W&M CSCI 780-02: Accelerating Deep Learning -- Introduction

Instructor: Jiajia Li, jli49@wm.edu

Fall 2021

3:30-4:50pm @ Washington Hall 310

WILLIAM & MARY

CHARTERED 1693

CSCI 780-02: Accelerating Deep Learning

- **Instructor**: Jiajia Li (http://jiajiali.org).
  Email: jli49@wm.edu
- **Class Schedule:** TR 3:30-4:50pm ET @ Washington Hall 310
- **Office hour:** Tue @ 2:00-3:00pm ET, MCGL 102 or Zoom
- **Online Discussion**: Slack (tw-courses.slack.com)
- **Course Website**:
  - https://longing-duckling-38b.notion.site/W-M-CSCI-780-02-Accelerating-Deep-Learning-Fall-2021-fb2fbf191cd04488a58eb45bf25afbcc

Agenda

- Introduction
- Course topics
- Course format
- Computing resources

Introduction

- Instructor
- Students

Instructor Introduction

- Education
  - 2013- 2018: Ph.D., Georgia Institute of Technology. Major: Computational Science & Engineering
  - 2008-2013: Ph.D., Institute of Computing Technology, Chinese Academy of Sciences. Major: CS
  - 2005-2008: BS, Dalian University of Technology. Major: Math
- Work experience
  - 2021/08 - : Assistant Professor, Computer Science, William & Mary
  - 2018/08-2021/08: Research Scientist, HPC Group, Pacific Northwest National Laboratory
  - 2016/05-08: IBM Thomas J. Watson Research Center, summer intern
  - 2015/05-07: Intel Parallel Computing Research Lab, summer intern

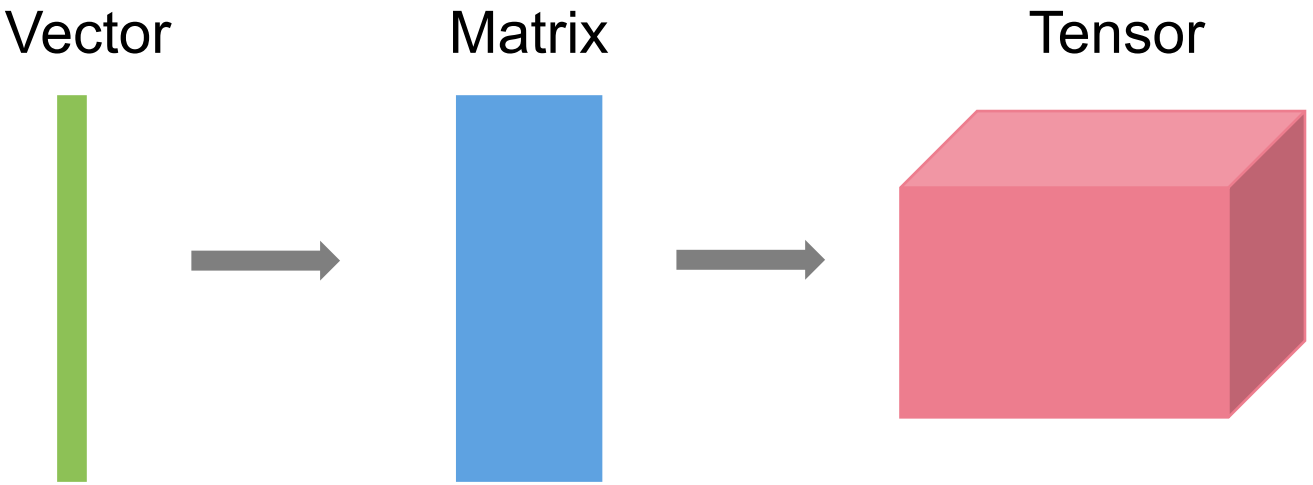
Website: www.jiajiali.org

Instructor Introduction -- Research

- Goal: obtain efficient computation and compression for sparse and multi-dimensional data

