Collectives Working Group Report

Torsten Hoefler and Andrew Lumsdaine

Open Systems Lab
Indiana University
Bloomington, USA

MPI Forum Meeting

Chicago, IL, USA

10th March 2008

Goals

Redesign the collective interface to the changed environment in HPC. Use gained experience to improve or correct current interface.

The approach

1. discuss
2. design
3. implement
4. propose

Were are we?

- stage 1/2/3
- everything in this talk is preliminary and open for discussion

Organizational

February Telecon

- 8 participants/7 institutions
- discussed group focus
- defined first opinions
- formed subgroups

Current Items

- new collectives
- non-blocking interface
- persistent interface
- sparse/topological collectives
- modularity/subsetting

New collective operations

MPI_Reduce_scatter_block()

- regular, non-verctor variant of MPI_Reduce_scatter()
- new MPI operations
- maybe MPI-2.2?

???

- new proposals welcome

Non-blocking and Persistet Interface

MPI_I<coll> for all

- very close to the original proposal for MPI-2 (see early mailinglist discussions)
- multiple outstanding requests
- but no matching between blocking non-blocking
- limited cancel/free functionality
- reference implementation available as open-source (LibNBC)

MPI_<coll>_init() for all

- cf. persistent send/recv
- might have performance benefits
- might also bloat interface

Sparse/Topological Collectives

- no consensus yet - two proposals under consideration -

1: MPI_<coll>_sparse(..., MPI_Group group)

- allows arbitrary processes to communicate
- arbitrary collectives
- might be hard to implement/optimize
- might bloat interface

2: Topological Collectives

- defined on cartesian/graph communicators
- graph communicator enables arbitrary processes to communicate (i.e., parse Alltoallv)
- implemented in open-source implementation (LibNBC)
- used in a real-world application (with arbitrary graphs)
- only 2/3 functions added to interface

Interaction with other Groups

Graph Topologies

- current interface is not scalable
- collaborate with existing group
- - or - make it topic of the group

Modularize collective Interface

- blocking collectives
- non-blocking collectives
- topological collectives

Coordination with Subsetting Group

- Torsten participates in subsetting discussions
- try to collaborate

Thanks

More Information:

- Mailinglist
- Wiki
- Teleconferences (announced on ML)
- LibNBC (http://www.unixer.de/NBC)