Programming Massively Parallel Processors A Hands On Approach Applications Of Gpu Computing Series 1st First Edition By David B Kirk Wen Mei W Hwu Published By Morgan Kaufmann 2010

This book constitutes the refereed proceedings of the 7th International Conference on Applied Parallel Computing, PARA 2004, held in June 2004. The 118 revised full papers presented together with five invited lectures and 15 contributed talks were carefully reviewed and selected for inclusion in the proceedings. The papers are organized in topical sections.

Parallel and High Performance Computing offers techniques guaranteed to boost your code's effectiveness. Summary Complex calculations, like training deep learning models or running large-scale simulations, can take an extremely long time. Efficient parallel programming can save hours—or even days—of computing time. Parallel and High Performance Computing shows you how to deliver faster run-times, greater scalability, and increased energy efficiency to your programs by mastering parallel techniques for multicore processor and GPU hardware. About the technology Write fast, powerful, energy efficient programs that scale to tackle huge volumes of data. Using parallel programming, your code spreads data processing tasks across multiple CPUs for radically better performance. With a little help, you can create software that maximizes both speed and efficiency. About the book Parallel and High Performance Computing offers techniques guaranteed to boost your code's effectiveness. You'll learn to evaluate hardware architectures and work with industry standard tools such as OpenMP and MPI. You'll master the data structures and algorithms best suited for high performance computing and learn techniques that save energy on handheld devices. You'll even run a massive tsunami simulation across a bank of GPUs. What's inside Planning a new parallel project Understanding differences in CPU and GPU architecture Addressing underperforming kernels and loops Managing applications with batch scheduling About the reader For experienced programmers proficient with a high-performance computing language like C, C++, or Fortran. About the author Robert Robey works at Los Alamos National Laboratory and has been active in the field of parallel computing for over 30 years. Yuliana Zamora is currently a PhD student and Siebel Scholar at the University of Chicago, and has lectured on programming modern hardware at numerous national conferences. Table of Contents PART 1 INTRODUCTION TO PARALLEL COMPUTING 1 Why parallel computing? 2 Planning for parallelization 3 Performance limits and profiling 4 Data design and performance models 5 Parallel algorithms and patterns PART 2 CPU: THE PARALLEL WORKHORSE 6 Vectorization: FLOPs for free 7 OpenMP that performs 8 MPI: The parallel backbone PART 3 GPUS: BUILT TO ACCELERATE 9 GPU architectures and concepts 10 GPU programming model 11 Directive-based GPU

programming 12 GPU languages: Getting down to basics 13 GPU profiling and tools PART 4 HIGH PERFORMANCE COMPUTING ECOSYSTEMS 14 Affinity: Truce with the kernel 15 Batch schedulers: Bringing order to chaos 16 File operations for a parallel world 17 Tools and resources for better code

This edited book aims to present the state of the art in research and development of the convergence of high-performance computing and parallel programming for various engineering and scientific applications. The book has consolidated algorithms, techniques, and methodologies to bridge the gap between the theoretical foundations of academia and implementation for research, which might be used in business and other real-time applications in the future.The book outlines techniques and tools used for emergent areas and domains, which include acceleration of large-scale electronic structure simulations with heterogeneous parallel computing, characterizing power and energy efficiency of a data-centric high-performance computing runtime and applications, security applications of GPUs, parallel implementation of multiprocessors on MPI using FDTD, particle-based fused rendering, design and implementation of particle systems for mesh-free methods with high performance, and evolving topics on heterogeneous computing. In the coming days the need to converge HPC, IoT, cloud-based applications will be felt and this volume tries to bridge that gap.

If you have a working knowledge of Haskell, this hands-on book shows you how to use the language's many APIs and frameworks for writing both parallel and concurrent programs. You'll learn how parallelism exploits multicore processors to speed up computation-heavy programs, and how concurrency enables you to write programs with threads for multiple interactions. Author Simon Marlow walks you through the process with lots of code examples that you can run, experiment with, and extend. Divided into separate sections on Parallel and Concurrent Haskell, this book also includes exercises to help you become familiar with the concepts presented: Express parallelism in Haskell with the Eval monad and Evaluation Strategies Parallelize ordinary Haskell code with the Par monad Build parallel array-based computations, using the Repa library Use the Accelerate library to run computations directly on the GPU Work with basic interfaces for writing concurrent code Build trees of threads for larger and more complex programs Learn how to build high-speed concurrent network servers Write distributed programs that run on multiple machines in a network

