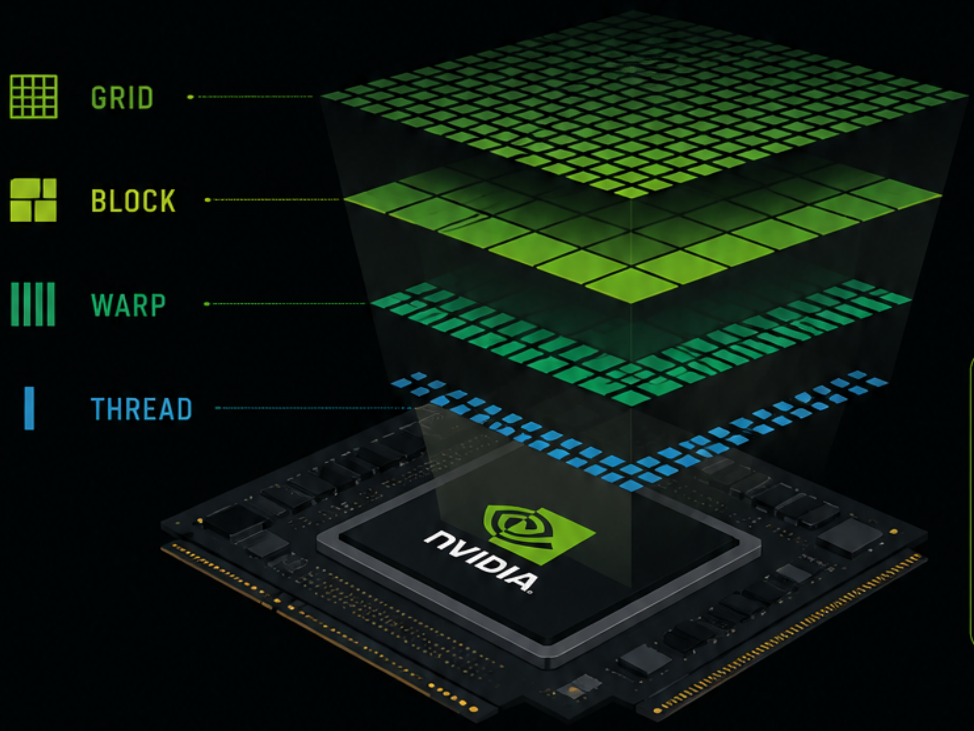
CUDA PROGRAMMING FROM SCRATCH

FROM FIRST PRINCIPLES TO PRODUCTION-GRADE GPU APPLICATIONS



COVERS MODERN CUDA THROUGH THE BLACKWELL ARCHITECTURE

WRITE EFFICIENT KERNELS

OPTIMIZE PERFORMANCE

PROFILE AND DEBUG LIKE A PRO

REAL-WORLD APPLICATIONS

AI, SCIENTIFIC COMPUTING, AND MORE

STEVE T.

CUDA Programming from Scratch

From First Principles to Production-Grade GPU Applications

Steve T. Team Publications

This book is available at https://leanpub.com/cudaprogrammingfromscratch

This version was published on 2026-07-01

Contents

CUDA Programming from Scratch

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/cudaprogrammingfromscratch.

From First Principles to Production-Grade GPU Applications

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/cudaprogrammingfromscratch.

Introduction: Why GPU Computing Matters Today

In 2006, NVIDIA unveiled a technology that would quietly reshape the trajectory of computing. CUDA (Compute Unified Device Architecture) was introduced alongside the GeForce 8800 GTX, marking the moment when graphics processors stopped being specialized rendering engines and became general-purpose parallel computers. The insight was deceptively simple: if you could expose the GPU's execution units directly to C-style code, developers could harness thousands of cores for tasks far beyond computer graphics.

