サイエンティフィック・システム研究会
科学技術計算分科会 2012 年度会合
2012 年 10 月 24 日

# スーパーコンピュータ「京」での MPI の実装と評価

三浦 健一
富士通株式会社 次世代テクニカルコンピューティング開発本部 ソフトウェア開発統括部

[アブストラクト]
スーパーコンピュータ「京」[1]の共用が 9/28 から開始された。「京」で使用されている MPI ライブラリでは 8 万ノード規模の並列計算に耐えられ、かつ高速通信ができるよう、様々な工夫が施されている。本発表においては、「京」における MPI ライブラリの特徴である、省メモリ通信、通信最適化、最適ランク配置、低レベル通信ライブラリの実装を説明し評価する。また併せて、アプリケーションへの適用事例に関して報告する。

[キーワード]
Tofu インターコネクト，MPI 通信，大規模並列，省メモリ，最適ランク配置

## 1 はじめに

スーパーコンピュータ「京」の共用が 9/28 から開始された。「京」は 8 万ノード超のノード群が Tofu(Torus fusion)と呼ばれる 6 次元メッシュ/トーラス型インターコネクトで結合された構成となっている。Tofu では 6 次元メッシュ/トーラスを生かした故障ノード回避や迂回通信、ジョブ単位に論理トーラス割当を行うことが可能となり高い運用性を実現している。また Tofu では 4 個の通信インターフェイスを用いた同時通信や高機能バリアを用いたハードウェア同期を行うことでより高速に通信を行うことが可能である。我々は「京」上で Tofu を用いて通信を行う MPI ライブラリの開発を行い、大規模環境での評価を行った。

## 2 MPI ライブラリの実装概要

「京」の MPI ライブラリは Open MPI をベースに拡張を行ったが、8 万ノード規模の Tofu インターコネクトへの適用にあたり、以下の点に注意して設計を行った。

- 8 万ノード規模の並列化に耐えうる実装
- Tofu の特性を生かした高速通信

[1] 「京」は理化学研究所の登録商標です。

- ユーザの使いやすさ

MPI 通信を行うためには通信相手毎に送受信バッファが必要となるが、8 万ノード分の送受信バッファを確保していたのでは MPI ライブラリ自体のメモリ使用量が膨大になってしまう。「京」の MPI ライブラリでは省メモリ通信モードと高速通信モードを使い分けることで、省メモリを維持しつつ高速に通信を行うことを狙っている。

Tofu の特性を生かした高速通信では、特に大規模並列計算で問題となる集団通信に関して、Tofu 向けアルゴリズムを新規開発することで、ハードウェア性能を最大限引出す事を狙っている。

またユーザの使いやすさとして、通信を行うランクを近くに配置することで性能チューニングを簡単に行うことができる RMATT(Rank Map Automatic Tuning Tool)、突き放し通信や 4 個の通信インターフェイスを使った並列通信を容易に記述することができる「拡張 RDMA インターフェイス」の設計・実装を行った。

## 3　MPI ライブラリの評価

省メモリ通信では 8 万ノード規模の並列処理においてデフォルトメモリ使用量を 400MB 程度まで抑えることができた。通信最適化においては集団通信の Tofu 専用アルゴリズムの採用により既存アルゴリズムに比べ 5～10 倍程度の性能向上が確認できた。RMATT を用いた最適ランク配置では、NAS parallel Benchmark に適用することで 7%の性能改善が確認できた。「拡張 RDMA インターフェイス」では 4 個の通信インターフェイスを用いた通信を行うことで性能改善の効果が確認できた。

アプリケーションへの適用事例としては昨年度ゴードン・ベル賞を獲得した RSDFT や 4 部門全てで 1 位を獲得した HPCC の Global FFT や Global RandomAccess の通信ライブラリとして使用され高効率の実行に貢献した。

## 4　まとめ

「京」でサポートしている MPI ライブラリの実装と評価に関して述べた。「京」の運用は始まったばかりであり、今後多くの技術者が「京」に触れて「京」の素晴らしさを体感していただけることを期待している。本 MPI ライブラリを用いることで大規模アプリケーションの開発・運用に少しでも貢献できれば幸いである。

[参考文献]

(1) 住元真司 et al., 「京」のための MPI 通信機構の設計(2012), SACSIS2012

(2) 今出広明 et al., 大規模計算環境のためのランク配置最適化手法 RMATT(2011), SACSIS2011

(3) T. Adachi et al., The design of ultra scalable MPI collective communication on the K computer(2012), ISC'12