Cuda By Example An Introduction To General Purpose Gpu Programming Portable Documents

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.

Still more useful techniques, tips, and tricks for harnessing the power of the new generation of powerful GPUs.

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.

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

A comprehensive overview of OpenMP, the standard application programming interface for shared memory parallel computing—a reference for students and professionals. "I hope that readers will learn to use the full expressibility and power of OpenMP. This book should provide an excellent introduction to beginners, and the performance section should help those with some experience who want to push OpenMP to its limits." —from the foreword by David J. Kuck, Intel Fellow, Software and Solutions Group, and Director, Parallel and Distributed Solutions, Intel Corporation OpenMP, a portable programming interface for shared memory parallel computers, was adopted as an informal standard in 1997 by computer scientists who wanted a unified model on which to base programs for shared memory systems. OpenMP is now used by many software developers; it offers significant advantages over both hand-threading and MPI. Using OpenMP offers a comprehensive introduction to parallel programming concepts and a detailed overview of OpenMP. Using OpenMP discusses hardware developments, describes where OpenMP is applicable, and compares OpenMP to other programming interfaces for shared and distributed memory parallel architectures. It introduces the individual features of OpenMP, provides many source code examples that demonstrate the use and functionality of the language constructs, and offers tips on writing an efficient OpenMP program. It describes how to use OpenMP in full-scale applications to achieve high performance on large-scale architectures, discussing several case studies in detail, and offers in-depth troubleshooting advice. It explains how OpenMP is translated into explicitly multithreaded code, providing a valuable behind-the-scenes account of OpenMP program performance. Finally, Using OpenMP considers trends likely to influence OpenMP development, offering a glimpse of the possibilities of a future OpenMP 3.0 from the vantage point of the current OpenMP 2.5. With multicore computer use increasing, the need for a comprehensive introduction and overview of the standard interface is clear. Using OpenMP provides an essential reference not only for students at both undergraduate and graduate levels but also for professionals who intend to parallelize existing codes or develop new parallel programs for shared memory computer architectures.

CUDA by ExampleAn Introduction to General-Purpose GPU Programming, Portable DocumentsAddison-Wesley Professional

CUDA Fortran for Scientists and Engineers shows how high-performance application developers can leverage the power of GPUs using Fortran, the familiar language of scientific computing and supercomputer performance benchmarking. The authors presume no prior parallel computing experience, and cover the basics along with best practices for efficient GPU computing using CUDA Fortran. To help you add CUDA Fortran to existing Fortran codes, the book explains how to understand the target GPU architecture, identify computationally intensive parts of the code, and modify the code to manage the data and parallelism and optimize performance. All of this is done in Fortran, without having to rewrite in another language. Each concept is illustrated with actual examples so you can immediately evaluate the performance of your code in comparison. Leverage the power of GPU computing with PGI's CUDA Fortran compiler Gain insights from members of the CUDA Fortran language development team Includes multi-GPU programming in CUDA Fortran, covering both peer-to-peer and message passing interface (MPI) approaches Includes full source code for all the examples and several case studies Download source code and slides from the book's companion website

The creation of ever more realistic 3-D images is central to the development of computer graphics. The ray tracing technique has become one of the most popular and powerful means by which photo-realistic images can now be created. The simplicity, elegance and ease of implementation makes ray tracing an essential part of understanding and exploiting state-of-the-art computer graphics. An Introduction to Ray Tracing develops from fundamental principles to advanced applications, providing "how-to" procedures as well as a detailed understanding of

the scientific foundations of ray tracing. It is also richly illustrated with four-color and black-and-white plates. This is a book which will be welcomed by all concerned with modern computer graphics, image processing, and computer-aided design. Provides practical "how-to" information Contains high quality color plates of images created using ray tracing techniques Progresses from a basic understanding to the advanced science and application of ray tracing

