Experimenting with LibMP

Thomas Hines

University of Tennessee Chattanooga

9/29/2022

LibMP Overview

- LibMP - a lightweight messaging library built on top of LibGDSync APIs to support GPUDirect asynchronous communication
- LibMP key features:
  - a thin layer built on top of IB Verbs and LibGDSync
  - MPI used to setup IB connections
  - No MPI calls are used for actual communications
  - Uses only point-to-point and one-sided communications (no collectives)
  - No tags, no wildcards, no data types
  - Could be used to combine GPUDirect Async with GPUDirect RDMA

Source: https://github.com/gpudirect/libmp

Unexpected Messages with LibMP

- If the receive is not posted before a message arrives then LibMP waits for the receive.
- 10x slower than if the receive is posted first

Benchmark - Pulse

- 3D stencil computation
  - Real game of life
  - Total life on board as a checksum
- Configurable
  - # variables
  - Halo width
  - Kernel execution time
- Posts receives one step before to avoid unexpected messages

Message Configuration

- Explicit – 26 sends
  - One for each of the faces, edges, and corners in the data cube
- Implicit – 6 sends, but ordered
  - One for each face, including the edges and corners
  - Send/Recv X faces, then Y faces, then Z faces

Packing/Unpacking

- Kernel for each message
- One fused kernel that packs/unpacks everything

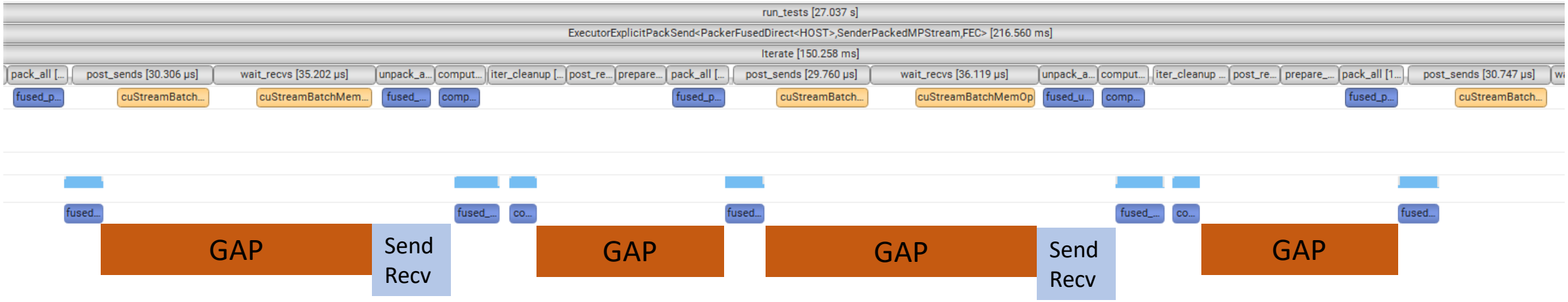
Packed Buffer Location

- Device memory
- Pinned host memory
  - Directly written to by packing kernel

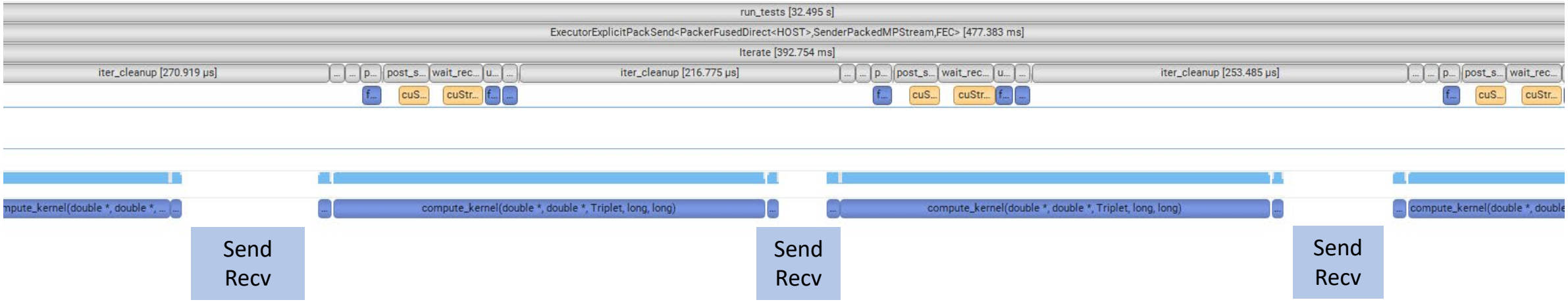
Send Modes

- Nonblocking (MPI_Isend)
- Persistent (MPI_Send_init/Start/Wait)
- LibMP CPU triggered (mp_isend)
- LibMP Stream triggered (mp_send_prepare/isend_post_on_stream)
- LibMP Kernel triggered – NYI
- LibMP Graph triggered - NYI