Authors Jim Jeffers and James Reinders spent two years helping educate customers about the prototype and pre-production hardware before Intel introduced the first Intel Xeon Phi coprocessor. They have distilled their own experiences coupled with insights from many expert customers, Intel Field Engineers, Application Engineers and Technical Consulting Engineers, to create this authoritative first book on the essentials of programming for this new architecture and these new products. This book is useful even before you ever touch a system with an Intel Xeon Phi coprocessor. To ensure that your applications run at maximum efficiency, the authors emphasize key techniques

for programming any modern parallel computing system whether based on Intel Xeon processors, Intel Xeon Phi coprocessors, or other high performance microprocessors. Applying these techniques will generally increase your program performance on any system, and better prepare you for Intel Xeon Phi coprocessors and the Intel MIC architecture. A practical guide to the essentials of the Intel Xeon Phi coprocessor Presents best practices for portable, high-performance computing and a familiar and proven threaded, scalar-vector programming model Includes simple but informative code examples that explain the unique aspects of this new highly parallel and high performance computational product Covers wide vectors, many cores, many threads and high bandwidth cache/memory architecture

Build real-world applications with Python 2.7, CUDA 9, and CUDA 10. We suggest the use of Python 2.7 over Python 3.x, since Python 2.7 has stable support across all the libraries we use in this book. Key Features Expand your background in GPU programming—PyCUDA, scikit-cuda, and Nsight Effectively use CUDA libraries such as cuBLAS, cuFFT, and cuSolver Apply GPU programming to modern data science applications Book Description Hands-On GPU Programming with Python and CUDA hits the ground running: you'll start by learning how to apply Amdahl's Law, use a code profiler to identify bottlenecks in your Python code, and set up an appropriate GPU programming environment. You'll then see how to "query" the GPU's features and copy arrays of data to and from the GPU's own memory. As you make your way through the book, you'll launch code directly onto the GPU and write full blown GPU kernels and device functions in CUDA C. You'll get to grips with profiling GPU code effectively and fully test and debug your code using Nsight IDE. Next, you'll explore some of the more well-known NVIDIA libraries, such as cuFFT and cuBLAS. With a solid background in place, you will now apply your new-found knowledge to develop your very own GPU-based deep neural network from scratch. You'll then explore advanced topics, such as warp shuffling, dynamic parallelism, and PTX assembly. In the final chapter, you'll see some topics and applications related to GPU programming that you may wish to pursue, including AI, graphics, and blockchain. By the end of this book, you will be able to apply GPU programming to problems related to data science and high-performance computing. What you will learn Launch GPU code directly from Python Write effective and efficient GPU kernels and device functions Use libraries such as cuFFT, cuBLAS, and cuSolver Debug and profile your code with Nsight and Visual Profiler Apply GPU programming to datascience problems Build a GPU-based deep neuralnetwork from scratch Explore advanced GPU hardware features, such as warp shuffling Who this book is for Hands-On GPU Programming with Python and CUDA is for developers and data scientists who want to learn the basics of effective GPU programming to improve performance using Python code. You should have an understanding of first-year college or university-level engineering mathematics and physics, and have some

experience with Python as well as in any C-based programming language such as C, C++, Go, or Java.

Advancements in microprocessor architecture, interconnection technology, and software development have fueled rapid growth in parallel and distributed computing. However, this development is only of practical benefit if it is accompanied by progress in the design, analysis and programming of parallel algorithms. This concise textbook provides, in one place, three mainstream parallelization approaches, Open MPP, MPI and OpenCL, for multicore computers, interconnected computers and graphical processing units. An overview of practical parallel computing and principles will enable the reader to design efficient parallel programs for solving various computational problems on state-of-the-art personal computers and computing clusters. Topics covered range from parallel algorithms, programming tools, OpenMP, MPI and OpenCL, followed by experimental measurements of parallel programs' run-times, and by engineering analysis of obtained results for improved parallel execution performances. Many examples and exercises support the exposition.

The contributions of a diverse selection of international hardware and software specialists are assimilated in this book's exploration of the development of massively parallel processing (MPP). The emphasis is placed on industrial applications and collaboration with users and suppliers from within the industrial community consolidates the scope of the publication. From a practical point of view, massively parallel data processing is a vital step to further innovation in all areas where large amounts of data must be processed in parallel or in a distributed manner, e.g. fluid dynamics, meteorology, seismics, molecular engineering, image processing, parallel data base processing. MPP technology can make the speed of computation higher and substantially reduce the computational costs. However, to achieve these features, the MPP software has to be developed further to create user-friendly programming systems and to become transparent for present-day computer software. Application of novel electro-optic components and devices is continuing and will be a key for much more general and powerful architectures. Vanishing of communication hardware limitations will result in the elimination of programming bottlenecks in parallel data processing. Standardization of the functional characteristics of a programming model of massively parallel computers will become established. Then efficient programming environments can be developed. The result will be a widespread use of massively parallel processing systems in many areas of application.

