACM Architectural Support for Programming Languages and Operating Systems (ASPLOS)*，第 577 – 591 页，2015 年。D OI：10.1145/2694344.2694391。72

John R. Allen、Ken Kennedy、Carrie Porterfield 和 Joe Warren。将控制依赖转换为数据 依赖。在 *Proc. of the ACM Symposium on Principles and Practices of Parallel Programming (PPoPP)*，第 177-189 页，1983 年。DOI：10.1145/567067.567085。16

Robert Alverson, David Callahan, Daniel Cummings, Brian Koblenz, Allan Porterfield, and Burton Smith. The tera computer system. In *Proc. of the ACM International Conference on Supercomputing (ICS)*, pages 1–6, 1990. DOI: 10.1145/77726.255132. 16

R700-Family Instruction Set Architecture. AMD, March 2009. 49

AMD Southern Islands Series Instruction Set Architecture. AMD, 1.1 ed., December 2012. 13, 17, 19, 20, 26, 54, 58, 74

A. Arunkumar, S. Y. Lee, and C. J. Wu. ID-cache: Instruction and memory divergence based cache management for GPUs. In *Proc. of the IEEE International Symposium on Workload Characterization (IISWC)*, 2016. DOI: 10.1109/iiswc.2016.7581276. 79

Akhil Arunkumar, Evgeny Bolotin, Benjamin Cho, Ugljesa Milic, Eiman Ebrahimi, Oreste Villa, Aamer Jaleel, Carole-Jean Wu, and David Nellans. MCM-GPU: Multi-chip-module GPUs for continued performance scalability. In *Proc. of the ACM/IEEE International Symposium on Computer Architecture (ISCA)*, pages 320–332, 2017. DOI: 10.1145/3079856.3080231. 84

Krste Asanovic, Stephen W. Keckler, Yunsup Lee, Ronny Krashinsky, and Vinod Grover. Convergence and scalarization for data-parallel architectures. In *Proc. of the ACM/IEEE International Symposium on Code Generation and Optimization (CGO)*, 2013. DOI: 10.1109/cgo.2013.6494995. 54, 56, 58

Rachata Ausavarungnirun, Saugata Ghose, Onur Kayran, Gabriel H. Loh, Chita R. Das, Mahmut T. Kandemir, and Onur Mutlu. Exploiting inter-warp heterogeneity to improve GPGPU performance. In *Proc. of the ACM/IEEE International Conference on Parallel Architecture and Compilation Techniques (PACT)*, 2015. DOI: 10.1109/pact.2015.38. 80

Ali Bakhoda, George L. Yuan, Wilson W. L. Fung, Henry Wong, and Tor M. Aamodt. Analyzing CUDA workloads using a detailed GPU simulator. In *Proc. of the IEEE Symposium of Performance and Analysis of Systems and Software, (ISPASS'09)*, pages 163–174, 2009. DOI: 10.1109/ispass.2009.4919648. 6, 14, 78

Ali Bakhoda, John Kim, and Tor M. Aamodt. Throughput-effective on-chip networks for many-core accelerators. In *Proc. of the ACM/IEEE International Symposium on Microarchitecture (MICRO)*, pages 421–432, 2010. DOI: 10.1109/micro.2010.50. 77

Ali Bakhoda, John Kim, and Tor M. Aamodt. Designing on-chip networks for throughput accelerators. *ACM Transactions on Architecture and Code Optimization (TACO)*, 10(3):21, 2013. DOI: 10.1145/2509420.2512429. 77

Markus Billeter, Ola Olsson, and Ulf Assarsson. Efficient stream compaction on wide SIMD many-core architectures. In *Proc. of the ACM Conference on High Performance Graphics*, pages 159–166, 2009. DOI: 10.1145/1572769.1572795. 45

Robert Alverson、David Callahan、Daniel Cummings、Brian Koblenz、Allan Porterfield 和 Burton Smith。Tera 计算机系统。在 *Proc. of the ACM International Conference on Supercomputing (ICS)*，第 1-6 页，1990 年。DOI：10.1145/77726.255132。16

R700-Family Instruction Set Architecture. AMD, 2009 年 3 月. 49

AMD Southern Islands Series Instruction Set Architecture。 AMD，1.1 版，2012 年 12 月。13、17、19、20、26、54、58、74

A. Arunkumar、S. Y. Lee 和 C. J. Wu。 ID-cache：GPU 的基于指令和内存分歧的缓存管理。在 *Proc. of the IEEE International Symposium on Workload Char- acterization (IISWC)*，2016 年。DOI：10.1109/iiswc.2016.7581276。 79

Akhil Arunkumar、Evgeny Bolotin、Benjamin Cho、Ugljesa Milic、Eiman Ebrahimi、Oreste Villa、Aamer Jaleel、Carole-Jean Wu 和 David Nellans。MCM-GPU：用于持续性能可扩展性的多芯片模块 GPU。在 *Proc. of the ACM/IEEE International Symposium on Computer Architecture (ISCA)*，第 320 – 332 页，2017 年。DOI：10.1145/3079856.3080231。84

Krste Asanovic、Stephen W. Keckler、Yunsup Lee、Ronny Krashinsky 和 Vinod Grover。数据并行架构的收敛和标量化。在 *Proc. of the ACM/IEEE International Symposium on Code Generation and Optimization (CGO)*，2013 年。DOI：10.1109/cgo.2013.6494995。54、56、58

Rachata Ausavarungnirun、Saugata Ghose、Onur Kayran、Gabriel H. Loh、Chita R. Das、Mahmut T. Kandemir 和 Onur Mutlu。利用扭曲间异构性来提高 GPGPU 性能。在 *Proc. of the ACM/IEEE International Conference on Parallel Architecture and Compilation Techniques (PACT)*，2015 年。DOI：10.1109/pact.2015.38。 80

Ali Bakhoda、George L. Yuan、Wilson W. L. Fung、Henry Wong 和 Tor M. Aamodt。使用详细的 GPU 模拟器分析 CUDA 工作负载。在 *Proc. of the IEEE Symposium of Performance and Analysis of Systems and Software, (ISPASS'09)* 中，第 163 – 174 页，2009 年。DOI：10.1109/ispass.2009.4919648。6、14、78

Ali Bakhoda、John Kim 和 Tor M. Aamodt。多核加速器吞吐量有效的片上网络。在 *Proc. of the ACM/IEEE International Symposium on Microarchitecture (MICRO)*，第 421 – 432 页，2010 年。DOI：10.1109/micro.2010.50。77

Ali Bakhoda、John Kim 和 Tor M. Aamodt。设计用于吞吐量加速器的片上网络。*ACM Transactions on Architecture and Code Optimization (TACO)*，10(3):21，2013 年。DOI：10.1145/2509420.2512429。77

Markus Billeter、Ola Olsson 和 Ulf Assarsson。在宽 SIMD 多核架构上实现高效的流压缩。在 *Proc. of the ACM Conference on High Performance Graphics*，第 159 – 166 页，2009 年。DOI：10.1145/1572769.1572795。45

Nicolas Brunie, Sylvain Collange, and Gregory Diamos. Simultaneous branch and warp interweaving for sustained GPU performance. In *Proc. of the ACM/IEEE International Symposium on Computer Architecture (ISCA)*, pages 49–60, 2012. DOI: 10.1109/isca.2012.6237005. 44, 53

Ian Buck, Tim Foley, Daniel Horn, Jeremy Sugerman, Kayvon Fatahalian, Mike Houston, and Pat Hanrahan. Brook for GPUs: Stream computing on graphics hardware. In *Proc. of the ACM International Conference on Computer Graphics and Interactive Techniques (SIGGRAPH)*, pages 777–786, 2004. DOI: 10.1145/1186562.1015800. 6

Brian Cabral. what is "SASS" short for? `https://stackoverflow.com/questions/9798258/what-is-sass-short-for`, November 2016. 14

J. F. Cantin, M. H. Lipasti, and J. E. Smith. Improving multiprocessor performance with coarse-grain coherence tracking. In *Proc. of the ACM/IEEE International Symposium on Computer Architecture (ISCA)*, June 2005. DOI: 10.1109/isca.2005.31. 100

Edwin Catmull. A subdivision algorithm for computer display of curved surfaces. *Technical Report*, DTIC Document, 1974. 72

Hassan Chafi, Jared Casper, Brian D. Carlstrom, Austen Mcdonald, Chi Cao, Minh Woongki Baek, Christos Kozyrakis, and Kunle Olukotun. A scalable, non-blocking approach to transactional memory. In *Proc. of the IEEE International Symposium on High-Performance Computer Architecture (HPCA)*, pages 97–108, 2007. DOI: 10.1109/hpca.2007.346189. 97

Guoyang Chen, Bo Wu, Dong Li, and Xipeng Shen. PORPLE: An extensible optimizer for portable data placement on GPU. In *Proc. of the ACM/IEEE International Symposium on Microarchitecture (MICRO)*, 2014a. DOI: 10.1109/micro.2014.20. 83

Xi E. Chen and Tor M. Aamodt. A first-order fine-grained multithreaded throughput model. In *Proc. of the IEEE International Symposium on High-Performance Computer Architecture (HPCA)*, pages 329–340, 2009. DOI: 10.1109/hpca.2009.4798270. 78

Xuhao Chen, Li-Wen Chang, Christopher I. Rodrigues, Jie Lv, Zhiying Wang, and Wen-Mei Hwu. Adaptive cache management for energy-efficient GPU computing. In *Proc. of the ACM/IEEE International Symposium on Microarchitecture (MICRO)*, 2014b. DOI: 10.1109/micro.2014.11. 81, 82

Sylvain Collange, David Defour, and Yao Zhang. Dynamic detection of uniform and affine vectors in GPGPU computations. In *Proc. of the European Conference on Parallel Processing (Euro-Par)*, 2010. DOI: 10.1007/978-3-642-14122-5_8. 59

Brett W. Coon and John Erik Lindholm. United States Patent #7,353,369: System and Method for Managing Divergent Threads in a SIMD Architecture (Assignee NVIDIA Corp.), April 2008. 26, 49, 50

Nicolas Brunie、Sylvain Collange 和 Gregory Diamos。同时进行分支和扭曲交织，以保持 GPU 性能。在 *Proc. of the ACM/IEEE International Symposium on Computer Architecture (ISCA)*，第 49 – 60 页，2012 年。DOI：10.1109/isca.2012.6237005。44、53

Ian Buck、Tim Foley、Daniel Horn、Jeremy Sugerman、Kayvon Fatahalian、Mike Houston 和 Pat Hanrahan。Brook for GPUs：图形硬件上的流计算。在 *Proc. of the ACM International Conference on Computer Graphics and Interactive Techniques (SIGGRAPH)*，第 777 – 786 页，2004 年。DOI：10.1145/1186562.1015800。6

Brian Cabral。“ SASS ”的缩写是什么？ `https://stackoverflow.com/questions/9798258/what-is-sass-short-for`，2016 年 11 月 14 日。

J. F. Cantin、M. H. Lipasti 和 J. E. Smith。通过粗粒度一致性跟踪提高多处理器性能。在 *Proc. of the ACM/IEEE International Symposium on Computer Architecture (ISCA)*，2005 年 6 月。DOI：10.1109/isca.2005.31。100

Edwin Catmull。一种用于计算机显示曲面的细分算法。*Technical Report*，DTIC 文档，1974 年。72

Hassan Chafi、Jared Casper、Brian D. Carlstrom、Austen Mcdonald、Chi Cao、Minh Woongki Baek、Christos Kozyrakis 和 Kunle Olukotun。一种可扩展、非阻塞事务内存方法。在 *Proc. of the IEEE International Symposium on High-Performance Computer Architecture (HPCA)*，第 97-108 页，2007 年。DOI：10.1109/hpca.2007.346189。97

Guoyang Chen, Bo Wu, Dong Li, and Xipeng Shen. PORPLE: An extensible optimizer for portable data placement on GPU. In *Proc. of the ACM/IEEE International Symposium on Microarchitecture (MICRO)*, 2014a. DOI: 10.1109/micro.2014.20. 83

Xi E. Chen 和 Tor M. Aamodt。一阶细粒度多线程吞吐量模型。在 *Proc. of the IEEE International Symposium on High-Performance Computer Architecture (HPCA)*，第 329 – 340 页，2009 年。DOI：10.1109/hpca.2009.4798270。78

Xuhao Chen、Li-Wen Chang、Christopher I. Rodrigues、Jie Lv、Zhiying Wang 和 Wen-Mei Hwu。用于节能 GPU 计算的自适应缓存管理。在 *Proc. of the ACM/IEEE International Symposium on Microarchitecture (MICRO)*，2014b。DOI：10.1109/micro.2014.11。81，82

