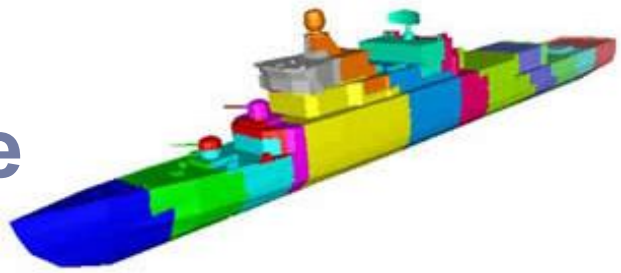
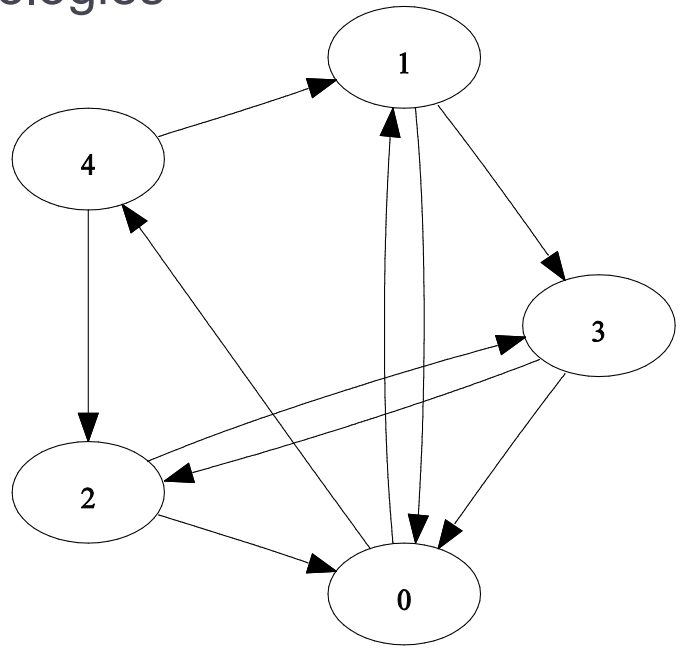
MPICH Birds of a Feather Portland, OR, November 2009

Torsten Hoefler

The Graph Topology Interface

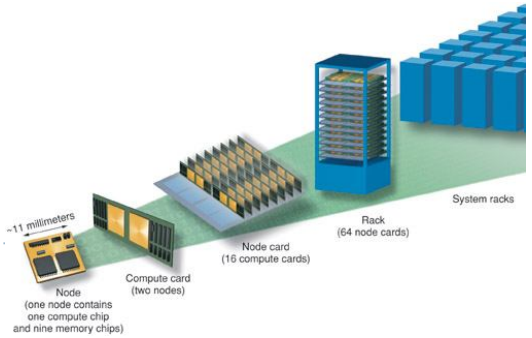


- The MPI Graph Topology in MPI-1
  - specify communication neighborhoods/topologies
  - specifies **full** graph at **each** process
  - process 5 knows neighbors of process 0
  - O(P^2) memory per process – O(P^3) total
  - → MPI-1 interface is non-scalable!
  - → it's rarely used



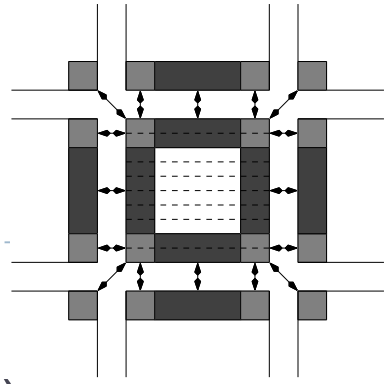
- Why should you use topologies?
  - enabling optimized process mapping
  - arrange neighborhood relations in a structured manner
  - give hints to the MPI library (where are messages sent to?)

Scalable Topologies in MPI-2.2



- MPI_Dist_graph_create()
  - each process can specify any edge in the graph
  - very helpful for ParMETIS partitions
- MPI_Dist_graph_create_adjacent()
  - each process specifies incoming and outgoing edges
  - each edge is specified twice (at src and tgt)
- The interface offers weights
  - MPI_UNWEIGHTED can be specified
  - semantics of weights can be defined by info object
- Neighbor queries are local only
  - requires communication for remote query (needed?)

Topological Collective Operations



- Topological Collectives
  - MPI_Neighbor_reduce(), MPI_Neighbor_alltoall(), MPI_Neighbor_gather()
  - Hoefler, Traeff: “Sparse Collective Operations for MPI”
  - We actively seek user-feedback! Talk to us!
- Streaming Collectives
  - react to data as it comes in
  - not decided yet, is there a need for this?
- Persistent Collectives
  - persistent P2P does not seem to be used much
  - would you like persistent collectives?

Nonblocking Collective Operations

- Nonblocking Collectives are accepted for MPI-3
  - MPI_Ibcast(&buf, 1, MPI_INT, 0, comm, &req)
  - /* compute */
  - MPI_Wait(&req, MPI_STATUS_IGNORE);
  - Concrete plans by MPI implementers
  - reference/preview implementation: LibNBC
- Three obvious use-cases:
  - overlapping communication and computation
  - relaxing synchronizations (load balance, OS noise)
  - new synchronization semantics (collective protocols)

Why's did they invite this guy?

- MPICH2 v1.2.1 fully supports MPI-2.2
  - scalable topology is implemented
  - creation as low as O(log P)

- Support for nonblocking collectives is planned
  - In MPICH version 3.0.x
  - works with LibNBC today (not optimized though)

- We're seeking feedback for the MPI Forum
  - talk to your favorite MPI implementer
  - or me ☺