100-Day Roadmap: GPU Programming

Days 1–20: Core GPU & CUDA Foundations

Goal: Build a strong foundation in GPU architecture and CUDA programming, essential for optimizing AI models.

Day	Topic	Hands-On / Mini-Project	Resources
1	GPU vs CPU architecture	Diagram SMs, cores, memory hierarchy	PMPP Ch. 1; CS149 Lecture: "Why Parallelism? Why Efficiency?" (Sep 26)
2	Install CUDA toolkit, NVCC flags	“Hello GPU” kernel	PMPP Ch. 2; NVIDIA CUDA Toolkit Docs
3	Host vs Device code (__global__, etc.)	Vector addition (single block)	PMPP Ch. 3
4	Thread hierarchy: grids, blocks, warps	Vector addition (multi-block)	PMPP Ch. 3; CS149 Lecture: "Multi-core Arch Part II + ISPC" (Oct 03)
5	Memory types: global vs shared vs registers	Vector copy with shared memory	PMPP Ch. 4
6	Synchronization (__syncthreads())	Tiled matrix multiply (naïve)	PMPP Ch. 5
7	Host–device transfers (cudaMemcpy)	Benchmark copy bandwidth	PMPP Ch. 2
8	Profiling intro: nvprof / Nsight Compute	Profile Day 6 kernel	NVIDIA Nsight Compute Docs
9	Warp divergence & branch control flow	Branchy kernel and divergence measurement	PMPP Ch. 6; CS149 Lecture: "GPU architecture and CUDA Programming" (Oct 17)
10	Occupancy & launch configuration	Tune block/thread sizes for Day 6 kernel	PMPP Ch. 6
11	Atomic ops & race conditions	Parallel reduction with atomicAdd	PMPP Ch. 5
12	Unified memory overview	Vector addition with cudaMallocManaged	PMPP Ch. 4
13	Streams & asynchronous copies	Overlap compute + H→D copy in two streams	PMPP Ch. 7
14	Events & timing	Precise timing of kernel + copy	PMPP Ch. 7
15	Shared-memory bank conflicts	Experiment with conflicting vs. non-conflict indices	PMPP Ch. 4
16	Constant & texture memory	Texture lookup kernel	PMPP Ch. 4
17	Dynamic parallelism	Kernel that launches sub-kernels	PMPP Ch. 8
18	CUDA Graphs basics	Capture and replay simple kernel sequence	NVIDIA CUDA Graphs Docs
19	Nsight Systems for end-to-end profiling	Profile host+kernel pipeline	NVIDIA Nsight Systems Docs
20	**Capstone #1: Optimized Matrix Multiply**	Shared memory + loop unrolling	PMPP Ch. 9; CS149 Assignment 1 (Oct 6)

Days 21–40: Intermediate CUDA + Math-Heavy Kernels for AI

Goal: Develop optimized kernels (e.g., GEMM, convolutions) critical for AI workloads, including VLMs and LLMs.

Day	Topic	Mini-Project	Resources
21	Tiled GEMM (shared + register tiling)	64×64 matrix multiply	PMPP Ch. 9

22	Loop unrolling & software pipelining	Unroll inner loops of Day 21 kernel	PMPP Ch. 6
23	Warp-level primitives (__shfl_*)	Warp-wide reduce	NVIDIA CUDA Programming Guide
24	Prefix sum (scan)	Parallel scan via Blelloch algorithm	PMPP Ch. 10; CS149 Lecture: "Data-Parallel Thinking" (Oct 19)
25	Histogram & binning	Atomic vs. shared memory approach	PMPP Ch. 11
26	2D convolution in CUDA	Image filter (edge detect)	PMPP Ch. 12; CS149 Lecture: "Efficiently Evaluating DNNs on GPUs" (Oct 26)
27	cuBLAS intro & tuning	Compare against Day 21 custom GEMM	NVIDIA cuBLAS Docs
28	cuFFT intro	1D FFT on sample signal	NVIDIA cuFFT Docs
29	Thrust library for high-level ops	Sort large array + reduction	NVIDIA Thrust Docs
30	Streaming large datasets	Double buffering with 3 streams	PMPP Ch. 7
31	Pinned vs pageable memory	Copy throughput benchmark	PMPP Ch. 4
32	Mixed precision (FP16) kernels	FP16 GEMM on Tensor Cores (WMMA)	NVIDIA Mixed Precision Docs
33	Tensor cores deep dive	WMMA API micro-benchmark	NVIDIA Tensor Core Programming Guide
34	CUDA Graphs advanced	Parameterized graph with loops	NVIDIA CUDA Graphs Docs
35	Multi-GPU peer-to-peer (P2P) copies	P2P memcpy between two GPUs	NVIDIA Multi-GPU Programming Docs
36	Cooperative groups	Grid-level sync reduce	NVIDIA Cooperative Groups Docs
37	Memory pools & cudaMallocAsync	Pool allocator micro-benchmark	NVIDIA CUDA Memory Management Docs
38	Profiling large kernels	Nsight Compute metrics analysis	NVIDIA Nsight Compute Docs
39	Kernel fusion techniques	Fuse GEMM + activation	Research papers on kernel fusion (e.g., arXiv)
40	**Capstone #2: High-Throughput Convolution**	Multi-stream + optimized memory	PMPP Ch. 12; CS149 Assignment 3 (Nov 8)

