Synopsis

Programming Massively Parallel Processors: A Hands-on Approach shows both student and professional alike 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. Topics of performance, floating-point format, parallel patterns, and dynamic parallelism are covered in depth. This best-selling guide to CUDA and GPU parallel programming has been revised with more parallel programming examples, commonly-used libraries such as Thrust, and explanations of the latest tools. With these improvements, the book retains its concise, intuitive, practical approach based on years of road-testing in the authors’ own parallel computing courses. 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.

Book Information

Paperback: 514 pages
Publisher: Morgan Kaufmann; 2 edition (December 28, 2012)
Language: English
ISBN-10: 0124159923
Product Dimensions: 7.5 x 1.2 x 9.2 inches
Shipping Weight: 1.8 pounds (View shipping rates and policies)
Average Customer Review: 3.8 out of 5 starsÂ – See all reviewsÂ (22 customer reviews)
Best Sellers Rank: #423,730 in Books (See Top 100 in Books) #39 inÂ Books > Computers & Technology > Hardware & DIY > Microprocessors & System Design > Microprocessor Design #41 inÂ Books > Computers & Technology > Programming > Parallel Programming #171 inÂ Books > Computers & Technology > Programming > Languages & Tools > C & C++ > C

Customer Reviews

This second edition of PMPP extends the table of contents of the first one, almost doubling the number of pages (in the 2nd ed. its ~500 pages. I have the paper version.) The book can be separated roughly in 4 parts: the first, and more important, deals with parallel programming using
Nvidia’s CUDA technology: this takes about the first 10 chapters and Ch. 20; the second slice shows a couple of important examples (MRI image reconstruction and molecular simulation and visualization, chapters 11 and 12); the 3rd important block of chapters (chapters 14 upto 19) deals with other parallel programming technologies and CUDA expansions: OpenCL, OpenACC, CUDA Fortran, Thrust, C++AMP, MPI. Finally, spread all over the book, there are several "outlier", but nevertheless important, chapters: Ch. 7 discusses floating-point issues and its impact in calculation's accuracy; Ch. 13, "PP and Computational Thinking", discusses broadly how to think when converting sequential algorithms to parallel; and Ch. 21 discusses the future of PP (using CUDA goggles :-) ). I've read about half of the book (I attended Coursera’s MOOC -"Heterogeneous Parallel Computing"- taught by one of the authors, Prof. W. Hwu, and waited until the 2nd edition was out to buy it), and browsed carefully the other half. Here are my...Comments----------(+++)

Pluses:# There are just a few typos, here and there, but they are easy to spot (the funniest is in line 5 of ch. 1 (!), where Giga corresponds to 10^12 and Tera to 10^15, according to the authors: of course Giga is 10^9 and Tera is 10^12 - this bug is browseable with 's "look inside" feature...).

"Programming Massively Parallel Processors (second edition)" by Kirk and Hwu is a very good second book for those interested in getting started with CUDA. A first must-read is "CUDA by Example: An Introduction to General-Purpose GPU Programming" by Jason Sanders. After reading all of Sanders work, feel free to jump right to chapters 8 and 9 of this Kirk and Hwu publication. In chapter 8, the authors do a nice job of explaining how to write an efficient convolution algorithm that is useful for smoothing and sharpening data sets. Their explanation of how shared memory can play a key role in improving performance is well written. They also handle the issue of "halo" data very well. Benchmark data would have served as a nice conclusion to this chapter. In chapter 9, the authors provide the best description of the Prefix Sum algorithm I have seen to date. It describes the problem being solved in terms that I can easily relate to - food. They write, "We can illustrate the applications of inclusive scan operations using an example of cutting sausage for a group of people." They first describe a simple algorithm, then a "work-efficient" algorithm, and then an extension for larger data sets. What puzzles me here is that the authors seem fixated on solving the problem with the least number of total operations (across all threads) as opposed to the least number of operations per thread. They do not mention that the "work-efficient" algorithm requires almost twice as many more operations for the longest-path thread than the simple algorithm. Actual performance benchmarks showing a net throughput gain would be required for a skeptical reader. Now before moving forward, lets back up a bit. Even though we have already read CUDA by
Example, it is worth reading chapter 6...

So much effort went into writing this text, but better guidance was needed. True, there are typos as many have said, but not as many as I find in most textbooks and research papers. Before I get started, I have one strong compliment for the text: exercises at the ends of the chapters, where they appear, are very well thought out and get right to the point. The worst part of this book is its index, which gets an unnegotiable F grade from me. So much is missing. Try looking for the gridDim built-in variable; it’s not there. blockDim is, but not gridDim. Or look up "constant memory"; that’s not there under "memory", "constant", or otherwise. In fact although most of parallel programming is about overcoming memory bottlenecks, the index doesn’t say much about memory in general.

Indexing books is very hard and is usually left to professional indexers; there really is such a trade. But for this book the authors needed to pay more attention to the end product. The overall arrangement is illogical, starting from an enormous abstraction and then adding detail. Better to begin with something small and concrete which can be understood, and build up. At several points a non-programmer was permitted to do the typing, which is like allowing a pilot to fly while waking out of general anesthesia. See Figure 10.4 at lines 1 and 5. Many concepts are never brought together in some kind of summary that one would ever turn to as a reference, such as special variables. One can also find seeming conflicts that need clarification, such as Table 5.1’s “local” memory for automatic array variables that are “stored into the global memory” on the next page. Both turn out to be true, but the presentation is not ideal.

*Download to continue reading...*