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 Discover how CUDA allows OpenCV to handle complex and rapidly growing image data processing in computer and machine vision by accessing the power of GPU Key Features Explore examples to leverage the GPU processing power with OpenCV and CUDA Enhance the performance of algorithms on embedded hardware platforms Discover C++ and Python libraries for GPU acceleration Book Description Computer vision has been revolutionizing a wide range of industries, and OpenCV is the most widely chosen tool for computer vision with its ability to work in multiple programming languages. Nowadays, in computer vision, there is a need to process large images in real time, which is difficult to handle for OpenCV on its own. This is where CUDA comes into the picture, allowing OpenCV to leverage powerful NVDIA GPUs. This book provides a detailed overview of integrating OpenCV with CUDA for practical applications. To start with, you'll understand GPU programming with CUDA, an essential aspect for computer vision developers who have never worked with GPUs. You'll then move on to exploring OpenCV acceleration with GPUs and CUDA by walking through some practical examples. Once you have got to grips with the core concepts, you'll familiarize yourself with deploying OpenCV applications on NVIDIA Jetson TX1, which is popular for computer vision and deep learning applications. The last chapters of the book explain PyCUDA, a Python library that leverages the power of CUDA and GPUs for accelerations and can be used by computer vision developers who use OpenCV with Python. By the end of this book, you'll have enhanced computer vision applications with the help of this book's hands-on approach. What you will learn Understand how to access GPU device properties and capabilities from CUDA programs Learn how to accelerate searching and sorting algorithms Detect shapes such as lines and circles in images Explore object tracking and detection with algorithms Process videos using different video analysis techniques in Jetson TX1 Access GPU device properties from the PyCUDA program Understand how kernel execution works Who this book is for This book is a go-to guide for you if you are a developer working with OpenCV and want to learn how to process more complex image data by exploiting GPU processing. A thorough understanding of computer vision concepts and programming languages such as C++ or Python is expected.

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

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

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

Extremely low-cost graphics cards now possess computational capabilities that were once limited to supercomputers. Using CUDA, you can to liberate the power of NVIDIA graphics cards for a wide spectrum of non-graphics applications. CUDA for Engineers is the first guide specifically focused on using CUDA to write high-performance engineering and scientific applications. Ideal for any scientist, engineer, or student with at least introductory programming experience, this tutorial presents examples and reusable C code to jumpstart a wide variety of applications. You'll walk through moving from serial to parallel computation; computing values of a function in parallel; understanding 2D parallelism; simulating dynamics in the phase plane; simulating heat conduction; interacting with 3D data; walking through a basic N-body simulation, and more. Written by a working engineer, this comfortable and conversational guide focuses on practical knowledge you need to solve real engineering and scientific problems with CUDA - at a small fraction of what it would have cost just a few years ago.

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

"GPU Programming in MATLAB" is intended for scientists, engineers, or students who develop or maintain applications in MATLAB and would like to accelerate their codes using GPU programming without losing the many benefits of MATLAB. The book starts with coverage of the Parallel Computing Toolbox and other MATLAB toolboxes for GPU Computing, which allow applications to be ported straightforwardly onto GPUs without specialize knowledge of GPU programming. Following sections cover built-in GPU-enabled features of MATLAB, including options to leverage GPUs across multicore or different computer systems. Finally, advanced material, including CUDA code in MATLAB and optimizing existing GPU applications are presented. Throughout the book, examples and source code illustrate every concept so that readers can immediately apply them to their own development. Provides in-depth, comprehensive coverage of GPUs with MATLAB, including the parallel computing toolbox and built-in features for other MATLAB toolboxesExplains how to accelerate computationally heavy applications in MATLAB without the need to re-write them in another language Presents case studies illustrating key concepts across multiple fieldsIncludes source code, sample datasets, and lecture slides

