

Contents

Preface	ix
Acknowledgments	xi
CHAPTER 1 Introduction	1
1.1 Motivational Example and Its Analysis	2
1.2 Parallelism Basics	10
1.3 HPC Trends and Rankings	16
1.4 Additional Exercises	18
CHAPTER 2 Theoretical Background	21
2.1 PRAM	22
2.2 Network Topologies	27
2.3 Amdahl's and Gustafson's Laws	31
2.4 Foster's Parallel Algorithm Design Methodology	37
2.5 Additional Exercises	41
References	45
CHAPTER 3 Modern Architectures	47
3.1 Memory Hierarchy	48
3.2 Levels of Parallelism	58
3.3 Additional Exercises	71
References	75
CHAPTER 4 C++11 Multithreading	77
4.1 Introduction to Multithreading (Hello World)	78
4.2 Handling Return Values (Fibonacci Sequence)	83
4.3 Scheduling Based on Static Distributions (Matrix Vector Multiplication)	92
4.4 Handling Load Imbalance (All-Pairs Distance Matrix)	106
4.5 Signaling Threads with Condition Variables (Ping Pong)	115
4.6 Parallelizing Over Implicitly Enumerable Sets (Thread Pool)	122
4.7 Additional Exercises	131
References	133
CHAPTER 5 Advanced C++11 Multithreading	135
5.1 Lock-Free Programming (Atomics, Compare-and-Swap)	136
5.2 Work-Sharing Thread Pool (Tree Traversal)	147
5.3 Parallel Graph Search (Binary Knapsack Problem)	151
5.4 Outlook	160
5.5 Additional Exercises	162
References	164

CHAPTER 6	OpenMP	165
6.1	Introduction to OpenMP (Hello World)	166
6.2	The <code>parallel for</code> Directive (Basic Linear Algebra)	169
6.3	Basic Parallel Reductions (Nearest-Neighbor Classifier)	179
6.4	Scheduling of Imbalanced Loops (Inner Products)	189
6.5	Advanced Reductions (Softmax Regression/AVX Reductions)	193
6.6	Task Parallelism (Tree Traversal)	209
6.7	SIMD Vectorization (Vector Addition)	213
6.8	Outlook	216
6.9	Additional Exercises	216
	References	223
CHAPTER 7	Compute Unified Device Architecture	225
7.1	Introduction to CUDA (Hello World)	226
7.2	Hardware Architecture of CUDA-Enabled GPUs	228
7.3	Memory Access Patterns (Eigenfaces)	234
7.4	Memory Hierarchy (Dynamic Time Warping)	256
7.5	Optimization Guidelines	281
7.6	Additional Exercises	282
	References	284
CHAPTER 8	Advanced CUDA Programming	287
8.1	Warp Ininsics and Atomic Operations (Parallel Reduction)	288
8.2	Utilizing Multiple GPUs and Streams (Newton Iteration)	296
8.3	Outlook	307
8.4	Additional Exercises	309
	References	312
CHAPTER 9	Message Passing Interface	315
9.1	Introduction to MPI	316
9.2	Basic Concepts (Hello World)	318
9.3	Point-to-Point Communication (Ping-Pong)	319
9.4	Nonblocking Communication (Ping-Pong in a Ring of Processes)	322
9.5	Collectives (Counting Primes)	325
9.6	Overlapping Computation and Communication (Jacobi Iteration)	331
9.7	Derived Datatypes (Matrix Multiplication With Submatrix Scattering)	341
9.8	Complex Communicators (Matrix Multiplication Using SUMMA)	348
9.9	Outlook	356
9.10	Additional Exercises	357
	References	363
CHAPTER 10	Unified Parallel C++	365
10.1	Introduction to PGAS and UPC++	365
10.2	Basic Concepts (Hello World)	368
10.3	Memory Affinity and Privatization (Vector Update)	368
10.4	Global Pointers and Collectives (Letter Count)	375
10.5	Locks (Image Histogramming)	382

10.6	Remote Function Invocation (Mandelbrot Sets)	387
10.7	Additional Exercises	394
	References	397
Index	399