Course Introduction

Prof. Seokin Hong

Fall 2021

New Giant in Semiconductor Industry

History of NVIDIA GPU

2000

GEFORCE 256
GPU Clock : 120MHz
Memory: 32 MB DDR
Bandwidth: 4.800GB/s

2001

GEFORCE 3 TI 500
GPU Clock : 240MHz
Memory: 64MB DDR
Bandwidth: 8.000 GB/s

2003

GEFORCE 4 TI 4800
GPU Clock : 300MHz
Memory: 128MB DDR
Bandwidth: 10.40 GB/s

2004

GEFORCE 6800
GPU Clock : 400MHz
Memory: 256MB DDR
Bandwidth: 35.20 GB/s

2006

GEFORCE 8800 GTX
GPU Clock : 576MHz
Memory: 768 MB DDR
Bandwidth: 86.40GB/s
Architecture: Tesla
First GPU built with CUDA architecture

2012

GEFORCE GTX 690
GPU Clock : 915MHz
Memory: 2GB GDDR5
Bandwidth: 192.3GB/s
Architecture: Kepler

2015

GEFORCE GTX 980 Ti
GPU Clock : 1076MHz
Memory: 6GB GDDR5
Bandwidth: 336.6 GB/s
Architecture: Maxwell

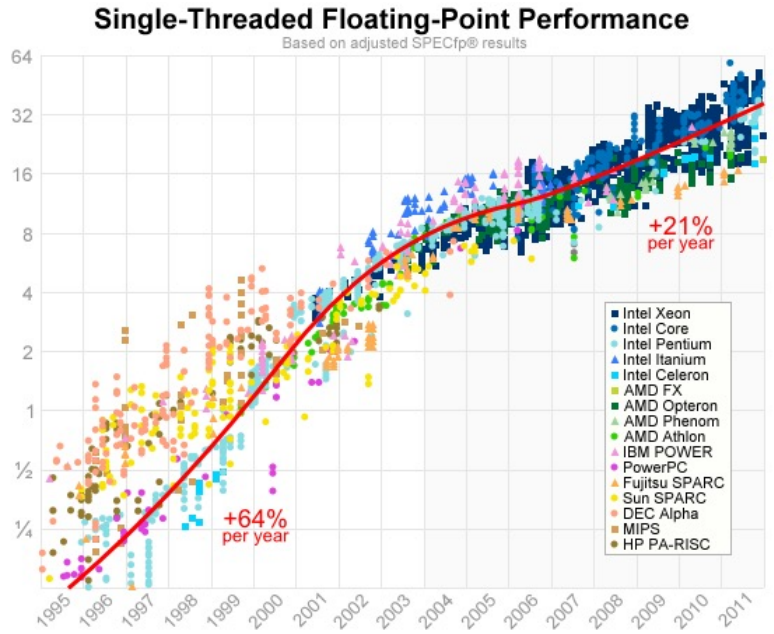
2017

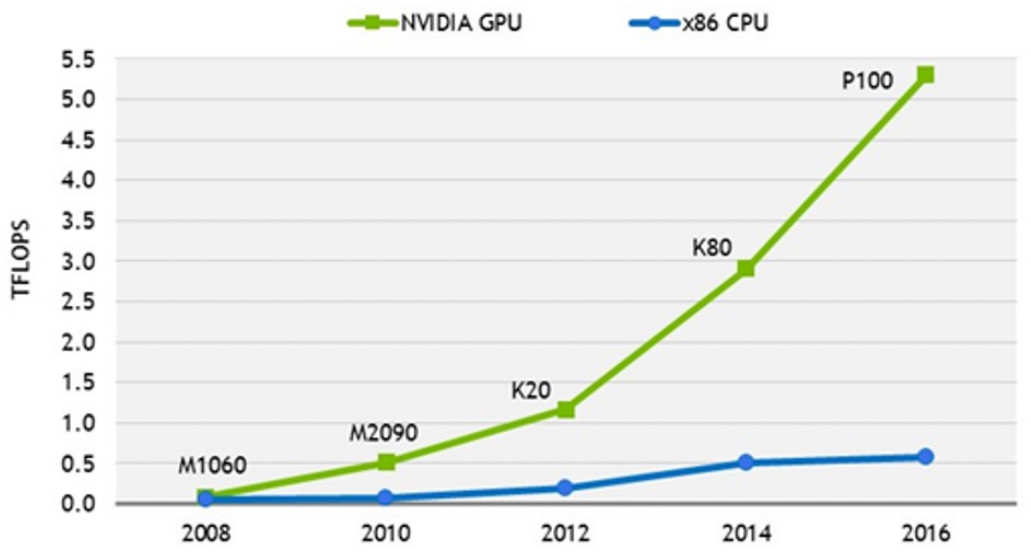
GEFORCE GTX 1080 Ti
GPU Clock : 1481MHz
Memory: 11GB GDDR5
Bandwidth: 484.4 GB/s
Architecture: Pascal

