Wisteria-O におけるノード間の電力ばらつきとその応用

三 輪 忍

電気通信大学大学院情報理工学研究科

１．はじめに

数千台（多いものでは数十万台）もの計算ノードによって構成されるスーパーコンピュータは膨大な電力を消費しており、その運用に必要とされる費用の大部分を電気料金が占めることで知られている。特に近年は電気料金が大幅に高騰したことにより、一部の計算機センターではシステムの運用方法そのものを見直さざるを得ない状況に追い込まれている。例えば、Wisteria-O では 2023 年 4 月からユーザの利用負担金を 1.5 倍値上げしたのに加えて、25%のノードを稼働停止する縮退運転を同年 7 月まで実施していた。また、「富岳」でも 3 分の 1 のノードを稼働停止する縮退運転が 2022 年度に実施されており、さらに 2023 年度からは、ユーザに対して省電力化を促すために消費電力削減量に応じてジョブを優先実行する仕組みを導入している。このように、多くの計算機センターにとってシステムの省電力化は喫緊の課題である。

一方、スーパーコンピュータを構成する各ノードには**電力ばらつき**が存在することが知られている。スーパーコンピュータの各計算ノードには同じ製品仕様の CPU や GPU が利用されることが多いが、これらの CPU や GPU は、理論上、同じアプリケーションに対して同じ消費電力を示すことが期待される。しかしながら、実際のシステムにおいては、LSI の製造ばらつき[1]やマシンルーム内の温度ばらつき[2]などの影響により、アプリケーションの実行に使用する CPU や GPU によって消費電力が異なることが報告されている（B. Rountree, et al. 2012; D. Hackenberg 2014; K. Yoshida, et al. 2022; P. Sinha, et al. 2022)。この現象を電力ばらつきと呼ぶ。なお、性能に関しては、一部のアプリケーションとシステムを除き、ノード間のばらつきはほとんどない[3]ことがわかっている。そのため、上記の性質をうまく利用すれば計算サービスの質を維持しつつシステムの消費電力を削減することが可能であり、電力ばらつきを考慮した電力管理手法の開発が行われている（Y. Inadomi, et al. 2015; E. Totoni, et al. 2015; B. Acun, et al. 2016; R. Sakamoto, et al. 2017; D. Chasapis, et al. 2019)。

我々の研究グループでは、2022 年 4 月から、Wisteria-O における電力ばらつきの調査と電力ばらつきを考慮した縮退運転方法の開発を行っている(草場智也ら 2023)。これまでの研究では、Wisteria-O を構成する 7,680 台のノードの中から無作為に抽出した 2,304 台のノードにおいて OpenMP 版の NAS Parallel Benchmark (NPB)を実行したところ、最大 26.4%の消費電力差を示す

[1] 製造工程において発生する LSI の不均質性。

[2] マシンルーム内の温度は一定の温度以下となるように管理されているが、空調の位置やラック内の位置によってコンピュータの周辺温度は異なる。また、LSI の静的電力はチップ温度に依存するため、同じ処理を行う場合でも、チップの周辺温度が異なる場合は消費電力が異なることが報告されている（池淵大輔ら 2010)。

[3] ノード間の性能ばらつきは CPU や GPU の消費電力が制限電力に到達した場合に自動的に周波数を低下させる機能が原因で発生するが、多くのシステムとアプリケーションの組み合わせでは CPU や GPU の消費電力が上記の制限電力に到達することはない。

項目	詳細
総理論演算性能	25.9 TFLOPS
総ノード数	7,680
総主記憶容量	240TB
インターコネクト	Tofu interconnect D

第 1 図: Wisteria-O のシステム構成。

項目	詳細
プロセッサ	A64FX (48+2 or 4 コア)
理論演算性能	3.3792 TFLOPS
製造プロセス	TSMC 7nm
メモリ	32GB, 1,024GB/s

