나란 스터디 **- GPU/CUDA** 프로그래밍 중급 스터디

강의 개요

북클럽 나란에서는 CUDA 초급 지식(스레드, 블록, 기본 메모리 관리)을 익힌 학습자를 대상으로 고급 GPU 병렬 프로그래밍을 함께 공부하는 스터디 그룹을 진행합니다. *Programming Massively Parallel Processors* (4th ed., Hwu/Kirk/El Hajj) 책과 *Udemy – Mastering GPU Parallel Programming with CUDA* 강의를 활용하여 워프 다이버전스, 고급 메모리 최적화, 멀티-GPU 프로그래밍, 라이브러리 활용(CUDA Thrust, cuBLAS) 등 심화된 CUDA 개념을 배우고 복잡한 병렬 애플리케이션을 구현합니다. NVIDIA의 최신 자료와 Nsight 도구를 활용하여 성능 최적화에 중점을 둡니다. 모든 토론과 협업은 북클럽 나란 (www.cyberseowon.com)과 Slack (krbookclub.slack.com)에서 진행됩니다. AI 프로그래밍을 배워보고 싶으신 분들 모두 환영합니다.

- 기간: 2025년 11월 15일(오리엔테이션) ~ 2026년 1월 31일 (8주, 추수감사절 및 크리스마스/새해 주 제외)
- 주당 시간: 9~11시간 (책 읽기 3시간, 강의 3~4시간, 코딩 1~2시간, 리뷰 2시간)
- 사전 요구 지식: CUDA 초급(스레드, 블록, 기본 커널 작성), C/C++ 중급, GPU 병렬 프로그래밍 기초

강의 자료

  - 책: *Programming Massively Parallel Processors* (4th ed., Hwu/Kirk/El Hajj)
  - Udemy: *Mastering GPU Parallel Programming with CUDA*
  - 도구: NVIDIA GPU (최소 Turing 이상 권장) 또는 Google Colab (T4/A100 GPU, 무료/Pro), CUDA Toolkit 12.x
  - 보조 자료: NVIDIA CUDA C++ Programming Guide (https://docs.nvidia.com), Nsight Compute Profiling Guide, CUDA Samples (https://github.com/NVIDIA/cuda-samples)

진행방식

- 주간 읽기, 강의 시청, 코딩 실습, Slack 토론, 주간 리뷰

참가비

- 책 구매 및 강의비: 대략 $150 (개인별 부담)
- 스터디 그룹 참가비: $50 (동기 부여를 위한 참가비로 모임을 끝까지 마무리하시면 환불해드립니다.)

학습 목표

- 워프 다이버전스 최소화 및 고급 메모리 최적화(코일레싱, 뱅크 충돌 방지) 이해
- CUDA 라이브러리(Thrust, cuBLAS, cuFFT) 활용
- 멀티-GPU 및 동적 병렬 처리를 포함한 복잡한 CUDA 커널 작성
- Nsight Systems/Compute를 활용한 성능 분석 및 최적화
- 최종 프로젝트(예: 병렬 행렬 곱셈 또는 이미지 처리)로 책과 강의 통합

주차별 계획

주차	날짜	책 챕터	**Udemy** 강의	주요 주제	활동	과제
오리엔테이션	2025.11.15	-	Introduction to the Nvidia GPUs hardware (1h)	스터디 소개, 환경 설정	스터디 목표 공유, CUDA Toolkit/Colab 설정, GitHub 저장소 생성, Slack 가입	환경 설정 완료, Slack 채널에서 인사
1	2025.11.22	1~2 (~50p)	Installing Cuda and other programs (0.5h) + Introduction to CUDA programming (2h)	CUDA 설치, 기본 커널 작성	챕터 1~2 읽기, 설치 및 기본 커널 실습	Slack 에서 설치 경험 토론
2	2025.12.6	3~4 (~60p)	Introduction to the Nvidia GPUs hardware (2h 22min)	CUDA 아키텍처, 워프 다이버전스	챕터 3~4 읽기, GPU 아키텍처 개요 및 워프 실습	워프 다이버전스 분석, Slack 공유

3	2025.12.13	5~6 (~60p)	Memory Optimization Techniques (3h)	메모리 코일레싱, L1/L2 캐시 활용	챕터 5~6 읽기, 행렬 전치 코딩, 코일레싱 실습	코일레싱 및 캐시 효과 분석, Slack 공유
4	2025.12.20	7 (~30p)	Shared Memory + Warp Divergence (0.5h) + Synchronization & Bank Conflicts (2h 30min)	공유 메모리, 동기화, 뱅크 충돌 최소화	챕터 7 읽기, 공유 메모리 및 동기화 실습	뱅크 충돌 및 동기화 성능 비교, Slack 토론
5	2026.1.3	8~9 (~60p)	CUDA Streams & Multi-GPU (2h) + Debugging tools (0.5h)	비동기 스트림, 멀티-GPU, 디버깅	챕터 8~9 읽기, 스트림 및 멀티-GPU 테스트	스트림 및 멀티-GPU 성능 비교, Slack 공유
6	2026.1.17	10~11 (~60p)	Vector Reduction (3h) + Roofline model (0.5h)	리덕션 최적화, Roofline 분석	챕터 10~11 읽기, 벡터 리덕션 코딩, Roofline 실습	리덕션 성능 분석, Slack 채널에서 토론
7	2026.1.24	12~13 (~50p)	Profiling (3h) + Performance analysis for the previous applications (0.5h)	Nsight를 통한 성능 분석	챕터 12~13 읽기, Nsight 프로파일링 실습	프로파일링 결과 공유, Slack에 업로드
8	2026.1.31	14 (~30p)	2D Indexing (1h) + Matrix Multiplication (Bonus) (3h 57min)	2D 인덱싱, 행렬 곱셈 최적화	챕터 14 읽기, 행렬 곱셈 최적화, Nsight 프로파일링	프로젝트 발표

참고: 2025년 11월 29일(추수감사절 주)와 2025년 12월 27일(크리스마스/새해 주)은 휴식 기간으로 제외되었습니다.

주간 활동

- 읽기 **(2**시간**)**: 책 챕터 읽고 주요 개념(예: “워프 다이버전스”) 요약.
- 강의 **(1~1.5**시간**)**: Udemy 강의 시청, 강의 제공 플레이그라운드 또는 Google Colab으로 실습.
- 코딩 **(2~3**시간**)**: 책 예제(예: 병렬 스캔)와 강의 실습(예: cuBLAS 행렬 곱셈) 구현, GitHub 공유.
- 리뷰 **(2**시간**):** 북클럽 나란 Slack 채널에서 해당 주차 내용 리뷰 발표

도구 및 설정

- 환경: CUDA Toolkit 12.x 설치 (https://developer.nvidia.com/cuda-downloads). GPU 없는 경우 Google Colab(T4/A100 GPU, 무료/Pro) 또는 AWS EC2 G4/P3(~$0.5~$3/시간) 사용.
- 협업**:** 북클럽 나란 Slack 채널에서 토론
- 프로파일링: Nsight Systems/Compute로 성능 분석 (https://developer.nvidia.com/nsight-systems).

연락처

- 스터디 리더**:** @바람 (admin@cyberseowon.com)
- 북클럽 나란**:** www.cyberseowon.com
- **Slack:** krbookclub.slack.com

참가 신청폼

- 참가를 원하시는 분들은 다음 참가 신청폼을 작성해주세요.
- https://bit.ly/4mhBEjl