Intel® Xeon PhiTM Coprocessor Architecture and Tools: The Guide for Application Developers provides developers a comprehensive introduction and in-depth look at the Intel Xeon Phi coprocessor architecture and the corresponding parallel data structure tools and algorithms used in the various technical computing applications for which it is suitable. It also examines the source code-level optimizations that can be performed to exploit the powerful features of the processor. Xeon Phi is at the heart of world's fastest commercial supercomputer,

which thanks to the massively parallel computing capabilities of Intel Xeon Phi processors coupled with Xeon Phi coprocessors attained 33.86 teraflops of benchmark performance in 2013. Extracting such stellar performance in real-world applications requires a sophisticated understanding of the complex interaction among hardware components, Xeon Phi cores, and the applications running on them. In this book, Rezaur Rahman, an Intel leader in the development of the Xeon Phi coprocessor and the optimization of its applications, presents and details all the features of Xeon Phi core design that are relevant to the practice of application developers, such as its vector units, hardware multithreading, cache hierarchy, and host-to-coprocessor communication channels. Building on this foundation, he shows developers how to solve real-world technical computing problems by selecting, deploying, and optimizing the available algorithms and data structure alternatives matching Xeon Phi's hardware characteristics. From Rahman's practical descriptions and extensive code examples, the reader will gain a working knowledge of the Xeon Phi vector instruction set and the Xeon Phi microarchitecture whereby cores execute 512-bit instruction streams in parallel.

The CUDA Handbook begins where CUDA by Example (Addison-Wesley, 2011) leaves off, discussing CUDA hardware and software in greater detail and covering both CUDA 5.0 and Kepler. Every CUDA developer, from the casual to the most sophisticated, will find something here of interest and immediate usefulness. Newer CUDA developers will see how the hardware processes commands and how the driver checks progress; more experienced CUDA developers will appreciate the expert coverage of topics such as the driver API and context migration, as well as the guidance on how best to structure CPU/GPU data interchange and synchronization. The accompanying open source code–more than 25,000 lines of it, freely available at www.cudahandbook.com–is specifically intended to be reused and repurposed by developers. Designed to be both a comprehensive reference and a practical cookbook, the text is divided into the following three parts: Part I, Overview, gives high-level descriptions of the hardware and software that make CUDA possible. Part II, Details, provides thorough descriptions of every aspect of CUDA, including Memory Streams and events Models of execution, including the dynamic parallelism feature, new with CUDA 5.0 and SM 3.5 The streaming multiprocessors, including descriptions of all features through SM 3.5 Programming multiple GPUs Texturing The source code accompanying Part II is presented as reusable microbenchmarks and microdemos, designed to expose specific hardware characteristics or highlight specific use cases. Part III, Select Applications, details specific families of CUDA applications and key parallel algorithms, including Streaming workloads Reduction Parallel prefix sum (Scan) N-body Image Processing These algorithms cover the full range of potential CUDA applications.

This collection of articles documents the design of one such computer, a single

instruction multiple data stream (SIMD) class supercomputer with 16,834 processing units capable of over 6 billion 8 bit operations per second.

A complete source of information on almost all aspects of parallel computing from introduction, to architectures, to programming paradigms, to algorithms, to programming standards. It covers traditional Computer Science algorithms, scientific computing algorithms and data intensive algorithms.

Innovations in hardware architecture, like hyper-threading or multicore processors, mean that parallel computing resources are available for inexpensive desktop computers. In only a few years, many standard software products will be based on concepts of parallel programming implemented on such hardware, and the range of applications will be much broader than that of scientific computing, up to now the main application area for parallel computing. Rauber and Rünger take up these recent developments in processor architecture by giving detailed descriptions of parallel programming techniques that are necessary for developing efficient programs for multicore processors as well as for parallel cluster systems and supercomputers. Their book is structured in three main parts, covering all areas of parallel computing: the architecture of parallel systems, parallel programming models and environments, and the implementation of efficient application algorithms. The emphasis lies on parallel programming techniques needed for different architectures. For this second edition, all chapters have been carefully revised. The chapter on architecture of parallel systems has been updated considerably, with a greater emphasis on the architecture of multicore systems and adding new material on the latest developments in computer architecture. Lastly, a completely new chapter on general-purpose GPUs and the corresponding programming techniques has been added. The main goal of the book is to present parallel programming techniques that can be used in many situations for a broad range of application areas and which enable the reader to develop correct and efficient parallel programs. Many examples and exercises are provided to show how to apply the techniques. The book can be used as both a textbook for students and a reference book for professionals. The material presented has been used for courses in parallel programming at different universities for many years.

