Systems for Machine Learning (CMSC828G)

Optimizing data movement

Abhinav Bhatele, Daniel Nichols

UNIVERSITY OF
MARYLAND

Annoucements

- Interim report due today
- Presentation slots: will be emailed soon
- Students who haven't presented in class yet: Submit a 5-minute video on a paper of your choice
  - Due date: May 1
  - If you want to do this in groups of two, that is okay
- Extra credit, will be posted on April 23 and due on May 7

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

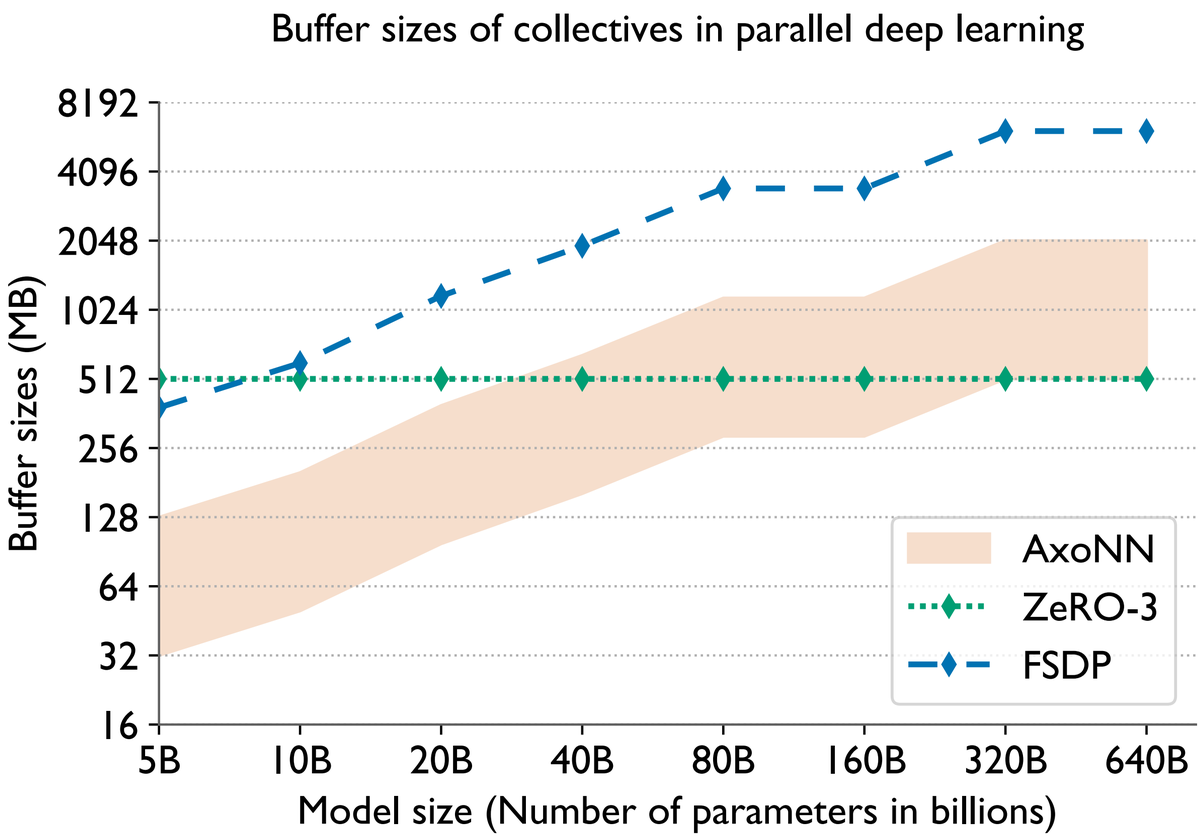
Strategies to optimize data movement

- On device:
  - Better data layouts, pre-fetching
  - Caching frequently used data: KV cache
- I/O
  - pre-fetching, overlapping using asynchronous I/O
  - using parallel data loaders
- Network communication
  - overlapping using asynchronous I/O
  - optimized collectives