Sylvain Collange、David Defour 和 Yao Zhang。GPGPU 计算中均匀和仿射向量的动态检测。在 *Proc. of the European Conference on Parallel Processing (Euro-Par)*，2010 年。DOI：10.1007/978-3-642-14122-5_8。59

Brett W. Coon 和 John Erik Lindholm。美国专利 #7,353,369：SIMD 架构中管理发散线程的系统和方法（受让人 NVIDIA Corp.），2008 年 4 月。26、49、50

Brett W. Coon, Peter C. Mills, Stuart F. Oberman, and Ming Y. Siu. United States Patent #7,434,032: Tracking Register Usage During Multithreaded Processing Using a Scorebard having Separate Memory Regions and Storing Sequential Register Size Indicators (Assignee NVIDIA Corp.), October 2008. 34

Brett W. Coon, John Erik Lindholm, Gary Tarolli, Svetoslav D. Tzvetkov, John R. Nickolls, and Ming Y. Siu. United States Patent #7,634,621: Register File Allocation (Assignee NVIDIA Corp.), December 2009. 35

Ron Cytron, Jeanne Ferrante, Barry K. Rosen, Mark N. Wegman, and F. Kenneth Zadeck. Efficiently computing static single assignment form and the control dependence graph. *ACM Transactions on Programming Languages and Systems (TOPLAS)*, 13(4):451–490, 1991. DOI: 10.1145/115372.115320. 16

Luke Dalessandro, Michael F. Spear, and Michael L. Scott. NOrec: Streamlining STM by abolishing ownership records. In *Proc. of the ACM Symposium on Principles and Practices of Parallel Programming (PPoPP)*, pages 67–78, 2010. DOI: 10.1145/1693453.1693464. 96, 99

R. H. Dennard, F. H. Gaensslen, and K. Mai. Design of ion-implanted MOSFET's with very small physical dimensions. *IEEE Journal of Solid-State Circuits*, October 1974. DOI: 10.1109/jssc.1974.1050511. 1

Gregory Diamos, Benjamin Ashbaugh, Subramaniam Maiyuran, Andrew Kerr, Haicheng Wu, and Sudhakar Yalamanchili. SIMD re-convergence at thread frontiers. In *Proc. of the ACM/IEEE International Symposium on Microarchitecture (MICRO)*, pages 477–488, 2011. DOI: 10.1145/2155620.2155676. 26, 51, 54

Gregory Frederick Diamos, Richard Craig Johnson, Vinod Grover, Olivier Giroux, Jack H. Choquette, Michael Alan Fetterman, Ajay S. Tirumala, Peter Nelson, and Ronny Meir Krashinsky. Execution of divergent threads using a convergence barrier, July 13, 2015. 27, 28, 29, 30

Roger Eckert. United States Patent #7,376,803: Page stream sorter for DRAM systems (Assignee: NVIDIA Corp.), May 2008. 73

Roger Eckert. United States Patent #9,195,618: Method and System for Scheduling Memory Requests (Assignee: NVIDIA Corp.), November 2015. 73

John H. Edmondson and James M. Van Dyke. United States Patent #7872657: Memory Addressing Scheme using Partition Strides, January 2011. 75

John H. Edmondson et al. United States Patent #8,464,001: Cache and Associated Method with Frame Buffer Managed Dirty Data Pull and High-Priority Clean Mechanism, June 2013. 75, 76

Brett W. Coon、Peter C. Mills、Stuart F. Oberman 和 Ming Y. Siu。美国专利号 7,434,032：使用具有独立内存区域和存储连续寄存器大小指示器的记分板跟踪多线程处理期间的寄存器使用情况（受让人 NVIDIA Corp.），2008 年 10 月。34

Brett W. Coon、John Erik Lindholm、Gary Tarolli、Svetoslav D. Tzvetkov、John R. Nickolls 和 Ming Y. Siu。美国专利号 7,634,621：寄存器文件分配（受让人 NVIDIA Corp.），2009 年 12 月。35

Ron Cytron、Jeanne Ferrante、Barry K. Rosen、Mark N. Wegman 和 F. Kenneth Zadeck。高效计算静态单分配形式和控制依赖图。*ACM Transactions on Programming Languages and Systems (TOPLAS)*，13(4):451 – 490，1991 年。DOI：10.1145/115372.115320。16

Luke Dalessandro、Michael F. Spear 和 Michael L. Scott。NOrec：通过废除所有权记录简化 STM。在 *Proc. of the ACM Symposium on Principles and Practices of Parallel Programming (PPoPP)*，第 67 – 78 页，2010 年。DOI：10.1145/1693453.1693464

R. H. Dennard、F. H. Gaensslen 和 K. Mai。设计具有极小物理尺寸的离子注入 MOSFET。*IEEE Journal of Solid-State Circuits*，1974 年 10 月。DOI：10.1109/jssc.1974.1050511。1

Gregory Diamos、Benjamin Ashbaugh、Subramaniam Maiyuran、Andrew Kerr、吴海城和 Sudhakar Yalamanchili。 SIMD 在线程边界重新收敛。载于 *Proc. of the ACM/IEEE International Symposium on Microarchitecture (MICRO)*，第 477 – 488 页，2011 年。DOI：10.1145/2155620.2155676。 26、51、54

Gregory Frederick Diamos、Richard Craig Johnson、Vinod Grover、Olivier Giroux、Jack H. Choquette、Michael Alan Fetterman、Ajay S. Tirumala、Peter Nelson 和 Ronny Meir Krashinsky。使用收敛屏障执行发散线程，2015 年 7 月 13 日。27、28、29、30

Roger Eckert。美国专利号 7,376,803：DRAM 系统的页面流分类器（受让人：NVIDIA Corp.），2008 年 5 月。73

Roger Eckert。美国专利号 9,195,618：用于调度内存请求的方法和系统（受让人：NVIDIA Corp.），2015 年 11 月。73

John H. Edmondson 和 James M. Van Dyke。美国专利号 7872657：使用分区步幅的内存寻址方案，2011 年 1 月。75

John H. Edmondson 等人，美国专利 #8,464,001：带有帧缓冲区管理脏数据提取和高优先级清理机制的缓存和相关方法，2013 年 6 月。75、76

Ahmed ElTantaway, Jessica Wenjie Ma, Mike O'Connor, and Tor M. Aamodt. A scalable multi-path microarchitecture for efficient GPU control flow. In *Proc. of the IEEE International Symposium on High-Performance Computer Architecture (HPCA)*, 2014. DOI: 10.1109/hpca.2014.6835936. 26, 28, 30, 31, 49, 52

Ahmed ElTantawy and Tor M. Aamodt. MIMD synchronization on SIMT architectures. In *Proc. of the ACM/IEEE International Symposium on Microarchitecture (MICRO)*, pages 1–14, 2016. DOI: 10.1109/micro.2016.7783714. 26, 27, 30, 32, 93

Ahmed ElTantawy and Tor M. Aamodt. Warp scheduling for fine-grained synchronization. In *Proc. of the IEEE International Symposium on High-Performance Computer Architecture (HPCA)*, pages 375–388, 2018. DOI: 10.1109/hpca.2018.00040. 93

Alexander L. Minken et al., United States Patent #7,649,538: Reconfigurable High Performance Texture Pipeline with Advanced Filtering (Assignee: NVIDIA Corp.), January 2010. 72

Wilson W. L. Fung. *GPU Computing Architecture for Irregular Parallelism*. Ph.D. thesis, University of British Columbia, January 2015. DOI: 10.14288/1.0167110. 41

Wilson W. L. Fung and Tor M. Aamodt. Thread block compaction for efficient SIMT control flow. In *Proc. of the IEEE International Symposium on High-Performance Computer Architecture (HPCA)*, pages 25–36, 2011. DOI: 10.1109/hpca.2011.5749714. 26, 28, 42, 43, 50, 91

Wilson W. L. Fung and Tor M. Aamodt. Energy efficient GPU transactional memory via space-time optimizations. In *Proc. of the ACM/IEEE International Symposium on Microarchitecture (MICRO)*, pages 408–420, 2013. DOI: 10.1145/2540708.2540743. 98

Wilson W. L. Fung, Ivan Sham, George Yuan, and Tor M. Aamodt. Dynamic warp formation and scheduling for efficient GPU control flow. In *Proc. of the ACM/IEEE International Symposium on Microarchitecture (MICRO)*, pages 407–420, 2007. DOI: 10.1109/micro.2007.4408272. 14, 23, 25, 42, 44, 49, 91

Wilson W. L. Fung, Inderpreet Singh, Andrew Brownsword, and Tor M. Aamodt. Hardware transactional memory for GPU architectures. In *Proc. of the ACM/IEEE International Symposium on Microarchitecture (MICRO)*, pages 296–307, 2011. DOI: 10.1145/2155620.2155655. 96, 97

Wilson Fung et al. Dynamic warp formation: Efficient MIMD control flow on SIMD graphics hardware. *ACM Transactions on Architecture and Code Optimization (TACO)*, 6(2):7:1–7:37, 2009. DOI: 10.1145/1543753.1543756. 42, 49

M. Gebhart, S. W. Keckler, and W. J. Dally. A compile-time managed multi-level register file hierarchy. In *Proc. of the ACM/IEEE International Symposium on Microarchitecture (MICRO)*, December 2011a. DOI: 10.1145/2155620.2155675. 63

Ahmed ElTantaway、Jessica Wenjie Ma、Mike O'Connor 和 Tor M. Aamodt。一种可扩展的多路径微架构，用于实现高效的 GPU 控制流。在 *Proc. of the IEEE International Symposium on High-Performance Computer Architecture (HPCA)* 中，2014 年。DOI：10.1109/h- pca.2014.6835936。26、28、30、31、49、52

Ahmed ElTantawy 和 Tor M. Aamodt。SIMT 架构上的 MIMD 同步。在 *Proc. of the ACM/IEEE International Symposium on Microarchitecture (MICRO)*，第 1-14 页，2016 年。DOI：10.1109/micro.2016.7783714。26、27、30、32、93

Ahmed ElTantawy 和 Tor M. Aamodt。用于细粒度同步的 Warp 调度。在 *Proc. of the IEEE International Symposium on High-Performance Computer Architecture (HPCA)*，第 375 – 388 页，2018 年。DOI：10.1109/hpca.2018.00040。93

Alexander L. Minken 等人，美国专利号 7,649,538：具有高级过滤功能的可重构高性能纹理管道（受让人：NVIDIA Corp.），2010 年 1 月。72

Wilson W. L. Fung。*GPU Computing Architecture for Irregular Parallelism*。博士论文，不列颠哥伦比亚大学，2015 年 1 月。DOI：10.14288/1.0167110。41

Wilson W. L. Fung 和 Tor M. Aamodt。线程块压缩以实现高效的 SIMT 控制流。在 *Proc. of the IEEE International Symposium on High-Performance Computer Architecture (HPCA)* 中，第 25 – 36 页，2011 年。DOI：10.1109/hpca.2011.5749714。26、28、42、43、50、91

Wilson W. L. Fung 和 Tor M. Aamodt。通过时空优化实现节能的 GPU 事务内存。在 *Proc. of the ACM/IEEE International Symposium on Microarchitecture (MICRO)*，第 408 – 420 页，2013 年。DOI：10.1145/2540708.2540743。98

Wilson W. L. Fung、Ivan Sham、George Yuan 和 Tor M. Aamodt。动态扭曲形成和调度以实现高效的 GPU 控制流。在 *Proc. of the ACM/IEEE International Symposium on Microarchitecture (MICRO)* 中，第 407 – 420 页，2007 年。DOI：10.1109/micro.2007.4408272。14、23、25、42、44、49、91

Wilson W. L. Fung、Inderpreet Singh、Andrew Brownsword 和 Tor M. Aamodt。GPU 架构的硬件事务内存。在 *Proc. of the ACM/IEEE International Symposium on Microarchitecture (MICRO)*，第 296 – 307 页，2011 年。DOI：10.1145/2155620.2155655。96、97

Wilson Fung 等人。动态扭曲形成：SIMD 图形硬件上的高效 MIMD 控制流。*ACM Transactions on Architecture and Code Optimization (TACO)*，6(2):7:1 – 7:37，2009 年。DOI：10.1145/1543753.1543756。42，49

M. Gebhart、S. W. Keckler 和 W. J. Dally。编译时管理的多级寄存器文件层次结构。在 *Proc. of the ACM/IEEE International Symposium on Microarchitecture (MICRO)*，2011 年 12 月。DOI：10.1145/2155620.2155675。63