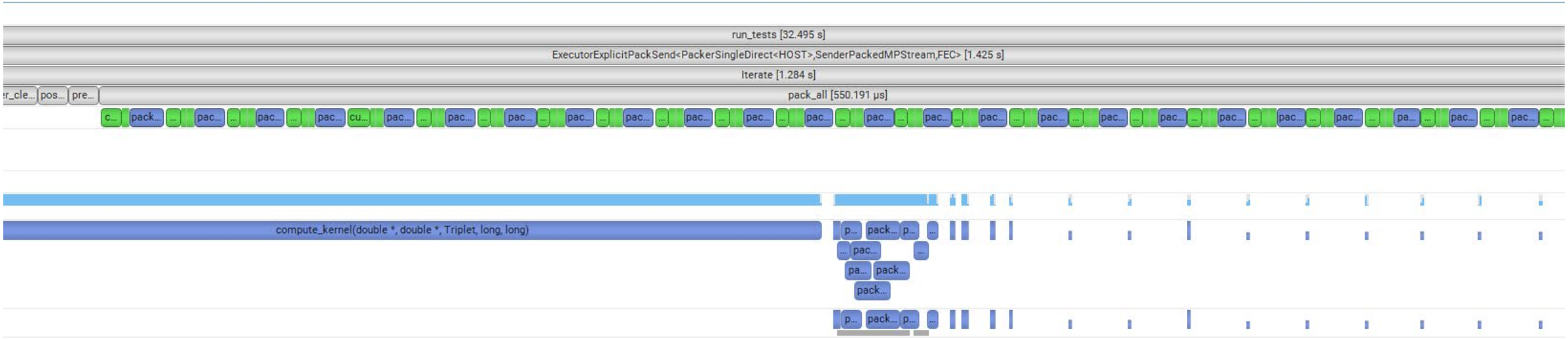
Short vs. Long Compute Kernel



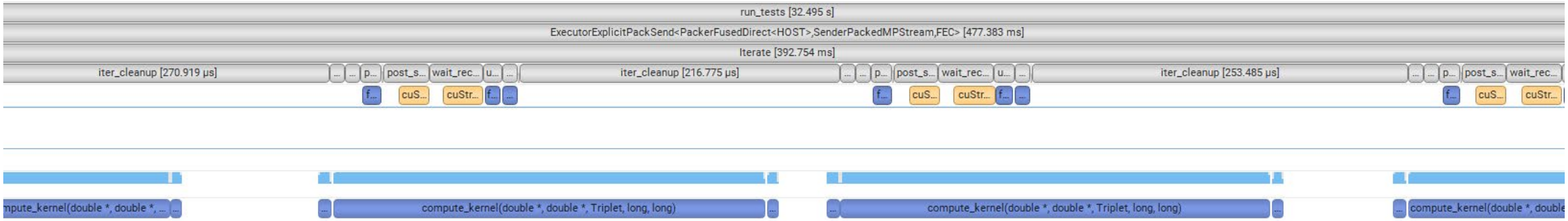
Short vs. Long Compute Kernel



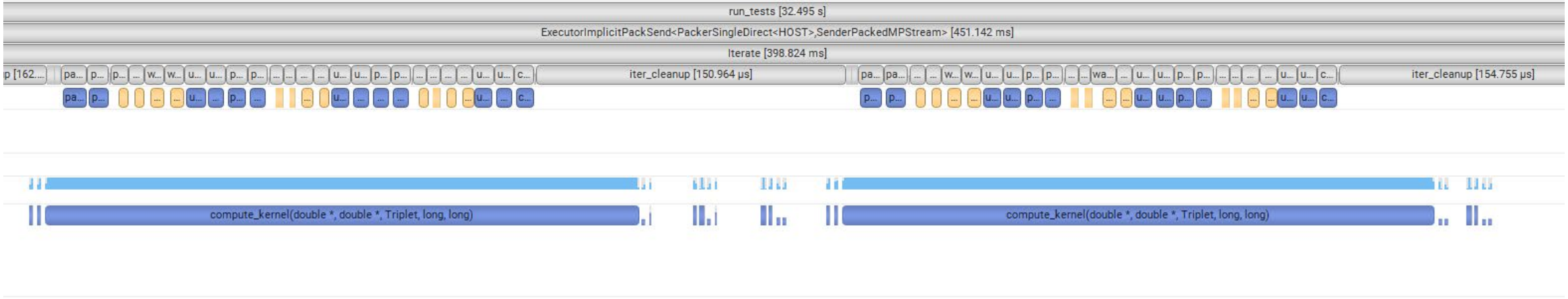
Single Kernel vs One per Message



Single Kernel vs One per Message

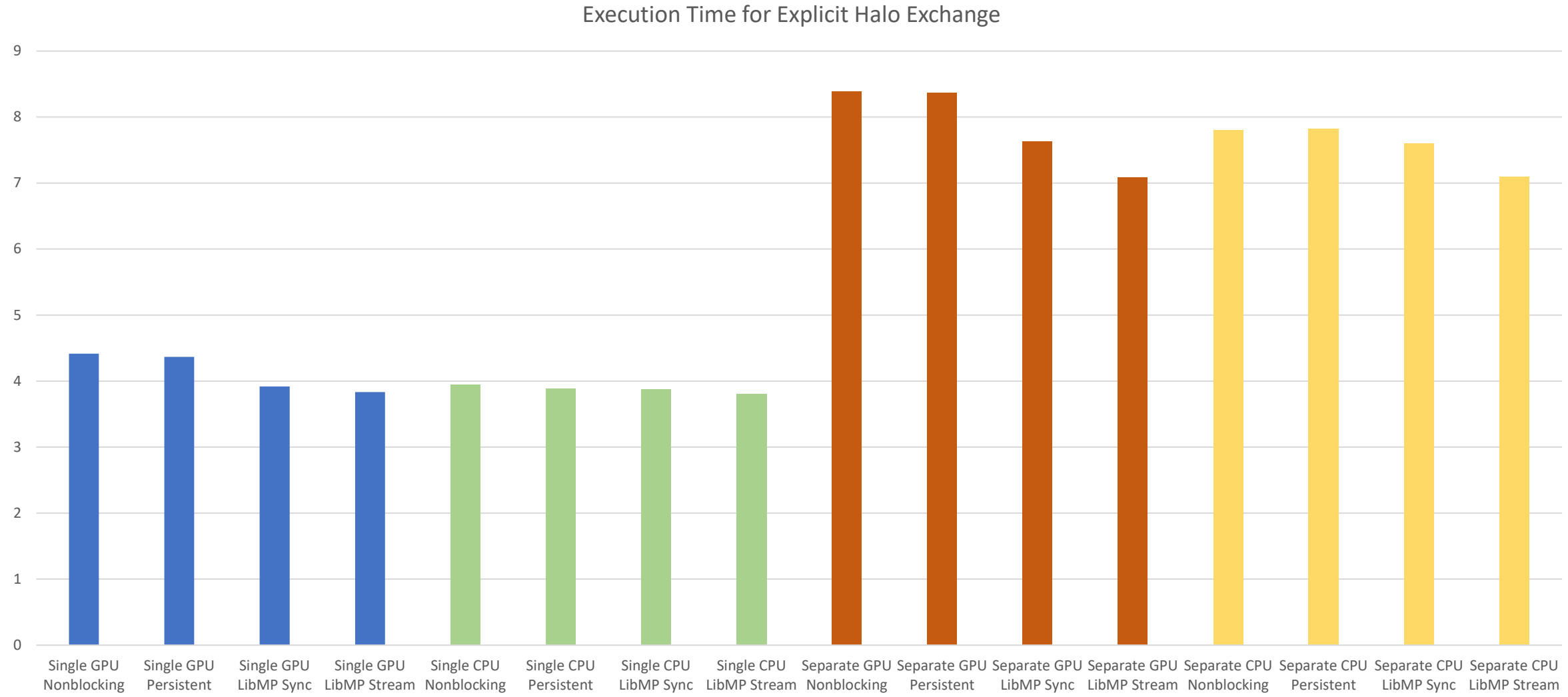


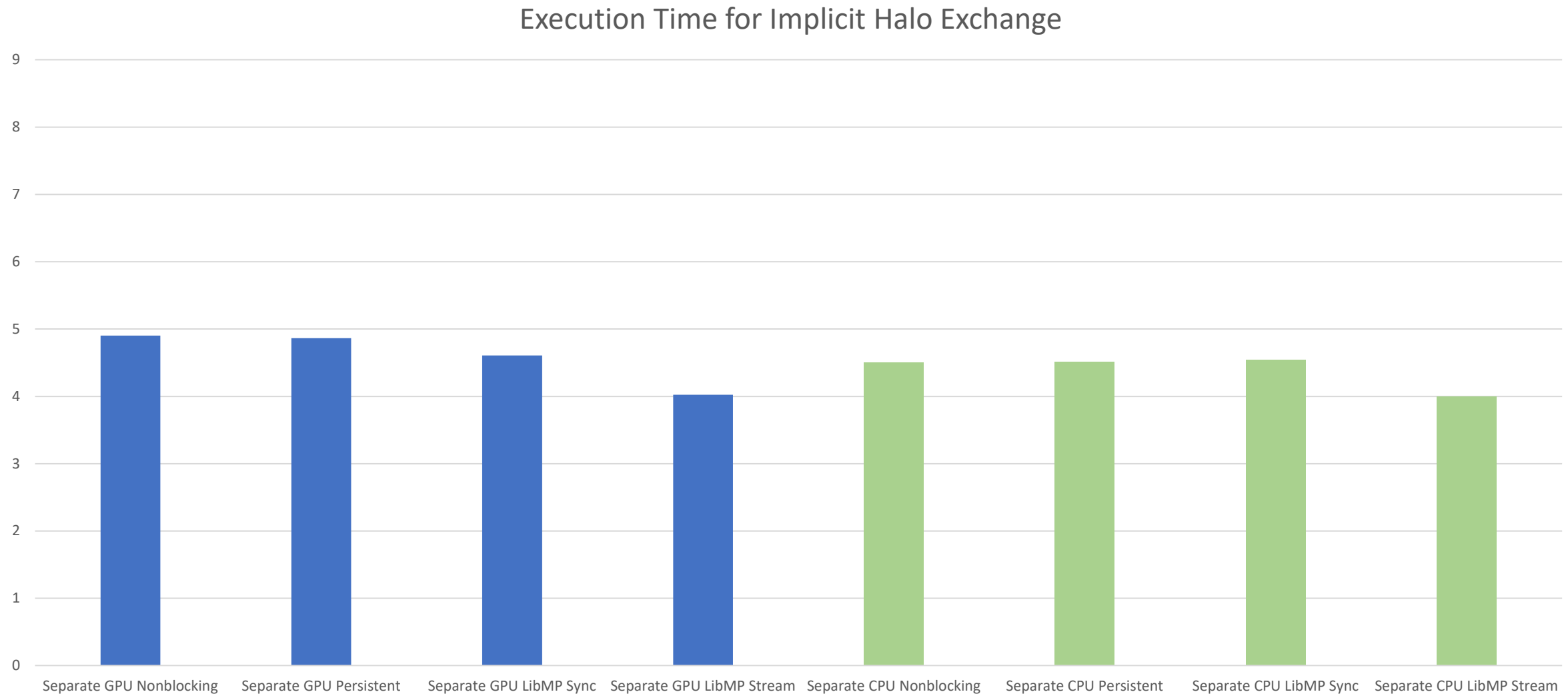
Explicit vs. Implicit



Experiment Details

- Lassen
- 50x50x50 per process
- 4x4x4 process grid (1 GPU per process)
- One variable
- Size one halo
- “Long” compute kernel


Execution Time for Explicit Halo Exchange
9
8
7
6
5
4
3
2
1
0
Single GPU Nonblocking
Single GPU Persistent
Single GPU LibMP Sync
Single GPU LibMP Stream
Single CPU Nonblocking
Single CPU Persistent
Single CPU LibMP Sync
Single CPU LibMP Stream
Separate GPU Nonblocking
Separate GPU Persistent
Separate GPU LibMP Sync
Separate GPU LibMP Stream
Separate CPU Nonblocking
Separate CPU Persistent
Separate CPU LibMP Sync
Separate CPU LibMP Stream


Execution Time for Implicit Halo Exchange
9
8
7
6
5
4
3
2
1
0
Separate GPU Nonblocking
Separate GPU Persistent
Separate GPU LibMP Sync
Separate GPU LibMP Stream
Separate CPU Nonblocking
Separate CPU Persistent
Separate CPU LibMP Sync
Separate CPU LibMP Stream

Ongoing Work

- Parallel Stream triggering
  - Hangs, working with Nvidia
- Kernel triggering
- Graph triggering
- Fused implicit message packing/unpacking

Future Work

- Put all this back into Comb
  - Will require some work
- Try on different systems
  - LibMP tricky to set up