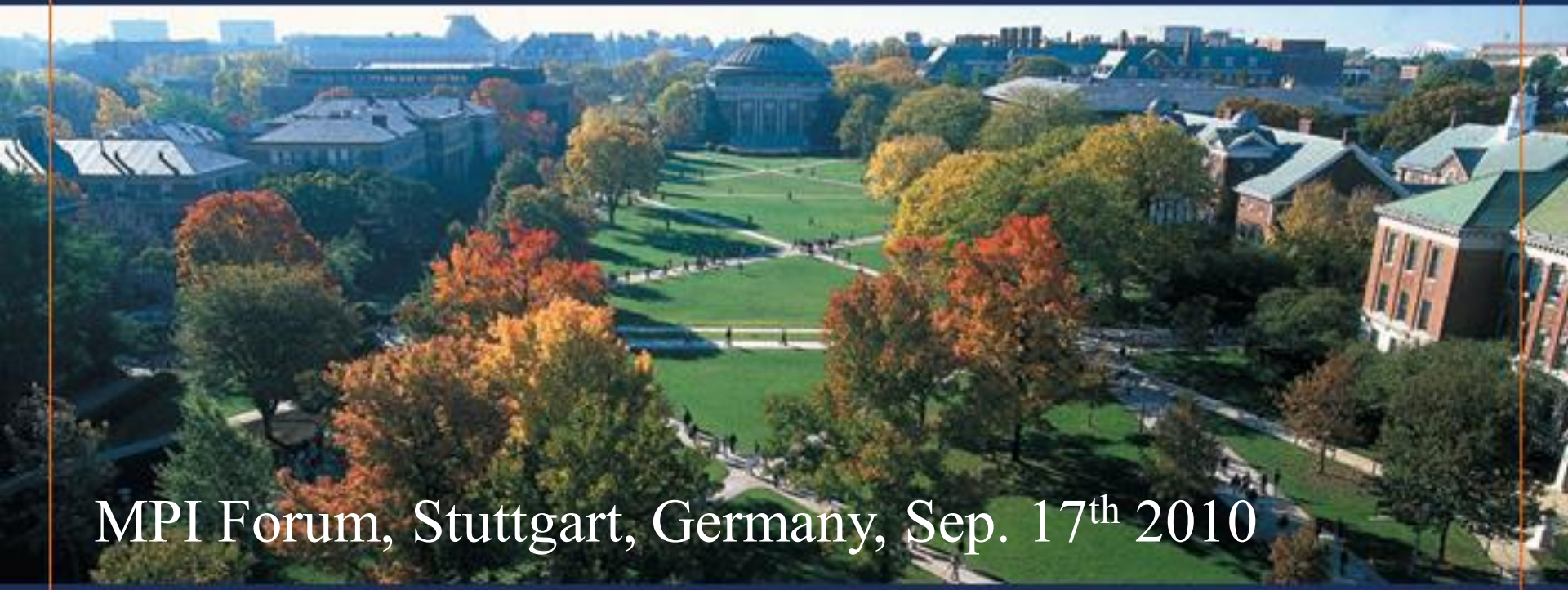
Collective Communications and Topologies Working Group

Torsten Hoefler



MPI Forum, Stuttgart, Germany, Sep. 17th 2010

UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

Roadmap Proposal Discussion

- NBC is on track (first vote at this meeting)
  - Left the WG focus
- Sparse Collectives are thought out
  - Proposal Straw Man exists – push forward to Forum?
- Persistent Collectives in flux
- One Sided Collectives
  - What are the benefits?
- New proposal from Hybrid WG
  - MPI_Comm_merge – discuss ticket (simple)

Sparse Collectives Status

- Reference implementation in LibNBC
  - Application examples exist (working on more)
  - EuroMPI'08: "Sparse Non-Blocking Collectives in Quantum Mechanical Calculations"
  - HIPS'09: Sparse Collective Operations for MPI
- IBM has very similar functionality in DCMF
  - Working with Sameer on prototype Sparse Collectives port

Sparse Collective Straw Man

- Add as extension to Chapter 7
  - 7.6 Sparse Collective Communication on Process Topologies
- Received and fixed some comments
  - Seems ready for release to Forum
- Got through proposal draft now!

SC: MPI_Neighbor_gather

- int MPI_Neighbor_gather(void* sendbuf,
  int sendcount, MPI_Datatype sendtype,
  void* recvbuf, int recvcount, MPI_Datatype recvtype,
  MPI_Comm comm)

- Sends the **same** data to each neighbor
  - Vector variant for receiving different sizes

SC: MPI_Neighbor_alltoall

- int MPI_Neighbor_alltoall(void* sendbuf,
  int sendcount, MPI_Datatype sendtype,
  void* recvbuf, int recvcount, MPI_Datatype recvtype,
  MPI_Comm comm)

- Sends **personalized** data to each neighbor
  - Vector variant for different size comms
  - W-variant for optimized DDT layouts

SC: MPI_Neighbor_reduce

- int MPI_Neighbor_reduce(void* sendbuf,
  int sendcount, MPI_Datatype sendtype,
  void* recvbuf, int recvcount, MPI_Datatype recvtype,
  MPI_Op op, MPI_Comm comm)

- Vector variant for overlapping neighborhoods
- It's not directly applicable to stencils
  - Clear use-case needs to be found
  - We might decide to exclude it for now (?)

One Sided Collectives

- Open Questions:
  - Are implementations/ideas mature enough to be standardized?
  - Is it MPI-ish enough?
  - How is it related to MPI-3 RMA
    - Which is itself in progress
  - What are the proposed interfaces
    - Learn from UPC? CAF?

MPI_Comm_merge

- Inherited from Hybrid WG
  - In-focus for our WG
- Discuss ticket here!

Questions/Discussion

Questions?/Discussion