Mark Gebhart, Daniel R. Johnson, David Tarjan, Stephen W. Keckler, William J. Dally, Erik Lindholm, and Kevin Skadron. Energy-efficient mechanisms for managing thread context in throughput processors. In *Proc. of the ACM/IEEE International Symposium on Computer Architecture (ISCA)*, 2011b. DOI: 10.1145/2000064.2000093. 63

Mark Gebhart, Daniel R. Johnson, David Tarjan, Stephen W. Keckler, William J. Dally, Erik Lindholm, and Kevin Skadron. Energy-efficient mechanisms for managing thread context in throughput processors. In *Proc. of the ACM/IEEE International Symposium on Computer Architecture (ISCA)*, pages 235–246, 2011c. DOI: 10.1145/2000064.2000093. 88

Isaac Gelado, John E. Stone, Javier Cabezas, Sanjay Patel, Nacho Navarro, and Wen-mei W. Hwu. An asymmetric distributed shared memory model for heterogeneous parallel systems. In *Proc. of the ACM Architectural Support for Programming Languages and Operating Systems (ASPLOS)*, pages 347–358, 2010. DOI: 10.1145/1736020.1736059. 3

Syed Zohaib Gilani, Nam Sung Kim, and Michael J. Schulte. Power-efficient computing for compute-intensive GPGPU applications. In *Proc. of the IEEE International Symposium on High-Performance Computer Architecture (HPCA)*, 2013. DOI: 10.1109/hpca.2013.6522330. 59, 61

David B. Glasco et al. United States Patent #8,135,926: Cache-Based Control of Atomic Operations in Conjunction with an External Alu Block, March 2012. 76

David B. Glasco et al. United States Patent #8,539,130: Virtual Channels for Effective Packet Transfer, September 2013. 75

Scott Gray. Assembler for NVIDIA Maxwell architecture. `https://github.com/NervanaSystems/maxas` 16, 40

Zvika Guz, Evgeny Bolotin, Idit Keidar, Avinoam Kolodny, Avi Mendelson, and Uri C. Weiser. Many-core vs. many-thread machines: Stay away from the valley. *IEEE Computer Architecture Letters*, 8(1):25–28, 2009. DOI: 10.1109/l-ca.2009.4. 5

Ziyad S. Hakura and Anoop Gupta. The design and analysis of a cache architecture for texture mapping. In *Proc. of the ACM/IEEE International Symposium on Computer Architecture (ISCA)*, pages 108–120, 1997. DOI: 10.1145/264107.264152. 73

Rehan Hameed, Wajahat Qadeer, Megan Wachs, Omid Azizi, Alex Solomatnikov, Benjamin C. Lee, Stephen Richardson, Christos Kozyrakis, and Mark Horowitz. Understanding sources of inefficiency in general-purpose chips. In *Proc. of the ACM/IEEE International Symposium on Computer Architecture (ISCA)*, pages 37–47, 2010. DOI: 10.1145/1815961.1815968. 1

Song Han, Xingyu Liu, Huizi Mao, Jing Pu, Ardavan Pedram, Mark A. Horowitz, and William J. Dally. EIE: efficient inference engine on compressed deep neural network. In

Mark Gebhart、Daniel R. Johnson、David Tarjan、Stephen W. Keckler、William J. Dally、Erik Lindholm 和 Kevin Skadron。用于管理吞吐量处理器中线程上下文的节能机制。在 *Proc. of the ACM/IEEE International Symposium on Computer Architecture (ISCA)*，2011b 年。DOI：10.1145/2000064.2000093。63 Mark Gebhart、Daniel R. Johnson、David Tarjan、Stephen W. Keckler、William J. Dally、Erik Lindholm 和 Kevin Skadron。用于管理吞吐量处理器中线程上下文的节能机制。在 *Proc. of the ACM/IEEE International Symposium on Computer Architecture (ISCA)*，第 235-246 页，2011c 年。DOI：10.1145/2000064.2000093。 88 Isaac Gelado、John E. Stone、Javier Cabezas、Sanjay Patel、Nacho Navarro 和 Wen-mei W. Hwu。异构并行系统的非对称分布式共享内存模型。在 *Proc. of the ACM Architectural Support for Programming Languages and Operating Systems (ASPLOS)*，第 347 – 358 页，2010 年。DOI：10.1145/1736020.1736059。3 Syed Zohaib Gilani、Nam Sung Kim 和 Michael J. Schulte。计算密集型 GPGPU 应用程序的节能计算。在 *Proc. of the IEEE International Symposium on High-Performance Computer Architecture (HPCA)*，2013 年。DOI：10.1109/hpca.2013.6522330。59、61 David B. Glasco 等人。美国专利 #8,135,926：与外部 Alu 块结合的基于缓存的原子操作控制，2012 年 3 月。76 David B. Glasco 等人。美国专利 #8,539,130：用于有效数据包传输的虚拟通道，2013 年 9 月。75 Scott Gray。NVIDIA Maxwell 架构的汇编程序。`https://github.com/NervanaSyst ems/maxas` 16, 40 Zvika Guz、Evgeny Bolotin、Idit Keidar、Avinoam Kolodny、Avi Mendelson 和 Uri C. Weiser。多核与多线程机器：远离低谷。*IEEE Computer Architecture Letters*，8(1):25 – 28，2009 年。DOI：10.1109/l-ca.2009.4。 5 Ziyad S. Hakura 和 Anoop Gupta。纹理映射缓存架构的设计和分析。在 *Proc. of the ACM/IEEE International Symposium on Computer Architecture (ISCA)*，第 108-120 页，1997 年。DOI：10.1145/264107.264152。73 Rehan Hameed、Wajahat Qadeer、Megan Wachs、Omid Azizi、Alex Solomatnikov、Benjamin C. Lee、Stephen Richardson、Christos Kozyrakis 和 Mark Horowitz。了解通用芯片的低效率来源。在 *Proc. of the ACM/IEEE International Symposium on Computer Architecture (ISCA)*，第 37-47 页，2010 年。DOI：10.1145/1815961.1815968。1 Song Han、Xingyu Liu、Huizi Mao、Jing Pu、Ardavan Pedram、Mark A. Horowitz 和 William J. Dally。 EIE：压缩深度神经网络上的高效推理引擎。

Proc. of the ACM/IEEE International Symposium on Computer Architecture (ISCA), pages 243–254, 2016. DOI: 10.1109/isca.2016.30. 6

Mark Harris. *An Easy Introduction to CUDA C and C++*. https://devblogs.nvidia.com/parallelforall/easy-introduction-cuda-c-and-c/, 2012. 11

Tim Harris, James Larus, and Ravi Rajwar. *Transactional Memory*, 2nd ed. Morgan & Claypool, 2010. DOI: 10.1201/b11417-16. 96

Steven J. Heinrich et al. United States Patent #9,595,075: Load/Store Operations in Texture Hardware (Assignee: NVIDIA Corp.), March 2017. 68, 73

John Hennessy and David Patterson. *Computer Architecture—A Quantitative Approach*, 5th ed. Morgan Kaufmann, 2011. 1, 10, 71, 72, 78

Maurice Herlihy and J. Eliot B. Moss. Transactional memory: Architectural support for lock-free data structures. In *Proc. of the ACM/IEEE International Symposium on Computer Architecture (ISCA)*, pages 289–300, 1993. DOI: 10.1109/isca.1993.698569. 96

Jared Hoberock, Victor Lu, Yuntao Jia, and John C. Hart. Stream compaction for deferred shading. In *Proc. of the ACM Conference on High Performance Graphics*, pages 173–180, 2009. DOI: 10.1145/1572769.1572797. 45

H. Peter Hofstee. Power efficient processor architecture and the cell processor. In *Proc. of the IEEE International Symposium on High-Performance Computer Architecture (HPCA)*, pages 258–262, 2005. DOI: 10.1109/hpca.2005.26. 68

Mark Horowitz, Elad Alon, Dinesh Patil, Samuel Naffziger, Rajesh Kumar, and Kerry Bernstein. Scaling, power, and the future of CMOS. In *IEEE International Electron Devices Meeting*, 2005. DOI: 10.1109/iedm.2005.1609253. 5

Homan Igehy, Matthew Eldridge, and Kekoa Proudfoot. Prefetching in a texture cache architecture. In *Proc. of the ACM SIGGRAPH/EUROGRAPHICS Workshop on Graphics hardware*, pages 133–ff, 1998. DOI: 10.1145/285305.285321. 72, 73, 74

Hyeran Jeon, Gokul Subramanian Ravi, Nam Sung Kim, and Murali Annavaram. GPU register file virtualization. In *Proc. of the ACM/IEEE International Symposium on Microarchitecture (MICRO)*, pages 420–432, 2015. DOI: 10.1145/2830772.2830784. 64

W. Jia, K. A. Shaw, and M. Martonosi. MRPB: Memory request prioritization for massively parallel processors. In *Proc. of the IEEE International Symposium on High-Performance Computer Architecture (HPCA)*, 2014. DOI: 10.1109/hpca.2014.6835938. 78, 79

Wenhao Jia, Kelly A Shaw, and Margaret Martonosi. Characterizing and improving the use of demand-fetched caches in GPUs. In *Proc. of the ACM International Conference on Supercomputing (ICS)*, pages 15–24, 2012. DOI: 10.1145/2304576.2304582. 78

Proc. of the ACM/IEEE International Symposium on Computer Architecture (ISCA)，第 243 – 254 页，2016 年。DOI：10.1109/isca.2016.30。 6

马克·哈里斯。*An Easy Introduction to CUDA C and C++*。https://devblogs.nvidia.com/parallelforall/easy-introduction-cuda-c-and-c/，2012 年。11

Tim Harris、James Larus 和 Ravi Rajwar。*Transactional Memory*，第二版。Morgan & Claypool，2010 年。DOI：10.1201/b11417-16。96

Steven J. Heinrich 等人。美国专利号 9,595,075：纹理硬件中的加载/存储操作（受让人：NVIDIA Corp.），2017 年 3 月。68、73

John Hennessy 和 David Patterson。*Computer Architecture—A Quantitative Approach*，第 5 版。Morgan Kaufmann，2011 年。1、10、71、72、78

Maurice Herlihy 和 J. Eliot B. Moss。事务内存：无锁数据结构的架构支持。在 *Proc. of the ACM/IEEE International Symposium on Computer Architec- ture (ISCA)*，第 289-300 页，1993 年。DOI：10.1109/isca.1993.698569。96

Jared Hoberock、Victor Lu、Yuntao Jia 和 John C. Hart。延迟着色的流压缩。在 *Proc. of the ACM Conference on High Performance Graphics*，第 173-180 页，2009 年。DOI：10.1145/1572769.1572797。45

H. Peter Hofstee。高效处理器架构和单元处理器。在 *Proc. of the IEEE International Symposium on High-Performance Computer Architecture (HPCA)*，第 258 – 262 页，2005 年。DOI：10.1109/hpca.2005.26。68

Mark Horowitz、Elad Alon、Dinesh Patil、Samuel Naffiziger、Rajesh Kumar 和 Kerry Bernstein。CMOS 的扩展、功率和未来。在 *IEEE International Electron Devices Meeting*，2005 年。DOI：10.1109/iedm.2005.1609253。5

Homan Igehy、Matthew Eldridge 和 Kekoa Proudfoot。纹理缓存架构中的预取。在 *Proc. of the ACM SIGGRAPH/EUROGRAPHICS Workshop on Graphics hardware*，第 133-ff 页，1998 年。DOI：10.1145/285305.285321。72、73、74

Hyeran Jeon、Gokul Subramanian Ravi、Nam Sung Kim 和 Murali Annavaram。GPU 寄存器文件虚拟化。在 *Proc. of the ACM/IEEE International Symposium on Microarchitecture (MICRO)*，第 420 – 432 页，2015 年。DOI：10.1145/2830772.2830784。64

W. Jia、K. A. Shaw 和 M. Martonosi。MRPB：大规模并行处理器的内存请求优先级。在 *Proc. of the IEEE International Symposium on High-Performance Computer Architecture (HPCA)*，2014 年。DOI：10.1109/hpca.2014.6835938。78，79

Wenhao Jia、Kelly A Shaw 和 Margaret Martonosi。描述和改进 GPU 中按需获取缓存的使用。在 *Proc. of the ACM International Conference on Supercom- puting (ICS)*，第 15-24 页，2012 年。DOI：10.1145/2304576.2304582。78

Adwait Jog, Onur Kayiran, Nachiappan Chidambaram Nachiappan, Asit K. Mishra, Mahmut T. Kandemir, Onur Mutlu, Ravishankar Iyer, and Chita R. Das. OWL: Cooperative thread array aware scheduling techniques for improving GPGPU performance. In *Proc. of the ACM Architectural Support for Programming Languages and Operating Systems (ASPLOS)*, 2013a. DOI: 10.1145/2451116.2451158. 90