Explore different GPU programming methods using libraries and directives, such as OpenACC, with extension to languages such as C, C++, and Python Key Features Learn parallel programming principles and practices and performance analysis in GPU computing Get to grips with distributed multi GPU programming and other approaches to GPU programming Understand how GPU acceleration in deep learning models can improve their performance Book Description Compute Unified Device Architecture (CUDA) is NVIDIA's GPU computing platform and application programming interface. It's designed to work with programming languages such as C, C++, and Python. With CUDA, you can leverage a GPU's parallel computing power for a range of high-performance computing applications in the fields of science, healthcare, and deep learning. Learn CUDA Programming will help you learn GPU parallel programming and understand its modern applications. In this book, you'll discover CUDA programming approaches for modern GPU architectures. You'll not only be guided through GPU features, tools, and APIs, you'll also learn how to analyze performance with sample parallel programming algorithms. This book will help you optimize the performance of your apps by giving insights into CUDA programming platforms with various libraries, compiler directives (OpenACC), and other languages. As you progress, you'll learn how additional computing power can be generated using multiple GPUs in a box or in multiple boxes. Finally, you'll explore how CUDA accelerates deep learning algorithms, including convolutional neural networks (CNNs) and recurrent neural networks (RNNs). By the end of this CUDA book, you'll be equipped with the skills you need to integrate the power of GPU computing in your applications. What you will learn Understand general GPU operations and programming patterns in CUDA Uncover the difference between GPU programming and CPU programming Analyze GPU application performance and implement optimization strategies Explore GPU programming, profiling, and debugging tools Grasp parallel programming algorithms and how to implement them Scale GPU-accelerated applications with multi-GPU and multi-nodes Delve into GPU programming platforms with accelerated libraries, Python, and OpenACC Gain insights into deep learning accelerators in CNNs and RNNs using GPUs Who this book is for This beginner-level book is for programmers who want to delve into parallel computing, become part of the high-performance computing community and build modern applications. Basic C and C++ programming experience is assumed. For deep learning enthusiasts, this book covers Python InterOps, DL libraries, and practical examples on performance estimation.

Master Powerful Off-the-Shelf Business Solutions for AI and Machine Learning Pragmatic AI will help you solve real-world problems with contemporary machine learning, artificial intelligence, and cloud computing tools. Noah Gift demystifies all the concepts and tools you need to get results—even if you don’t have a strong background in math or data science. Gift illuminates powerful off-the-shelf cloud offerings from Amazon, Google, and Microsoft, and demonstrates proven techniques using the Python data science ecosystem. His workflows and examples help you streamline and simplify every step, from deployment to production, and build exceptionally scalable solutions. As you learn how machine language (ML) solutions work, you’ll gain a more intuitive understanding of what you can achieve with them and how to maximize their value. Building on these fundamentals, you’ll walk step-by-step through building cloud-based AI/ML applications to address realistic issues in sports marketing, project management, product pricing, real estate, and beyond. Whether you’re a business professional, decision-maker, student, or programmer, Gift’s expert guidance and wide-ranging case studies will prepare you to solve data science problems in virtually any environment. Get and configure all the tools you’ll need Quickly review all the Python you need to start building machine learning applications Master the AI and ML toolchain and project lifecycle Work with Python data science tools such as IPython, Pandas, Numpy, Juypter Notebook, and Sklearn Incorporate a pragmatic feedback loop that continually improves the efficiency of your workflows and systems Develop cloud AI solutions with Google Cloud Platform, including TPU, Colaboratory, and Datalab services Define Amazon Web Services cloud AI workflows, including spot instances, code pipelines, boto, and more Work with Microsoft Azure AI APIs Walk through building six real-world AI applications, from start to finish Register your book for convenient access to downloads, updates, and/or corrections as they become available. See inside book for details.

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

Parallel Programming with OpenACC is a modern, practical guide to implementing dependable computing systems. The book explains how anyone can use OpenACC to quickly ramp-up application performance using high-level code directives called pragmas. The OpenACC directive-based programming model is designed to provide a simple, yet powerful, approach to accelerators without significant programming effort. Author Rob Farber, working with a team of expert contributors, demonstrates how to turn existing applications into portable GPU accelerated programs that demonstrate immediate speedups. The book also helps users get the most from the latest NVIDIA and AMD GPU plus multicore CPU architectures (and soon for Intel® Xeon PhiTM as well). Downloadable example codes provide hands-on OpenACC experience for common problems in scientific, commercial, big-data, and real-time systems. Topics include writing reusable code, asynchronous capabilities, using libraries, multicore clusters, and much more. Each chapter explains how a specific aspect of OpenACC technology fits, how it works, and the pitfalls to avoid. Throughout, the book demonstrates how the use of simple working examples that can be adapted to solve application needs. Presents the simplest way to leverage GPUs to achieve application speedups Shows how OpenACC works, including working examples that can be adapted for application needs Allows readers to download source code and slides from the book's companion web page