Twenty years later, that insight defines the modern computing landscape. GPUs are now the dominant accelerator across artificial intelligence, scientific simulation, financial modeling, and high-performance computing. The NVIDIA B200 Blackwell GPU, released in 2024, delivers up to 9 petaflops of dense FP4 tensor performance with 180 gigabytes of HBM3e memory and 8 terabytes per second of memory bandwidth [1]. These are numbers that would have been science fiction in the CUDA launch year.

Yet for all the hardware sophistication, writing high-performance CUDA code remains a craft that demands deep understanding of execution models, memory hierarchies, synchronization primitives, and profiling-driven optimization. The gap between a naive kernel and a production-ready one can be measured in orders of magnitude. This book exists to close that gap.

The central thesis of this book is straightforward but demanding: CUDA mastery comes from understanding the hardware execution model at its core and applying systematic optimization grounded in empirical profiling data. You will not learn CUDA by memorizing APIs. You will learn it by building kernels from scratch, breaking them intentionally, measuring their behavior with professional profilers, and iteratively improving them until they extract maximum performance from the silicon.

This book is structured as a progressive journey. We begin with the hardware — what GPUs actually are, how they differ fundamentally from CPUs, and why NVIDIA's design philosophy has made them the engine of modern AI

and HPC. Then we move to the programming model: threads, warps, blocks, grids, and the SIMT execution paradigm that governs everything. From there, we dive deep into memory — every memory type in the GPU hierarchy, its scope, latency, bandwidth, and optimal use cases.

The middle chapters cover kernel design and optimization: tiling strategies, bank conflict avoidance, warp-level primitives for fine-grained collective operations, cooperative groups for flexible synchronization, and asynchronous execution patterns using streams, events, and CUDA Graphs. We then explore the modern memory abstractions — unified memory with its page migration engine, pinned memory, and prefetch APIs — before tackling the hardware features that power AI at scale: tensor cores, mixed-precision computing, and the Transformer Engine.

The advanced chapters cover Hopper's innovations (the Tensor Memory Accelerator, asynchronous barriers, pipelines), CUDA libraries that you should leverage before writing your own kernels, dynamic parallelism for recursive workloads, multi-GPU programming with NVLink and NCCL, and finally the complete toolkit for profiling, debugging, and performance engineering.

Throughout, every concept is grounded in working code. You will find complete kernel implementations, performance benchmarks comparing optimization stages, and mini-projects that synthesize multiple concepts into real applications. The code targets modern CUDA (through CUDA 13.3) and reflects the Blackwell architecture's capabilities while remaining compatible with earlier architectures where appropriate.

No prior GPU experience is assumed beyond familiarity with C/C++ and basic parallel programming concepts. If you can write a loop, allocate memory, and call a function in C++, this book will teach you to think in parallel at the scale of millions of threads.

Chapter 1: The GPU Revolution — Architecture and History

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/cudaprogrammingfromscratch.

From Graphics to General-Purpose Computing

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/cudaprogrammingfromscratch.

GPU vs CPU: Divergent Design Philosophies

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/cudaprogrammingfromscratch.

The CUDA Platform Ecosystem

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/cudaprogrammingfromscratch.

GPU Architecture Roadmap: Fermi through Blackwell

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/cudaprogrammingfromscratch.

Chapter 2: The CUDA Programming Model — Threads, Blocks, and Warps

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/cudaprogrammingfromscratch.

SIMT Execution: Single Instruction, Multiple Threads

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/cudaprogrammingfromscratch.

Thread Hierarchy: Threads, Warps, Blocks, Grids, and Clusters

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/cudaprogrammingfromscratch.

Kernel Launch Syntax and Configuration

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/cudaprogrammingfromscratch.

The Grid-Stride Loop Pattern

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/cudaprogrammingfromscratch.

Occupancy: Theory, Calculation, and the Occupancy API

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/cudaprogrammingfromscratch.

Chapter 3: Memory Models — The GPU Memory Hierarchy

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/cudaprogrammingfromscratch.

Registers and Local Memory

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/cudaprogrammingfromscratch.