第 2 図: Wisteria-O のノード構成。

ことを確認している。さらに、上記の電力ばらつきを考慮して縮退運転時の稼働停止ノードを選択した場合には、システムの消費電力削減量が最大 10.2%増加することを確認している。しかしながら、Wisteria-O の全 7,680 ノードにおける電力ばらつきと、電力ばらつきを考慮して稼働停止ノードを全 7,680 ノードの中から選択した場合の縮退運転の効果は未確認であった。

このような背景から、2023 年 7 月に実施した大規模 HPC チャレンジにおいて、我々の研究グループは Wisteria-O の全 7,680 ノードを対象に電力ばらつきの調査を実施した。より具体的には、Wisteria-O の各ノードにおいて計 8 種類のアプリケーションを実行した時の消費電力を計測し、分析を行った。また、上記の実験によって求めた各ノードの消費電力データをもとに、縮退運転時に稼働停止するノードを変更した場合の消費電力削減効果の見積もりも行った。その結果、1) Wisteria-O のノードには大きな電力ばらつきが存在すること (平均消費電力の差は最大 42%)、2) ノードの消費電力の大小関係はアプリケーションによって差がほとんどないこと (あるアプリケーションを実行した時の消費電力が大きいノードは、他のアプリケーションを実行した時の消費電力も大きい)、3) 電力ばらつきを考慮して稼働停止ノードを選択することで縮退運用時の消費電力削減量が最大 25.7%増加すること、が確認できた。本稿では、その詳細について報告する。

本稿の構成は以下の通りである。まず次章では、実験環境である Wisteria-O のアーキテクチャについて述べる。続く 3 章では今回行った実験の方法について説明し、4 章では実験結果を示す。最後 5 章でまとめと今後の展望を述べる。

2. Wisteria-O

Wisteria-O は、東京大学情報基盤センターが 2021 年 8 月に正式運用を開始したスーパーコンピュータ Wisteria/BDEC-01 のサブシステムである。Wisteria-O のシステム構成を第 1 図に示す。Wisteria-O は「FUJITSU Supercomputer PRIMEHPC FX1000」20 ラックから構成されており、後述する Fujitsu A64FX プロセッサによって構成されたノードを計 7,680 台搭載している。各ノードは 6 次元メッシュ／トーラス結合の Tofu interconnect D によって接続されており、総理論演算性能は 25.9PFLOPS、総主記憶容量は 240TB である。

Wisteria-O のノードの諸元を第 2 図に示す。各ノードは CPU として SVE (Scalable Vector Extension) を実装した A64FX プロセッサを 1 台搭載する。SVE は Armv8-A 64 ビットアーキテクチャのスーパーコンピュータ向け拡張である。A64FX は TSMC 7nm 半導体プロセスを用いて製造されており、約 90 億個のトランジスタによって構成されている。CPU あたり計算コアを 48 個、アシスタントコアを最大 4 個搭載しており、ピーク演算性能は倍精度浮動小数点演算で 3.3792TFLOPS である。メインメモリは 4 スタックの HBM2 で構成されており、容量は 32GB、バンド幅は 1,024GB/s である。

アプリ名	最大 (W)	最小 (W)	電力差 (W)
CG	217	177	40 (18.4%)
EP	142	127	15 (10.6%)
FT	202	167	35 (17.3%)
IS	134	119	15 (11.2%)
LU	174	150	24 (13.8%)
MG	248	198	50 (20.2%)
DGEMM	239	177	62 (25.9%)
STREAM	237	191	46 (19.4%)

第 3 図: GIO ノードの電力ばらつき。

アプリ名	最大 (W)	最小 (W)	電力差 (W)
CG	199	153	46 (23.1%)
EP	119	104	15 (12.6%)
FT	183	144	39 (21.3%)
IS	112	95	17 (15.2%)
LU	155	127	28 (18.1%)
MG	230	175	55 (23.9%)
DGEMM	222	153	69 (31.1%)
STREAM	219	169	50 (22.8%)