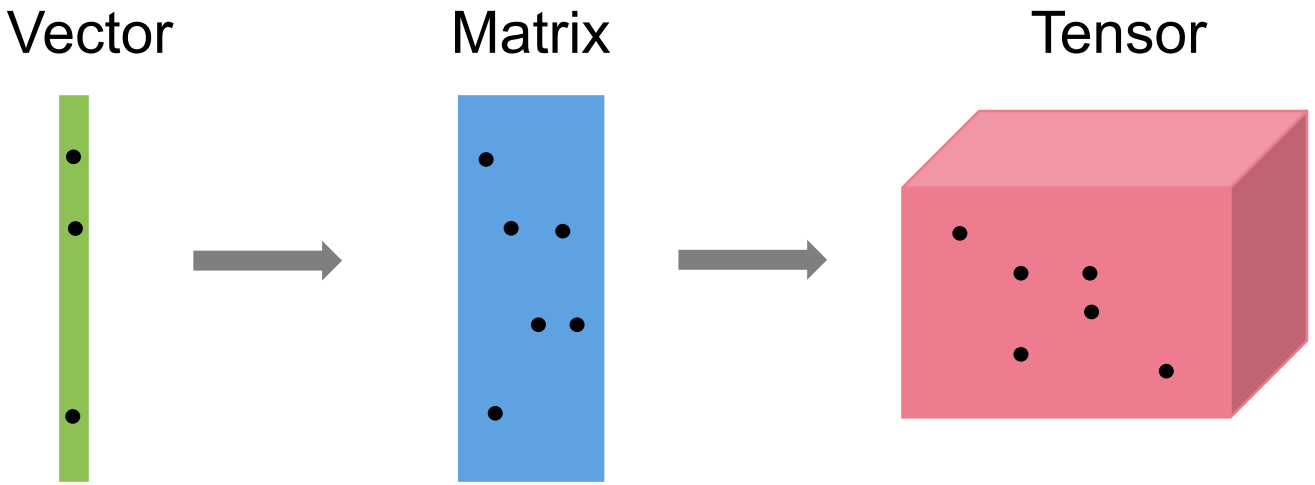
Instructor Introduction -- Research

- Goal: obtain efficient computation and compression for sparse and multi-dimensional data

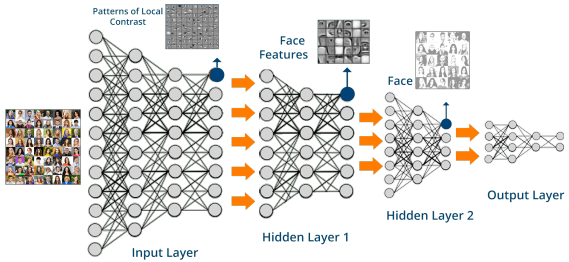


Instructor Introduction -- Research

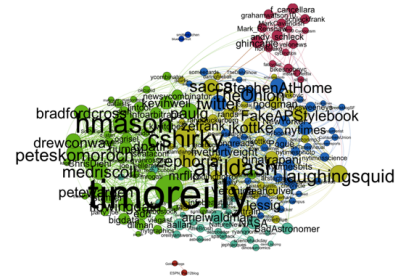
- Goal: obtain efficient computation and compression for sparse and multi-dimensional data