Adwait Jog, Onur Kayiran, Asit K. Mishra, Mahmut T. Kandemir, Onur Mutlu, Ravishankar Iyer, and Chita R. Das. Orchestrated scheduling and prefetching for GPGPUs. In *Proc. of the ACM/IEEE International Symposium on Computer Architecture (ISCA)*, 2013b. DOI: 10.1145/2508148.2485951. 90

Norman P. Jouppi, Cliff Young, Nishant Patil, David Patterson, Gaurav Agrawal, Raminder Bajwa, Sarah Bates, Suresh Bhatia, Nan Boden, Al Borchers, et al. In-datacenter performance analysis of a tensor processing unit. In *Proc. of the ACM/IEEE International Symposium on Computer Architecture (ISCA)*, 2017. DOI: 10.1145/3079856.3080246. 2

David R. Kaeli, Perhaad Mistry, Dana Schaa, and Dong Ping Zhang. *Heterogeneous Computing with OpenCL 2.0*. Morgan Kaufmann, 2015. 10

Ujval J. Kapasi et al. Efficient conditional operations for data-parallel architectures. In *Proc. of the ACM/IEEE International Symposium on Microarchitecture (MICRO)*, pages 159–170, 2000. DOI: 10.1109/micro.2000.898067. 45

O. Kayiran, N. C. Nachiappan, A. Jog, R. Ausavarungnirun, M. T. Kandemir, G. H. Loh, O. Mutlu, and C. R. Das. Managing GPU concurrency in heterogeneous architectures. In *Proc. of the ACM/IEEE International Symposium on Microarchitecture (MICRO)*, 2014. DOI: 10.1109/micro.2014.62. 99

Onur Kayiran, Adwait Jog, Mahmut T. Kandemir, and Chita R. Das. Neither more nor less: Optimizing thread-level parallelism for GPGPUs. In *Proc. of the ACM/IEEE International Conference on Parallel Architecture and Compilation Techniques (PACT)*, 2013. 86

S. W. Keckler, W. J. Dally, B. Khailany, M. Garland, and D. Glasco. GPUs and the future of parallel computing. *Micro, IEEE*, 31(5):7–17, September 2011. DOI: 10.1109/mm.2011.89. 53, 58

Shane Keil and John H. Edmondson. United States Patent #8,195,858: Managing Conflicts on Shared L2 Bus, June 2012. 77

Shane Keil et al. United States Patent #8,307,165: Sorting Requests to the Dram for High Page Locality, November 2012. 77

Farzad Khorasani, Rajiv Gupta, and Laxmi N. Bhuyan. Efficient warp execution in presence of divergence with collaborative context collection. In *Proc. of the ACM/IEEE International Symposium on Microarchitecture (MICRO)*, 2015. DOI: 10.1145/2830772.2830796. 45

Adwait Jog、Onur Kayiran、Nachiappan Chidambaram Nachiappan、Asit K. Mishra、Mah mut T. Kandemir、Onur Mutlu、Ravishankar Iyer 和 Chita R. Das。 OWL：协作线程阵列感知调度技术，用于提高 GPGPU 性能。在*Proc. of the ACM Architectural Support for Programming Languages and Operating Systems (ASPLOS)*，2013a。 DOI：10.1145/2451116.2451158。 90

Adwait Jog、Onur Kayiran、Asit K. Mishra、Mahmut T. Kandemir、Onur Mutlu、Ravisha nkar Iyer 和 Chita R. Das。协调 GPGPU 的调度和预取。在 *Proc. of the ACM/IEEE International Symposium on Computer Architecture (ISCA)*，2013b。 DOI ：10.1145/2508148.2485951。 90

Norman P. Jouppi、Cliffi Young、Nishant Patil、David Patterson、Gaurav Agrawal、Rami nder Bajwa、Sarah Bates、Suresh Bhatia、Nan Boden、Al Borchers 等人。张量处理单元的数据中心内性能分析。在 *Proc. of the ACM/IEEE International Symposium on Computer Architecture (ISCA)*，2017 年。DOI：10.1145/3079856.3080246。2

David R. Kaeli、Perhaad Mistry、Dana Schaa 和 Dong Ping Zhang。*Heterogeneous Computing with OpenCL 2.0*。Morgan Kaufmann，2015 年。10

Ujval J. Kapasi 等人。数据并行架构的高效条件操作。在 *Proc. of the ACM/IEEE International Symposium on Microarchitecture (MICRO)*，第 159 – 170 页，200 0 年。DOI：10.1109/micro.2000.898067。45

O. Kayiran、N. C. Nachiappan、A. Jog、R. Ausavarungnirun、M. T. Kandemir、G. H. Loh 、O. Mutlu 和 C. R. Das。管理异构架构中的 GPU 并发。在 *Proc. of the ACM/IEEE International Symposium on Microarchitecture (MICRO)*，2014 年。D OI：10.1109/micro.2014.62。 99

Onur Kayiran、Adwait Jog、Mahmut T. Kandemir 和 Chita R. Das。不多不少：优化 GPG PU 的线程级并行性。在 *Proc. of the ACM/IEEE International Conference on Parallel Architecture and Compilation Techniques (PACT)*，2013 年。86

S. W. Keckler、W. J. Dally、B. Khailany、M. Garland 和 D. Glasco。GPU 和并行计算的未来。*Micro, IEEE*，31(5):7 – 17，2011 年 9 月。DOI：10.1109/mm.2011.89。53、58

Shane Keil 和 John H. Edmondson。美国专利号 8,195,858：管理共享 L2 总线上的冲突，2012 年 6 月。77

Shane Keil 等人，美国专利号 8,307,165：对 Dram 的请求进行排序以实现高页面局部性，2012 年 11 月。77

Farzad Khorasani、Rajiv Gupta 和 Laxmi N. Bhuyan。在存在分歧的情况下通过协作上下文收集实现高效的扭曲执行。在 *Proc. of the ACM/IEEE International Symposium on Microarchitecture (MICRO)*，2015 年。DOI：10.1145/2830772.2830796。45

J. Y. Kim and C. Batten. Accelerating irregular algorithms on GPGPUs using fine-grain hardware worklists. In *Proc. of the ACM/IEEE International Symposium on Microarchitecture (MICRO)*, 2014. DOI: 10.1109/micro.2014.24. 93

Ji Kim, Christopher Torng, Shreesha Srinath, Derek Lockhart, and Christopher Batten. Microarchitectural mechanisms to exploit value structure in SIMT architectures. In *Proc. of the ACM/IEEE International Symposium on Computer Architecture (ISCA)*, 2013. DOI: 10.1145/2508148.2485934. 57, 58, 59, 60, 61, 62

Sangman Kim, Seonggu Huh, Xinya Zhang, Yige Hu, Amir Wated, Emmett Witchel, and Mark Silberstein. GPUnet: Networking abstractions for GPU programs. In *Proc. of the USENIX Symposium on Operating Systems Design and Implementation*, pages 6–8, 2014. DOI: 10.1145/2963098. 2

David B. Kirk and W. Hwu Wen-Mei. *Programming Massively Parallel Processors: A Hands-on Approach*. Morgan Kaufmann, 2016. DOI: 10.1016/c2011-0-04129-7. 9

John Kloosterman, Jonathan Beaumont, D. Anoushe Jamshidi, Jonathan Bailey, Trevor Mudge, and Scott Mahlke. Regless: Just-in-time operand staging for GPUs. In *Proc. of the ACM/IEEE International Symposium on Microarchitecture (MICRO)*, pages 151–164, 2017. DOI: 10.1145/3123939.3123974. 65

R. Krashinsky, C. Batten, M. Hampton, S. Gerding, B. Pharris, J. Casper, and K. Asanovic. The vector-thread architecture. In *Proc. of the ACM/IEEE International Symposium on Computer Architecture (ISCA)*, pages 52–63, June 2004. DOI: 10.1109/isca.2004.1310763. 52

Ronny M. Krashinsky. United States Patent Application #20130042090 A1: Temporal SIMT Execution Optimization, August 2011. 53, 58

David Kroft. Lockup-free instruction fetch/prefetch cache organization. In *Proc. of the ACM/IEEE International Symposium on Computer Architecture (ISCA)*, pages 81–87, 1981. DOI: 10.1145/285930.285979. 33, 71

Jens Krüger and Rüdiger Westermann. Linear algebra operators for GPU implementation of numerical algorithms. In *ACM Transactions on Graphics (TOG)*, V. 22, pages 908–916, 2003. DOI: 10.1145/882262.882363. 6

Junjie Lai and André Seznec. Performance upper bound analysis and optimization of SGEMM on Fermi and Kepler GPUs. In *Proc. of the ACM/IEEE International Symposium on Code Generation and Optimization (CGO)*, pages 1–10, 2013. DOI: 10.1109/cgo.2013.6494986. 16

Nagesh B. Lakshminarayana and Hyesoon Kim. Effect of instruction fetch and memory scheduling on GPU performance. In *Workshop on Language, Compiler, and Architecture Support for GPGPU*, 2010. 88

J. Y. Kim 和 C. Batten。使用细粒度硬件工作列表加速 GPGPU 上的不规则算法。在 *Proc. of the ACM/IEEE International Symposium on Microarchitecture (MI- CRO)*，2014 年。DOI：10.1109/micro.2014.24。93

Ji Kim、Christopher Torng、Shreesha Srinath、Derek Lockhart 和 Christopher Batten。微架构机制利用 SIMT 架构中的值结构。在 *Proc. of the ACM/IEEE International Symposium on Computer Architecture (ISCA)*，2013 年。DOI：10.1145/2508148.2485934。57、58、59、60、61、62

Sangman Kim、Seonggu Huh、Xinya Zhang、Yige Hu、Amir Wated、Emmett Witchel 和 Mark Silberstein。GPUnet：GPU 程序的网络抽象。在 *Proc. of the USENIX Symposium on Operating Systems Design and Implementation*，第 6-8 页，2014 年。DOI：10.1145/2963098。2

大卫·柯克 (David B. Kirk) 和胡文梅 (W. Hwu Wen-Mei)。*Programming Massively Parallel Processors: A Hands-onApproach*。摩根考夫曼，2016。DOI：10.1016/c2011-0-04129-7。9

John Kloosterman、Jonathan Beaumont、D. Anoushe Jamshidi、Jonathan Bailey、Trevor Mudge 和 Scott Mahlke。Regless：GPU 的即时操作数暂存。在 *Proc. of the ACM/IEEE International Symposium on Microarchitecture (MICRO)*，第 151 – 164 页，2017 年。DOI：10.1145/3123939.3123974。65

R. Krashinsky、C. Batten、M. Hampton、S. Gerding、B. Pharris、J. Casper 和 K. Asanovic。矢量线程架构。在 *Proc. of the ACM/IEEE International Symposium on Computer Architecture (ISCA)*，第 52 – 63 页，2004 年 6 月。DOI：10.1109/isca.2004.1310763。52

Ronny M. Krashinsky。美国专利申请 #20130042090 A1：时间 SIMT 执行优化，2011 年 8 月。53、58

David Kroft。无锁定指令提取/预取缓存组织。在 *Proc. of the ACM/IEEE International Symposium on Computer Architecture (ISCA)*，第 81-87 页，1981 年。DOI：10.1145/285930.285979。33、71

Jens Krüger 和 Rüdiger Westermann。用于 GPU 实现数值算法的线性代数运算符。在 *ACM Transactions on Graphics (TOG)*，卷 22，第 908-916 页，2003 年。DOI：10.1145/882262.882363。6

Junjie Lai 和 André Seznec。SGEMM 在 Fermi 和 Kepler GPU 上的性能上限分析和优化。在 *Proc. of the ACM/IEEE International Symposium on Code Generation and Optimization (CGO)*，第 1-10 页，2013 年。DOI：10.1109/cgo.2013.6494986。16

Nagesh B. Lakshminarayana 和 Hyesoon Kim。指令提取和内存调度对 GPU 性能的影响。在 *Workshop on Language, Compiler, and Architecture Support for GPGPU*，2010 年。88

Ahmad Lashgar, Ebad Salehi, and Amirali Baniasadi. A case study in reverse engineering GPGPUs: Outstanding memory handling resources. *ACM SIGARCH Computer Architecture News*, 43(4):15–21, 2016. DOI: 10.1145/2927964.2927968. 34

C. L. Lawson, R. J. Hanson, D. R. Kincaid, and F. T. Krogh. Basic linear algebra subprograms for fortran usage. *ACM Transactions on Mathematical Software*, 5(3):308–323, September 1979. DOI: 10.1145/355841.355848. 10