Days 41–60: Deep Learning Framework Internals + Multimodal AI

Goal: Extend frameworks like PyTorch with custom CUDA ops, focusing on components of VLMs and multimodal models.

Day	Topic	Mini-Project	Resources
41	PyTorch extension scaffold (cpp_extension)	Hello from C++ + CUDA into Python	PyTorch C++ Extension Docs
42	Custom CUDA op: ReLU / GELU	Integrate and test in a PyTorch model	PyTorch Custom Ops Tutorial
43	autograd.Function for custom backward	GELU with custom backward	PyTorch Autograd Docs
44	Softmax kernel + numerical stability	Log-sum-exp trick	Research papers on stable softmax (e.g., arXiv)
45	LayerNorm kernel	Batch vs. layer norm	Research papers on Layer Normalization
46	Attention (QKV) CUDA kernel	Single-head scaled dot-product	Research papers on Attention; CS149 Lecture: "Efficiently Evaluating DNNs"
47	Multi-head attention & grouping	Merge heads + optimize memory reuse	Research papers on Multi-head Attention
48	FlashAttention techniques	Tiled attention with shared memory	FlashAttention paper (arXiv:2205.14135)

49	Profile end-to-end transformer block	Combine Day 46+Day 45 kernels	NVIDIA Nsight Systems Docs
50	TensorRT integration & custom plugins	Export small model + optimize	NVIDIA TensorRT Docs
51	ONNX export & quant-aware graph	Convert PyTorch model → ONNX → TensorRT	ONNX Docs
52	Triton (OpenAI) intro	Write simple matmul in Triton	OpenAI Triton Docs
53	Compare Triton vs. CUDA kernel	Perf benchmark on small GEMM	Benchmarking Triton and CUDA
54	TVM or XLA auto-tuning overview	Try simple schedule on a kernel	Apache TVM Docs
55	JIT compilation in PyTorch 2.0	torch.compile on custom op	PyTorch 2.0 Docs
56	Profiling frameworks (TensorBoard Profiler)	Trace and visualize ML pipeline	TensorBoard Profiler Docs
57	Memory fragmentation & defragmentation	Simulate large tensor allocations	Research papers on memory management
58	Data-loader bottlenecks (pin_memory)	Optimize DataLoader + prefetch	PyTorch DataLoader Docs
59	Mixed-precision training with AMP	Train small CNN with autocast	NVIDIA AMP Docs
60	**Capstone #3: End-to-end multimodal block (e.g., CLIP)**	Integrate vision + language in PyTorch	Research papers on CLIP (arXiv:2103.00020)

Days 61–80: Model Compression & Quantization for Large Models

Goal: Master techniques to compress VLMs, LLMs, and multimodal models (e.g., pruning, quantization).