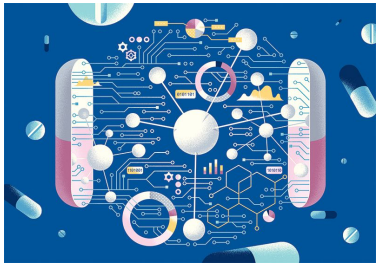
Applications



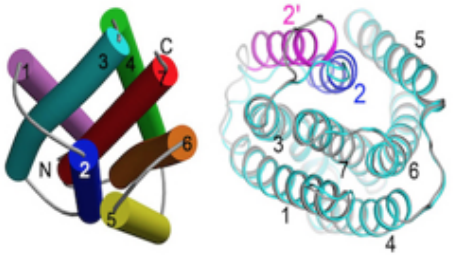
Deep Learning



Social Networks



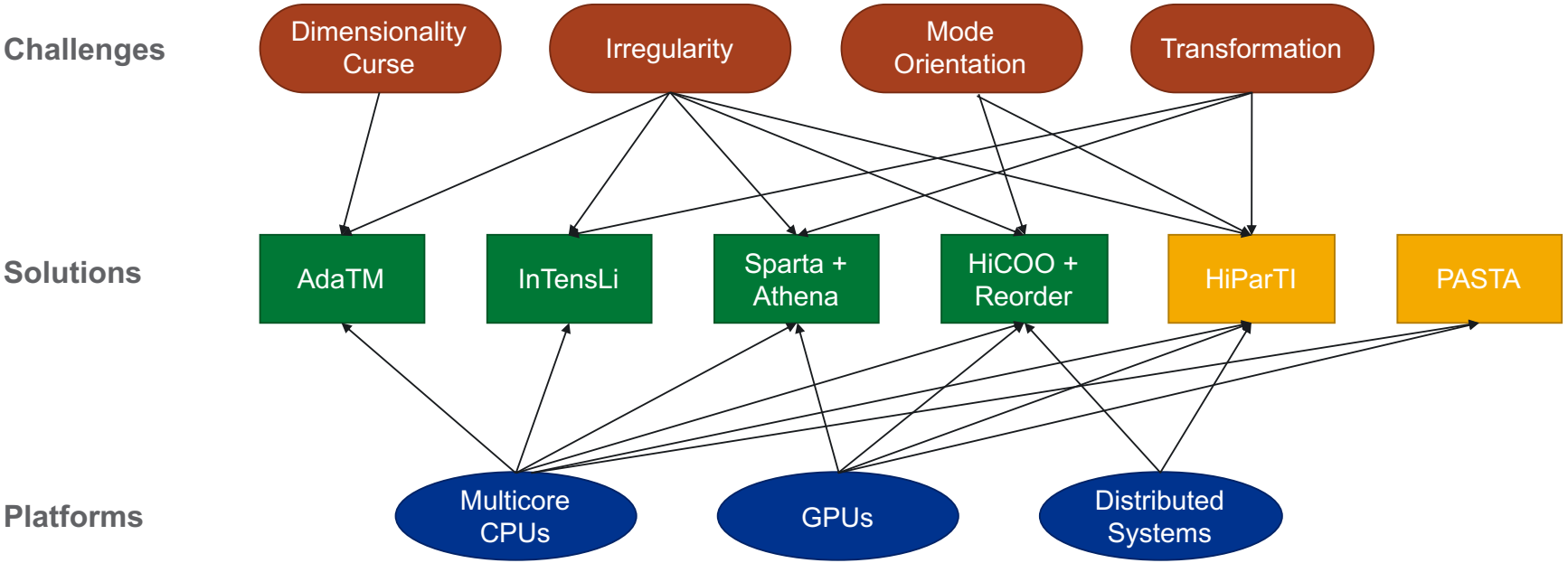
Healthcare



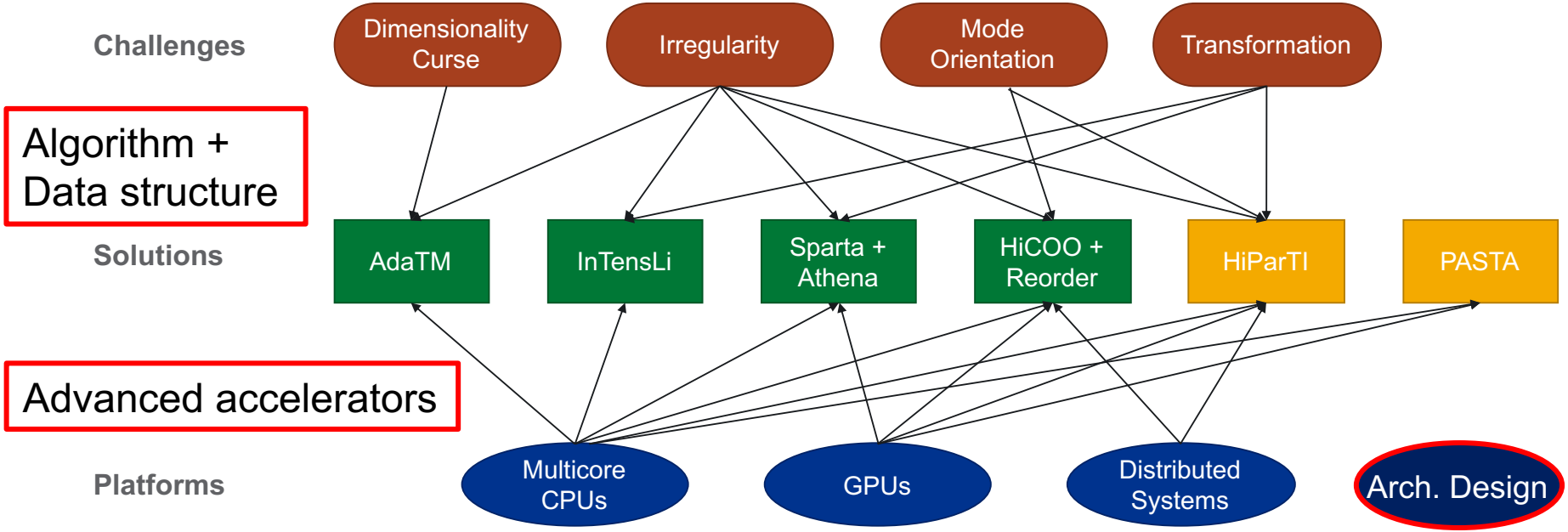
Quantum Chemistry

Figure sources: https://www.slideshare.net/AntonioMaccioni/finding-all-maximal-cliques-in-very-large-social-networks

My Prior Work
-- Multi-dimensional Data



My Prior Work



More info: CSCI-666 Sep. 27 11:00-12:30 @ McGL 002

Student Introduction

Agenda

- Introduction
- Course topics
- Course format
- Computing resources

Course Description

