

Contents

Foreword	xiii
Preface	xv
Acknowledgments	xvii
About the Authors	xix
1 WHY CUDA? WHY NOW?	1
<hr/>	
1.1 Chapter Objectives	2
1.2 The Age of Parallel Processing	2
1.2.1 Central Processing Units	2
1.3 The Rise of GPU Computing	4
1.3.1 A Brief History of GPUs	4
1.3.2 Early GPU Computing	5
1.4 CUDA	6
1.4.1 What Is the CUDA Architecture?	7
1.4.2 Using the CUDA Architecture	7
1.5 Applications of CUDA	8
1.5.1 Medical Imaging	8
1.5.2 Computational Fluid Dynamics	9
1.5.3 Environmental Science	10
1.6 Chapter Review	11

2	GETTING STARTED	13
2.1	Chapter Objectives	14
2.2	Development Environment	14
2.2.1	CUDA-Enabled Graphics Processors	14
2.2.2	NVIDIA Device Driver	16
2.2.3	CUDA Development Toolkit	16
2.2.4	Standard C Compiler	18
2.3	Chapter Review	19
3	INTRODUCTION TO CUDA C	21
3.1	Chapter Objectives	22
3.2	A First Program	22
3.2.1	Hello, World!	22
3.2.2	A Kernel Call	23
3.2.3	Passing Parameters	24
3.3	Querying Devices	27
3.4	Using Device Properties	33
3.5	Chapter Review	35
4	PARALLEL PROGRAMMING IN CUDA C	37
4.1	Chapter Objectives	38
4.2	CUDA Parallel Programming	38
4.2.1	Summing Vectors	38
4.2.2	A Fun Example	46
4.3	Chapter Review	57

5	THREAD COOPERATION	59
5.1	Chapter Objectives	60
5.2	Splitting Parallel Blocks	60
5.2.1	Vector Sums: Redux	60
5.2.2	GPU Ripple Using Threads	69
5.3	Shared Memory and Synchronization	75
5.3.1	Dot Product	76
5.3.1	Dot Product Optimized (Incorrectly)	87
5.3.2	Shared Memory Bitmap	90
5.4	Chapter Review	94
6	CONSTANT MEMORY AND EVENTS	95
6.1	Chapter Objectives	96
6.2	Constant Memory	96
6.2.1	Ray Tracing Introduction	96
6.2.2	Ray Tracing on the GPU	98
6.2.3	Ray Tracing with Constant Memory	104
6.2.4	Performance with Constant Memory	106
6.3	Measuring Performance with Events	108
6.3.1	Measuring Ray Tracer Performance	110
6.4	Chapter Review	114
7	TEXTURE MEMORY	115
7.1	Chapter Objectives	116
7.2	Texture Memory Overview	116

7.3	Simulating Heat Transfer	117
7.3.1	Simple Heating Model	117
7.3.2	Computing Temperature Updates	119
7.3.3	Animating the Simulation	121
7.3.4	Using Texture Memory	125
7.3.5	Using Two-Dimensional Texture Memory	131
7.4	Chapter Review	137
8	GRAPHICS INTEROPERABILITY	139
<hr/>		
8.1	Chapter Objectives	140
8.2	Graphics Interoperation	140
8.3	GPU Ripple with <i>Graphics Interoperability</i>	147
8.3.1	The GPUAnimBitmap Structure	148
8.3.2	GPU Ripple Redux	152
8.4	Heat Transfer with Graphics Interop	154
8.5	DirectX Interoperability	160
8.6	Chapter Review	161
9	ATOMICS	163
<hr/>		
9.1	Chapter Objectives	164
9.2	Compute Capability	164
9.2.1	The Compute Capability of NVIDIA GPUs	164
9.2.2	Compiling for a Minimum Compute Capability	167
9.3	Atomic Operations Overview	168
9.4	Computing Histograms	170
9.4.1	CPU Histogram Computation	171
9.4.2	GPU Histogram Computation	173
9.5	Chapter Review	183

10	STREAMS	185
10.1	Chapter Objectives	186
10.2	Page-Locked Host Memory	186
10.3	CUDA Streams	192
10.4	Using a Single CUDA Stream	192
10.5	Using Multiple CUDA Streams	198
10.6	GPU Work Scheduling	205
10.7	Using Multiple CUDA Streams Effectively	208
10.8	Chapter Review	211
11	CUDA C ON MULTIPLE GPUS	213
11.1	Chapter Objectives	214
11.2	Zero-Copy Host Memory	214
11.2.1	Zero-Copy Dot Product	214
11.2.2	Zero-Copy Performance	222
11.3	Using Multiple GPUs	224
11.4	Portable Pinned Memory	230
11.5	Chapter Review	235
12	THE FINAL COUNTDOWN	237
12.1	Chapter Objectives	238
12.2	CUDA Tools	238
12.2.1	CUDA Toolkit	238
12.2.2	CUFFT	239
12.2.3	CUBLAS	239
12.2.4	NVIDIA GPU Computing SDK	240

12.2.5	NVIDIA Performance Primitives	241
12.2.6	Debugging CUDA C	241
12.2.7	CUDA Visual Profiler	243
12.3	Written Resources	244
12.3.1	Programming Massively Parallel Processors: A Hands-On Approach	244
12.3.2	CUDA U	245
12.3.3	NVIDIA Forums	246
12.4	Code Resources	246
12.4.1	CUDA Data Parallel Primitives Library	247
12.4.2	CULAtools	247
12.4.3	Language Wrappers	247
12.5	Chapter Review	248
A	ADVANCED ATOMICS	249
<hr/>		
A.1	Dot Product Revisited	250
A.1.1	Atomic Locks	251
A.1.2	Dot Product Redux: Atomic Locks	254
A.2	Implementing a Hash Table	258
A.2.1	Hash Table Overview	259
A.2.2	A CPU Hash Table	261
A.2.3	Multithreaded Hash Table	267
A.2.4	A GPU Hash Table	268
A.2.5	Hash Table Performance	276
A.3	Appendix Review	277
	Index	279