Multicore and GPU Programming offers broad coverage of the key parallel computing skillsets: multicore CPU programming and manycore "massively parallel" computing. Using threads, OpenMP, MPI, and CUDA, it teaches the design and development of software capable of taking advantage of today’s computing platforms incorporating CPU and GPU hardware and explains how to transition from sequential programming to a parallel computing paradigm. Presenting material refined over more than a decade of teaching parallel computing, author Gerassimos Barlas minimizes the challenge with multiple examples, extensive case studies, and full source code. Using this book, you can develop programs that run over distributed memory machines using MPI, create multi-threaded applications with either libraries or directives, write optimized

applications that balance the workload between available computing resources, and profile and debug programs targeting multicore machines. Comprehensive coverage of all major multicore programming tools, including threads, OpenMP, MPI, and CUDA Demonstrates parallel programming design patterns and examples of how different tools and paradigms can be integrated for superior performance Particular focus on the emerging area of divisible load theory and its impact on load balancing and distributed systems Download source code, examples, and instructor support materials on the book's companion website Programming Massively Parallel Processors: A Hands-on Approach, Second Edition, teaches students how to program massively parallel processors. It offers a detailed discussion of various techniques for constructing parallel programs. Case studies are used to demonstrate the development process, which begins with computational thinking and ends with effective and efficient parallel programs. This guide shows both student and professional alike the basic concepts of parallel programming and GPU architecture. Topics of performance, floating-point format, parallel patterns, and dynamic parallelism are covered in depth. This revised edition contains more parallel programming examples, commonly-used libraries such as Thrust, and explanations of the latest tools. It also provides new coverage of CUDA 5.0, improved performance, enhanced development tools, increased hardware support, and more; increased coverage of related technology, OpenCL and new material on algorithm patterns, GPU clusters, host programming, and data parallelism; and two new case studies (on MRI reconstruction and molecular visualization) that explore the latest applications of CUDA and GPUs for scientific research and high-performance computing. This book should be a valuable resource for advanced students, software engineers, programmers, and hardware engineers. New coverage of CUDA 5.0, improved performance, enhanced development tools, increased hardware support, and more Increased coverage of related technology, OpenCL and new material on algorithm patterns, GPU clusters, host programming, and data parallelism Two new case studies (on MRI reconstruction and molecular visualization) explore the latest applications of CUDA and GPUs for scientific research and high-performance computing.

The Parallel Programming Guide for Every Software Developer From grids and clusters to next-generation game consoles, parallel computing is going mainstream. Innovations such as Hyper-Threading Technology, HyperTransport Technology, and multicore microprocessors from IBM, Intel, and Sun are accelerating the movement's growth. Only one thing is missing: programmers with the skills to meet the soaring demand for parallel software. That's where Patterns for Parallel Programming comes in. It's the first parallel programming guide written specifically to serve working software developers, not just computer scientists. The authors introduce a complete, highly accessible pattern language that will help any experienced developer "think parallel"-and start writing effective parallel code almost immediately. Instead of formal theory, they deliver proven

