XC40 計算ノードにおける消費電力のばらつき評価とその活用研究

深沢 圭一郎 [1], 疋田 淳一 [2], 當山 達也 [2], 島袋 友里 [2]

1) 京都大学 学術情報メディアセンター
2) 京都大学 情報部

fukazawa@media.kyoto-u.ac.jp

Evaluation of Power Consumption Heterogeneity on XC40 Compute Node and Study of its Application

Keiichiro Fukazawa [1], Junichi Hikita [2], Tatsuya Tohyama [2], Yuri Shimabukuro [2]

1) Academic Center for Computing and Media Studies, Kyoto Univ.
2) Information Department, Kyoto Univ.

概要

大学などのスーパーコンピュータは数百から数千計算ノードから構成されているが、この計算ノードでは CPU 等の製造ばらつきに起因した消費電力のばらつきが存在する。本研究では、ベンチマークアプリケーションを利用して京都大学のスーパーコンピュータである XC40 における消費電力ばらつきを評価し、そのばらつきを活用した消費電力低減に繋がる研究について議論する。

1 はじめに

近年スーパーコンピュータ（スパコン）システムの開発において、消費電力の増大は大きな課題となっている。これは、スパコンシステムに限らず増加しているデータセンターにおいても同様の課題となっている。特に、データセンターの電力消費量は、2020 年では世界の総電力使用量の 2%に達すると推定されている[1]。

スパコンシステムでは、リプレイスにおいて高い性能の CPU を搭載した計算ノードを多数導入する場合、その消費電力量が大きすぎ、導入できる量に制限がかかることも起きている。例えば、スパコンに導入されるような高コアの 1 世代前の Intel 製 Xeon と最新の Xeon を比べると TDP が約 100W 上昇しており、NVIDIA 製 GPU であれば、A100 と最新世代である H100 では TDP が 300W 上昇している。世代更新に伴う演算性能も上昇しているので、Flops/W としては向上しているが、多数ノードの導入は難しくなっている。このような消費電力を抑えるために、システムのスペックを制限したり、ピーク電力に制限をかけたりするような施策を行う必要があるが、同時に計算性能も制限されてしまう。

一方、近年はスパコン専用 CPU ではなく、x86 系 CPU である Xeon や EPYC が多く利用され、GPU では NVIDIA や AMD が多くのシステムで利用されている。これらは富岳や SX-Aurora TSUBASA などの CPU と比べ、性能ばらつきが多く存在する。一般的には、半導体の製造では、周波数の高い製品は少なく、周波数の低い製品が多く作られるため、スパコン専用品である場合には、この中でも厳しい基準を満たす製品が利用される。汎用品の場合は、性能に余裕を持たせたある一定の性能基準を満たす必要がある。この結果、例えば、PC に使われる CPU で行われるオーバークロックにおいて、オーバークロック耐性の高い CPU と低い CPU が表れる。これは周波数を一定にした場合、消費電力が多い CPU と少ない CPU があるということに繋がる。

HPC 向けである Xeon 系では性能ばらつきは比較的少ないと考えられるが、この消費電力のばらつきを理解しておくことで、縮退運転をする際にどのノードを停止させるかなど、消費電力と演算性能を考慮したスパコンシステム運用が可能となると考えられる。そのため。本研究では、京都大学学術情報メディアセンター（京大メディアセンター）のスパコンシステムである XC40 においてベンチマークアプリケーションを利用して計算ノードの消費電力ばらつきを評価し、そのばらつきを利用したスパコンシステム全体の消費電力削減に繋がる手法を議論する。

2　計算ノードの評価手法

京大メディアセンターではスパコンの利用サービスとして全ノード利用は行っておらず、また利用者の利用時間を最大限確保する運用思想から、スパコンの停止を伴うメンテナンスを少なくし、また停止を伴うメンテにおいても可能な限り停止時間を短くする努力をしている。このため、京大メディアセンターのスパコンである XC40 の全ノード（1,800 ノード）で同時に同じベンチマークを走らせ、ばらつきを評価することができない。そこで、シングルノードジョブを大量に投入し、割り当てられたノード情報を確認し、多数の異なるノードを評価する手法をとった。