- Not a DL course, emphasizing on “Accelerating”.
  - Take DL methods as our running examples to show how to accelerate them in different directions.
- Goals:
  - Learn knowledge &
  - Taste of research &
  - Aware of diverse accelerating approaches &
  - Obtain good GRADES!

Pre-requisites

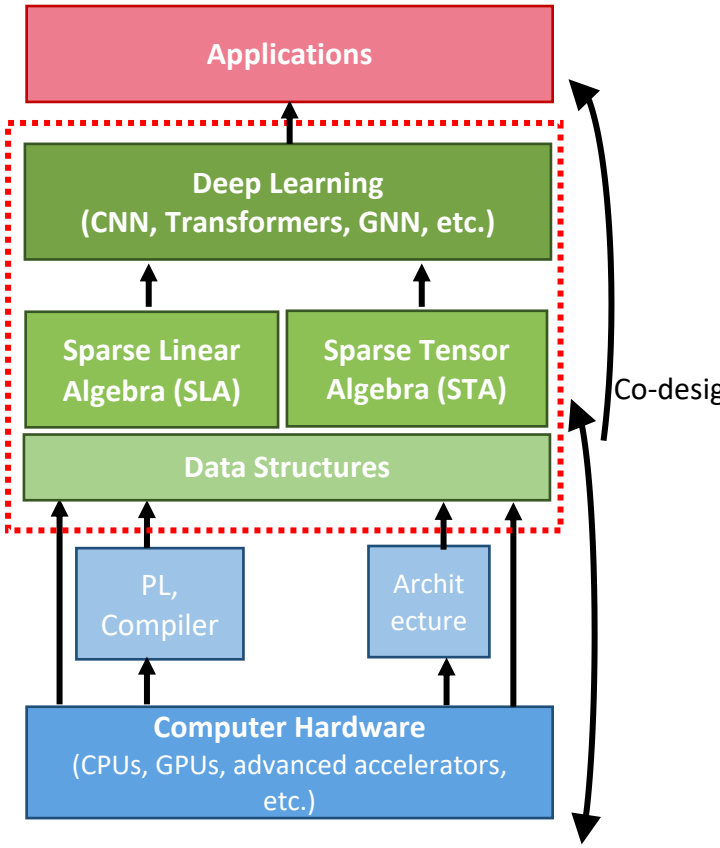
- CSCI 416/516: Intro to Machine Learning (not required for graduate students)
- Or CSCI 520: Intro to Neural Networks or Neural Networks Machine Learning (not required for graduate students)
- CSCI 303: Algorithms (not required for graduate students)