Global Memory and Coalesced Access

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/cudaprogrammingfromscratch.

Shared Memory: Scope, Latency, and Bank Conflicts

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/cudaprogrammingfromscratch.

Bank Conflicts

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/cudaprogrammingfromscratch.

Constant and Texture Memory

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/cudaprogrammingfromscratch.

L1/L2 Cache Architecture

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/cudaprogrammingfromscratch.

Memory Alignment and Vectorized Access

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/cudaprogrammingfromscratch.

Chapter 4: Writing and Optimizing Kernels

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/cudaprogrammingfromscratch.

Your First CUDA Kernels: Vector Addition, Matrix Multiply

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/cudaprogrammingfromscratch.

Tiling and Shared Memory Optimization

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/cudaprogrammingfromscratch.

Avoiding Bank Conflicts: Padding and Swizzling

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/cudaprogrammingfromscratch.

Warp-Specialized Kernels and Producer-Consumer Patterns

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/cudaprogrammingfromscratch.

Common Pitfalls: Divergence, Race Conditions, Out-of-Bounds

Chapter 5: Synchronization — From Warps to Grids

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/cudaprogrammingfromscratch.

Warp-Level Synchronization (Implicit and Explicit)

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/cudaprogrammingfromscratch.

Block-Level Barriers (syncthreads)

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/cudaprogrammingfromscratch.

Cooperative Groups: Thread Block Tiles, Cluster Groups, Grid Groups

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/cudaprogrammingfromscratch.

Thread Block Groups

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/cudaprogrammingfromscratch.

Tiled Partitions and Warp-Level Collectives

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/cudaprogrammingfromscratch.

Grid Groups and Cross-Block Synchronization

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/cudaprogrammingfromscratch.

Scoped Atomics and Thread Scopes

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/cudaprogrammingfromscratch.

Asynchronous Barriers and cuda::barrier (Hopper+)

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/cudaprogrammingfromscratch.

Chapter 6: Warp-Level and Intrinsics Programming

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/cudaprogrammingfromscratch.

Warp Shuffle Primitives: __shfl_sync, __shfl_down_sync, __shfl_up_sync, __shfl_xor_sync

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/cudaprogrammingfromscratch.

Warp-Level Parallel Reduction

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/cudaprogrammingfromscratch.

Vote and Mask Operations: __ballot_sync, __any_sync, __all_sync, __activemask

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/cudaprogrammingfromscratch.

Warp-Level Reductions and Scans

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/cudaprogrammingfromscratch.

Inline PTX Assembly for Performance-Critical Code

When to Use Warp Primitives vs. Cooperative Groups

Chapter 7: Asynchronous Execution — Streams, Events, and Overlap

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/cudaprogrammingfromscratch.

CUDA Streams: Default and User-Created

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/cudaprogrammingfromscratch.

Events for Synchronization and Timing

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/cudaprogrammingfromscratch.

Overlapping Data Transfers with Computation

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/cudaprogrammingfromscratch.

Multi-Stream Pipelining Patterns

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/cudaprogrammingfromscratch.

CUDA Graphs: Capture, Replay, and Constant-Time Launch

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/cudaprogrammingfromscratch.

Chapter 8: Unified Memory and Advanced Memory Management

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/cudaprogrammingfromscratch.

The Problem with Explicit Host-Device Transfers

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/cudaprogrammingfromscratch.

cudaMallocManaged and Page Migration Engine

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/cudaprogrammingfromscratch.

Pinned (Page-Locked) Memory

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/cudaprogrammingfromscratch.

Unified Memory Performance: When It Works, When It Doesn't

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/cudaprogrammingfromscratch.

Chapter 9: Tensor Cores and Mixed-Precision Computing

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/cudaprogrammingfromscratch.

Evolution of Tensor Cores: Volta through Blackwell

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/cudaprogrammingfromscratch.