これらのノードにおいて、CrayPat という性能プロファイラーツールを利用し、消費電力と実行時間を測定した。以降の評価結果では、測定結果の平均値を利用するため、3 回評価できなかったノードは除外している。

2.1 計算機システム

京大メディアセンターのスパコンである XC40 は、2016 年から 2022 年に運用されていたシステムであり、2022 年 7 月 28 日に運用を停止した。XC40 は Xeon Phi KNL を搭載したメニーコアシステムであり、構成の詳細を表 1 にまとめている。このスパコンを利用し、前述のようにジョブを投入し、ベンチマークアプリケーションが動作したノードを評価した。また、評価を行う場合には、ハイパースレッディングはオフにしている。

2.2 評価ベンチマークアプリケーション

ノードを評価するベンチマークアプリケーションとして、STREAM[2]と HPCG[3]を利用した。STREAM はメモリ操作を行い、メモリバンド幅を評価するベンチマークアプリケーションであり、スパコンの性能評価ではよく利用されている。本評価では STREAM の MPI 版を利用し、1 ノード内

表 1　XC40 の諸元

機種名	CRAY XC40	
計算ノード	CPU	Intel Xeon Phi KNL7250 × 1 /node
	コア数	68 cores /CPU
	周波数	1.4 GHz
	理論性能	3.05 TFlops /node (**倍精度**)
	メモリ	MCDRAM: 16GB/node DDR4: 96 GB /node
	Bandwidth	MCDRAM: **約** 490GB/s DDR4: 115.2 GB/s
	B/F	MCDRAM: 0.16 DDR4: 0.038
総ノード数	1,800 nodes	
総理論性能	6.91 PFlops	
ノード間接続	CRAY Aries (15.75 GB/sec)	

68 プロセス（Xeon Phi KNL のコア数は 68 コア）の Flat MPI として、実行した。

HPCG は Linpack に比べて実アプリケーションに近いベンチマークとして開発されており、疎行列の線形ソルバーである。つまりベクトル演算性能を評価している。HPCG では 1 ノード 64 プロセスとして実行した。

本研究では、STREAM を 6,000 ジョブ、HPCG を 4,600 ジョブ投入した結果、STREAM では 623 ノード（3 回以上評価できたノード数は 414 ノード）、HPCG では 430 ノード（同 327 ノード）の異なるノードを評価することができた。

