Problem Details

Communication performance has not improved at the same rate as computational performance on GPU-based machines. The production LLNL rad-hydro code Ares was ported to GPUs but has lower communication performance than reported. Relevant characteristics:
- 2/3D semi-structured meshes
- parallelized by mesh splitting
- halo-exchange communication

Communication Benchmark

The Comb halo-exchange communication performance benchmarking tool was written to find ways to improve communication performance.
- Explore memory spaces, execution methods, and communication methods
- Comb steps:
  1. Placeholder computation
  2. Halo-exchange communication
  3. Pack and Send
  4. Wait Recv and Unpack
  5. Wait Send
  6. Placeholder computation

Communication Options

Look at communication performance on Sierra by combining various resources and techniques.
- Memory spaces:
  - Host based memory
  - Host pinned memory
  - Device based memory
  - CUDA device memory
  - CUDA managed memory
- Execution methods:
  - Serial
  - OpenMP host threading
  - CUDA kernels
  - CUDA graphs
  - Manual CUDA kernel fusion
- CUDA Aware MPI datatypes
- Communication staging methods:
  - pack and send one message at a time
  - pack messages and send messages in groups
  - Pack all and send all messages
- Communication libraries and methods:
  - MPI
  - CUDA Aware MPI (GPU direct)
  - Littigroup (GPU direct async)

Future Work

More fully examine:
- performance of MPI with one rank per core. CPU results here use OpenMP to utilize all CPU cores.
- parallelizing groups of messages using CUDA streams to parallelize packing and potentially increase packing and communication overlap.
- automatic CUDA kernel fusion to reduce amount of communication code rewrite while gaining the benefits of kernel fusion.
- stream triggered communication (GPU direct async) to eliminate device synchronization overhead, and potentially increase packing and communication overlap.
- CUDA graphs as a means of combining automatic CUDA kernel fusion and stream triggered communication.

Acknowledgements

LLNL
- Lou Ellison
- Arjun Gambhir
- Olga Pearce
- Brian Pauline
- Brian Ryujin

Halo Exchange Performance on the Sierra Supercomputer

Jason Burmark

LLNL

LLNL-POST-884617