Collective communication over a group of processors is an integral and time consuming component in many high performance computing applications. We present novel techniques to perform collective operations over a subset of processors in a torus network obtaining over 5x speedup over current best.

### Motivation

- **Why Collectives?**
  Performance of MPI collectives is critical for improved scalability and efficiency, for large messages network contention need to be avoided.

- **Why Sub-communicators?**
  Many applications involve collective operations on sub partitions (i.e., sub-communicators in MPI).

- **Why a new approach?**
  The algorithms developed so far are not contention free for a random sub-communicator.

### Common Sub-communicators

- **Random Communicator**
  Our algorithm achieves a single link throughput performance.

- **Single Missing Node**
  The tree is constructed using a pseudo root that is at least one hop away from the missing node in every dimension. The missing node is by-passed during the tree construction and the directions of the links are adjusted according to the user specified root.

- **Random subset of parallel planes**
  Phase 1 and 2 proceeds as described in [1] assuming that the missing planes are in the Z direction. For phase 3, the data is transferred along the Z dimension, with the missing planes in Z dimension being bypassed to transfer the data to the nodes in next 2D plane.

### Results

All the experiments were performed on Blue Gene/P.

**Implemented versions**
1. Optimized version which uses lower level API (SPI)
2. MPI version by modifying the MPI stack

**Results**
Our MPI implementation obtains a speedup of 1.5-3x over IBM’s product MPI for a single random communicator. Our SPI implementation achieves a speedup of 1.7-5.8x for special sub-communicators. For multiple sub-comm.s, we observe a speedup of 1.6-2.6x.

---

[1] Faraj et al., MPI collective communications on the Blue Gene/P supercomputer: algorithms and optimizations*. In ICS 2009