

# 情報爆発時代へ向けた不均一アーキテクチャにおける スーパーコンピューティング

## Supercomputing on Heterogeneous Architecture toward the Information Explosion Era

遠藤 敏夫<sup>†, \*</sup>      松岡 聡<sup>†, ‡, \*</sup>  
Toshio Endo      Satoshi Matsuoka

東京工業大学<sup>†</sup>      国立情報学研究所<sup>‡</sup>      科学技術振興機構\*  
Tokyo Institute of      National Institute of      JST  
Technology      Informatics      CREST

### 1. はじめに

大規模計算機システムを構築する上で、近年のプロセッサの消費電力の上昇が大きな問題となっている。情報爆発時代の大規模なアプリケーションからの要求に答えるために、用途を特化した省電力型プロセッサを用いた並列アーキテクチャが注目されている。一方、多様なアプリケーションへの対応を確保する上では汎用 CPU も重要であり続けるであろう。東京工業大学のスーパーコンピュータ TSUBAME は、10000 以上の汎用 CPU である Opteron と、600 以上の省電力型プロセッサである ClearSpeed SIMD アクセラレータを持つ、不均一型アーキテクチャとしては現在世界最大の規模を持つシステムである。

本稿では TSUBAME の(ほぼ)全ての CPU とアクセラレータを用いた並列計算について報告する。Cell プロセッサや GPGPU などのアクセラレータへの近年の注目は顕著であり、CPU とアクセラレータの併用についても報告されているが[1]、スーパーコンピュータ規模での報告は依然ない。

対象とする並列計算は、Top500 スーパーコンピュータランキング ([www.top500.org](http://www.top500.org)) でも用いられている High performance Linpack (HPL) [2] である。各計算プロセスは密に結合し、均一性能であることが仮定されているため、不均一環境へ対応する技法が必要となる。更には、CPU とアクセラレータの性能特性を考慮した負荷分散およびチューニングが必要となる。

結果として TSUBAME 全体で Linpack 性能 56.43TFlops が得られた。これは CPU のみの場合の 38.18TFlops に比べ 48%改善されており、さらに 2007 年 11 月の Top500 において、不均一なプロセッサを混在させた性能としては世界一となっている。

### 2. TSUBAME と ClearSpeed アクセラレータ

東工大 TSUBAME は 2006 年から稼動しているシ

ステムであり、655 ノードが InfiniBand により結合されている。各ノードは 8 個の 2.4GHz Dual core Opteron (計 16core) と 32GB の共有メモリ (一部例外あり) を持っている。

さらに、ClearSpeed SIMD アクセラレータ [3] が、各ノードの PCI-X バス上に装着されている。アクセラレータは、SIMD プロセッサとオンボードメモリを持っている。また、専用の数値演算ライブラリが提供されており、本研究ではそれを用いる。TSUBAME の理論性能は、Opteron 50TFlops, ClearSpeed 52TFlops で、合計 102TFlops であり、現在日本最速である。

### 3. High performance Linpack

HPL は密行列連立一次方程式を直接法により解く並列プログラムであり、詳細は文献 [2] を参照されたい。本稿で重要となる性質は、(1) 計算時間のほとんどは行列積演算で占められること、(2) プロセス間通信が発生するため、各プロセスの性能が均一である必要があることである。

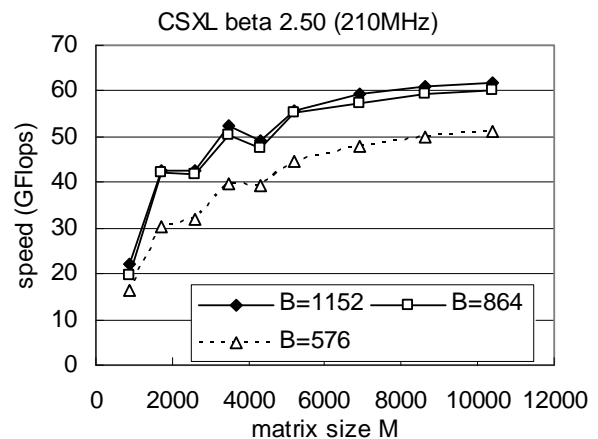


図 1. ClearSpeed による行列積性能。  
M\*B 行列と B\*M 行列の積を示す。

図 1 に ClearSpeed 一台による行列積性能を示す。ピークでは 60GFlops 程度を達成する一方、

その性能は行列サイズに大きく影響を受ける。このピーク性能も、汎用 CPU の場合と大きく異なっている。

#### 4. 不均一プロセッサの効率的な利用

均一環境用に設計された並列プログラムを、不均一環境上で効率的に動作させる手法を示す。基本的な方針は、計算に参加するプロセス群と物理プロセッサの間のマッピングを調整するというものである。図 2 にその様子を示す。なお、提案手法では、一部のノードのみアクセラレータを持つ場合にも対応し、図では右の 2 ノードのみが持っているとする。