第 4 図: BIO ノードの電力ばらつき。

Wisteria-O のノードには、通常の計算処理に加えて特殊な処理を行うノードが存在する。1 つは FEFS ファイルシステムとの間の I/O 処理を行うグローバル IO ノード (GIO ノード)、もう 1 つは各計算ノードのブートディスクを保持しブート処理を行うブート IO ノード (BIO ノード) である。また、通常の計算処理のみを行うノードに関しても、Tofu インターフェースであるアクティブ光ケーブルの電力を担うノード (AOC ノード) と担わないノード (非 AOC ノード) の 2 種類がある。これらのノードは同じアプリケーションを実行した場合でも消費電力が大きく異なることから、本研究では電力ばらつきをノードの種類ごとに分析する。

3. 実験方法

実験には OpenMP 版の NPB に含まれるアプリケーション 6 種類 (CG, EP, FT, IS, LU, MG) に、計算/メモリバウンドなアプリケーションとして DGEMM と STREAM の 2 種類を加えた計 8 種類のアプリケーションを使用する。NPB の問題サイズは C とし、OpenMP のスレッド数はいずれのアプリケーションも 48 とした。

消費電力の計測は Power API (Sandia National Laboratories 2023) を用いて行う。Power API で測定可能な電力には推定電力と実測電力の 2 種類がある。推定電力は CPU の命令発行数などに基づいて推定した電力量であり、1 ミリ秒ごとに更新される。推定電力にはノードの個体差が反映されておらず、同一アプリケーションを実行した場合には常に同じ値を示す。一方、実測電力はノード内の電力計測素子によって計測された電力量であり、5 ミリ秒間隔で更新される。実測電力にはノードの個体差が反映されていることから、本研究では実測電力を使用する。

なお、Power API を用いた電力計測ではアプリケーションのカーネル部分の消費電力の計測も可能であるが、本研究で使用したアプリケーションはカーネル以外の部分の計算時間が全体の実行時間に占める割合が小さいことから、アプリケーション全体の消費電力を計測対象とした。また、測定間のばらつきを排除するために、各アプリケーションはそれぞれのノードで 40 回ずつ実行し、その平均消費電力を求めた。

4. 実験結果

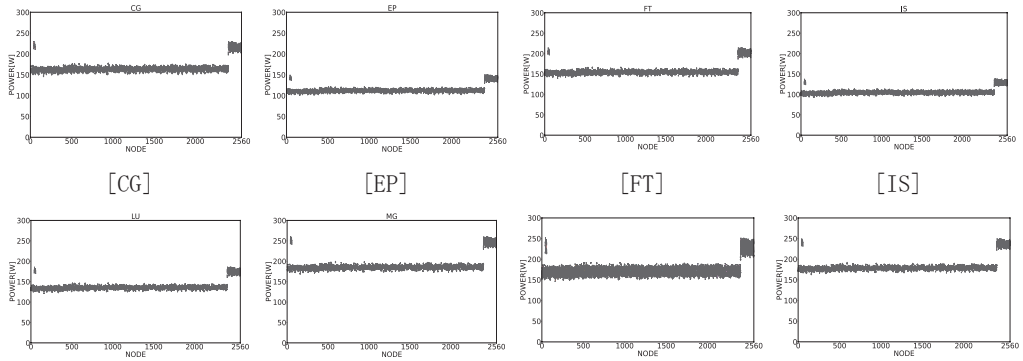
GIO ノードと BIO ノードの測定結果を、それぞれ、第 3 図と第 4 図に示す。表の第 2 列と第 3 列は、それぞれ、当該アプリケーションの実行において最も大きな/小さな電力を消費したノードの平均消費電力を表す。また、表の第 4 列は消費電力が最大のノードと最小のノードとの間の