HyoukJoong Lee, Kevin J. Brown, Arvind K. Sujeeth, Tiark Rompf, and Kunle Olukotun. Locality-aware mapping of nested parallel patterns on GPUs. In *Proc. of the ACM/IEEE International Symposium on Microarchitecture (MICRO)*, 2014a. DOI: 10.1109/micro.2014.23. 94

J. Lee and H. Kim. TAP: A TLP-aware cache management policy for a CPU-GPU heterogeneous architecture. In *Proc. of the IEEE International Symposium on High-Performance Computer Architecture (HPCA)*, 2012. DOI: 10.1109/hpca.2012.6168947. 100

S. Y. Lee and C. J. Wu. Ctrl-C: Instruction-aware control loop based adaptive cache bypassing for GPUs. In *Proc. of the IEEE International Conference on Computer Design (ICCD)*, pages 133–140, 2016. DOI: 10.1109/iccd.2016.7753271. 80

Sangpil Lee, Keunsoo Kim, Gunjae Koo, Hyeran Jeon, Won Woo Ro, and Murali Annavaram. Warped-compression: Enabling power efficient GPUs through register compression. In *Proc. of the ACM/IEEE International Symposium on Computer Architecture (ISCA)*, 2015. DOI: 10.1145/2749469.2750417. 59, 60, 61

Shin-Ying Lee and Carole-Jean Wu. CAWS: Criticality-aware warp scheduling for GPGPU workloads. In *Proc. of the ACM/IEEE International Conference on Parallel Architecture and Compilation Techniques (PACT)*, 2014. DOI: 10.1145/2628071.2628107. 93

Victor W. Lee, Changkyu Kim, Jatin Chhugani, Michael Deisher, Daehyun Kim, Anthony D. Nguyen, Nadathur Satish, Mikhail Smelyanskiy, Srinivas Chennupaty, Per Hammarlund, et al. Debunking the 100X GPU vs. CPU myth: An evaluation of throughput computing on CPU and GPU. In *Proc. of the ACM/IEEE International Symposium on Computer Architecture (ISCA)*, pages 451–460, 2010. DOI: 10.1145/1815961.1816021. 2

Yunsup Lee, Rimas Avizienis, Alex Bishara, Richard Xia, Derek Lockhart, Christopher Batten, and Krste Asanović. Exploring the tradeoffs between programmability and efficiency in data-parallel accelerators. In *Proc. of the ACM/IEEE International Symposium on Computer Architecture (ISCA)*, pages 129–140, 2011. DOI: 10.1145/2000064.2000080. 53

Yunsup Lee, Vinod Grover, Ronny Krashinsky, Mark Stephenson, Stephen W. Keckler, and Krste Asanović. Exploring the design space of SPMD divergence management on data-

Ahmad Lashgar、Ebad Salehi 和 Amirali Baniasadi。GPG-PU 逆向工程案例研究：出色的内存处理资源。*ACM SIGARCH Computer Architecture News*，43(4):15 – 21，2016 年。DOI：10.1145/2927964.2927968。34 C. L. Lawson、R. J. Hanson、D. R. Kincaid 和 F. T. Krogh。Fortran 使用的基本线性代数子程序。*ACM Transactions on Mathematical Software*，5(3):308 – 323，1979 年 9 月。DOI：10.1145/355841.355848。10

HyoukJoong Lee、Kevin J. Brown、Arvind K. Sujeeth、Tiark Rompf 和 Kunle Olukotun。GPU 上嵌套并行模式的局部感知映射。在 *Proc. of the ACM/IEEE International Symposium on Microarchitecture (MICRO)* 中，2014a。DOI：10.1109/micro.2014.23。94

J. Lee 和 H. Kim。TAP：一种用于 CPU-GPU 异构架构的 TLP 感知缓存管理策略。在 *Proc. of the IEEE International Symposium on High-Performance Computer Architecture (HPCA)* 中，2012 年。DOI：10.1109/hpca.2012.6168947。100 S. Y. Lee 和 C. J. Wu。Ctrl-C：基于指令感知控制循环的 GPU 自适应缓存旁路。在 *Proc. of the IEEE International Conference on Computer Design (ICCD)*，第 133 – 140 页，2016 年。DOI：10.1109/iccd.2016.7753271。80 Sangpil Lee、Keunsoo Kim、Gunjae Koo、Hyeran Jeon、Won Woo Ro 和 Murali Annavaram。扭曲压缩：通过寄存器压缩实现节能 GPU。在 *Proc. of the ACM/IEEE International Symposium on Computer Architecture (ISCA)*，2015 年。DOI：10.1145/2749469.2750417。59、60、61 Shin-Ying Lee 和 Carole-Jean Wu。CAWS：GPGPU 工作负载的关键感知扭曲调度。在 *Proc. of the ACM/IEEE International Conference on Parallel Architecture and Compilation Techniques (PACT)*，2014 年。DOI：10.1145/2628071.2628107。93 Victor W. Lee、Changkyu Kim、Jatin Chhugani、Michael Deisher、Daehyun Kim、Anthony D. Nguyen、Nadathur Satish、Mikhail Smelyanskiy、Srinivas Chennupaty、Per Hammarlund 等人。揭穿 GPU 与 CPU 100X 神话：对 CPU 和 GPU 吞吐量计算的评估。在 *Proc. of the ACM/IEEE International Symposium on Computer Architecture (ISCA)*，第 451-460 页，2010 年。DOI：10.1145/1815961.1816021。2

Yunsup Lee、Rimas Avizienis、Alex Bishara、Richard Xia、Derek Lockhart、Christopher Bat-ten 和 Krste Asanovi ć。探索数据并行加速器中可编程性和效率之间的权衡。在 *Proc. of the ACM/IEEE International Symposium on Computer Architecture (ISCA)*，第 129 – 140 页，2011 年。DOI：10.1145/2000064.2000080。53

Yunsup Lee、Vinod Grover、Ronny Krashinsky、Mark Stephenson、Stephen W. Keckler 和 Krste Asanovi ć。探索 SPMD 散度管理在数据上的设计空间

parallel architectures. In *Proc. of the ACM/IEEE International Symposium on Microarchitecture (MICRO)*, 2014b. DOI: 10.1109/micro.2014.48. 55, 56

Jingwen Leng, Tayler Hetherington, Ahmed ElTantawy, Syed Gilani, Nam Sung Kim, Tor M. Aamodt, and Vijay Janapa Reddi. GPUWattch: Enabling energy optimizations in GPGPUs. In *Proc. of the ACM/IEEE International Symposium on Computer Architecture (ISCA)*, pages 487–498, 2013. DOI: 10.1145/2508148.2485964. 6

Adam Levinthal and Thomas Porter. Chap—A SIMD graphics processor. In *Proc. of the ACM International Conference on Computer Graphics and Interactive Techniques (SIGGRAPH)*, pages 77–82, 1984. DOI: 10.1145/800031.808581. 50

Dong Li, Minsoo Rhu, Daniel R. Johnson, Mike O'Connor, Mattan Erez, Doug Burger, Donald S. Fussell, and Stephen W. Redder. Priority-based cache allocation in throughput processors. In *Proc. of the IEEE International Symposium on High-Performance Computer Architecture (HPCA)*, 2015. DOI: 10.1109/hpca.2015.7056024. 82

E. Lindholm, J. Nickolls, S. Oberman, and J. Montrym. NVIDIA Tesla: A unified graphics and computing architecture. *Micro, IEEE*, 28(2):39–55, March–April 2008a. DOI: 10.1109/mm.2008.31. 9

Erik Lindholm, Mark J. Kilgard, and Henry Moreton. A user-programmable vertex engine. In *Proc. of the ACM International Conference on Computer Graphics and Interactive Techniques (SIGGRAPH)*, pages 149–158, 2001. DOI: 10.1145/383259.383274. 6, 21

John Erik Lindholm, Ming Y. Siu, Simon S. Moy, Samuel Liu, and John R. Nickolls. United States Patent #7,339,592: Simulating Multiported Memories Using Lower Port Count Memories (Assignee NVIDIA Corp.), March 2008b. 35, 38

Erik Lindholm et al. United States Patent #9,189,242: Credit-Based Streaming Multiprocessor Warp Scheduling (Assignee NVIDIA Corp.), November 2015. 33, 41

John S. Liptay. Structural aspects of the system/360 model 85, II: The cache. *IBM Systems Journal*, 7(1):15–21, 1968. DOI: 10.1147/sj.71.0015. 70

Z. Liu, S. Gilani, M. Annavaram, and N. S. Kim. G-Scalar: Cost-effective generalized scalar execution architecture for power-efficient GPUs. In *Proc. of the IEEE International Symposium on High-Performance Computer Architecture (HPCA)*, 2017. DOI: 10.1109/hpca.2017.51. 60, 61, 62

Samuel Lui, John Erik Lindholm, Ming Y. Siu, Brett W. Coon, and Stuart F. Oberman. United States Patent Application 11/555,649: Operand Collector Architecture (Assignee NVIDIA Corp.), May 2008. 35

并行架构。在 *Proc. of the ACM/IEEE International Symposium on Microarchitecture (MICRO)*，2014b。DOI：10.1109/micro.2014.48。55、56 Jingwen Leng、Tayler Hetherington、Ahmed ElTantawy、Syed Gilani、Nam Sung Kim、Tor M. Aamodt 和 Vijay Janapa Reddi。GPUWatchch：在 GPG-PU 中实现能量优化。在 *Proc. of the ACM/IEEE International Symposium on Computer Architecture (ISCA)*，第 487-498 页，2013 年。DOI：10.1145/2508148.2485964。6 Adam Levinthal 和 Thomas Porter。章节 — SIMD 图形处理器。在 *Proc. of the ACM International Conference on Computer Graphics and Interactive Techniques (SIGGRAPH)*，第 77-82 页，1984 年。DOI：10.1145/800031.808581。 50 Dong Li、Minsoo Rhu、Daniel R. Johnson、Mike O’Connor、Mattan Erez、Doug Burger、Don- ald S. Fussell 和 Stephen W. Redder。吞吐量处理器中基于优先级的缓存分配。在 *Proc. of the IEEE International Symposium on High-Performance Computer Architecture (HPCA)*，2015 年。DOI：10.1109/hpca.2015.7056024。82 E. Lindholm、J. Nickolls、S. Oberman 和 J. Montrym。NVIDIA Tesla：统一的图形和计算架构。*Micro, IEEE*，28(2):39 – 55，2008 年 3 月至 4 月。DOI：10.1109/mm.2008.31。9 Erik Lindholm、Mark J. Kilgard 和 Henry Moreton。用户可编程的顶点引擎。在 *Proc. of the ACM International Conference on Computer Graphics and Interactive Techniques (SIGGRAPH)*，第 149 – 158 页，2001 年。DOI：10.1145/383259.383274。6、21 John Erik Lindholm、Ming Y. Siu、Simon S. Moy、Samuel Liu 和 John R. Nickolls。美国专利 #7,339,592：使用低端口数内存模拟多端口内存（受让人 NVIDIA Corp.），2008 年 3 月。35、38 Erik Lindholm 等人。美国专利 #9,189,242：基于信用的流式多处理器 Warp 调度（受让人 NVIDIA Corp.），2015 年 11 月。33、41 John S. Liptay。系统/360 模型 85 的结构方面，II：缓存。 *IBM Systems Journal*，7(1):15 – 21，1968 年。DOI：10.1147/sj.71.0015。70 Z. Liu、S. Gilani、M. Annavaram 和 N. S. Kim。G-Scalar：适用于节能 GPU 的经济高效的通用标量执行架构。在 *Proc. of the IEEE International Symposium on High-Performance Computer Architecture (HPCA)*，2017 年。DOI：10.1109/hpca.2017.51。60、61、62 Samuel Lui、John Erik Lindholm、Ming Y. Siu、Brett W. Coon 和 Stuart F. Oberman。美国专利申请 11/555,649：操作数收集器架构（受让人 NVIDIA Corp.），2008 年 5 月。35

Michael D. McCool, Arch D. Robison, and James Reinders. *Structured Parallel Programming: Patterns for Efficient Computation*. Elsevier, 2012. 10

Jiayuan Meng, David Tarjan, and Kevin Skadron. Dynamic warp subdivision for integrated branch and memory divergence tolerance. In *Proc. of the ACM/IEEE International Symposium on Computer Architecture (ISCA)*, pages 235–246, 2010. DOI: 10.1145/1815961.1815992. 30, 48, 90