Day	Topic	Mini-Project	Resources
61	Pruning theory: structured vs. unstructured	Magnitude pruning on small MLP	Research papers on pruning; CS149 Lecture: "Performance Optimization" (Oct 10)
62	Implement weight pruning kernel	Zero out pruned weights in CUDA	PyTorch Pruning Tutorial
63	Knowledge distillation overview	Train student from teacher model	Research papers on knowledge distillation (e.g., arXiv:1503.02531)
64	Implement distillation loss in CUDA	MSE + KL-divergence kernel	Custom loss function in PyTorch
65	PTQ workflow with BitsAndBytes	8-bit quant of small BERT	BitsAndBytes Docs
66	Calibration & min-max vs. percentile	Compare calibration methods	Research papers on calibration
67	QAT workflow in PyTorch	Simulate quant noise in training	PyTorch Quantization Docs
68	4-bit quant basics	Per-tensor vs. per-channel scaling	Research papers on 4-bit quantization (e.g., arXiv:2106.08295)
69	Build simple 4-bit linear kernel	Integrate into custom PyTorch op	Custom quantization kernel
70	Error analysis: activation vs. weight	Plot histograms + spikes	Research papers on quantization error
71	Dynamic quantization strategy (Unsloth style)	Layer-sensitivity measure + skip list	Unsloth Docs
72	Implement dynamic skip logic in your 4-bit kernel	Conditional bit-skipping	Custom kernel with skip logic
73	Mixed-bitwidth inference	4-bit + 8-bit hybrid GEMM	Research papers on mixed precision

74	Benchmark quantized vs. FP16/FP32	Memory, latency, accuracy trade-offs	Benchmarking tools
75	**Capstone #4: Dynamic 4-bit quantized VLM**	Full forward pass + profiling	Combine previous days' work
76	Post-training accuracy recovery (fine-tuning)	QLoRA-style finetune on small LM	QLoRA paper (arXiv:2305.14314)
77	Mixed precision + quantization pipelining	Integrate AMP with quantized kernels	Research papers on mixed precision training
78	ONNX + TensorRT INT8/4 plugins	Export and serve quantized model	NVIDIA TensorRT Docs
79	Real-world benchmark (e.g., MMLU, GLUE, VQA)	Evaluate quantized VLM on tasks	Benchmark datasets (MMLU, GLUE, VQA)
80	Distillation + quantization hybrid	Tiny student model in 4-bit	Research papers on combined techniques

Days 81–100: Distributed Training & Deployment for Large Models

Goal: Scale training/inference of VLMs and LLMs across GPUs and deploy efficiently.

Day	Topic	Mini-Project	Resources
81	Fundamentals of NCCL & MPI	All-reduce on two GPUs	NVIDIA NCCL Docs; MPI Docs
82	PyTorch DDP & FSDP	DataParallel vs. FullyShardedParallel	PyTorch Distributed Docs; CS149 Lecture: "Distributed Data-Parallel" (Oct 24)
83	DeepSpeed ZeRO & Offloading	Zero-offload config on small model	DeepSpeed Docs
84	Megatron-LM model parallelism	Split layers across 2 GPUs	Megatron-LM Docs
85	Gradient checkpointing	Save memory in long transformers	Research papers on gradient checkpointing
86	Horovod integration	Simple MNIST training across GPUs	Horovod Docs
87	Fault tolerance & elastic training	Simulate node failure with DDP	Research papers on fault tolerance
88	CUDA MPS & Multi-process service	Share GPU among CPU processes	NVIDIA MPS Docs
89	Kubernetes + GPU scheduling (intro)	Dockerfile + simple k8s GPU pod	Kubernetes GPU Scheduling Docs
90	CI/CD for ML (GitHub Actions + CUDA)	Build, test, deploy custom op	GitHub Actions Docs
91	Monitoring & telemetry (Prometheus, Grafana)	GPU metrics dashboard	Prometheus and Grafana Docs
92	Profiling distributed jobs	Nsight Systems on multi-GPU	NVIDIA Nsight Systems Docs
93	Serving inference at scale (TorchServe, Triton)	Dockerized Triton server	TorchServe and Triton Docs
94	A/B testing & canary deploys	Two versions of quantized model	Research papers on deployment strategies
95	Security: sandboxing CUDA kernels	User-code isolation	Research papers on kernel security
96	Optimizing latency vs. throughput	Batch size tuning	Research papers on optimization
97	Cost optimization on cloud GPUs	Spot instances, GPU families	Cloud provider docs (e.g., AWS, GCP)
98	Write technical blog series on your journey	Publish on Medium or personal blog	Blogging platforms (Medium, GitHub Pages)
99	Mock interviews: system design + CUDA trivia	Solve CUDA whiteboard questions	System design resources (e.g., "Designing Data-Intensive Applications")

