Argonne NATIONAL LABORATORY

Non-Collective Communicator Creation

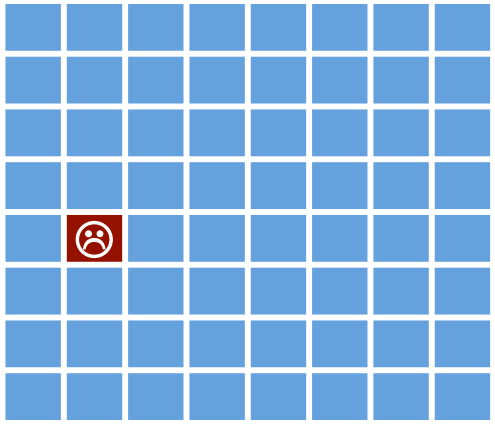
Tickets #286 and #305

int MPI_Comm_create_group(MPI_Comm comm,
MPI_Group group, int tag, MPI_Comm *newcomm)

Non-Collective Communicator Creation in MPI. Dinan, et al., Euro MPI '11.

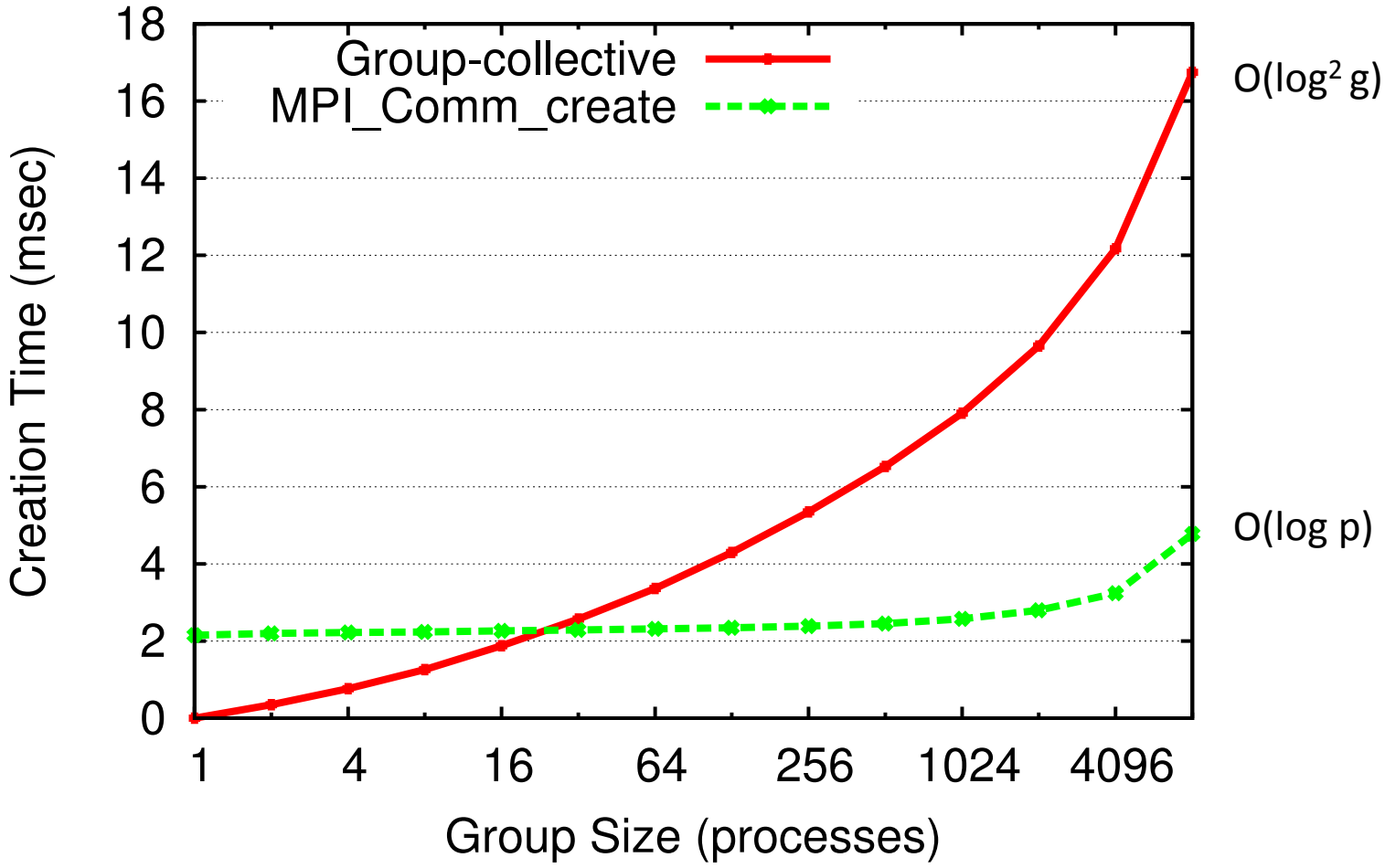
Non-Collective Communicator Creation

- Current: Collective on parent
- Proposed: Create communicator collectively only on new members



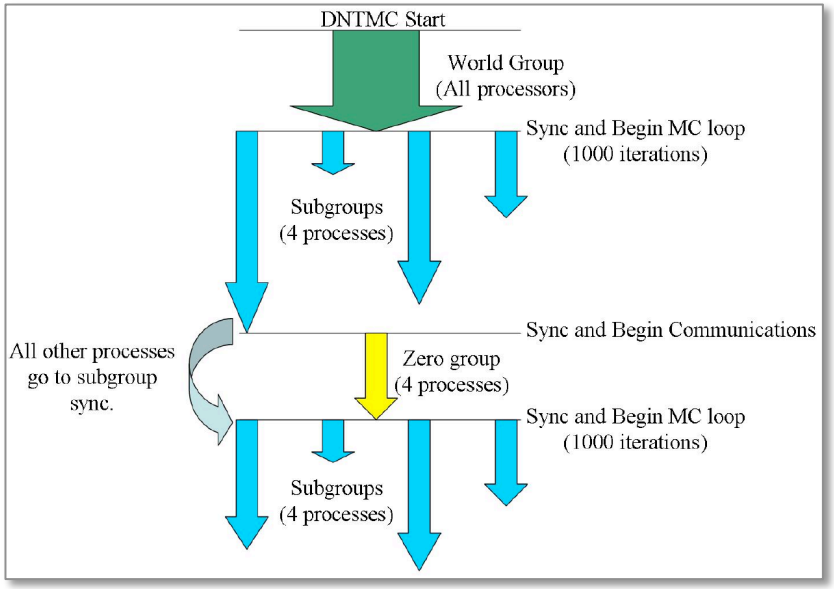
1. Avoid coarse-grain synchronization
   - Load balancing: Reassign processes from idle groups to active groups
2. Reduce overhead
   - Multi-level parallelism, create small communicators
3. Recovery from failures
   - Not all ranks in parent can participate
4. Compatibility with Global Arrays
   - Past: collectives using MPI Send/Recv

Evaluation: Microbenchmark



Case Study: Markov Chain Monte Carlo

- Dynamical nucleation theory Monte Carlo (DNTMC)
  - Markov chain Monte Carlo
  - Part of NWChem
- Multiple levels of parallelism
  - Multi-node "Walker" groups
  - Walker: N random samples
- Load imbalance across groups
  - Regroup idle processes into active group
  - 37% decrease in total application execution time
  - *Load Balancing of Dynamical Nucleation Theory Monte Carlo Simulations Through Resource Sharing Barriers* [IPDPS '12]



T L Windus *et al* 2008 *J. Phys.: Conf. Ser.* **125** 012017

Update: Tagged Collectives Tag Space

- Define two tag spaces:
  - Point-to-point tag space
  - Tagged collectives tag space
- Spaces share the same semantics
  - MPI_TAG_UB
  - MPI_TAG_ANY
  - …
- Tags match only within the same space
- Ticket #305: Extend MPI_Intercomm_create(…) to use TCTS
  - Backward compatible, much easier to use