solutions to the challenges faced by parallel programmers, and pragmatic guidance for using today's parallel APIs in the real world. Coverage includes: Understanding the parallel computing landscape and the challenges faced by parallel developers Finding the concurrency in a software design problem and decomposing it into concurrent tasks Managing the use of data across tasks Creating an algorithm structure that effectively exploits the concurrency you've identified Connecting your algorithmic structures to the APIs needed to implement them Specific software constructs for implementing parallel programs Working with today's leading parallel programming environments: OpenMP, MPI, and Java Patterns have helped thousands of programmers master object-oriented development and other complex programming technologies. With this book, you will learn that they're the best way to master parallel programming too.
GPUs can be used for much more than graphics processing. As opposed to a CPU, which can only run four or five threads at once, a GPU is made up of hundreds or even thousands of individual, low-powered cores, allowing it to perform thousands of concurrent operations. Because of this, GPUs can tackle large, complex problems on a much shorter time scale than CPUs. Dive into parallel programming on NVIDIA hardware with CUDA by Chris Rose, and learn the basics of unlocking your graphics card. This updated and expanded second edition of Book provides a user-friendly introduction to the subject, Taking a clear structural framework, it guides the reader through the subject's core elements. A flowing writing style combines with the use of illustrations and diagrams throughout the text to ensure the reader understands even the most complex of concepts. This succinct and enlightening overview is a required reading for all those interested in the subject . We hope you find this book useful in shaping your future career & Business.
Intended to anyone interested in numerical computing and data science: students, researchers, teachers, engineers, analysts, hobbyists... Basic knowledge of Python/NumPy is recommended. Some skills in mathematics will help you understand the theory behind the computational methods.
Programming is now parallel programming. Much as structured programming revolutionized traditional serial programming decades ago, a new kind of structured programming, based on patterns, is relevant to parallel programming today. Parallel computing experts and industry insiders Michael McCool, Arch Robison, and James Reinders describe how to design and implement maintainable and efficient parallel algorithms using a pattern-based approach. They present both theory and practice, and give detailed concrete examples using multiple programming models. Examples are primarily given using two of the most popular and cutting edge programming models for parallel programming: Threading Building Blocks, and Cilk Plus. These architecture-independent models enable easy integration into existing applications, preserve investments in existing code, and speed the development of parallel applications. Examples from realistic contexts illustrate patterns and themes in parallel

algorithm design that are widely applicable regardless of implementation technology. The patterns-based approach offers structure and insight that developers can apply to a variety of parallel programming models Develops a composable, structured, scalable, and machine-independent approach to parallel computing Includes detailed examples in both Cilk Plus and the latest Threading Building Blocks, which support a wide variety of computers

Integrating associative processing concepts with massively parallel SIMD technology, this volume explores a model for accessing data by content rather than abstract address mapping.

Machine generated contents note: 1. How to think in CUDA 2. Tools to build, debug and profile 3. The GPU performance envelope 4. The CUDA memory subsystems 5. Exploiting the CUDA execution grid 6. MultiGPU applications and scaling 7. Numerical CUDA, libraries and high-level language bindings 8. Mixing CUDA with rendering 9. High Performance Machine Learning 10. Scientific Visualization 11. Multimedia with OpenCV 12. Ultra Low-power Devices: Tegra.

CUDA for Engineers gives you direct, hands-on engagement with personal, high-performance parallel computing, enabling you to do computations on a gaming-level PC that would have required a supercomputer just a few years ago. The authors introduce the essentials of CUDA C programming clearly and concisely, quickly guiding you from running sample programs to building your own code. Throughout, you'll learn from complete examples you can build, run, and modify, complemented by additional projects that deepen your understanding. All projects are fully developed, with detailed building instructions for all major platforms. Ideal for any scientist, engineer, or student with at least introductory programming experience, this guide assumes no specialized background in GPU-based or parallel computing. In an appendix, the authors also present a refresher on C programming for those who need it. Coverage includes Preparing your computer to run CUDA programs Understanding CUDA's parallelism model and C extensions Transferring data between CPU and GPU Managing timing, profiling, error handling, and debugging Creating 2D grids Interoperating with OpenGL to provide real-time user interactivity Performing basic simulations with differential equations Using stencils to manage related computations across threads Exploiting CUDA's shared memory capability to enhance performance Interacting with 3D data: slicing, volume rendering, and ray casting Using CUDA libraries Finding more CUDA resources and code Realistic example applications include Visualizing functions in 2D and 3D Solving differential equations while changing initial or boundary conditions Viewing/processing images or image stacks Computing inner products and centroids Solving systems of linear algebraic equations Monte-Carlo computations

Today all computers, from tablet/desktop computers to super computers, work in parallel. A basic knowledge of the architecture of parallel computers and how to program them, is thus, essential for students of computer science and IT professionals. In its second edition, the book retains the lucidity of the first edition

and has added new material to reflect the advances in parallel computers. It is designed as text for the final year undergraduate students of computer science and engineering and information technology. It describes the principles of designing parallel computers and how to program them. This second edition, while retaining the general structure of the earlier book, has added two new chapters, ‘Core Level Parallel Processing’ and ‘Grid and Cloud Computing’ based on the emergence of parallel computers on a single silicon chip popularly known as multicore processors and the rapid developments in Cloud Computing. All chapters have been revised and some chapters are re-written to reflect the emergence of multicore processors and the use of MapReduce in processing vast amounts of data. The new edition begins with an introduction to how to solve problems in parallel and describes how parallelism is used in improving the performance of computers. The topics discussed include instruction level parallel processing, architecture of parallel computers, multicore processors, grid and cloud computing, parallel algorithms, parallel programming, compiler transformations, operating systems for parallel computers, and performance evaluation of parallel computers.

