スーパーコンピュータにおける電力性能最適化フレームワークの評価

九州大学大学院システム情報科学研究院　井上弘士

【電力制約下での性能向上を目的とした電力バジェット配分フレームワークの評価】

概要：カタログスペックが同じ計算ノードであってもその消費電力特性が異なっており，電力制約を一律に適用した場合には演算性能にばらつきが生じる．これは，半導体の微細化に伴い製造ばらつきが顕著になったことに起因する．そこで我々は，トータル電力バジェット一定の下で各計算ノードの消費電力特性に応じて電力制約を適用することで，プロセス間の実行時間ばらつきを軽減し電力制約時の並列性能を改善する電力バジェット配分フレームワークを提案してきた．本手法では，演算性能が動作周波数に比例することを考慮して，CPU への電力キャッピング，または，CPU 動作周波数制御を行う．これにより，CPU 間の処理性能を均等に保ちながら消費電力を制御することで，電力制約下での並列アプリの実効性能を改善する．これまでの研究では，最大 32 計算ノードでの性能／消費電力評価しか行えていなかった．そこで，平成 26 年度先端的計算科学研究プロジェクトにおいて，HA8000 の 960 ノードを用いた大規模性能評価を実施した．その結果，提案するフレームワークを適用することで大幅な性能向上を達成可能であることが分かった．

実験内容：CPU 消費電力制約や CPU 動作周波数制約を適用して各種並列 HPC アプリケーションプログラム（アプリ）実行時の CPU 消費電力を測定した．CPU 消費電力測定ならびに CPU 電力制約値の設定は，Sandy Bridge 以降のインテルプロセッサに搭載されている RAPL (Running Average Power Limit)[1]を用いた．我々は RAPL を利用するためのライブラリを作成して，各アプリにライブラリ利用のための関数を挿入してコンパイルし，電力測定や制約を行った．また，CPU 動作周波数の制御には cpufreqlibs ライブラリを用いた．今回実験に用いたのは，特性が異なると思われる 7 種類のアプリ（*DGEMM，*STREAM-Triad，NPB-BT，NPB-SP，MHD，mVMC-mini，Modylas-mini）を用いた．*DGEMM と*STREAM-Triad は HPC challenge[2]ベンチマークに含まれる embarrassingly parallel プログラムであり，NPB-BT, NPB-SP は NAS parallel benchmark suite[3]に含まれるブロック行列，帯行列に対する連立方程式ソルバである．MHD[4]は電磁流体シミューレションプログラムでステンシル計算型の並列アプリである．vVMC-mini ならびに Modylas-mini はそれぞれ変分モンテカルロ計算と分子動力学シミュレーションを行うプログラムのカーネルを取り出して作成されたベンチマークコードであり，理化学研究所からミニアプリとして公開されている[5]．本実験では，1,920 プロセス

（Modylas のみ 1,024 プロセス）×12 スレッド（1 プロセス/1 CPU(12 コア)）で実行し電力測定等を行った．バイナリ作成には Intel C/C++, Fortran コンパイラ(version 15.0.1)を用い, MPI ライブラリ, 数値演算ライブラリとして, それぞれ Intel MPI(version 5.0), Intel Math Kernel Library(MKL, version 11.2.1)を利用した．

結果：電力制約を適用しない（通常の）状態でアプリの実行時間，ならびに，CPU 消費電力を測定したところ，実行時間にはプロセス（CPU）間での差がほとんど見られないものの，消費電力には最大 30%程度のばらつきがあることを確認した．このように CPU 間での電力消費特性が異なるシステムを用いた電力制約下での並列アプリ実行時に全プロセス一律の CPU 消費電力制約を適用すると, 消費電力は各 CPU でほぼ等しくなるものの，実行時間に CPU 間でのばらつきが生じた．このような実行時間のばらつきが見られるのは，CPU 電力制約条件を満たす（CPU 消費電力を制約値に保つ）ために CPU 動作周波数が調整され，消費電力特性が異なる各 CPU の平均動作周波数（≒処理性能）に差を生じたからである．HPC アプリには，一般に並列性能向上のために各 CPU での処理性能が均等であることを前提とした負荷均等化が適用されているが，そのようなアプリに対して一律 CPU 電力制約を適用するとプロセス間での処理時間に不均衡が生じるため，並列性能低下が予想される．一方，CPU 動作周波数を制御することで間接的に CPU 消費電力を抑制すると，消費電力は CPU 間でばらつくが，各 CPU の処理性能をほぼ一定に保つことができた．評価の結果，計算ノード間ばらつきを考慮しないナイーブな電力配分法と比較して，最大で 5.4 倍，平均で 1.8 倍の性能向上を実現した．

全ノード実行時に生じた問題：Intel MPI と hakozaki のジョブスケージューラとの組み合わせでは，ジョブキャンセルに長時間（場合によっては 1 時間以上）かかった. 全ノードジョブ実行中に異常が発生した場合にはジョブキャンセルで対処したが, ジョブキャンセル待ちによって実験が滞ることが何度かあった. 本件はベンダによる原因調査中であるが（今のところ原因は不明との連絡あり），現状では準備して頂いたジョブキャンセル専用スクリプトで対応している．

【電力制約下での性能向上を目的とした適応的電力ジューリング・フレームワークの評価】

概要 : 将来的な高性能計算機システムでは，消費電力がシステム設計や実効性能を制約する最大の要因になると考えられており，供給電力，あるいは熱設計電力制約の中でハードウェア資源を投入し, 運用時のピーク消費電力が制約を超えないことを保証

する従来の設計思想では, 将来的にシステムを有効利用することは難しくなると考えられる．そこで，限られた電力資源を各アプリケーションに適応的に配分し，システムの電力効率を最適化することが必須となる．この実現に向け，HPC 計算機システム全体の電力供給と使用状況, およびアプリケーションのカテゴリに応じた適応的電力スケジューリングを行うシステムソフトウェアを開発中である. 本ソフトウェアの開発には，まず電力監視・制御に伴うオーバーヘッドやスケーラビリティを評価することが重要であり，先端的計算科学研究プロジェクトにおいて PRIMERGY CX400 の全系を利用し，それら基本データの取得と低遅延な通信方式の開発を行った．その結果，1,000 ノード規模のシステムで，実用的な制御遅延・オーバーヘッドで電力スケジューリングが行えることを確認した.

参考文献

[1] Intel Corporation: Intel 64 and IA{32 Architectures Software Developer_s Manual Volume 3(3A, 3B & 3C):System Programming Guide (2012).

[2] Luszczek, P., Bailey, D., Dongarra, J. et al.: HPC Challenge, http://icl.cs.utk.edu/hpcc/index.html.

[3] NASA Advanced Supercomputing Division, NAS Parallel Benchmark Suite v3.3. 2006. http: /www.nas.nasa.gov/Resources/Software/npb.html.

[4] T. Ogino, R. J. Walker, and M. Ashour-Abdalla. A Global Magnetohydrodynamic Simulation of the Magnetopause when the Interplanetary Magnetic Field is Northward. IEEE Transaction on Plasma Science, 20:817{828, December 1992.

[5] N. Maruyama, S. Suzuki, K. Mikami, Y. Komuro, S. Takizawa, and M. Matsuda. Fiber Miniapp Suite. fiber-miniapp.github.io.