1. Record Nr. UNINA9910966759003321

Autore	Cheng John
Titolo	Professional CUDA C Programming
Pubbl/distr/stampa	Hoboken, : Wiley, 2014
ISBN	9781118739273 1118739272
Edizione	[1st ed.]
Descrizione fisica	1 online resource (527 p.)
Altri autori (Persone)	GrossmanMax McKercherTy
Disciplina	004.35 004/.35
Soggetti	Computer architecture Multiprocessors Parallel processing (Electronic computers) Parallel programming (Computer science) Engineering & Applied Sciences Computer Science
Lingua di pubblicazione	Inglese
Formato	Materiale a stampa
Livello bibliografico	Monografia
Note generali	Description based upon print version of record.
Nota di contenuto	Cover; Title Page; Copyright; Contents; Chapter 1 Heterogeneous Parallel Computing with CUDA; Parallel Computing; Sequential and Parallel Programming; Parallelism; Computer Architecture; Heterogeneous Computing; Heterogeneous Architecture; Paradigm of Heterogeneous Computing; CUDA: A Platform for Heterogeneous Computing; Hello World from GPU; Is CUDA C Programming Difficult?; Summary; Chapter 2 CUDA Programming Model; Introducing the CUDA Programming Model; CUDA Programming Structure; Managing Memory; Organizing Threads; Launching a CUDA Kernel; Writing Your Kernel; Verifying Your Kernel Handling ErrorsCompiling and Executing; Timing Your Kernel; Timing with CPU Timer; Timing with nvprof; Organizing Parallel Threads; Indexing Matrices with Blocks and Threads; Summing Matrices with a 2D Grid and 2D Blocks; Summing Matrices with a 1D Grid and 1D Blocks; Summing Matrices with a 2D Grid and 1D Blocks; Managing Devices; Using the Runtime API to Query GPU Information; Determining

the Best GPU; Using nvidia-smi to Query GPU Information; Setting Devices at Runtime; Summary; Chapter 3 CUDA Execution Model; Introducing the CUDA Execution Model; GPU Architecture Overview
The Fermi ArchitectureThe Kepler Architecture; Profile-Driven Optimization; Understanding the Nature of Warp Execution; Warps and Thread Blocks; Warp Divergence; Resource Partitioning; Latency Hiding; Occupancy; Synchronization; Scalability; Exposing Parallelism; Checking Active Warps with nvprof; Checking Memory Operations with nvprof; Exposing More Parallelism; Avoiding Branch Divergence; The Parallel Reduction Problem; Divergence in Parallel Reduction; Improving Divergence in Parallel Reduction; Reducing with Interleaved Pairs; Unrolling Loops; Reducing with Unrolling
Reducing with Unrolled WarpsReducing with Complete Unrolling; Reducing with Template Functions; Dynamic Parallelism; Nested Execution; Nested Hello World on the GPU; Nested Reduction; Summary; Chapter 4 Global Memory; Introducing the CUDA Memory Model; Benefits of a Memory Hierarchy; CUDA Memory Model; Memory Management; Memory Allocation and Deallocation; Memory Transfer; Pinned Memory; Zero-Copy Memory; Unified Virtual Addressing; Unified Memory; Memory Access Patterns; Aligned and Coalesced Access; Global Memory Reads; Global Memory Writes; Array of Structures versus Structure of Arrays
Performance TuningWhat Bandwidth Can a Kernel Achieve?; Memory Bandwidth; Matrix Transpose Problem; Matrix Addition with Unified Memory; Summary; Chapter 5 Shared Memory and Constant Memory; Introducing CUDA Shared Memory; Shared Memory; Shared Memory Allocation; Shared Memory Banks and Access Mode; Configuring the Amount of Shared Memory; Synchronization; Checking the Data Layout of Shared Memory; Square Shared Memory; Rectangular Shared Memory; Reducing Global Memory Access; Parallel Reduction with Shared Memory; Parallel Reduction with Unrolling
Parallel Reduction with Dynamic Shared Memory

Sommario/riassunto

Break into the powerful world of parallel GPU programming with this down-to-earth, practical guide Designed for professionals across multiple industrial sectors, Professional CUDA C Programming presents CUDA -- a parallel computing platform and programming model designed to ease the development of GPU programming -- fundamentals in an easy-to-follow format, and teaches readers how to think in parallel and implement parallel algorithms on GPUs. Each chapter covers a specific topic, and includes workable examples that demonstrate the development process, allowing readers to explore both the "