100	**Final Capstone: End-to-end pipeline for VLM/LLM**	Train → Quantize → Serve a multimodal API	Combine all learned skills; CS149 Assignment 4 (Dec 4)

Additional Resources & Tips

- **PMPP Book**: Use as your primary reference for GPU programming (chapters listed above).
- **CS149 Lectures & Assignments**: Follow the Fall 2023 schedule and assignments for parallel computing insights (specific dates provided).
- **Research Papers**: Access via arXiv or Google Scholar for cutting-edge techniques (e.g., FlashAttention, QLoRA, CLIP).
- **NVIDIA Documentation**: Essential for CUDA, TensorRT, Nsight, and other tools (links provided).
- **PyTorch Ecosystem**: Leverage official docs for extensions, quantization, and distributed training.

Youtube playlists

GPU Mode https://www.youtube.com/@gpumode

Notes https://christianjmills.com/series/notes/cuda-mode-notes.html

codes https://github.com/gpu-mode/lectures/tree/main/

for intuition

https://youtube.com/playlist?list=PL5XwKDZZlwaY7t0M5OLprpkJUlrF8Lc9j

https://www.youtube.com/playlist?list=PLU0zjpa44nPXddA_hWV1U8oO7AevFgXnT

https://www.youtube.com/playlist?list=PLRRuQYjFhpmubuwx-w8X964ofVkW1T8O4

Complete course

CS149 Course https://gfxcourses.stanford.edu/cs149/fall23 https://www.youtube.com/playlist?list=PLoROMvodv4rMp7MTFr4hQsDEcX7Bx6Odp

George Hotz(American Hacker) Youtube channel(Have so much stuff Like MOE Transformers tiny grad)

Deep learning Course which may assist in flash attention transformers etc https://course.fast.ai/

Github Repo links

Learning GPU Programming (30 days plan)(Have good ML Tiny exercises)

```
https://github.com/hkproj/100-days-of-gpu/blob/main/CUDA.md
```

Leaderboard of discord server(Post your github code link daily they will add you to the leaderboard)

```
https://github.com/hkproj/100-days-of-gpu/blob/main/CUDA.md
```

https://github.com/rkinas/cuda-learning?tab=readme-ov-file(Have Many Resources Including exceptional optimization)

A-hamdi the man who completed 100 days of GPU Programming https://github.com/a-hamdi/GPU

Some other guys

```
https://github.com/1y33/100Days
```

```
https://github.com/JungHoyoun/100days-gpu-challenge
```

For who wanted to get into HPC

```
https://github.com/AdepojuJeremy/CUDA-120-DAYS--CHALLENGE
```

Andrej Karpathy The man Behind the Tesla and OpenAI Dont know pytorch wrote entire LLM Training on C https://github.com/karpathy/llm.c

Similarly another guy wrote a script llama.cpp to run an large LLM locally on phone or laptop https://andrewkchan.dev/posts/yalm.html https://github.com/ggml-org/llama.cpp

https://github.com/kmohan321/LLMs/tree/master Have implementation of bert llm etc

Blogs

https://salykova.github.io/sgemm-gpu(This guy Beating cuBLAS in Single-Precision General Matrix Multiplication)

https://minami.bearblog.dev/gpu/?s=09

Follow https://unsloth.ai/blog(the guys behind every LLM VLM Quantization) https://github.com/unslothai

Challenges Unsloth https://colab.research.google.com/drive/1JqKqA1XWeLHvnYAc0wzrR4JBCnq43H

Codes by different LLMs benchmark (Have max every code)

https://scalingintelligence.stanford.edu/KernelBenchLeaderboard/

Platform to code

https://leetgpu.com/

Leetcode for GPU programming

https://tensara.org/problems

This guy is like god of kernels

https://github.com/youkaichao

Principal engineer at apple writes so much about GPU programming

https://www.linkedin.com/in/yidewang?trk=public_post_feed-actor-name

Training LLM on GPU Kernels

https://huggingface.co/spaces/nanotron/ultrascale-playbook?section=first_steps:_training_on_one_gpu