The Architecture of Intelligence: A Complete Guide to AI Accelerators, GPUs, and HPC Systems

PART I — FOUNDATIONS OF ACCELERATOR ARCHITECTURE

Chapter 1 — Fundamentals of Parallel Computing

- Flynn’s Taxonomy (SISD, SIMD, MIMD, SIMT)
- The Parallel Programming Model Hierarchy (Data, Task, Pipeline)
- The Role of GPUs in High-Performance and AI Workloads
- GPU vs. CPU vs. TPU vs. ASIC vs. FPGA: Comparative Overview

Chapter 2 — Digital Logic and Microarchitecture Refresher

- From Transistor to Logic Gate
- Pipelining, Superscalar, and Out-of-Order Concepts
- Instruction Set Architecture (ISA) Design Principles
- Power, Area, and Performance Trade-offs
- Clock Distribution and Synchronization

PART II — GPU AND ACCELERATOR ARCHITECTURE IN DEPTH

Chapter 3 — The GPU Execution Model

- SIMT Execution and CUDA Thread Hierarchy
- Streaming Multiprocessors: Internal Structure
- Thread Divergence vs. Control Flow Divergence
- Occupancy, Latency Hiding, and Warp Scheduling

Chapter 4 — Memory Hierarchy and Data Movement

- GPU Memory Pyramid: Registers → Shared → L1 → L2 → DRAM → Host
- Coalescing Global Memory Accesses
- L2 Cache Persistence and Prefetching Techniques
- High Bandwidth Memory and Sparse Workloads

Chapter 5 — Tensor Cores and Mixed-Precision Arithmetic

- Evolution of Tensor Cores (Volta → Hopper)
- Internal Flow of Tensor Core Operations
- Mixed Precision and Energy Efficiency
- Advanced Formats: FP8 and Dynamic Range Management

PART III — ALTERNATE ARCHITECTURES: TPU, ASIC, FPGA

Chapter 6 — Google TPU Architecture

- TPU Systolic Array Fundamentals
- Dataflow vs. Control Flow Architectures
- Instruction Scheduling and Weight Stationarity
- Comparison with GPU SM Architecture

Chapter 7 — AI Accelerator Design Principles

- Dataflow Architectures: Stationary Designs
- Systolic vs. Spatial vs. Temporal Models
- Designing a MAC Array from Scratch
- On-Chip Interconnects and Compiler–Hardware Co-Design

PART IV — PARALLEL TRAINING AND SCALABILITY

Chapter 8 — Distributed Deep Learning

- Scaling to 100B+ Parameter Models
- Data, Tensor, Pipeline, and Expert Parallelism
- Collective Communication Patterns
- Overlap of Computation and Communication

Chapter 9 — Inter-GPU and Inter-Node Communication

- NVLink, NVSwitch, PCIe, and Infiniband Overview
- Network Topologies and Collective Communication
- Profiling Tools: Nsight and NCCL
- Diagnosing Bandwidth vs. Latency Bottlenecks

PART V — PERFORMANCE OPTIMIZATION AND NUMERICAL TECHNIQUES

Chapter 10 — CUDA Kernel Optimization

- Kernel Launch Configuration and Warp-Level Primitives
- Register Pressure and Occupancy Tuning
- Memory Coalescing and Shared Memory Optimization
- Synchronization and Roofline Analysis

Chapter 11 — Quantization and Low-Precision Inference

- Post-Training Static Quantization
- Dynamic vs. Static Quantization
- Non-Uniform and Bimodal Distributions
- Advanced Calibration Techniques and INT8 Fidelity

Chapter 12 — Advanced Kernel Algorithms

- FlashAttention: Streaming Softmax Optimization
- Memory Fusion and DRAM Write Reduction
- Custom GEMM Kernels and CUTLASS Templates
- Hand-Tuned PTX and SASS Optimization

PART VI — SYSTEM-LEVEL CO-DESIGN AND FUTURE DIRECTIONS

Chapter 13 — Energy Efficiency and Thermal Design

- Power Efficiency Metrics
- Thermal Constraints and Cooling Architectures
- Voltage-Frequency Scaling
- 3D-Stacked Memory and Chiplets

Chapter 14 — Software and Compiler Stack

- CUDA, ROCm, MLIR Overview
- TensorRT, cuDNN, and CUTLASS Libraries
- XLA, TVM, and Triton Internals
- Operator Fusion and Graph Optimization

Chapter 15 — Designing Your Own AI Chip

- Defining Target Workloads
- Choosing a Dataflow and Compute Primitive
- Sizing Compute Arrays vs. Memory Bandwidth
- RTL Flow, Verification, and Tape-out

Chapter 16 — Future of AI Compute

- Analog, Photonic, and Neuromorphic Computing
- Edge AI Acceleration Trends
- Quantum-Assisted Architectures