2019

GEFORCE RTX 2080 Ti
GPU Clock : 1340MHz
Memory: 11GB GDDR5
Bandwidth: 616.0GB/s
Architecture: Turing

Why GPU?





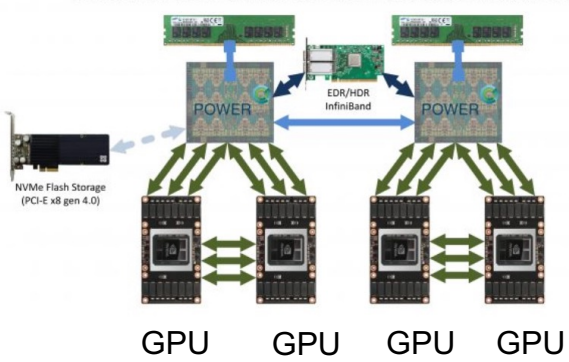
Source: NVIDIA

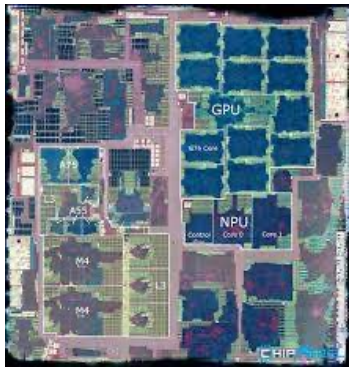
Source:https://preshing.com/20120208/a-look-back-at-single-threaded-cpu-performance/

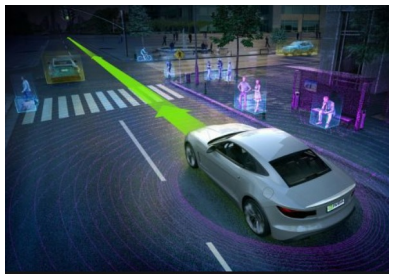
In machine learning and big data era, GPUs are essential to meet the required computing power.

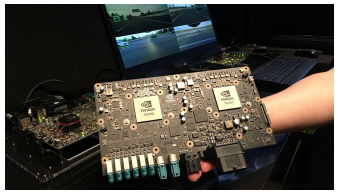
GPU Acceleration is Everywhere for Everything

- 3D Rendering
- Big Data & Data Mining
- Bioinformatics & Genomics
- Business Intelligence & Analytics
- Climate, Weather and Ocean Modeling
- Computational Fluid Dynamics
- Database
- Electronic Design Automation
- Game
- Machine Learning
- …….









Course Objectives

- You will learn about
  - Increasingly important role of parallel computing
  - Multicore CPU and GPU architectures
  - Parallel Programming with PThread, OpenMP, and CUDA
  - Important patterns of parallel computation
  - Case study of accelerating machine learning

Course Schedule (Tentative)