Textbooks

- Deep Learning, by Ian Goodfellow, Yoshua Bengio, and Aaron Courville (Optional)
- Multicore and GPU Programming: An Integrated Approach (1st Edition), by Gerassimos Barlas (Optional)

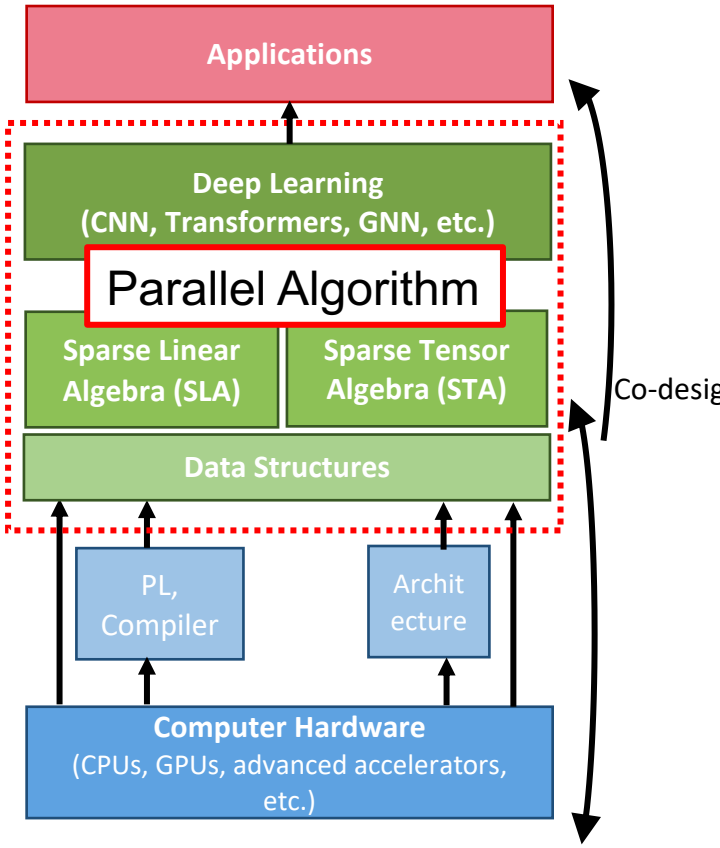
Topics

- **Topic 0**: Introduction
  - Deep Learning
  - Parallel Computing
  - Deep Learning Systems