計算に参加するプロセスのうち、CPU プロセスは CPU のみを用い、SIMD プロセスは、行列積演算をアクセラレータへ依頼する。

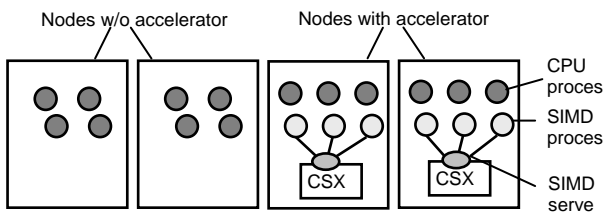


図 2. 各ノードにおけるプロセス構成

アクセラレータを含む環境の並列性能は、計算粒度などのパラメータに大きく依存する。以下、チューニング項目の一部を示す。チューニングによるトレードオフや、詳細な点については、文献[4]を参照されたい。

- ブロックサイズ: 図 1 の B に相当する。アクセラレータの性能を向上させるためには、通常のケースより大きくする必要がある。CPU のみの実験では 240 程度が最適だったが、アクセラレータを含む場合には 864 とした。
- プロセス粒度: 各 CPU プロセスが利用する CPU core 数も調節可能である。予備実験により 4core とした。そしてそれをもとに各ノード(アクセラレータのあるノード、ないノード)のプロセス数を決定した。

#### 5. 性能評価

TSUBAME 全体を用いて本手法の評価を行った。なお多数ユーザにより共有されているシステムなので、システムメンテナンスの時期に実験を行った。HPL のソースコードの一部を変更し、前節の対応を行った。また、Voltaire MPI, GotoBLAS, ClearSpeed による BLAS の各ライブラリを利用した。

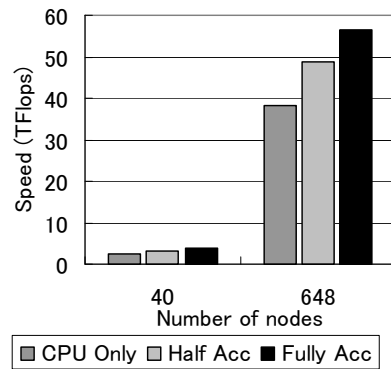


図 3. 不均一な TSUBAME での Linpack 性能

図 3 に、CPU のみの場合、アクセラレータを含む場合(Fully Acc)、さらに半分のノードのみがアクセラレータを持つ場合(Half Acc)の性能を示す。648 ノードの場合、Fully Acc では 56.43TFlops を達成し、38.18TFlops である CPU only に比較し、48%の向上が見られる。また、Half Acc も、アクセラレータ数にほぼ比例した良好な性能が得られている。

#### 6. おわりに

多数の汎用 CPU と SIMD アクセラレータを用いた環境において、均一環境用に設計された並列プログラムである HPL を、小さい改造で効率的に動作させる手法を提案した。アクセラレータの特性を考慮したチューニングは必須であり、それにより不均一な環境における Linpack 性能としては世界最高の速度性能を得た。

**謝辞** 本研究の一部は科学研究費補助金特定領域研究(18049028)および JST-CREST の援助による。

#### 参考文献

- [1] 大島聡史, 吉瀬謙二, 片桐孝洋, and 弓場敏嗣. CPU と GPU を用いた並列 GEMM 演算の提案と実装. In 先進的計算基盤システムシンポジウム SACSIS2006 論文集, pages 41.50, 2006.
- [2] A. Petitet, R. Whaley, J. Dongarra, and A. Cleary. HPL - a portable implementation of the high-performance Linpack benchmark for distributed-memory computers. [www.netlib.org/benchmark/hpl/](http://www.netlib.org/benchmark/hpl/).
- [3] ClearSpeed Inc. [www.clearspeed.com](http://www.clearspeed.com).
- [4] Toshio Endo and Satoshi Matsuoka. Massive Supercomputing Coping with Heterogeneity of Modern Accelerators. In Proc of IEEE IPDPS, 2008 (to be appeared).