This book follows an example-driven, simplified, and practical approach to using OpenCL for general purpose GPU programming. If you are a beginner in parallel programming and would like to quickly accelerate your algorithms using OpenCL, this book is perfect for you! You will find the diverse topics and case studies in this book interesting and informative. You will only require a good knowledge of C programming for this book, and an understanding of parallel implementations will be useful, but not necessary.

Explore GPU-enabled programmable environment for machine learning, scientific applications, and gaming using PuCUDA, PyOpenGL, and Anaconda Accelerate Key Features Understand effective synchronization strategies for faster processing using GPUs Write parallel processing scripts with PyCuda and PyOpenCL Learn to use the CUDA libraries like CuDNN for deep learning on GPUs Book Description GPUs are proving to be excellent general purpose-parallel computing solutions for high performance tasks such as deep learning and scientific computing. This book will be your guide to getting started with GPU computing. It will start with introducing GPU computing and explain the architecture and programming models for GPUs. You will learn, by example, how to perform GPU programming with Python, and you’ll look at using integrations such as PyCUDA, PyOpenCL, CuPy and Numba with Anaconda for various tasks such as machine learning and data mining. Going further, you will get to grips with GPU work flows, management, and deployment using modern containerization solutions. Toward the end of the book, you will get familiar with the principles of distributed computing for training machine learning models and enhancing efficiency and performance. By the end of this book, you will be able to set up a GPU ecosystem for running complex applications and data models that demand great processing capabilities, and be able to efficiently manage memory to compute your application effectively and quickly. What you will learn Utilize Python libraries and frameworks for GPU acceleration Set up a GPU-enabled programmable machine learning environment on your system with Anaconda Deploy your machine learning system on cloud containers with illustrated examples Explore PyCUDA and PyOpenCL and compare them with platforms such as CUDA, OpenCL and ROCm. Perform data mining tasks with machine learning models on GPUs Extend your knowledge of GPU computing in scientific applications Who this book is for Data Scientist, Machine Learning enthusiasts and professionals who wants to get started with GPU computation and perform the complex tasks with low-latency. Intermediate knowledge of Python programming is assumed.

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.

Thought-provoking and accessible in approach, this updated and expanded second edition of the CUDA by Example: An Introduction to General-Purpose GPU Programming 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 advanced graduate-level students. We hope you find this book useful in shaping your future career. Feel free to send us your enquiries related to our publications to info@risepress.pw Rise Press

More useful techniques, tips, and tricks for harnessing the power of the new generation of powerful GPUs.

Cg is a complete programming environment for the fast creation of special effects and real-time cinematic quality experiences on multiple platforms. This text provides a guide to the Cg graphics language.

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.

Based on a course developed by the author, Introduction to High Performance Scientific Computing introduces methods for adding parallelism to numerical methods for solving differential equations. It contains exercises and programming projects that facilitate learning as well as examples and discussions based on the C programming language, with additional comments for those already familiar with C++. The text provides an overview of concepts and algorithmic techniques for modern scientific computing and is divided into six self-contained parts that can be assembled in any order to create an introductory course using available computer hardware. Part I introduces the C programming language for those not already familiar with programming in a compiled language. Part II describes parallelism on shared memory architectures using OpenMP. Part III details parallelism on computer clusters using MPI for coordinating a computation. Part IV demonstrates the use of graphical programming units (GPUs) to solve problems using the CUDA language for NVIDIA graphics cards. Part V addresses programming on GPUs for non-NVIDIA graphics cards using the OpenCL framework. Finally, Part VI contains a brief discussion of numerical methods and applications, giving the reader an opportunity to test the methods on typical computing problems.