This book outlines a set of issues that are critical to all of parallel architecture--communication latency, communication bandwidth, and coordination of cooperative work (across modern designs). It describes the set of techniques available in hardware and in software to address each issues and explore how the various techniques interact.

Programming Massively Parallel Processors discusses the basic concepts of parallel programming and GPU architecture. Various techniques for constructing parallel programs are explored in detail. Case studies demonstrate the development process, which begins with computational thinking and ends with effective and efficient parallel programs. This book describes computational thinking techniques that will enable students to think about problems in ways that are amenable to high-performance parallel computing. It utilizes CUDA (Compute Unified Device Architecture), NVIDIA's software development tool created specifically for massively parallel environments. Studies learn how to achieve both high-performance and high-reliability using the CUDA programming model as well as OpenCL. This book is recommended for advanced students, software engineers, programmers, and hardware engineers. Teaches computational thinking and problem-solving techniques that facilitate high-performance parallel computing. Utilizes CUDA (Compute Unified Device Architecture), NVIDIA's software development tool created specifically for massively parallel environments. Shows you how to achieve both high-performance and high-reliability using the CUDA programming model as well as OpenCL.

Programming Massively Parallel Processors: A Hands-on Approach, Second Edition, teaches students how to program massively parallel processors. It offers a detailed discussion of various techniques for constructing parallel programs. Case studies are used to demonstrate the development process, which begins with computational thinking and ends with effective and efficient parallel programs. This guide shows both student and professional alike the basic concepts of parallel programming and GPU

architecture. Topics of performance, floating-point format, parallel patterns, and dynamic parallelism are covered in depth. This revised edition contains more parallel programming examples, commonly-used libraries such as Thrust, and explanations of the latest tools. It also provides new coverage of CUDA 5.0, improved performance, enhanced development tools, increased hardware support, and more; increased coverage of related technology, OpenCL and new material on algorithm patterns, GPU clusters, host programming, and data parallelism; and two new case studies (on MRI reconstruction and molecular visualization) that explore the latest applications of CUDA and GPUs for scientific research and high-performance computing. This book should be a valuable resource for advanced students, software engineers, programmers, and hardware engineers. New coverage of CUDA 5.0, improved performance, enhanced development tools, increased hardware support, and more Increased coverage of related technology, OpenCL and new material on algorithm patterns, GPU clusters, host programming, and data parallelism Two new case studies (on MRI reconstruction and molecular visualization) explore the latest applications of CUDA and GPUs for scientific research and high-performance computing

CUDA is a computing architecture designed to facilitate the development of parallel programs. In conjunction with a comprehensive software platform, the CUDA Architecture enables programmers to draw on the immense power of graphics processing units (GPUs) when building high-performance applications. GPUs, of course, have long been available for demanding graphics and game applications. CUDA now brings this valuable resource to programmers working on applications in other domains, including science, engineering, and finance. No knowledge of graphics programming is required—just the ability to program in a modestly extended version of C. CUDA by Example, written by two senior members of the CUDA software platform team, shows programmers how to employ this new technology. The authors introduce each area of CUDA development through working examples. After a concise introduction to the CUDA platform and architecture, as well as a quick-start guide to CUDA C, the book details the techniques and trade-offs associated with each key CUDA feature. You'll discover when to use each CUDA C extension and how to write CUDA software that delivers truly outstanding performance. Major topics covered include Parallel programming Thread cooperation Constant memory and events Texture memory Graphics interoperability Atomics Streams CUDA C on multiple GPUs Advanced atomics Additional CUDA resources All the CUDA software tools you'll need are freely available for download from NVIDIA. http://developer.nvidia.com/object/cuda-by-example.html

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 "hard" and "soft" aspects of GPU programming. Computing architectures are experiencing a fundamental shift toward scalable parallel computing motivated by application requirements in industry and science. This book demonstrates

the challenges of efficiently utilizing compute resources at peak performance, presents modern techniques for tackling these challenges, while increasing accessibility for professionals who are not necessarily parallel programming experts. The CUDA programming model and tools empower developers to write high-performance applications on a scalable, parallel computing platform: the GPU. However, CUDA itself can be difficult to learn without extensive programming experience. Recognized CUDA authorities John Cheng, Max Grossman, and Ty McKercher guide readers through essential GPU programming skills and best practices in Professional CUDA C Programming, including: CUDA Programming Model GPU Execution Model GPU Memory model Streams, Event and Concurrency Multi-GPU Programming CUDA Domain-Specific Libraries Profiling and Performance Tuning The book makes complex CUDA concepts easy to understand for anyone with knowledge of basic software development with exercises designed to be both readable and high-performance. For the professional seeking entrance to parallel computing and the high-performance computing community, Professional CUDA C Programming is an invaluable resource, with the most current information available on the market.