Alexander L. Minken and Oren Rubinstein. United States Patent #6,629,188: Circuit and Method for Prefetching Data for a Texture Cache (Assignee: NVIDIA Corp.), September 2003. 72

Alexander L. Minkin et al. United States Patent #8,266,383: Cache Miss Processing Usinga Defer/Replay Mechanism (Assignee: NVIDIA Corp.), September 2012. 68, 69, 71

Alexander L. Minkin et al. United States Patent #8,595,425: Configurable Cache for Multiple Clients (Assignee: NVIDIA Corp.), November 2013. 68

Michael Mishkin, Nam Sung Kim, and Mikko Lipasti. Write-after-read hazard prevention in GPGPUSIM. In *Workshop on Deplicating, Deconstructing, and Debunking (WDDD)*, June 2016. 39, 40

John Montrym and Henry Moreton. The geforce 6800. *IEEE Micro*, 25(2):41–51, 2005. DOI: 10.1109/mm.2005.37. 1

Veynu Narasiman, Michael Shebanow, Chang Joo Lee, Rustam Miftakhutdinov, Onur Mutlu, and Yale N. Patt. Improving GPU performance via large warps and two-level warp scheduling. In *Proc. of the ACM/IEEE International Symposium on Microarchitecture (MICRO)*, pages 308–317, 2011. DOI: 10.1145/2155620.2155656. 33, 38, 43, 88, 91

John R. Nickolls and Jochen Reusch. Autonomous SIMD flexibility in the MP-1 and MP-2. In *Proc. of the ACM Symposium on Parallel Algorithms and Architectures (SPAA)*, pages 98–99, 1993. DOI: 10.1145/165231.165244. 9

Cedric Nugteren, Gert-Jan Van den Braak, Henk Corporaal, and Henri Bal. A detailed GPU cache model based on reuse distance theory. In *Proc. of the IEEE International Symposium on High-Performance Computer Architecture (HPCA)*, pages 37–48, 2014. DOI: 10.1109/hpca.2014.6835955. 79

NVIDIA's Next Generation CUDA Compute Architecture: Fermi. NVIDIA, 2009. 16, 46

Nvidia. *NVIDIA tesla V100 GPU architecture*. 2017. 27, 31

NVIDIA Corp. Pascal l1 cache. `https://devtalk.nvidia.com/default/topic/1006066/pascal-l1-cache/?offset=6` 70

Michael D. McCool、Arch D. Robison 和 James Reinders。*Structured Parallel Programming: Patterns for Efficient Computation*。Elsevier，2012 年。10

Jiayuan Meng、David Tarjan 和 Kevin Skadron。动态扭曲细分以实现集成分支和内存发散容差。在 *Proc. of the ACM/IEEE International Symposium on Computer Architecture (ISCA)*，第 235 – 246 页，2010 年。DOI：10.1145/1815961.1815992。30、48、90

Alexander L. Minken 和 Oren Rubinstein。美国专利号 6,629,188：纹理缓存预取数据的电路和方法（受让人：NVIDIA Corp.），2003 年 9 月。72

Alexander L. Minkin 等人，美国专利号 8,266,383：使用延迟/重放机制的缓存未命中处理（受让人：NVIDIA Corp.），2012 年 9 月。68、69、71

Alexander L. Minkin 等人，美国专利号 8,595,425：可配置多客户端缓存（受让人：NVIDIA Corp.），2013 年 11 月。68

Michael Mishkin、Nam Sung Kim 和 Mikko Lipasti。GPGPUSIM 中的“读后写”风险预防。在 *Workshop on Deplicating, Deconstructing, and Debunking (WDDD)* 中，2016 年 6 月。39、40

John Montrym 和 Henry Moreton。 GeForce 6800。*IEEE Micro*，25(2):41 – 51，2005 年。DOI：10.1109/mm.2005.37。1

Veynu Narasiman、Michael Shebanow、Chang Joo Lee、Rustam Miftakhutdinov、Onur Mutlu 和 Yale N. Patt。通过大型 Warp 和两级 Warp 调度提高 GPU 性能。在 *Proc. of the ACM/IEEE International Symposium on Microarchitecture (MICRO)*，第 308 – 317 页，2011 年。DOI：10.1145/2155620.2155656。33、38、43、88、91

John R. Nickolls 和 Jochen Reusch。MP-1 和 MP-2 中的自主 SIMD 灵活性。在 *Proc. of the ACM Symposium on Parallel Algorithms and Architectures (SPAA)*，第 98-99 页，1993 年。DOI：10.1145/165231.165244。9

Cedric Nugteren、Gert-Jan Van den Braak、Henk Corporaal 和 Henri Bal。基于复用距离理论的详细GPU缓存模型。载于 *Proc. of the IEEE International Symposium on High-Performance Computer Architecture (HPCA)*，第 37-48 页，2014 年。DOI：10.1109/h-pca.2014.6835955。 79

NVIDIA's Next Generation CUDA Compute Architecture: Fermi。NVIDIA，2009 年。16，46

Nvidia。*NVIDIA tesla V100 GPU architecture*。2017 年。27、31

NVIDIA Corp. Pascal l1 缓存。`https://devtalk.nvidia.com/default/topic/1006066/pascal-l1-cache/?offset=6` 70

NVIDIA Corp. Inside volta: The world's most advanced data center GPU. https://devblogs.nvidia.com/parallelforall/inside-volta/, May 2017. 1, 17, 26

NVIDIA Corporation. *NVIDIA's Next Generation CUDA Compute Architecture: Kepler TM GK110*, a. 13

NVIDIA Corporation. *NVIDIA GeForce GTX 680*, b. 16

NVIDIA Corporation. *CUDA Binary Utilities*, c. 16

Parallel Thread Execution ISA (Version 6.1). NVIDIA Corporation, CUDA Toolkit 9.1 ed., November 2017. 14

Lars Nyland et al. United States Patent #8,086,806: Systems and Methods for Coalescing Memory Accesses of Parallel Threads (Assignee: NVIDIA Corp.), December 2011. 71

Marek Olszewski, Jeremy Cutler, and J. Gregory Steffan. JudoSTM: A dynamic binary-rewriting approach to software transactional memory. In *Proc. of the ACM/IEEE International Conference on Parallel Architecture and Compilation Techniques (PACT)*, pages 365–375, 2007. DOI: 10.1109/pact.2007.4336226. 96

Jason Jong Kyu Park, Yongjun Park, and Scott Mahlke. Chimera: Collaborative preemption for multitasking on a shared GPU. In *Proc. of the ACM Architectural Support for Programming Languages and Operating Systems (ASPLOS)*, 2015. DOI: 10.1145/2694344.2694346. 92

David A. Patterson and John L. Hennessy. *Computer Organization and Design: The Hardware/-Software Interface*. 2013. 78

Gennady Pekhimenko, Vivek Seshadri, Onur Mutlu, Phillip B. Gibbons, Michael A. Kozuch, and Todd C. Mowry. Base-delta-immediate compression: Practical data compression for on-chip caches. In *Proc. of the ACM/IEEE International Conference on Parallel Architecture and Compilation Techniques (PACT)*, 2012. DOI: 10.1145/2370816.2370870. 60

J. Power, A. Basu, J. Gu, S. Puthoor, B. M. Beckmann, M. D. Hill, S. K. Reinhardt, and D. A. Wood. Heterogeneous system coherence for integrated CPU-GPU systems. In *Proc. of the ACM/IEEE International Symposium on Microarchitecture (MICRO)*, December 2013a. DOI: 10.1145/2540708.2540747. 100

Jason Power, Arkaprava Basu, Junli Gu, Sooraj Puthoor, Bradford M. Beckmann, Mark D. Hill, Steven K. Reinhardt, and David A. Wood. Heterogeneous system coherence for integrated CPU-GPU systems. In *Proc. of the ACM/IEEE International Symposium on Microarchitecture (MICRO)*, pages 457–467, 2013b. DOI: 10.1145/2540708.2540747. 4

NVIDIA Corp. Inside volta：全球最先进的数据中心 GPU。https://devblogs.nvidia.com/parallelforall/inside-volta/，2017 年 5 月。1、17、26 NVIDIA Corporation。*NVIDIA's Next Generation CUDA Compute Architecture: Kepler TM GK110*，a。13 NVIDIA Corporation。*NVIDIA GeForce GTX 680*，b。16 NVIDIA Corporation。*CUDA Binary Utilities*，c。16 *Parallel Thread Execution ISA (Version 6.1)*。NVIDIA Corporation，CUDA Toolkit 9.1 ed.，2017 年 11 月。14 Lars Nyland 等人。美国专利 #8,086,806：用于合并并行线程内存访问的系统和方法（受让人：NVIDIA Corp.），2011 年 12 月。71 Marek Olszewski、Jeremy Cutler 和 J. Gregory Steffan。JudoSTM：一种动态二进制重写软件事务内存的方法。在 *Proc. of the ACM/IEEE International Conference on Parallel Architecture and Compilation Techniques (PACT)*，第 365 – 375 页，2007 年。DOI：10.1109/pact.2007.4336226。96 Jason Jong Kyu Park、Yongjun Park 和 Scott Mahlke。Chimera：共享 GPU 上多任务的协作抢占。在 *Proc. of the ACM Architectural Support for Programming Languages and Operating Systems (ASPLOS)*，2015 年。DOI：10.1145/2694344.2694346。92 David A. Patterson 和 John L. Hennessy。*Computer Organization and Design: The Hardware/- Software Interface*。2013 年。78 Gennady Pekhimenko、Vivek Seshadri、Onur Mutlu、Phillip B. Gibbons、Michael A. Kozuch 和 Todd C. Mowry。基本增量即时压缩：片上缓存的实用数据压缩。在 *Proc. of the ACM/IEEE International Conference on Parallel Architecture and Compilation Techniques (PACT)*，2012 年。DOI：10.1145/2370816.2370870。60 J. Power、A. Basu、J. Gu、S. Puthoor、B. M. Beckmann、M. D. Hill、S. K. Reinhardt 和 D. A. Wood。集成 CPU-GPU 系统的异构系统一致性。在 *Proc. of the ACM/IEEE International Symposium on Microarchitecture (MICRO)*，2013 年 12 月。DOI：10.1145/2540708.2540747。100 Jason Power、Arkaprava Basu、Junli Gu、Sooraj Puthoor、Bradford M. Beckmann、Mark D. Hill、Steven K. Reinhardt 和 David A. Wood。集成 CPU-GPU 系统的异构系统一致性。在 *Proc. of the ACM/IEEE International Symposium on Microarchitecture (MICRO)*，第 457 – 467 页，2013b。DOI：10.1145/2540708.2540747。4

M. K. Qureshi and Y. N. Patt. Utility-based cache partitioning: A low-overhead, high-performance, runtime mechanism to partition shared caches. In *Proc. of the ACM/IEEE International Symposium on Microarchitecture (MICRO)*, pages 423–432, 2006. DOI: 10.1109/micro.2006.49. 101

Xiaowei Ren and Mieszko Lis. Efficient sequential consistency in GPUs via relativistic cache coherence. In *Proc. of the IEEE International Symposium on High-Performance Computer Architecture (HPCA)*, pages 625–636, 2017. DOI: 10.1109/hpca.2017.40. 72

Minsoo Rhu and Mattan Erez. CAPRI: Prediction of compaction-adequacy for handling control-divergence in GPGPU architectures. In *Proc. of the ACM/IEEE International Symposium on Computer Architecture (ISCA)*, pages 61–71, 2012. DOI: 10.1109/isca.2012.6237006. 44

Minsoo Rhu and Mattan Erez. The dual-path execution model for efficient GPU control flow. In *Proc. of the IEEE International Symposium on High-Performance Computer Architecture (HPCA)*, pages 591–602, 2013a. DOI: 10.1109/hpca.2013.6522352. 49

Minsoo Rhu and Mattan Erez. Maximizing SIMD resource utilization in GPGPUs with SIMD lane permutation. In *Proc. of the ACM/IEEE International Symposium on Computer Architecture (ISCA)*, 2013b. DOI: 10.1145/2485922.2485953. 46

Scott Rixner, William J. Dally, Ujval J. Kapasi, Peter Mattson, and John D. Owens. Memory access scheduling. In *Proc. of the ACM/IEEE International Symposium on Computer Architecture (ISCA)*, pages 128–138, 2000. DOI: 10.1109/isca.2000.854384. 77

James Roberts et al. United States Patent #8,234,478: Using Data Cache Array as a Dram Load/Store Buffer, July 2012. 75

Timothy G. Rogers, Mike O'Connor, and Tor M. Aamodt. Cache-conscious wavefront scheduling. In *Proc. of the ACM/IEEE International Symposium on Microarchitecture (MICRO)*, 2012. DOI: 10.1109/micro.2012.16. 33, 78, 79, 88