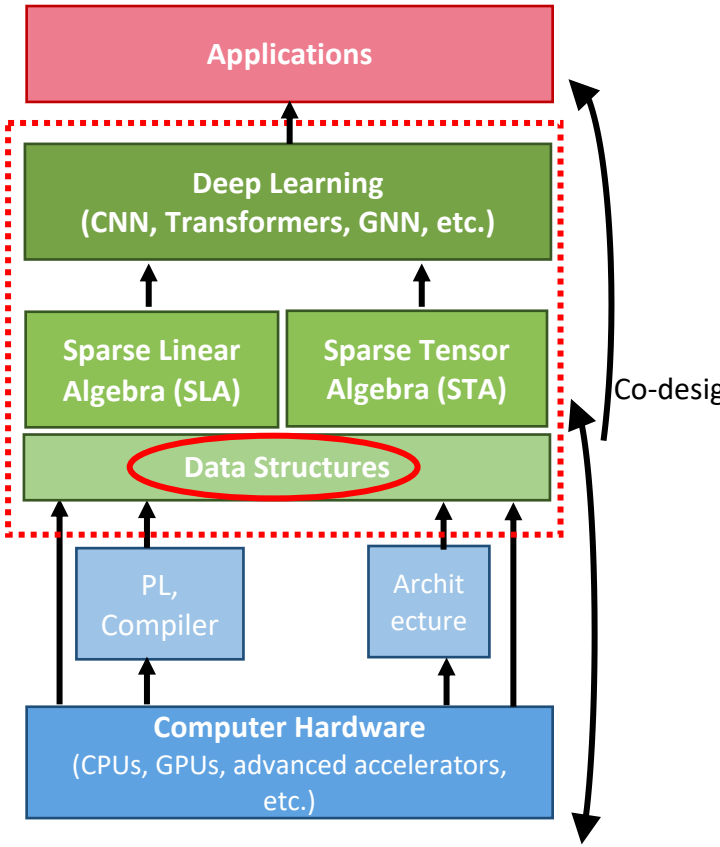
Topics

- **Topic 0**: Introduction
  - Deep Learning
  - Parallel Computing
  - Deep Learning Systems
- **Topic 1**: Parallel Deep Learning
  - Parallelism
  - Efficient Communication
  - Asynchronization
  - Load Balancing



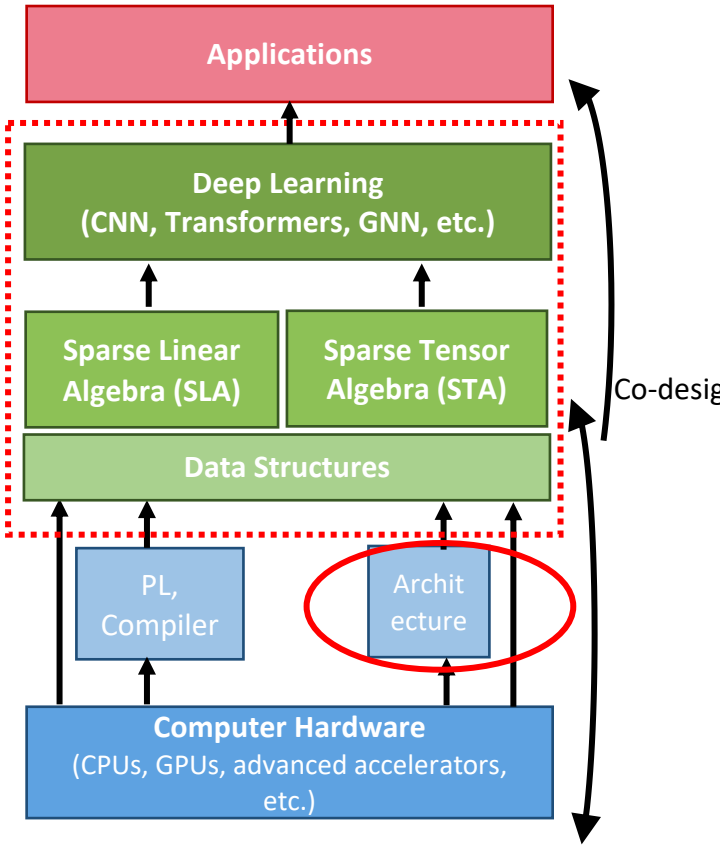
Topics

- **Topic 0**: Introduction
  - Deep Learning
  - Parallel Computing
  - Deep Learning Systems
- **Topic 1**: Parallel Deep Learning
  - Parallelism
  - Efficient Communication
  - Asynchronization
  - Load Balancing
- **Topic 2**: Lossy Compression
- **Topic 3**: Sparsity