Matrix Multiply-Accumulate (MMA) Operations

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/cudaprogrammingfromscratch.

Data Precision Formats

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/cudaprogrammingfromscratch.

Writing Tensor Core Kernels with WGMMA

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/cudaprogrammingfromscratch.

The Transformer Engine and Dynamic Scaling

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/cudaprogrammingfromscratch.

Chapter 10: Hopper Innovations — TMA, Barriers, and Pipelines

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/cudaprogrammingfromscratch.

Tensor Memory Accelerator (TMA): Architecture and Programming Model

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/cudaprogrammingfromscratch.

cuda::memcpy_async and Asynchronous Data Copies

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/cudaprogrammingfromscratch.

CUDA Pipelines: Producer-Consumer Patterns with Multi-Buffering

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/cudaprogrammingfromscratch.

Warp-Specialized Kernels for Maximum Utilization

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/cudaprogrammingfromscratch.

Cluster-Sized Thread Blocks and GPC-Level Scheduling

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/cudaprogrammingfromscratch.

Chapter 11: CUDA Libraries — Building on NVIDIA's Foundation

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/cudaprogrammingfromscratch.

Linear Algebra: cuBLAS, cuSOLVER, cuSPARSE

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/cudaprogrammingfromscratch.

Signal Processing: cuFFT

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/cudaprogrammingfromscratch.

Parallel Primitives: CUB and Thrust

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/cudaprogrammingfromscratch.

Random Numbers: cuRAND

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/cudaprogrammingfromscratch.

Image and Video: NPP, nvJPEG, nvCodec

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/cudaprogrammingfromscratch.

When to Use Libraries vs. Custom Kernels

Chapter 12: Dynamic Parallelism and Multi-GPU Programming

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/cudaprogrammingfromscratch.

Dynamic Parallelism: Child Kernels, Nested Launches

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/cudaprogrammingfromscratch.

Multi-GPU Architecture: PCIe vs NVLink

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/cudaprogrammingfromscratch.

Peer-to-Peer Memory Access (GPUDirect P2P)

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/cudaprogrammingfromscratch.

NCCL: Collective Communication Primitives

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/cudaprogrammingfromscratch.

Multi-GPU Design Patterns and Scaling Considerations

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/cudaprogrammingfromscratch.

Chapter 13: Profiling, Debugging, and Performance Engineering

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/cudaprogrammingfromscratch.

CUDA-GDB and Nsight Debugger

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/cudaprogrammingfromscratch.

Compute Sanitizer: memcheck, racecheck, synccheck, initcheck

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/cudaprogrammingfromscratch.

Nsight Systems: System-Wide Profiling

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/cudaprogrammingfromscratch.

Nsight Compute: Kernel-Level Metrics and Analysis

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/cudaprogrammingfromscratch.

The APOD Framework (Assess, Parallelize, Optimize, Deploy)

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/cudaprogrammingfromscratch.

Performance Engineering Case Studies

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/cudaprogrammingfromscratch.

Case Study 1: Matrix Multiplication Optimization (4096×4096 on A100)

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/cudaprogrammingfromscratch.

Case Study 2: Image Convolution (512×512 RGB image, 3×3 kernel)

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/cudaprogrammingfromscratch.

Chapter 14: Real-World Applications — AI, HPC, and Scientific Computing

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/cudaprogrammingfromscratch.

Convolutional Kernels for Image Processing

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/cudaprogrammingfromscratch.

Matrix Multiplication at Scale: From Naive to Tensor Core Optimized

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/cudaprogrammingfromscratch.

Sparse Linear Algebra for Scientific Computing

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/cudaprogrammingfromscratch.

AI Training and Inference Pipelines

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/cudaprogrammingfromscratch.

Mini-Project: GPU-Accelerated Particle Simulation

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/cudaprogrammingfromscratch.

Conclusion: The Future of GPU Computing

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/cudaprogrammingfromscratch.

References

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/cudaprogrammingfromscratch.