2P03

InfiniBand ループ接続を用いた並列計算システム

○北山 清章[1]、猪俣 健輔[1]、二川 潤[2]、善甫 康成[3]
[1]株式会社シミュラティオ(〒222-0033 神奈川県横浜市港北区新横浜 1-14-20)
[2]インテグシステム(〒359-1111 埼玉県所沢市緑町 2-7-18)
[3]法政大学情報学科(〒184-8584 東京都小金井市梶野町 3-7-2)

【緒言】

InfiniBand は、転送速度や信頼性の高さといった点でイーサネットより優れており、HPC の構築においてよく採用されるインターコネクトである。我々は今回、この InfiniBand を用いた IPoIB リングネットワークと、Lustre ファイルシステムによって小規模クラスタを構築し、実用試験を行った。Lustre ファイルシステムは膨大な計算データを非常に効率良く処理する事ができ、しばしば大規模クラスタに導入されるが、今回の様な小規模クラスタに用いても十分に意義がある。本クラスタでは、リングネットワークにスイッチを用いず、ノード間を直接 InfiniBand で接続し、リング形状の接続となっている。この方法は安価にシンプルなクラスタを構築できる事が最大の利点である。我々はこのシステムでいくつかの並列プログラムを実行した。その結果小規模システムとして遜色ない計算が行える事が判明した。なおスイッチを用いない分、その機能を代替するソフトウェアが必要である。つまりスイッチが持つ InfiniBand マネージャー(サブネットマネージャー)を各ノードに持たせる必要がある事に注意しなければならない。また、リングネットワークでは 1 ノードでも停止すると Lustre ファイルシステムが正常に動作しなくなってしまう難点がある。

【方法】

今回構築したクラスタは、ノード 4 台(コア数はそれぞれ 4,4,4,8)を直接 InfiniBand DDR(20Gbps)で接続したリングネットワーク構造であり、ルーティングは図 1 の様に設定した。サブネットマネージャーは OpenSM[1]を用いた。

実用試験には負荷の軽い計算プログラムとして SIESTA を、負荷の重い計算プログラムとして実時間・実空間 TDDFT プログラム[2]を選択した。それぞれ用いるノード・コア数毎に実用試験を行い、計算時間を time コマンドにより複数回測定して、平均値を算出した。

図 1. ルーティング
ただし、隣接するノードならば、逆方向の送信も許可している。

【結果】

実用試験の一部結果を図 2 に示す。図 2 は SIESTA プログラムによる GaAs(64 原子)の計算結果である。交換相関汎関数として PBE を用いる。1 ノード・4 コアでの計算時間は 129.01 秒であり、この計算速度を 1 とした時の計算速度の向上率をプロットしたものである。非常にリニアな関係が得られており、問題なく効率的に並列計算ができている事がわかる。

詳細についてはポスターにて報告するが、より大きなクラスタによる応用例についても報告する予定である。

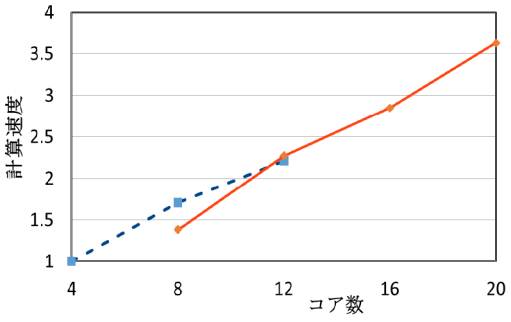

図 2. 実用試験結果(SIESTA)
実線(橙)は 8 コアのノードを含む計算結果を、点線(青)は 8 コアのノードを含まない計算結果を示す。

[1] Mellanox OFED ソフトウェアスタックに含まれる。
http://jp.mellanox.com/page/products_dyn?product_family=26

[2] 2P09 田中志歩、遠越光輝、善甫康成(法大情報)