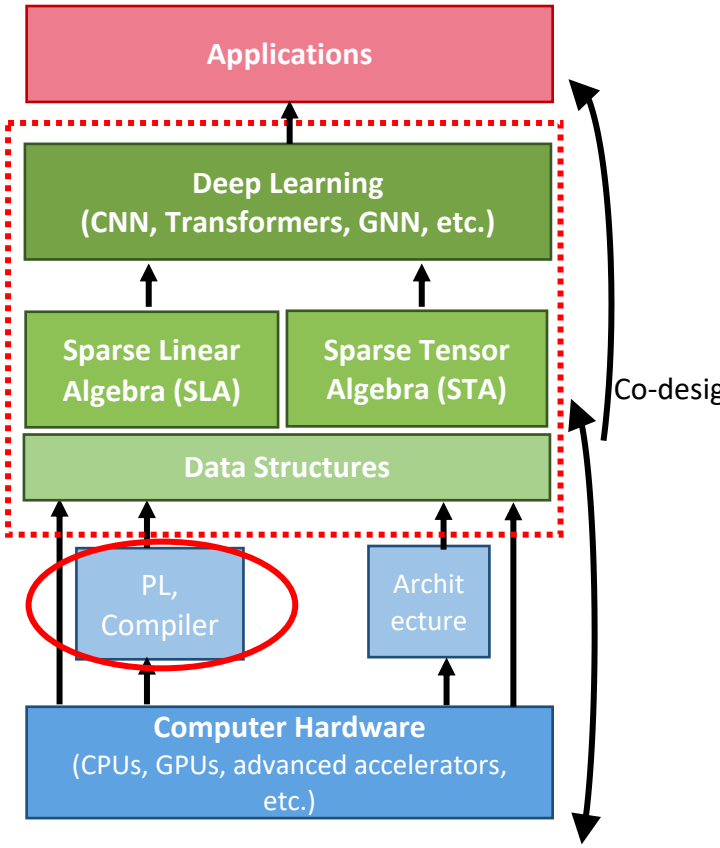
Topics

- **Topic 4**: Parallel Architecture
  - Memory Systems
  - AI Accelerators
  - Accelerator Designs



Topics

- **Topic 4**: Parallel Architecture
  - Memory Systems
  - AI Accelerators
  - Accelerator Designs
- **Topic 5**: Intro to Compiler
  - Domain-specific language
  - Compilation Techniques
  - Auto-Tuning