アプリ名	最大 (W)	最小 (W)	電力差 (W)
CG	230	147	83 (36.1%)
EP	150	100	50 (33.3%)
FT	214	139	75 (35.0%)
IS	138	93	45 (32.6%)
LU	186	121	65 (34.9%)
MG	261	169	92 (35.2%)
DGEMM	252	147	105 (41.7%)
STREAM	250	164	86 (34.4%)

第 5 図: AOC ノードの電力ばらつき。

アプリ名	最大 (W)	最小 (W)	電力差 (W)
CG	217	138	79 (36.4%)
EP	138	92	46 (33.3%)
FT	202	130	72 (35.6%)
IS	126	84	42 (33.3%)
LU	174	113	61 (35.1%)
MG	249	160	89 (35.7%)
DGEMM	239	139	100 (41.8%)
STREAM	238	155	83 (34.9%)

第 6 図: 非 AOC ノードの電力ばらつき。



[CG] [EP] [FT] [IS]

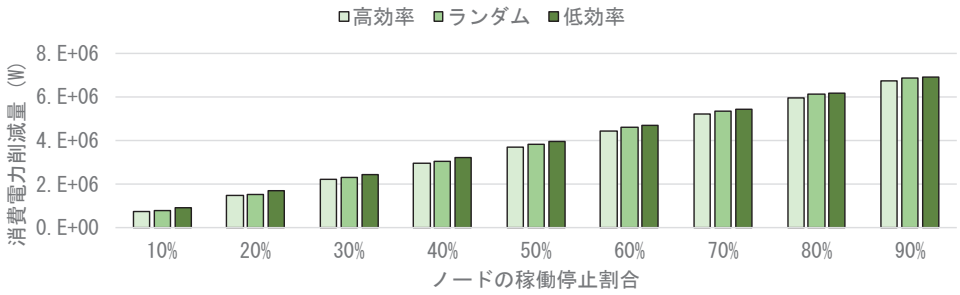
[LU] [MG] [DGEMM] [STREAM]

第 7 図: AOC ノードの消費電力。

消費電力差を表しており、括弧内の数字はその差を最大消費電力（第 2 列の値）で割った値である。表より、いずれのアプリケーションにおいても電力ばらつきが存在し、電力ばらつきの大きさはアプリケーションによって異なることが確認できた。また、いずれのノードにおいても DGEMM が最大の電力ばらつきを示しており、その大きさは GIO ノードでは 62W (25.9%)、BIO ノードでは 69W (31.1%) であった。

AOC ノードと非 AOC ノードの消費電力測定結果を、それぞれ、第 5 図と第 6 図に示す. 表より、GIO ノードや BIO ノードと比べて、AOC ノードと非 AOC ノードの電力ばらつきは全体的に大きいことが確認できる。特に DGEMM に関しては、AOC ノードでは最大 105W (41.7%)、非 AOC ノードでは最大 100W (41.8%) の消費電力差が存在することがわかった。この結果は先行研究 (Y. Inadomi, et al. 2015) で報告されている CPU の電力ばらつき (最大 23%) と比べて明らかに大きい。その理由として、Power API が報告する消費電力量は A64FX プロセッサだけでなく HBM2 の消費電力量も含んでおり、HBM2 にも電力ばらつきが存在することが原因と考えられる。

AOC ノードの全ノードの消費電力を第 7 図に示す。GIO、BIO、非 AOC の全ノードの消費電力データは AOC ノードと同様の傾向を示したため割愛する。各グラフの横軸はノード番号を表しており、縦軸は消費電力を表す。グラフより、ノードの消費電力の大小関係はアプリケーションによってほとんど変わらないことが確認できる。すなわち、あるアプリケーションに対して相対的に大きな (小さな) 消費電力を示すノードは、他のアプリケーションに対しても相対的に大きな (小さな) 消費電力を示す。