3　評価結果

3.1 STREAM ベンチマーク

図 1 に XC40 で STREAM を実行した際の各ノードの消費電力と実行時間を示す。グラフの横軸はノードを表すが、ノード ID 順ではなく、計測が

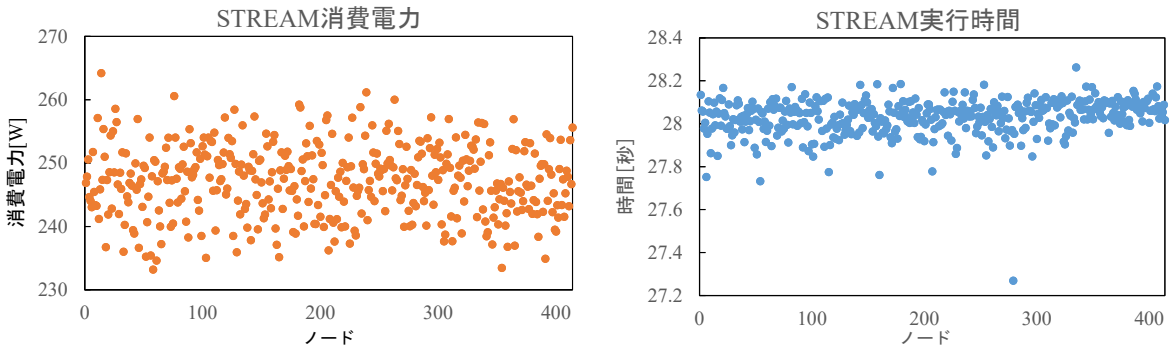

図 1　XC40 ノードにおける STREAM ベンチマーク評価結果

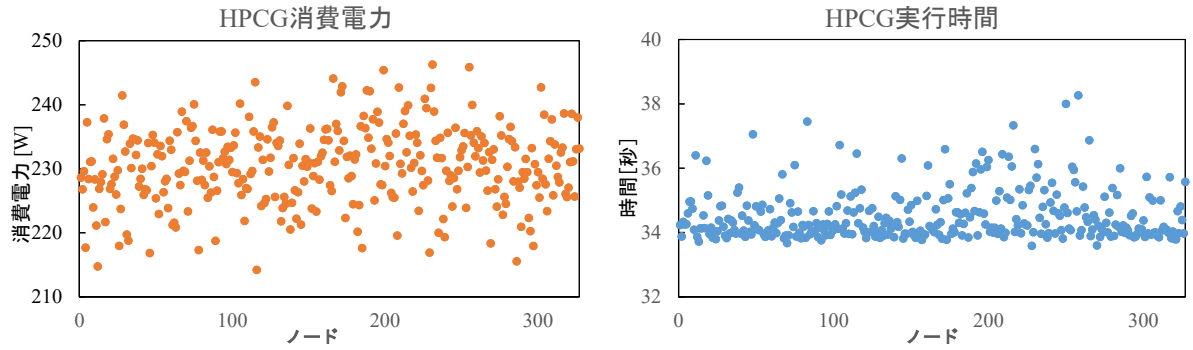

図 2　XC40 ノードにおける HPCG ベンチマーク評価結果

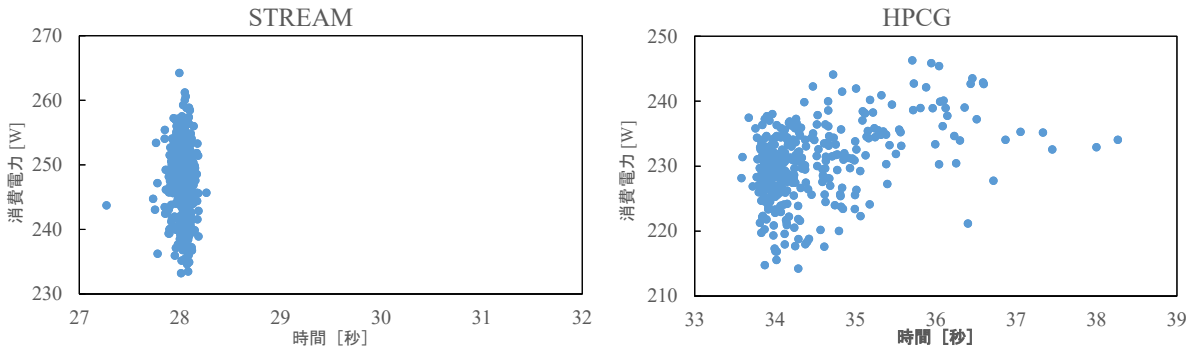

図 3　STREAM と HPCG の消費電力と実行時間の関係

できた順であり、その並び自体に意味は無い。また、消費電力と実行時間で横軸は同じノード順である。実行時間では、ほとんどの結果が 27.8～28.2 秒に収まっている。一方で、消費電力は 235～260W となっており、ノード毎の結果の変動が大きいことが明らかに分かる。詳細に解析すると、実行時間は最大で 3.7%の変動があるが、70%以上のノードは実行時間の変動が 3%以下となっている。消費電力は最大で 13.5%の変動があり、50%以上のノードで 6～12%の変動があった。STREAM は CPU ではなく、メモリに対して負荷の高いベンチマークであるがため、メモリ性能には大きなばらつきはなく、消費電力には大きなばらつきがあることが分かる。