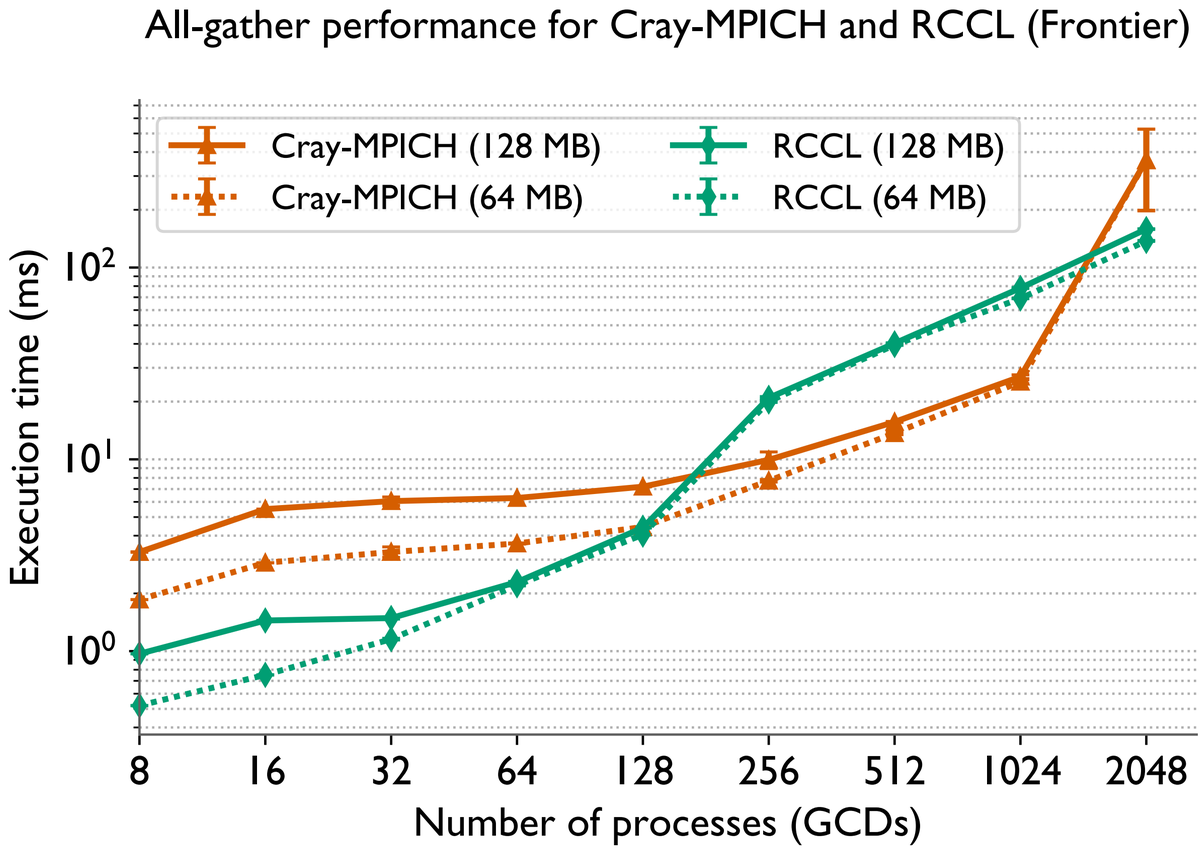
Strategies to optimize data movement

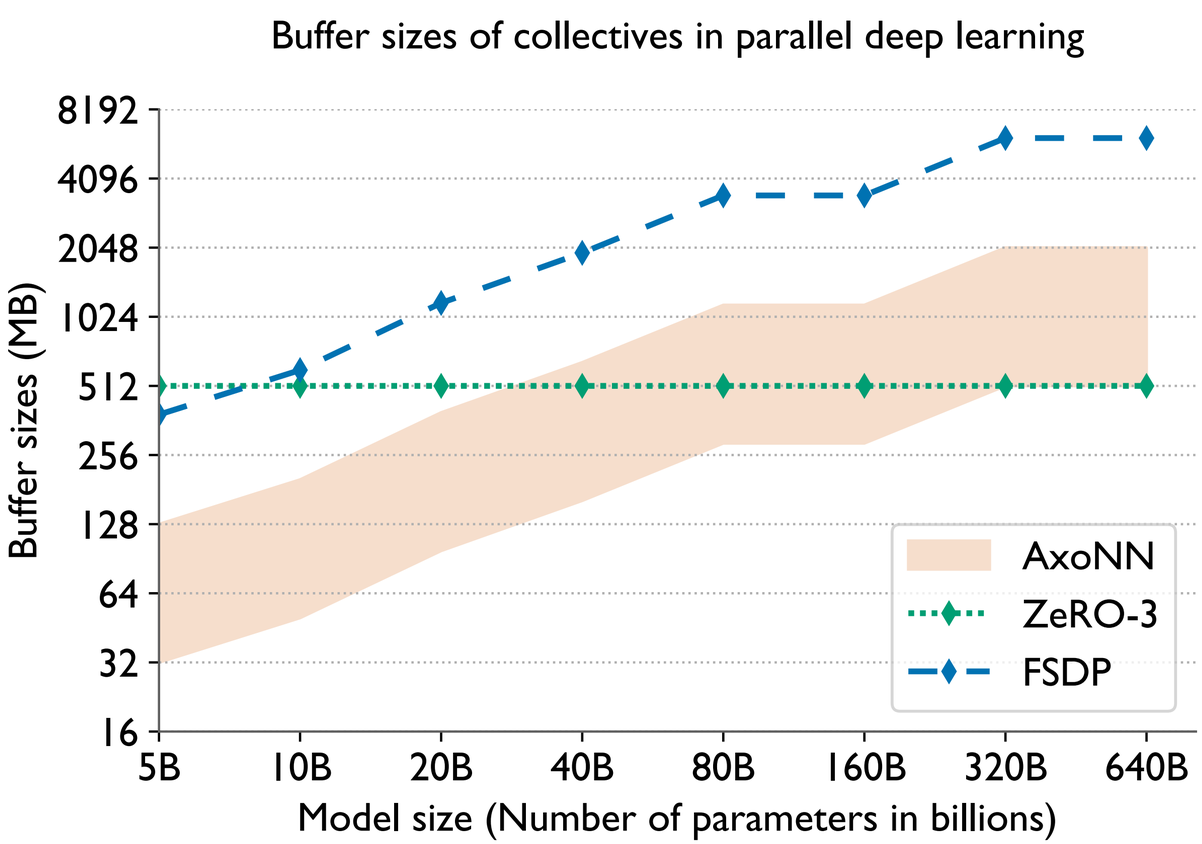
- Send data in reduced precision
- Only send non-zeros
- Other approximation techniques

Why optimize collectives?



Why optimize collectives?





UNIVERSITY OF
MARYLAND