Discover the essential building blocks of the most common forms of deep belief networks. At each step this book provides intuitive motivation, a summary of the most important equations relevant to the topic, and concludes with highly commented code for threaded computation on modern CPUs as well as massive parallel processing on computers with CUDA-capable video display cards. The first of three in a series on C++ and CUDA C deep learning and belief nets, Deep Belief Nets in C++ and CUDA C: Volume 1 shows you how the structure of these elegant models is much closer to that of human brains than traditional neural networks; they have a thought process that is capable of learning abstract concepts built from simpler primitives. As such, you'll see that a typical deep belief net can learn to recognize complex patterns by optimizing millions of parameters, yet this model can still be resistant to overfitting. All the routines and algorithms presented in the book are available in the code download, which also contains some libraries of related routines. What You Will Learn Employ deep learning using C++ and CUDA C Work with supervised feedforward networks Implement restricted Boltzmann machines Use generative samplings Discover why these are important Who This Book Is For Those who have at least a basic knowledge of neural networks and some prior programming experience, although some C++ and CUDA C is recommended.

Every other day we hear about new ways to put deep learning to good use: improved medical imaging, accurate credit card fraud detection, long range weather forecasting, and more. PyTorch puts these superpowers in your hands, providing a comfortable Python experience that gets you started quickly and then grows with you as you—and your deep learning skills—become more sophisticated. Deep Learning with PyTorch will make that journey engaging and fun. Summary Every other day we hear about new ways to put deep learning to good use: improved medical imaging, accurate credit card fraud detection, long range weather forecasting, and more. PyTorch puts these superpowers

in your hands, providing a comfortable Python experience that gets you started quickly and then grows with you as you—and your deep learning skills—become more sophisticated. Deep Learning with PyTorch will make that journey engaging and fun. Foreword by Soumith Chintala, Cocreator of PyTorch. Purchase of the print book includes a free eBook in PDF, Kindle, and ePub formats from Manning Publications. About the technology Although many deep learning tools use Python, the PyTorch library is truly Pythonic. Instantly familiar to anyone who knows PyData tools like NumPy and scikit-learn, PyTorch simplifies deep learning without sacrificing advanced features. It’s excellent for building quick models, and it scales smoothly from laptop to enterprise. Because companies like Apple, Facebook, and JPMorgan Chase rely on PyTorch, it’s a great skill to have as you expand your career options. It’s easy to get started with PyTorch. It minimizes cognitive overhead without sacrificing the access to advanced features, meaning you can focus on what matters the most - building and training the latest and greatest deep learning models and contribute to making a dent in the world. PyTorch is also a snap to scale and extend, and it partners well with other Python tooling. PyTorch has been adopted by hundreds of deep learning practitioners and several first-class players like FAIR, OpenAI, FastAI and Purdue. About the book Deep Learning with PyTorch teaches you to create neural networks and deep learning systems with PyTorch. This practical book quickly gets you to work building a real-world example from scratch: a tumor image classifier. Along the way, it covers best practices for the entire DL pipeline, including the PyTorch Tensor API, loading data in Python, monitoring training, and visualizing results. After covering the basics, the book will take you on a journey through larger projects. The centerpiece of the book is a neural network designed for cancer detection. You'll discover ways for training networks with limited inputs and start processing data to get some results. You'll sift through the unreliable initial results and focus on how to diagnose and fix the problems in your neural network. Finally, you'll look at ways to improve your results by training with augmented data, make improvements to the model architecture, and perform other fine tuning. What's inside Training deep neural networks Implementing modules and loss functions Utilizing pretrained models from PyTorch Hub Exploring code samples in Jupyter Notebooks About the reader For Python programmers with an interest in machine learning. About the author Eli Stevens had roles from software engineer to CTO, and is currently working on machine learning in the self-driving-car industry. Luca Antiga is cofounder of an AI engineering company and an AI tech startup, as well as a former PyTorch contributor. Thomas Viehmann is a PyTorch core developer and machine learning trainer and consultant. consultant based in Munich, Germany and a PyTorch core developer. Table of Contents PART 1 - CORE PYTORCH 1 Introducing deep learning and the PyTorch Library 2 Pretrained networks 3 It starts with a tensor 4 Real-world data representation using tensors 5 The mechanics of learning 6 Using a neural network to fit the data 7 Telling birds from airplanes: Learning from images 8 Using convolutions to generalize PART 2 - LEARNING FROM IMAGES IN THE REAL WORLD: EARLY DETECTION OF LUNG CANCER 9 Using PyTorch to fight cancer 10 Combining data sources into a unified dataset 11 Training a classification model to detect suspected tumors 12 Improving training with metrics and augmentation 13 Using segmentation to find suspected nodules 14 End-to-end nodule analysis, and where to go next PART 3 - DEPLOYMENT 15 Deploying to production