3.2 HPCG ベンチマーク

次に HPCG の評価結果を示す。STREAM の結果と同様に評価結果を図 2 に表す。ノード数が STREAM と異なるのは、STREAM と HPCG で 3 回実行できたノード数に違いがあるからである。HPCG では消費電力は STREAM の場合と同様に変動が大きい。一方で、実行時間は多くが下限の 34 秒付近に分布し、一部実行時間が遅くなるノードがあることが分かる。詳細には、78%のノードが 4%以内の実行時間変動となっており、最大で 14%の変動があった。STREAM に比べて変動が大きいことが分かる。消費電力は 50%のノードで 6～12%の変動があった。

3.3 STREAM と HPCG の評価比較

STREAM と HPCG の評価結果を比べると、変動の振る舞い自体は似ているが、変動の幅は違いがあった。HPCG では実行時間の変動の幅が消費電力と同じ程度であり、アプリケーションにより、異なることが分かる。ここで、STREAM と HPCG の評価結果から実行時間と消費電力の関係を図 3 に示す。この図から、明らかに HPCG の実行時間の変動幅が大きいことが分かる。一方で、消費電力の大きさ自体は STREAM の方が大きくなっているが、消費電力の変動幅は両ベンチマークとも 30W 程度となっている。ここから、消費電力のばらつきはある程度どのアプリケーションでも共通の可能性がある。

3.4 他のシステムのばらつき評価

ノード消費電力の変動が XC40 以外にも現れることを確認するため、九州大学のスーパーコンピュータである ITO-A を利用し、ノードの消費電力を計測した。ITO-A の構成詳細は、九州大学情報基盤研究開発センターの Web ページを参照されたい[4]。ノード消費電力に関係する CPU は Skylake

