発表概要

京 Tofu における MPI-3.0 隣接集団通信の実装と評価

畑中 正行[1,a)]　堀 敦史[1]　石川 裕[1,2]

2014年7月30日発表

本発表では，京 Tofu インターコネクトに最適化された MPI 隣接集団通信プリミティブの実装と評価について説明する．隣接集団通信プリミティブは袖通信のような，隣接プロセス間のデータ交換を最適化するために，MPI 3.0 仕様で導入されている．このプリミティブによって与えられる通信パターンの事前知識に基いて，MPI_Neighbor_alltoallw を実装した．本実装では複数の RDMA エンジンおよびネットワークリンクを有する Tofu インターコネクト用の RDMA 転送スケジューラと組み合わせた．RDMA 転送スケジューラは RDMA エンジン間の負荷不均衡およびネットワーク資源の競合を軽減するために設計されている．本発表では実際のアプリケーションに基いたベンチマークプログラムの評価結果から得られた重要なスケジューリングの課題と対策について説明する．

Implementation and Evaluation of MPI-3.0 Neighborhood Collectives in Tofu Interconnect

Masayuki Hatanaka[1,a)]　Atsushi Hori[1]　Yutaka Ishikawa[1,2]

Presented: July 30, 2014

In this presentation, we describe the implementation and evaluation results of MPI Neighborhood collective communication primitives in Tofu interconnect. These neighborhood primitives are introduced in MPI 3.0 specification to optimize data exchange among neighboring processes such as ghost region updates. Based on a priori knowledge regarding communication pattern given by a neighborhood primitive, we developed the MPI_Neighbor_alltoallw implementation combined with RDMA transfer scheduler for Tofu interconnect, which has multiple RDMA engines and network links. The RDMA transfer scheduler is designed to mitigate the load imbalance of RDMA engines and network resource contentions. We shows the major scheduling issues and its solutions obtained from the benchmark results based on real applications.

[1] 理化学研究所計算科学研究機構
RIKEN Advanced Institute for Computational Science, Kobe, Hyogo 650–0047, Japan
[2] 東京大学情報科学科
Department of Computer Science, University of Tokyo, Bunkyo, Tokyo 113–8656, Japan
[a)] mhatanaka@riken.jp