岸本芳典，市川周一："不均一クラスタ上での実行時間予測モデルとその評価，"
SACSIS2003 論文集，pp. 167--168，情報処理学会 (2003).

不均一クラスタ上での実行時間予測モデルとその評価

岸 本 芳 典† 市 川 周 一†

均一環境のための応用を不均一クラスタで実行すると，負荷不均衡により性能上の問題を生ずる．また，一部の PE には仕事を割り当てないほうが全体の実行時間が短縮できる場合がある．本研究では，高速な要素プロセッサ (PE) 上で複数のプロセスを起動することにより，全体の実行時間を短縮する方法を検討する．さらに，各 PE 上の実行時間を実測値からモデル化し，得られた予測モデルを用いて最適な PE 構成およびマルチプロセス数を予測することを試みる．HPL について N=400〜6400 の測定値からモデルを実装し，N=3200〜9600 に対して最適な実行方法をモデルから予測した．得られた構成の実行時間は，真に最適な実行時間から 0%〜9%の誤差であった．

The Execution Time Estimation Model for Heterogeneous Clusters and Its Evaluation

Yoshinori Kishimoto† and Shuichi Ichikawa†

A heterogeneous cluster can incur the performance degradation caused by the load unbalance in executing the application for homogeneous cluster. Also, the total execution time can be improved by neglecting some of the PEs because communication time is reduced. This study examines to invoke multiple processes on fast processing elements (PEs) to avoid load unbalance. In this study, the execution time of each PE is firstly modeled from measurement results. Then, the derived model is used to estimate the optimal PE configuration and process configuration. We implemented the models for HPL (N = 400–6400), and estimated the optimal configuration for N = 3200–9600. The error of the estimated execution time was 0%–9% of the actual optimal execution time.

1. 不均一クラスタ上での実行時間最適化

要素プロセッサ (PE) の構成・演算性能などが同一でないクラスタを，不均一クラスタという．一般には均一なクラスタが多く利用されるが，不均一クラスタには (1) 手持ちの計算機を任意に組合せて構築できる，(2) 最新の高速プロセッサを随時追加できる，など多くの利点がある．しかし，均一環境を想定した (各 PE に同量の負荷を割り当てる) 応用を不均一クラスタで実行すると，PE 間の負荷不均衡のため性能が発揮できない．

負荷不均衡は負荷の割当量を可変にすれば解決するが，そのためには個別の応用プログラムの修正が必要になる．一方，高速なプロセッサ上にプロセスを複数立ち上げて使う方法 (マルチプロセス法) では，多少の性能オーバヘッドは発生するが，応用プログラムの変更は不要である．マルチプロセス法は，均一な PE を前提とした既存の応用を変更なしに利用できる点で魅力的である．そこで本研究では，並列処理応用をマルチプロセス法で最適化する方法を検討する．応用の例としては，HPL (High Performance Linpack)[2] を扱う．

HPL の不均一クラスタ上での実行には複数の先行研究がある．例えば笹生ら[3] は HPL を修正して負荷の均衡化を図っているが，マルチプロセス法はオーバヘッドが大きいとして棄却している．しかし我々は，HPL をマルチプロセス法で実行しても性能オーバヘッドは 3 割程度に留まることを示した[1]．

負荷を完全に均衡化しても，必要以上に多くの PE を使用すると，通信時間が増加して全体の実行時間が増大する場合がある (過剰な負荷分散)．実行時間を最小化するには，通信時間を考慮して適切な PE(群) を選択する必要がある．本研究では，(1) 使用する PE 構成と各 PE 上のマルチプロセス数から予測実行時間 T を出力する予測モデルを作成し，(2) 作成した予測モデルを利用して，実行時間を最小化する PE 構成とマルチプロセス数を決定する．

† 豊橋技術科学大学 知識情報工学系
Department of Knowledge-based Information Engineering, Toyohashi University of Technology

表 1 HPL 実行環境

OS	RedHat Linux7.0J (kernel 2.4.2)
コンパイラ	gcc 2.96, -fomit-frame-pointer -O3 -funroll-loops -W -Wall
ライブラリ	MPICH–1.2.5, ATLAS 3.2.1

表 2 HPL 測定時のクラスタ構成パラメータ

モデル構築時	N=400〜6400, Athlon(P_1:1,M_1:1〜6), Pentium2(P_2:1〜8,M_2:1〜6)
評価用実測時	N=3200〜9600, Athlon(P_1:0〜1,M_1:1〜6), Pentium2(P_2:0〜8,M_2:1)

