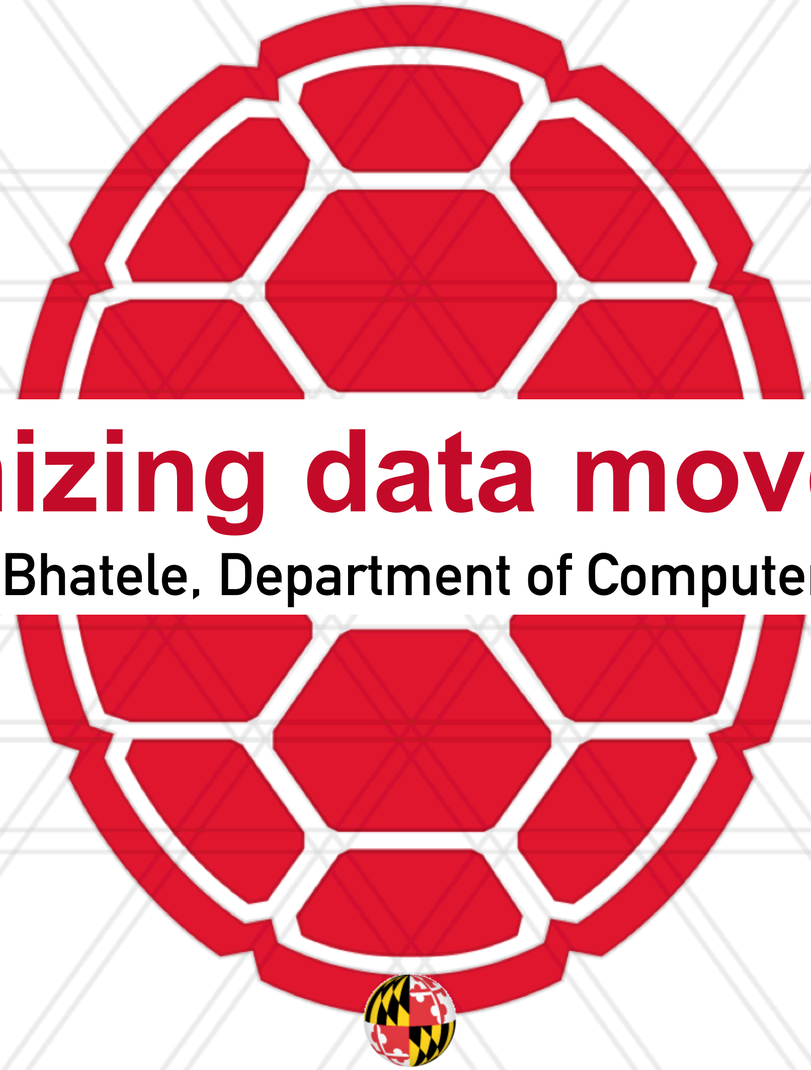
Systems for Machine Learning (CMSC828G)



Optimizing data movement

Abhinav Bhatele, Department of Computer Science

UNIVERSITY OF MARYLAND

Annoucements

- Interim report due tomorrow
- Project presentation slots: all filled now
- Students who haven't presented in class yet: Submit a 5-minute video recording on a paper of your choice (from the assigned readings)
  - Due date: May 1
  - If you want to do this in groups of two, that is okay
  - Presentation uploaded to gradescope should have a link to the video recording (Youtube, Google drive, …)

Various types of data movement

- Between CPU and GPU (host-device transfers)
- Within the GPU memory hierarchy
- Between storage and memory (disk I/O)
- Between devices in parallel training (network communication)
- These can impact: computation time, scaling, and energy efficiency

Strategies to optimize data movement

- On device:
  - Better data layouts
  - Caching frequently used data: KV cache
- I/O
  - using parallel data loaders
- Network communication
  - optimized collectives
- Pre-fetching data
- Overlapping asynchronous data movement with compute

Strategies to optimize data movement

- Send data in reduced precision
- Only send non-zeros: exploit sparsity
- Other approximation techniques to reduce size of tensors

Disaggregated systems

- Separate CPU, memory and storage into independent pooled resources connected via high-speed networks
- Different users can request different amounts of resources from pools
- Research question: how do you optimize DL applications running on such systems

UNIVERSITY OF
MARYLAND

Abhinav Bhatele

5218 Brendan Iribe Center (IRB) / College Park, MD 20742

phone: 301.405.4507 / e-mail: bhatele@cs.umd.edu