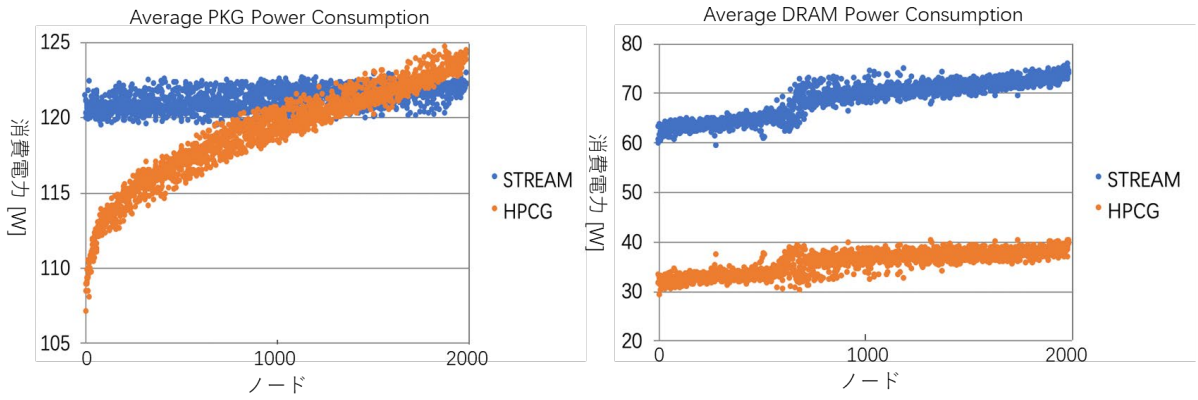

図 4　ITO-A における STREAM と HPCG の CPU（PKG）と DRAM 消費電力

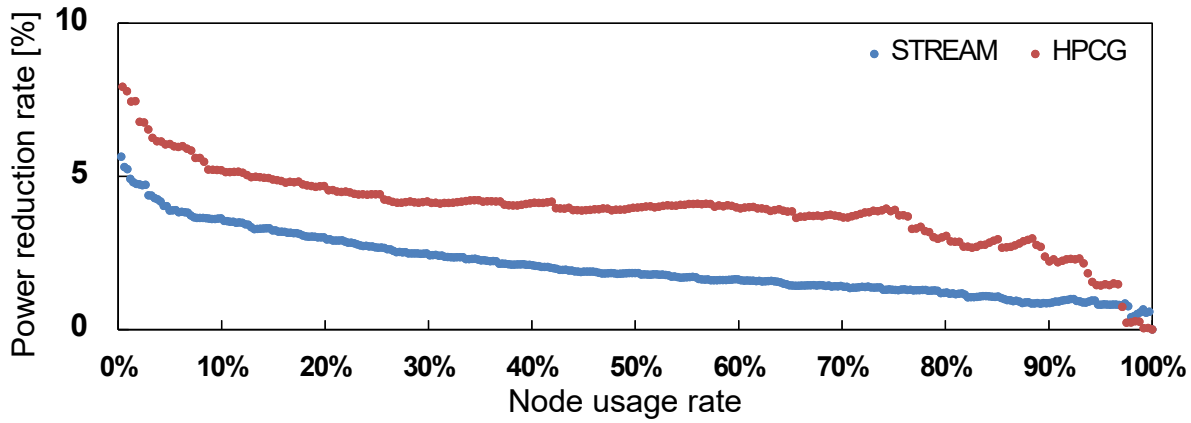

図 5　ノード利用率を変化させた場合の消費電力効率の良いノード優先ジョブ割当による消費電力削減割合

世代の Xeon を 2CPU/ノード搭載し、DDR4-2666×12/ノードとなっている。この ITO-A を全ノード（2,000 ノード）利用し、消費電力の変動を調べた。ITO-A では RAPL が利用できたため、CPU 電力（PKG）と DRAM 電力を CPU カウンターから取得し、消費電力とした。図 4 に ITO-A における STREAM と HPCG 実行時のノード消費電力を示す。ここでは、HPCG に PKG 消費電力が低い順にノードを並べている。HPCG は、CPU（PKG）消費電力で大きな変動（約 15W）があるが、STREAM では 3W 程度と変動幅が小さくなっている。DRAM の消費電力では、STREAM で 15W 程度と DRAM 消費電力から考えると大きな変動が見える。HPCG でも 10W 程度の変動はあり、DRAM 消費電力が 30～40W と考えると変動幅の割合は大きい。このように XC40 ではない、他システムでも消費電力の変動自体は存在することが確認できた。

4 消費電力のばらつきの活用

同一システム内のノードにおける実行時間のばらつきや消費電力の変動を理解しておくことは、スパコンセンターの運用としては重要であるが、この消費電力ばらつきを上手く利用することで、スパコン全体の消費電力の削減に繋がることが考えられる。ノードのジョブへの割当はスパコンによってルールは異なるが、一般的にノードの消費電力を考慮した割当は行われていない。今回の評価では、ノードの消費電力は変動がある、つまり、消費電力が少ないノードや多いノードがあることが分かった。この情報を利用し、消費電力の少ないノードを優先してジョブに割り当てれば、スパコンを運用する際にスパコン全体の消費電力を削減できると考えられる。

XC40 の評価結果から、平均消費電力より少ない消費電力となるノードの割合は約 10%であり、その 10%のノードの中にも平均消費電力に近いノードから最も消費電力が少ないノードまでが含まれる（図 1、2 を参照）。ここで、非現実的だが、スパコン上で 1 ノードしか使われない運用状況を考える。最も消費電力が少ないノードが割り当てられ続ける状況と、様々な消費電力のノードが割

り当てられる状況（結果的に平均的なノード消費電力になる）を比べると、運用結果として消費電力が削減される。この場合、ノードの平均消費電力に対する最低消費電力の差の量だけ削減され、XC40 の評価から計算すると、STREAM で 5.6%、HPCG では 7.7%の消費電力削減となる。