Parallel Programming: Concepts and Practice provides an upper level introduction to parallel programming. In addition to covering general parallelism concepts, this text teaches practical programming skills for both shared memory and distributed memory architectures. The authors' open-source system for automated code evaluation provides easy access to parallel computing resources, making the book particularly suitable for classroom settings. Covers parallel programming approaches for single computer nodes and HPC clusters: OpenMP, multithreading, SIMD vectorization, MPI, UPC++ Contains numerous practical parallel programming exercises Includes access to an automated code evaluation tool that enables students the opportunity to program in a web browser and receive immediate feedback on the result validity of their program Features an example-based teaching of concept to enhance learning outcomes

Programming Massively Parallel ProcessorsA Hands-On ApproachMorgan Kaufmann

GPU Parallel Program Development using CUDA teaches GPU programming by showing the differences among different families of GPUs. This approach prepares the reader for the next generation and future generations of GPUs. The book emphasizes concepts that will remain relevant for a long time, rather than concepts that are platform-specific. At the same time, the book also provides platform-dependent explanations that are as valuable as generalized GPU concepts. The book consists of three separate parts; it starts by explaining parallelism using CPU multi-threading in Part I. A few simple programs are used to demonstrate the concept of dividing a large task into multiple parallel sub-tasks and mapping them to CPU threads. Multiple ways of parallelizing the same task are analyzed and their pros/cons are studied in terms of both core and memory operation. Part II of the book introduces GPU massive parallelism. The same programs are parallelized on multiple Nvidia GPU platforms and the same performance analysis is repeated. Because the core and memory structures of CPUs and GPUs are different, the results differ in interesting ways. The end goal is to make programmers aware of all the good ideas, as well as the bad ideas, so readers can apply the good ideas and avoid the bad ideas in their own programs. Part III of the book provides pointer for readers who want to expand their horizons. It provides a brief introduction to popular CUDA libraries (such as cuBLAS, cuFFT, NPP, and Thrust),the

OpenCL programming language, an overview of GPU programming using other programming languages and API libraries (such as Python, OpenCV, OpenGL, and Apple's Swift and Metal,) and the deep learning library cuDNN.
Heterogeneous Computing with OpenCL 2.0 teaches OpenCL and parallel programming for complex systems that may include a variety of device architectures: multi-core CPUs, GPUs, and fully-integrated Accelerated Processing Units (APUs). This fully-revised edition includes the latest enhancements in OpenCL 2.0 including: • Shared virtual memory to increase programming flexibility and reduce data transfers that consume resources • Dynamic parallelism which reduces processor load and avoids bottlenecks • Improved imaging support and integration with OpenGL Designed to work on multiple platforms, OpenCL will help you more effectively program for a heterogeneous future. Written by leaders in the parallel computing and OpenCL communities, this book explores memory spaces, optimization techniques, extensions, debugging and profiling. Multiple case studies and examples illustrate high-performance algorithms, distributing work across heterogeneous systems, embedded domain-specific languages, and will give you hands-on OpenCL experience to address a range of fundamental parallel algorithms. Updated content to cover the latest developments in OpenCL 2.0, including improvements in memory handling, parallelism, and imaging support Explanations of principles and strategies to learn parallel programming with OpenCL, from understanding the abstraction models to thoroughly testing and debugging complete applications Example code covering image analytics, web plugins, particle simulations, video editing, performance optimization, and more
Programming Massively Parallel Processors: A Hands-on Approach, Third Edition shows both student and professional alike the basic concepts of parallel programming and GPU architecture, exploring, in detail, various techniques for constructing parallel programs. Case studies demonstrate the development process, detailing computational thinking and ending with effective and efficient parallel programs. Topics of performance, floating-point format, parallel patterns, and dynamic parallelism are covered in-depth. For this new edition, the authors have updated their coverage of CUDA, including coverage of newer libraries, such as CuDNN, moved content that has become less important to appendices, added two new chapters on parallel patterns, and updated case studies to reflect current industry practices. Teaches computational thinking and problem-solving techniques that facilitate high-performance parallel computing Utilizes CUDA version 7.5, NVIDIA's software development tool created specifically for massively parallel environments Contains new and updated case studies Includes coverage of newer libraries, such as CuDNN for Deep Learning.
An Introduction to Parallel Programming, Second Edition presents a tried-and-true tutorial approach that shows students how to develop effective parallel programs with MPI, Pthreads and OpenMP. As the first undergraduate text to directly address compiling and running parallel programs on multi-core and cluster architecture, this second edition carries forward its clear explanations for designing, debugging and evaluating the performance of distributed and shared-memory programs while adding coverage of accelerators via new content on GPU programming and heterogeneous programming. New and improved user-friendly exercises teach students how to compile, run and modify example programs. Takes a tutorial approach, starting with small programming examples and building progressively to more challenging examples