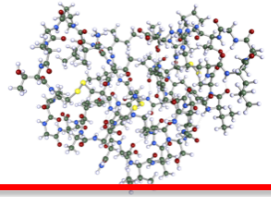


Objective

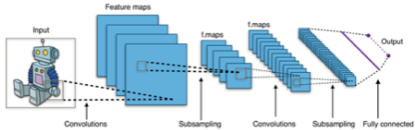
Applications

Healthcare

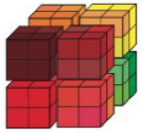
Quantum Chemistry



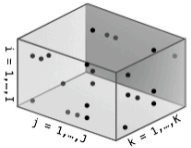
Deep Learning



Software

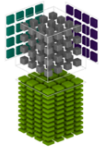


CO-DESIGN



Hardware

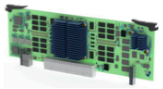
NVIDIA Tensor Cores



DL accelerators

Google TPU

EMU Chick



Agenda

- Introduction
- Course topics
- Course format
- Computing resources

Format

- Paper review + presentation
- Project
  - Presentation
  - Report + Code

- ALL DUE: 3:00pm EST

Paper Review + Presentations

- Paper bidding (2 rounds. 1st: today; 2nd: 10/14).
  - FCFS (no overwriting on others)
- Online paper review. (25%)
  - Use EasyChair system (https://easychair.org/conferences/?conf=adl21)
  - Everyone can see other students' reviews after submitted his/her own to get more insights.
  - Including evaluation score, reviewer's confidence, summary, strengths, weaknesses, things you learned or could be used in your research.
- Discussion in class. (25% + 10%)
  - Every student will give a paper presentation in turn.
  - Each paper presentation: 25 min for presentation + 15 mins for discussion.
  - Paper presentation load: 2 / student.
- Bonus (<=10%)
  - I'll pick interesting points from online reviews to discuss more in class.
  - If your questions got chosen, BOOM, you earn **bonus**!
  - 1 point per question.

Project

- Project could be single person or at most two students teamed up. For a team work project, more work will be expected.
  - Proposal (1 page). (no grade, only for topic chosen and discussion purpose)
    - **DUE**: 10/12, 3:00pm ET
  - Presentation (15 min presentation + 5 mins Q&A). (10%)
    - **DUE**: 12/02 - 12/09, 3:00pm ET
  - Final Report (4 pages, excluding references) + Code. (30%)
    - **DUE**: 12/16, 3:00pm ET
- We will use BlackBoard to submit project proposal, slides, final report+code.

Project Topics

- Related to DL
- Related to high performance computing
  - Not limited to the approaches
- Expected Outcomes (Optional)
  - Publication in different levels journal or conferences

Grading

- **Evaluation and Grading**
  - Paper review: 25%
  - Paper presentation: 25%
  - Final project presentation: 10%
  - Final Project: 30%
  - Class Participation/Online Discussion: 10%
  - **Bonus**: up to 10%
- BlackBoard will be used for grading.
- A will be the highest score, no A+.

Student Accessibility Services

Please feel free to find me for any Qs.

William & Mary accommodates students with disabilities in accordance with federal laws and university policy. Any student who feels they may need an accommodation based on the impact of a learning, psychiatric, physical, or chronic health diagnosis should contact Student Accessibility Services staff at 757-221-2512 or at sas@wm.edu to determine if accommodations are warranted and to obtain an official letter of accommodation. For more information, please see www.wm.edu/sas.

Honor Code

All work in this course is subject to the College's Honor Code. Cheating cases involving projects in CS courses are typically Level III violations of the Honor Code.

COVID

- Mask required (with nose covered)
- No food or drink in class.

Agenda

- Introduction
- Course topics
- Course format
- Computing resources

Computing Resources

- GPU machines
  - bg1 (3xRTX 2080Ti), bg2 (4xGTX 1080Ti), bg6 (3xTitan RTX)
  - th121-{23-24} (GTX 1090Ti)
- CPU machines
  - (w/ more cores):
    - bg7 (72c), bg8 (44c), bg{9-12} (64c)
  - Other normal ones:
    - th121-{13-18}, th121-{1-12}
    - bg{1-6}

Demo

- Apply a department account:
  - https://accounts.cs.wm.edu/newuser_template
- ``ssh [xx].cs.wm.edu``
- Play with that.
  - Install packages
  - Compilation
  - Test Running

My Computing Resources (Upcoming)

- AMD CPU + GPU
- ARM (A64FX)
- NVIDIA Tesla A100 GPU
- Potential cutting-edge new accelerators

HW

- Preparation stuff
  - Accept my invitation & register EasyChair
  - Accept my invitation to Slack
  - Fill out Survey
  - Choose presentation papers

- DUE: 09/07 3:00 ET

Q & A

- **Office hours:** Tue @ 2:00-3:00pm ET, MCGL 102 or Zoom
- Slack

- Welcome for any suggestions or feedback!