The authors introduce the core function of the Message Printing Interface (MPI). This edition adds material on the C++ and Fortran 90 binding for MPI.

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

Learn how to accelerate C++ programs using data parallelism. This open access book enables C++ programmers to be at the forefront of this exciting and important new development that is helping to push computing to new levels. It is full of practical advice, detailed explanations, and code examples to illustrate key topics. Data parallelism in C++ enables access to parallel resources in a modern heterogeneous system, freeing you from being locked into any particular computing device. Now a single C++ application can use any combination of devices—including GPUs, CPUs, FPGAs and AI ASICs—that are suitable to the problems at hand. This book begins by introducing data parallelism and foundational topics for effective use of the SYCL standard from the Khronos Group and Data Parallel C++ (DPC++), the open source compiler used in this book. Later chapters cover advanced topics including error handling, hardware-specific programming, communication and synchronization, and memory model considerations. Data Parallel C++ provides you with everything needed to use SYCL for programming heterogeneous systems. What You'll Learn Accelerate C++ programs using data-parallel programming Target multiple device types (e.g. CPU, GPU, FPGA) Use SYCL and SYCL compilers Connect with computing’s heterogeneous future via Intel’s oneAPI initiative Who This Book Is For Those new data-parallel programming and computer programmers interested in data-parallel programming using C++.

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.

Heterogeneous Computing Architectures: Challenges and Vision provides an updated vision of the state-of-the-art of heterogeneous computing systems, covering all the aspects related to their design: from the architecture and programming models to hardware/software integration and orchestration to real-time and security requirements. The transitions from multicore processors, GPU computing, and Cloud computing are not separate trends, but aspects of a single trend-mainstream; computers from desktop to smartphones are being permanently transformed into heterogeneous supercomputer clusters. The reader will get an organic perspective of modern heterogeneous systems and their future evolution.

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.

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.

Many of today’s complex scientific applications now require a vast amount of computational power. General purpose graphics processing units (GPGPUs) enable researchers in a variety of fields to benefit from the computational power of all the cores available inside graphics cards. Understand the Benefits of Using GPUs for Many Scientific Applications Designing Scientific Applications on GPUs shows you how to use GPUs for applications in diverse scientific fields, from physics and mathematics to computer science. The book explains the methods necessary for designing or porting your scientific application on GPUs. It will improve your knowledge about image processing, numerical applications, methodology to design efficient applications, optimization methods, and much more. Everything You Need to Design/Port Your Scientific Application on GPUs The first part of the book introduces the GPUs and Nvidia’s CUDA programming model, currently the most widespread environment for designing GPU applications. The second part focuses on significant image processing applications on GPUs. The third part presents general methodologies for software development on GPUs and the fourth part describes the use of GPUs for addressing several optimization problems. The fifth part covers many numerical applications, including obstacle problems, fluid simulation, and atomic physics models. The last part illustrates agent-based simulations, pseudorandom number generation, and the solution of large sparse linear systems for integer factorization. Some of the codes presented in the book are available online.

Parallel Computing for Data Science: With Examples in R, C++ and CUDA is one of the first parallel computing books to concentrate exclusively on parallel data structures, algorithms, software tools, and applications in data science. It includes examples not only from the classic "n observations, p variables" matrix format but also from time series,