Date	Topic	Reading	Note
2021-08-31	Course Introduction		
2021-09-02	Intoroduction to Parallel Computing		
2021-09-07	Multicore Architecture		
2021-09-09	Multicore Architecture		
2021-09-14	Parallel Programming Models		
2021-09-16	Parallel Programming Basics		
2021-09-21	Parallel Programming Basics		
2021-09-23	Pthread		
2021-09-28	OpenMP		
2021-09-30	OpenMP		
2021-10-05	Synchronization		
2021-10-07	Cache Coherence		
2021-10-12	Cache Coherence		
2021-10-14	Memory Consistency		
2021-10-19	Interconnection Networks		
2021-10-21	Midterm exam		
2021-10-26	Heterogeneous Parallel Computing		
2021-10-28	GPU Architecture		
2021-11-02	Fundamentals of CUDA1		
2021-11-04	Fundamentals of CUDA2		
2021-11-09	CUDA Threads 1		
2021-11-11	CUDA Threads 2		
2021-11-16	CUDA Memory Model 1		
2021-11-18	CUDA Memory Model 2		
2021-11-23	Performance Considerations		
2021-11-25	Parallel Algorithm: Convolution		
2021-11-30	Parallel Algorithm: Sparse Matrix Computation		
2021-12-02	Parallel Algorithm: Graph Search		
2021-12-07	Application case study: Deep Learning		
2021-12-09	Final exam		

Phase1: Multicore CPU

Phase2: GPU

Who am I?

- Seokin Hong, Ph.D.
- seokin@skku.edu
- Office : #400526, Semiconductor Bldg.
- Office Hour : 14:00~16:00, Friday
- compasslab.skku.edu
- Research Interests
  - Computer architecture
  - Memory and storage systems
  - Near-data processing for big data analytics and machine learning
  - Domain-specific Accelerators (e.g., Hardwares for AI)

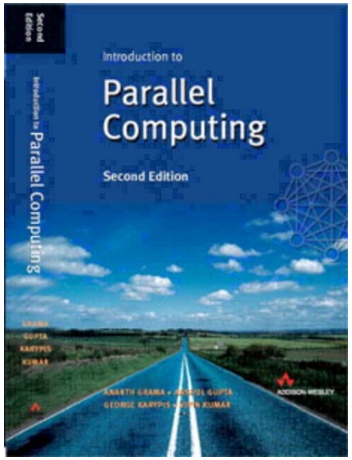
Teaching Assistants

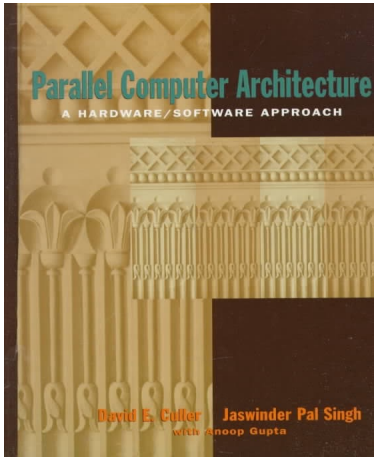
- **변광은 (Byeon, Kwang Eun)**
  - puzzlebook26@gmail.com

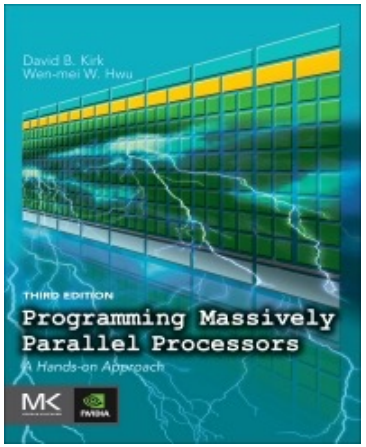
- **김용준 (Kim, Yongjun)**
  - yongjunkim@skku.edu

Textbooks (Reference)

- Introduction to Parallel Computing, 2nd Edition
- Parallel Computer Architecture: A Hardware/Software Approach,
- Programming Massively Parallel Processors, 3rd Edition.







Course Information

- **Meet Tuesday 15:00-16:15, Thursday 16:30-17:45**
- **Lecture style: Pre-recorded video (lecture) + Realtime video streaming (Q&A, every Thursday)**
- **WebEx Info.: https://skku-ict.webex.com/meet/seokin**
- **Website: i-campus**
  - Announcement, Lecture notes, Homework, Course schedule, Q&A

How Will You Be Evaluated?

- Midterm : 30%
- Final : 30%
- Attendance : 5%
- Homework : 20%
- Project (individual): 15%

- Grade will be determined solely by the score (non-negotiable)

Next Time

- Introduction to Parallel Computing