Timothy G. Rogers, Mike O'Connor, and Tor M. Aamodt. Divergence-aware warp scheduling. In *Proc. of the ACM/IEEE International Symposium on Microarchitecture (MICRO)*, 2013. DOI: 10.1145/2540708.2540718. 79, 90

Timothy G. Rogers, Daniel R. Johnson, Mike O'Connor, and Stephen W. Keckler. A variable warp size architecture. In *Proc. of the ACM/IEEE International Symposium on Computer Architecture (ISCA)*, 2015. DOI: 10.1145/2749469.2750410. 53

Sangmin Seo, Gangwon Jo, and Jaejin Lee. Performance characterization of the NAS parallel benchmarks in OpenCL. In *Proc. of the IEEE International Symposium on Workload Characterization (IISWC)*, pages 137–148, 2011. DOI: 10.1109/iiswc.2011.6114174. 10

M. K. Qureshi 和 Y. N. Patt。基于实用程序的缓存分区：一种低开销、高性能的运行时机制，用于对共享缓存进行分区。在 *Proc. of the ACM/IEEE International Symposium on Microarchitecture (MICRO)*，第 423 – 432 页，2006 年。DOI：10.1109/micro.2006.49。101

任晓伟和 Mieszko Lis。通过相对论缓存一致性实现 GPU 中的高效顺序一致性。在 *Proc. of the IEEE International Symposium on High-Performance Computer Architecture (HPCA)* 中，第 625 – 636 页，2017 年。DOI：10.1109/hpca.2017.40。 72

Minsoo Rhu 和 Mattan Erez。CAPRI：预测 GPGPU 架构中处理控制发散的压缩充分性。在 *Proc. of the ACM/IEEE International Sympo- sium on Computer Architecture (ISCA)*，第 61-71 页，2012 年。DOI：10.1109/isca.2012.6237006。44

Minsoo Rhu 和 Mattan Erez。高效 GPU 控制流的双路径执行模型。在 *Proc. of the IEEE International Symposium on High-Performance Computer Architecture (HPCA)*，第 591 – 602 页，2013a 年。DOI：10.1109/hpca.2013.6522352。49

Minsoo Rhu 和 Mattan Erez。使用 SIMD 通道置换最大化 GPGPU 中的 SIMD 资源利用率。在 *Proc. of the ACM/IEEE International Symposium on Computer Architecture (ISCA)*，2013b。DOI：10.1145/2485922.2485953。46 Scott Rixner、William J. Dally、Ujval J. Kapasi、Peter Mattson 和 John D. Owens。内存访问调度。在 *Proc. of the ACM/IEEE International Symposium on Computer Architecture (ISCA)*，第 128-138 页，2000 年。DOI：10.1109/isca.2000.854384。77 James Roberts 等人。美国专利 #8,234,478：使用数据缓存阵列作为 Dram 加载/存储缓冲区，2012 年 7 月。75

Timothy G. Rogers、Mike O’Connor 和 Tor M. Aamodt。缓存意识波前调度。在 *Proc. of the ACM/IEEE International Symposium on Microarchitecture (MICRO)*，2012 年。DOI：10.1109/micro.2012.16。33、78、79、88

Timothy G. Rogers、Mike O’Connor 和 Tor M. Aamodt。发散感知扭曲调度。在 *Proc. of the ACM/IEEE International Symposium on Microarchitecture (MICRO)*，2013 年。DOI：10.1145/2540708.2540718。79，90

Timothy G. Rogers、Daniel R. Johnson、Mike O’Connor 和 Stephen W. Keckler。可变扭曲尺寸架构。在 *Proc. of the ACM/IEEE International Symposium on Computer Architecture (ISCA)*，2015 年。DOI：10.1145/2749469.2750410。53

Sangmin Seo、Gangwon Jo 和 Jaejin Lee。OpenCL 中 NAS 并行基准的性能表征。在 *Proc. of the IEEE International Symposium on Workload Charac- terization (IISWC)*，第 137 – 148 页，2011 年。DOI：10.1109/iiswc.2011.6114174。10

A. Sethia, D. A. Jamshidi, and S. Mahlke. Mascar: Speeding up GPU warps by reducing memory pitstops. In *Proc. of the IEEE International Symposium on High-Performance Computer Architecture (HPCA)*, pages 174–185, 2015. DOI: 10.1109/hpca.2015.7056031. 91, 92

Ankit Sethia and Scott Mahlke. Equalizer: Dynamic tuning of GPU resources for efficient execution. In *Proc. of the ACM/IEEE International Symposium on Microarchitecture (MICRO)*, 2014. DOI: 10.1109/micro.2014.16. 86

Ryan Shrout. AMD ATI radeon HD 2900 XT review: R600 arrives. *PC Perspective*, May 2007. 75

Mark Silberstein, Bryan Ford, Idit Keidar, and Emmett Witchel. GPUfs: Integrating a file system with GPUs. In *Proc. of the ACM Architectural Support for Programming Languages and Operating Systems (ASPLOS)*, pages 485–498, 2013. DOI: 10.1145/2451116.2451169. 2

Inderpreet Singh, Arrvindh Shriraman, Wilson W. L. Fung, Mike O'Connor, and Tor M. Aamodt. Cache coherence for GPU architectures. In *Proc. of the IEEE International Symposium on High-Performance Computer Architecture (HPCA)*, pages 578–590, 2013. DOI: 10.1109/hpca.2013.6522351. 72

Michael F. Spear, Virendra J. Marathe, William N. Scherer, and Michael L. Scott. Conflict detection and validation strategies for software transactional memory. In *Proc. of the EATCS International Symposium on Distributed Computing*, pages 179–193, Springer-Verlag, 2006. DOI: 10.1007/11864219_13. 99

Michael F. Spear, Maged M. Michael, and Christoph Von Praun. RingSTM: Scalable transactions with a single atomic instruction. In *Proc. of the ACM Symposium on Parallel Algorithms and Architectures (SPAA)*, pages 275–284, 2008. DOI: 10.1145/1378533.1378583. 97

Michael Steffen and Joseph Zambreno. Improving SIMT efficiency of global rendering algorithms with architectural support for dynamic micro-kernels. In *Proc. of the ACM/IEEE International Symposium on Microarchitecture (MICRO)*, pages 237–248, 2010. DOI: 10.1109/micro.2010.45. 45

Ivan E. Sutherland. *Sketchpad a Man-machine Graphical Communication System*. Ph.D. thesis, 1963. DOI: 10.1145/62882.62943. 6

David Tarjan and Kevin Skadron. On demand register allocation and deallocation for a multithreaded processor, June 30, 2011. U.S. Patent App. 12/649,238. 64

Sean J. Treichler et al. United States Patent #9,098,383: Consolidated Crossbar that Supports a Multitude of Traffic Types, August 2015. 75

A. Sethia、D. A. Jamshidi 和 S. Mahlke。Mascar：通过减少内存中断来加速 GPU 扭曲。在 *Proc. of the IEEE International Symposium on High-Performance Computer Architecture (HPCA)*，第 174 – 185 页，2015 年。DOI：10.1109/hpca.2015.7056031。91

Ankit Sethia 和 Scott Mahlke。均衡器：动态调整 GPU 资源以实现高效执行。在 *Proc. of the ACM/IEEE International Symposium on Microarchitecture (MICRO)* 中，2014 年。DOI：10.1109/micro.2014.16。86

Ryan Shrout。AMD ATI radeon HD 2900 XT 评论：R600 上市。*PC Perspective*，2007 年 5 月。75

Mark Silberstein、Bryan Ford、Idit Keidar 和 Emmett Witchel。GPUfs：将文件系统与 GPU 集成。在 *Proc. of the ACM Architectural Support for Programming Languages and Operating Systems (ASPLOS)*，第 485 – 498 页，2013 年。DOI：10.1145/2451116.2451169。2 Inderpreet Singh、Arrvindh Shriraman、Wilson W. L. Fung、Mike O’Connor 和 Tor M. Aamodt。GPU 架构的缓存一致性。在 *Proc. of the IEEE International Symposium on High-Performance Computer Architecture (HPCA)*，第 578 – 590 页，2013 年。DOI：10.1109/hpca.2013.6522351。72 Michael F. Spear、Virendra J. Marathe、William N. Scherer 和 Michael L. Scott。软件事务内存的冲突检测和验证策略。在 *Proc. of the EATCS International Symposium on Distributed Computing*，第 179 – 193 页，Springer-Verlag，2006 年。DOI：10.1007/11864219_13。99 Michael F. Spear、Maged M. Michael 和 Christoph Von Praun。RingSTM：具有单个原子指令的可扩展事务。在 *Proc. of the ACM Symposium on Parallel Algorithms and Architectures (SPAA)*，第 275 – 284 页，2008 年。DOI：10.1145/1378533.1378583。97 Michael Steffin 和 Joseph Zambreno。通过对动态微内核的架构支持提高全局渲染算法的 SIMT 效率。在 *Proc. of the ACM/IEEE Inter- national Symposium on Microarchitecture (MICRO)*，第 237 – 248 页，2010 年。DOI：10.1109/micro.2010.45。45 Ivan E. Sutherland。*Sketchpad a Man-machine Graphical Communication System*。博士论文，1963 年。DOI：10.1145/62882.62943。6 David Tarjan 和 Kevin Skadron。多线程处理器的按需寄存器分配和释放，2011 年 6 月 30 日。美国专利申请 12/649,238。64 Sean J. Treichler 等人。美国专利 #9,098,383：支持多种流量类型的整合交叉开关，2015 年 8 月。75

Dean M. Tullsen, Susan J. Eggers, Joel S. Emer, Henry M. Levy, Jack L. Lo, and Rebecca L. Stamm. Exploiting choice: Instruction fetch and issue on an implementable simultaneous multithreading processor. In *Proc. of the ACM/IEEE International Symposium on Computer Architecture (ISCA)*, 1996. DOI: 10.1145/232973.232993. 88

Rafael Ubal, Byunghyun Jang, Perhaad Mistry, Dana Schaa, and David Kaeli. Multi2Sim: A simulation framework for CPU-GPU computing. In *Proc. of the ACM/IEEE International Conference on Parallel Architecture and Compilation Techniques (PACT)*, pages 335–344, 2012. DOI: 10.1145/2370816.2370865. 17

Aniruddha S. Vaidya, Anahita Shayesteh, Dong Hyuk Woo, Roy Saharoy, and Mani Azimi. SIMD divergence optimization through intra-warp compaction. In *Proc. of the ACM/IEEE International Symposium on Computer Architecture (ISCA)*, pages 368–379, 2013. DOI: 10.1145/2485922.2485954. 44, 46

Wladimir J. van der Lann. Decuda. `http://wiki.github.com/laanwj/decuda/` 14

Jin Wang and Sudhakar Yalamanchili. Characterization and analysis of dynamic parallelism in unstructured GPU applications. In *Proc. of the IEEE International Symposium on Workload Characterization (IISWC)*, pages 51–60, 2014. DOI: 10.1109/iiswc.2014.6983039. 95

Jin Wang, Norm Rubin, Albert Sidelnik, and Sudhakar Yalamanchili. Dynamic thread block launch: A lightweight execution mechanism to support irregular applications on GPUs. In *Proc. of the ACM/IEEE International Symposium on Computer Architecture (ISCA)*, pages 528–540, 2016a. DOI: 10.1145/2749469.2750393. 95

Jin Wang, Norm Rubin, Albert Sidelnik, and Sudhakar Yalamanchili. Laperm: Locality aware scheduler for dynamic parallelism on GPUs. In *Proc. of the ACM/IEEE International Symposium on Computer Architecture (ISCA)*, 2016b. DOI: 10.1109/isca.2016.57. 96

Kai Wang and Calvin Lin. Decoupled affine computation for SIMT GPUs. In *Proc. of the ACM/IEEE International Symposium on Computer Architecture (ISCA)*, 2017. DOI: 10.1145/3079856.3080205. 58, 61, 62

D. Wong, N. S. Kim, and M. Annavaram. Approximating warps with intra-warp operand value similarity. In *Proc. of the IEEE International Symposium on High-Performance Computer Architecture (HPCA)*, 2016. DOI: 10.1109/hpca.2016.7446063. 60, 61

X. Xie, Y. Liang, Y. Wang, G. Sun, and T. Wang. Coordinated static and dynamic cache bypassing for GPUs. In *Proc. of the IEEE International Symposium on High-Performance Computer Architecture (HPCA)*, 2015. DOI: 10.1109/hpca.2015.7056023. 81