Explains how to develop parallel programs using MPI, Pthreads and OpenMP programming models A robust package of online ancillaries for instructors and students includes lecture slides, solutions manual, downloadable source code, and an image bank New to this edition: New chapters on GPU programming and heterogeneous programming New examples and exercises related to parallel algorithms
Programming Massively Parallel Processors: A Hands-on Approach, Third Edition shows both student and professional alike the basic concepts of parallel programming and GPU architecture, exploring, in detail, various techniques for constructing parallel programs. Case studies demonstrate the development process, detailing computational thinking and ending with effective and efficient parallel programs. Topics of performance, floating-point format, parallel patterns, and dynamic parallelism are covered in-depth. For this new edition, the authors have updated their coverage of CUDA, including coverage of newer libraries, such as CuDNN, moved content that has become less important to appendices, added two new chapters on parallel patterns, and updated case studies to reflect current industry practices. Teaches computational thinking and problem-solving techniques that facilitate high-performance parallel computing Utilizes CUDA version 7.5, NVIDIA's software development tool created specifically for massively parallel environments Contains new and updated case studies Includes coverage of newer libraries, such as CuDNN for Deep Learning
If you need to learn CUDA but don't have experience with parallel computing, CUDA Programming: A Developer's Introduction offers a detailed guide to CUDA with a grounding in parallel fundamentals. It starts by introducing CUDA and bringing you up to speed on GPU parallelism and hardware, then delving into CUDA installation. Chapters on core concepts including threads, blocks, grids, and memory focus on both parallel and CUDA-specific issues. Later, the book demonstrates CUDA in practice for optimizing applications, adjusting to new hardware, and solving common problems. Comprehensive introduction to parallel programming with CUDA, for readers new to both Detailed instructions help readers optimize the CUDA software development kit Practical techniques illustrate working with memory, threads, algorithms, resources, and more Covers CUDA on multiple hardware platforms: Mac, Linux and Windows with several NVIDIA chipsets Each chapter includes exercises to test reader knowledge
Foreword by Bjarne Stroustrup Software is generally acknowledged to be the single greatest obstacle preventing mainstream adoption of massively-parallel computing. While sequential applications are routinely ported to platforms ranging from PCs to mainframes, most parallel programs only ever run on one type of machine. One reason for this is that most parallel programming systems have failed to insulate their users from the architectures of the machines on which they have run. Those that have been platform-independent have usually also had poor performance. Many researchers now believe that object-oriented languages may offer a solution. By hiding the architecture-specific constructs required for high performance inside platform-independent abstractions, parallel object-oriented programming systems may be able to combine the speed of massively-parallel computing with the comfort of sequential programming. Parallel Programming Using C++ describes fifteen parallel programming systems based on C++, the most popular object-oriented language of today. These systems cover the whole spectrum of parallel programming paradigms, from data parallelism through dataflow and distributed shared memory to message-passing control parallelism. For

the parallel programming community, a common parallel application is discussed in each chapter, as part of the description of the system itself. By comparing the implementations of the polygon overlay problem in each system, the reader can get a better sense of their expressiveness and functionality for a common problem. For the systems community, the chapters contain a discussion of the implementation of the various compilers and runtime systems. In addition to discussing the performance of polygon overlay, several of the contributors also discuss the performance of other, more substantial, applications. For the research community, the contributors discuss the motivations for and philosophy of their systems. As well, many of the chapters include critiques that complete the research arc by pointing out possible future research directions. Finally, for the object-oriented community, there are many examples of how encapsulation, inheritance, and polymorphism can be used to control the complexity of developing, debugging, and tuning parallel software.