現実的な状況を考えると、XC40 を 1 ノードだけの利用から 1,800 ノード利用する状況まで考えて、同様の見積もりを行うと、図 5 のようなノードの利用率と消費電力削減率が求まる。ここでは、理想的に 1 ノードジョブの STREAM または HPCG のみが動くことを仮定している。利用率が 100%になれば、消費電力の低いノード優先割当は全く行えないので、消費電力削減率は 0 となる。また、利用率が上がるほど消費電力が高いノードも割り当てられるため、特に HPCG では利用率が 75%を超えると消費電力削減率が低くなる。更に消費電力が平均より良いノードは全体の 10%程度だったため、利用率が 10%を超えると、消費電力削減率が低くなる傾向が見える。70%や 80%の利用率では、消費電力が 5%も削減されず、2、3%の削減となっているが、消費電力量が大きいスパコンにとっては、意味のある削減量と考えられる。

この見積もりは、消費電力の良いノードから利用していき（N_0, N_1,…, N_n）、その場合のノード消費電力の和（$P_0+P_1+\ldots+P_n$）を平均的な消費電力ノードの和（$P_{avg}\times n$）で割ることで求めており、ごく簡単な見積もりとなっている。しかしながらノード消費電力の変動を活用することにより、スパコン全体として消費電力が削減可能だという 1 つの例となる。より現実的な、マルチノードジョブや実際のスパコン運用から取り出したワークロードを用いた、ノードの消費電力のばらつきを利用した消費電力の削減に関しては、別の研究として報告されている[5]。

5 まとめ

本研究では、京都大学学術情報メディアセンターのスパコンである XC40 におけるノード消費電力のばらつきを評価した。半導体製造上の性能ばらつきは一定量存在することは知られているが、製品として規定された性能を満たすために、そのばらつきが消費電力に変換されている。今回の評価では、ベンチマークの実行時間のばらつきよりも消費電力のばらつきの方が大きく、また、消費電力のばらつき自体を持つノードも多かった。最大最小消費電力の差は 30W にもなり、スパコンの運用上このようなばらつきを把握しておくことは重要だと考えられる。また、実行時間のばらつきはアプリケーションに依存することが予想され、今後スパコンの運用を行う上で、実行時間に影響のあるアプリケーションの把握やユーザへの情報提供も必要となる可能性が高い。

消費電力のばらつきは運用上好ましくないが、そのばらつきを活用することで、消費電力の削減に繋がる可能性もある。消費電力のいばらつきの中でも消費電力が低いノード、消費電力効率の良いノードに優先してジョブを割り当てることを考えると、単純な見積もり計算では、ノード利用率が 80%程度であっても 2～3%程度の消費電力削減が見込める結果となった。この見積もりは想定条件が現実的では無いが、うまくばらつきを利用することで、スパコンの運用に良い影響を与える可能性が示されたと考えられる。

今後は、新システム導入時や全ノード利用時にノードの特性を把握しておくことができるように、従来には行われていない評価を容易に行えるツールの整備を考えている。また、ノード消費電力のばらつきを利用したスパコンの消費電力削減が可能なノードアロケーション手法を開発していく。

参考文献

[1] Andrew Younge, Gregor von Laszewski, Lizhe Wang, and Geoffrey Fox. Providing a Green Framework for Cloud Based Data Centers. 01 2011.

[2] STREAM Benchmark (https://www.cs.virginia.edu/stream/)

[3] HPCG Benchmark (https://hpcg-benchmark.org/)

[4] 九州大学情報基盤研究開発センター研究用計算機システム (https://www.cc.kyushu-u.ac.jp/scp/)

[5] K. Fukazawa, J. Zhou, and H. Nakashima, Energy Aware Scheduler of Single/Multi-node Jobs Considering CPU Node Heterogeneity, IGSC2022 Proceedings, accepted.