Dean M. Tullsen、Susan J. Eggers、Joel S. Emer、Henry M. Levy、Jack L. Lo 和 Rebecca L. Stamm。利用选择：可实现同步多线程处理器上的指令获取和发出。在 *Proc. of the ACM/IEEE International Symposium on Computer Architecture (ISCA)* 中，1996 年。DOI：10.1145/232973.232993。88 Rafael Ubal、Byunghyun Jang、Perhaad Mistry、Dana Schaa 和 David Kaeli。Multi2Sim：用于 CPU-GPU 计算的模拟框架。在 *Proc. of the ACM/IEEE International Conference on Parallel Architecture and Compilation Techniques (PACT)* 中，第 335-344 页，2012 年。DOI：10.1145/2370816.2370865。 17 Aniruddha S. Vaidya、Anahita Shayesteh、Dong Hyuk Woo、Roy Saharoy 和 Mani Az- imi。通过内部 warp 压缩实现 SIMD 发散优化。在 *Proc. of the ACM/IEEE International Symposium on Computer Architecture (ISCA)*，第 368 – 379 页，2013 年。DOI：10.1145/2485922.2485954。44、46 Wladimir J. van der Laan。Decuda。`http://wiki.github.com/laanwj/decuda/` 14 Jin Wang 和 Sudhakar Yalamanchili。非结构化 GPU 应用中动态并行性的特征和分析。在 *Proc. of the IEEE International Symposium on Workload Characterization (IISWC)*，第 51 – 60 页，2014 年。DOI：10.1109/iiswc.2014.6983039。 95 Jin Wang、Norm Rubin、Albert Sidelnik 和 Sudhakar Yalamanchili。动态线程块启动：一种轻量级执行机制，用于支持 GPU 上的不规则应用程序。在 *Proc. of the ACM/IEEE International Symposium on Computer Architecture (ISCA)*，第 528-540 页，2016a。DOI：10.1145/2749469.2750393。95 Jin Wang、Norm Rubin、Albert Sidelnik 和 Sudhakar Yalamanchili。Laperm：用于 GPU 上动态并行的局部感知调度程序。在 *Proc. of the ACM/IEEE International Sympo- sium on Computer Architecture (ISCA)*，2016b。DOI：10.1109/isca.2016.57。96 Kai Wang 和 Calvin Lin。SIMT GPU 的解耦仿射计算。在 *Proc. of the ACM/IEEE International Symposium on Computer Architecture (ISCA)*，2017 年。DOI：10.1145/3079856.3080205。58、61、62 D. Wong、N. S. Kim 和 M. Annavaram。使用内部扭曲操作数值相似性近似扭曲。在 *Proc. of the IEEE International Symposium on High-Performance Computer Architecture (HPCA)*，2016 年。DOI：10.1109/hpca.2016.7446063。60、61 X. Xie、Y. Liang、Y. Wang、G. Sun 和 T. Wang。协调 GPU 的静态和动态缓存旁路。在 *Proc. of the IEEE International Symposium on High-Performance Computer Architecture (HPCA)*，2015 年。DOI：10.1109/hpca.2015.7056023。81

Yunlong Xu, Rui Wang, Nilanjan Goswami, Tao Li, Lan Gao, and Depei Qian. Software transactional memory for GPU architectures. In *Proc. of the ACM/IEEE International Symposium on Code Generation and Optimization (CGO)*, pages 1:1–1:10, 2014. DOI: 10.1145/2581122.2544139. 99

Y. Yang, P. Xiang, M. Mantor, N. Rubin, L. Hsu, Q. Dong, and H. Zhou. A case for a flexible scalar unit in SIMT architecture. In *Proc. of the IEEE International Parallel and Distributed Processing Symposium (IPDPS)*, 2014. DOI: 10.1109/ipdps.2014.21. 47

George L. Yuan, Ali Bakhoda, and Tor M. Aamodt. Complexity effective memory access scheduling for many-core accelerator architectures. In *Proc. of the ACM/IEEE International Symposium on Microarchitecture (MICRO)*, pages 34–44, 2009. DOI: 10.1145/1669112.1669119. 77

Hou Yunqing. Assembler for NVIDIA FERMI. `https://github.com/hyqneuron/asfermi` 16

Eddy Z. Zhang, Yunlian Jiang, Ziyu Guo, and Xipeng Shen. Streamlining GPU applications on the fly: Thread divergence elimination through runtime thread-data remapping. In *Proc. of the ACM International Conference on Supercomputing (ICS)*, pages 115–126, 2010. DOI: 10.1145/1810085.1810104. 45

William K. Zuravleff and Timothy Robinson. U.S. patent #5,630,096: Controller for a synchronous dram that maximizes throughput by allowing memory requests and commands to be issued out of order, May 13, 1997. 77

Yunlong Xu, Rui Wang, Nilanjan Goswami, Tao Li, Lan Gao, and Depei Qian. Soft- ware tra nsactional memory for GPU architectures. In *Proc. of the ACM/IEEE Interna- tional Symposium on Code Generation and Optimization (CGO)*, pages 1:1 – 1:10, 2014. DOI: 1 0.1145/2581122.2544139. 99

Y. Yang、P. Xiang、M. Mantor、N. Rubin、L. Hsu、Q. Dong 和 H. Zhou。SIMT 架构中灵活标量单元的案例。在 *Proc. of the IEEE International Parallel and Distributed Processing Symposium (IPDPS)*，2014 年。DOI：10.1109/ipdps.2014.21。47

George L. Yuan、Ali Bakhoda 和 Tor M. Aamodt。多核加速器架构的复杂度有效内存访问调度。在 *Proc. of the ACM/IEEE International Symposium on Microarchitecture (MICRO)*，第 34-44 页，2009 年。DOI：10.1145/1669112.1669119。77

侯云清。NVIDIA FERMI 汇编程序。`https://github.com/hyqneuron/asfermi` 16 Eddy Z. Zhang、Yunlian Jiang、Ziyu Guo 和 Xipeng Shen。在运行中简化 GPU 应用程序：通过运行时线程数据重新映射消除线程发散。在 *Proc. of the ACM International Conference on Supercomputing (ICS)*，第 115-126 页，2010 年。D OI：10.1145/1810085.1810104。45 William K. Zuravleffi 和 Timothy Robinson。美国专利 #5,630,096：同步 DRAM 控制器，通过允许无序发出内存请求和命令来最大化吞吐量，1997 年 5 月 13 日。77

Authors' Biographies

TOR M. AAMODT

Tor M. Aamodt is a Professor in the Department of Electrical and Computer Engineering at the University of British Columbia, where he has been a faculty member since 2006. His current research focuses on the architecture of general-purpose GPUs and energy-efficient computing, most recently including accelerators for machine learning. Along with students in his research group, he developed the widely used GPGPU-Sim simulator. Three of his papers have been selected as "Top Picks" by *IEEE Micro Magazine*, a fourth was selected as a "Top Picks" honorable mention. One of his papers was also selected as a "Research Highlight" in *Communications of the ACM*. He is in the MICRO Hall of Fame. He served as an Associate Editor for *IEEE Computer Architecture Letters* from 2012–2015 and the *International Journal of High Performance Computing Applications* from 2012–2016, was Program Chair for ISPASS 2013, General Chair for ISPASS 2014, and has served on numerous program committees. He was a Visiting Associate Professor in the Computer Science Department at Stanford University from 2012–2013. He was awarded an NVIDIA Academic Partnership Award in 2010, a NSERC Discovery Accelerator for 2016–2019, and a 2016 Google Faculty Research Award.

Tor received his BASc (in Engineering Science), MASc, and Ph.D. at the University of Toronto. Much of his Ph.D. work was done while he was an intern at Intel's Microarchitecture Research Lab. Subsequently, he worked at NVIDIA on the memory system architecture ("framebuffer") of GeForce 8 Series GPU—the first NVIDIA GPU to support CUDA.

Tor is registered as a Professional Engineer in the province of British Columbia.

WILSON WAI LUN FUNG

Wilson Wai Lun Fung is an architect in Advanced Computing Lab (ACL) as part of Samsung Austin R & D Center (SARC) at Samsung Electronics, where he contributes to the development of a next generation GPU IP. He is interested in both theoretical and practical aspects of computer architecture. Wilson is a winner of the NVIDIA Graduate Fellowship, the NSERC Postgraduate Scholarship, and the NSERC Canada Graduate Scholarship. Wilson was one of the main contributors to the widely used GPGPU-Sim simulator. Two of his papers were selected as a "Top Pick" from computer architecture by *IEEE Micro Magazine*. Wilson received his BASc (in Computer Engineering), MASc, and Ph.D. at the University of British Columbia. During his Ph.D., Wilson interned at NVIDIA.

作者简介

TOR M. AAMODT

Tor M. Aamodt 是不列颠哥伦比亚大学电气与计算机工程系的教授，自 2006 年起担任该系教员。他目前的研究重点是通用 GPU 的架构和节能计算，最近又研究了机器学习加速器。他与研究小组的学生一起开发了广泛使用的 GPGPU-Sim 模拟器。他的三篇论文被 *IEEE Micro Magazine* 选为“ 最佳论文 ”，第四篇被选为“ 最佳论文 ”荣誉奖。他的一篇论文还被 *Communications of the ACM.* 选为“ 研究亮点 ”。他已入选 MICRO 名人堂。他曾于 2012 年至 2015 年担任 *IEEE Computer Architecture Letters* 副主编，并于 2012 年至 2016 年担任 *International Journal of High Performance Computing Applications* 副主编，还曾担任 ISPASS 2013 计划主席、ISPASS 2014 总主席，并曾担任多个计划委员会成员。他于 2012 年至 2013 年担任斯坦福大学计算机科学系客座副教授。他曾于 2010 年获得 NVIDIA 学术合作伙伴奖，2016 年至 2019 年获得 NSERC 发现加速器奖，并于 2016 年获得 Google 教师研究奖。

Tor 在多伦多大学获得了理学学士（工程科学）、理学硕士和博士学位。他的大部分博士研究工作都是在英特尔微架构研究实验室实习期间完成的。随后，他在 NVIDIA 工作，负责 GeForce 8 系列 GPU 的内存系统架构（“ 帧缓冲区 ”）——这是第一款支持 CUDA 的 NVIDIA GPU。

Tor 是一名不列颠哥伦比亚省注册的专业工程师。

冯伟伦

Wilson Wai Lun Fung 是三星电子奥斯汀研发中心 (SARC) 高级计算实验室 (ACL) 的架构师，他为下一代 GPU IP 的开发做出了贡献。他对计算机架构的理论和实践方面都很感兴趣。Wilson 是 NVIDIA 研究生奖学金、NSERC 研究生奖学金和 NSERC 加拿大研究生奖学金的获得者。Wilson 是广泛使用的 GPGPU-Sim 模拟器的主要贡献者之一。他的两篇论文被 *IEEE Micro Magazine.* 选为计算机架构“ 首选 ”Wilson 在英属哥伦比亚大学获得了学士（计算机工程专业）、硕士和博士学位。在攻读博士学位期间，Wilson 在 NVIDIA 实习。

TIMOTHY G. ROGERS

Timothy G. Rogers is an Assistant Professor in the Electrical and Computer Engineering department at Purdue University, where his research focuses on massively multithreaded processor design. He is interested in exploring computer systems and architectures that improve both programmer productivity and energy efficiency. Timothy is a winner of the NVIDIA Graduate Fellowship and the NSERC Alexander Graham Bell Canada Graduate Scholarship. His work has been selected as a "Top Pick" from computer architecture by *IEEE Micro Magazine* and as a "Research Highlight" in *Communications of the ACM.* During his Ph.D., Timothy interned at NVIDIA Research and AMD Research. Prior to attending graduate school, Timothy worked as a software engineer at Electronic Arts and received his BEng in Electrical Engineering from McGill University.

蒂莫西·G·罗杰斯

蒂莫西·G·罗杰斯 (Timothy G. Rogers) 是普渡大学电气与计算机工程系的助理教授，他的研究重点是大规模多线程处理器设计。他对探索能够提高程序员工作效率和能源效率的计算机系统和架构很感兴趣。蒂莫西是 NVIDIA 研究生奖学金和 NSERC Alexander Graham Bell 加拿大研究生奖学金的获得者。他的作品被 *IEEE Micro Magazine* 选为计算机架构类别的“首选”，并被 *Communications of the ACM.* 选为“研究亮点”。在攻读博士学位期间，蒂莫西在 NVIDIA 研究中心和 AMD 研究中心实习。在进入研究生院之前，蒂莫西曾在 Electronic Arts 担任软件工程师，并在麦吉尔大学获得电气工程学士学位。