2. HPL の実行時間予測モデル

HPL の PE_i 上での実行時間は，計算時間 Ta_i と通信時間 Tc_i からなるが，現実には計算と通信の重畳などもあって個別の測定は簡単でない．本研究では，HPL のコンパイル時に HPL_DETAILED_TIMING を define し，更にソースコードに手を入れて計算時間や通信時間の概略を測定した．得られた測定値から，PE_i の予測実行時間 T_i の式を求める．まず HPL のアルゴリズムから Ta_i は 3 次式，Tc_i は 2 次式と仮定して，実測値を最小二乗法でフィッティングし，近似式を求める．さらに全体の予測実行時間 T は，$T = \max_i T_i = \max_i(Ta_i + Tc_i)$ であると仮定する．

本研究では，Athlon 1.33GHz (1 台) と Pentium2 400MHz (8 台) を 100base-TX で接続した不均一クラスタを測定に用いる．Pentium2 PE (8 台) は全く同一の構成をとるため予測モデルは共通とし，実行時のマルチプロセス数も同一とする．HPL は表 1 の環境で実行し，プロセス格子は横一列に固定した．

予測モデル作成時および評価用実測時のクラスタ構成パラメータを表 2 に示す．P_1 は使用する Athlon の台数 (0 か 1)，M_1 は Athlon 上のプロセス数を表す．Athlon と Pentium2 の単体演算性能比は約 4:1 なので，M_1 の値は 4 前後と考え，$1 \le M_1 \le 6$ とした．P_2 は使用する Pentium2 の数 (0〜8)，M_2 は Pentium2 上のプロセス数である ($1 \le M_2 \le 6$)．HPL は，不均一クラスタ上の $P_1M_1 + P_2M_2$ 個のプロセスに対して均等に負荷分散されて実行される．

3. 実行時間予測モデルの評価

作成した予測モデルを評価するために，表 2 下段の範囲内で可能な全てのクラスタ構成について，実行時間を予測した．さらに，予測結果と比較するため，可能な全ての構成で実際に HPL の実行時間を測定した．

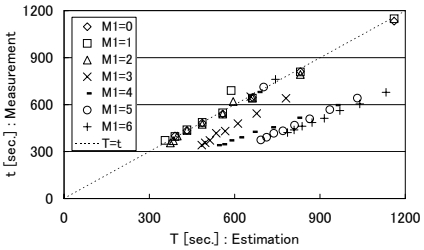

図 1 $T - t$ 相関図 ($N = 9600$)

$N = 3200, 6400, 9600$ について，予測による最良クラスタ構成と実測による最良構成を，表 3 にまとめた．予測モデルで最良とされた構成の予測実行時間を τ，その構成の実測実行時間を $\hat{T}$ で示した．$\hat{t}$ は，実測で最良だった構成の実行時間である．$N = 3200$ では問題サイズが小さいため，Athlon 1 台での実行が最良と予測されており，その結果は実測でも確認されている (過剰な負荷分散が回避された)．$N = 6400$ では最適構成を予測できなかったが，最良の実行時間に対して誤差 4%という準最適な構成を予測している．$N = 9600$ では誤差が増大し，最適構成に対して実行時間 9%増の構成を予測している．

サイズの大きい計算で誤差が増大する原因を調べるため，$N = 9600$ について，可能な全構成で予測実行時間 T と実測時間 t の相関を調べた (図 1)．$0 \le M_1 \le 2$ では非常に良く一致しているが，$M_1 \ge 3$ で T を過大に見積もっていることが解る．$N = 9600$ では計算量が多いため，Athlon 上のプロセス数 M_1 を 2 以上にすると実行時間が短縮される．しかし $M_1 \ge 3$ でモデルの誤差が大きいため，最適構成の予測では M_1 の小さい構成が選ばれてしまう．結果として Athlon が有効に利用されない．図 1 では，各 M_1 毎に $T - t$ 間に綺麗な相関が認められるので，比較的簡単な改良により正確な予測が可能になると考えられる．

表 3 N に対する予測値・実測値の最良値と誤差

サイズ	予測による最良構成			実測による最良構成		誤差	
N	P_1, M_1, P_2, M_2	τ	$\hat{T}$	P_1, M_1, P_2, M_2	$\hat{t}$	$(\tau - \hat{t})/\hat{t}$	$(\hat{T} - \hat{t})/\hat{t}$
3200	1,1,0,0	20.0	20.4	1,1,0,0	20.4	-0.019	0.000
6400	1,1,8,1	129.7	129.8	1,2,8,1	125.2	0.036	0.037
9600	1,1,8,1	355.4	371.7	1,4,8,1	340.9	0.043	0.090

参 考 文 献

1) 岸本芳典, 市川周一: 不均一クラスタ上での並列 Linpack の性能に関する検討, 並列処理シンポジウム JSPP2002, pp. 177–178 (2002).
2) Petitet, A., et al.: HPL – A Portable Implementation of the High-Performance Linpack Benchmark for Distributed-Memory Computers. http://www.netlib.org/benchmark/hpl/.
3) 笹生健, 松岡聡, 建部修見: ヘテロなクラスタ環境における並列 LINPACK の最適化, 情処研報 2001–HPC–86, pp. 49–54 (2001).