第 8 図：STREAM の消費電力データを用いて稼働停止ノードを選択した場合の消費電力削減量

また同グラフより、いずれのアプリケーションにおいても、2400 番台のノードが他のノードと比べて著しく大きな電力を消費していることが確認できる。これらのノードでは何らかの不具合が発生している可能性があり、このような測定結果となった原因については東京大学情報基盤センターに現在問い合わせ中である。

上記の実験結果を踏まえて、縮退運転時に稼働停止するノードを変更した場合の消費電力削減量の見積もりを行う。まず前提として、PRIMEHPC FX1000 は 1 ラック(384 ノード)に GIO ノードが 8 台、BIO ノードが 24 台、AOC ノードが 160 台、非 AOC ノードが 192 台存在する。そのため、稼働停止ノードの選択は上記のノード比を保った状態で行う。また、PRIMEHPC FX1000 では非 AOC ノードと GIO/BIO/AOC のいずれかのノードが対となり、1 つの CMU（ボード）を構成している。実際に Wisteria-O のノードの一部を停止させる場合は Tofu のトポロジーを考慮して停止ノードを選択する必要があるが、本研究では CMU 単位でノードを（ボードごと）交換した上で Tofu のトポロジー的な制約を満たすようにノードを停止することを想定し、稼働停止ノードを CMU 単位で選択する。

結果を第 8 図に示す。図は、STREAM 実行時の各ノードの消費電力データを用いて稼働停止ノードを選択した上で、全ノードで STREAM を実行した場合の消費電力削減量である。ノードの選択に使用するアプリケーションを変更した場合の消費電力削減量は、ほぼ同様の結果が得られたため省略する。グラフの横軸は Wisteria-O の全ノードに対するノードの稼働停止割合を表しており、縦軸は稼働ノード数が 100%の状態において全ノードで STREAM を実行した場合の消費電力に対する縮退運転時の全稼働ノードで同アプリケーションを実行した場合の消費電力削減量を表す。稼働停止割合ごとの 3 つの棒グラフは、左から順に、電力効率の高いノードから順にノードを停止した場合、停止ノードをランダムに選択した場合、電力効率の低いノードから順にノードを停止した場合を表す。

グラフより、電力ばらつきを考慮して稼働停止ノードを選択する方法は、ノードの稼働停止割合が小さい時ほど効果が高い。特にノードの稼働停止割合が 10%, 20%, 30%の場合において、電力効率の低いノードから順に停止する方法は電力効率の高いノードから順に停止する方法に対して、それぞれ、25.7%, 15.3%, 11.2%、ランダムにノードを停止する方法に対して、それぞれ、18.7%, 10.3%, 7.0%の電力を追加で削減できることがわかった。

５．まとめと今後の展望

本稿では，2023 年 7 月の大規模 HPC チャレンジにて行った，Wisteria-O におけるノードの電

力ばらつきの実験結果について報告した。今回の実験では，Wisteria-O の全 7,680 台のノードには有意な電力ばらつきが存在しており、ノードの電力の大小関係はどのアプリケーションを実行した場合でもほとんど同じであることが確認できた。また、縮退運転時には、今回の実験で明らかになった電力効率の低いノードから順に停止させることによって、システムのさらなる省電力化が見込めることも確認した。

今後は、電力ばらつきの計測に使用するアプリケーションを増やして実際のワークロードに近づけることで、運用中のシステムの消費電力に対する電力ばらつきの影響を評価したいと考えている。また、「富岳」でも同様の実験を行うことにより、Wisteria-O だけでなく「富岳」にも応用可能な省電力運用技術を開発する予定である。

謝　辞

本研究は、東京大学情報基盤センター「Wisteria/BDEC-01 大規模 HPC チャレンジ」における採択課題「Wisteria-O における CPU の電力性能ばらつきと縮退運転への応用に関する研究」によって行われた。実験とデータ分析に関しては、電気通信大学の草場智也君に大いに協力していただいた。また、本研究を進めるにあたり、電気通信大学の本多弘樹教授、八巻隼人准教授、および、東京大学の塙敏博教授には非常に有意義なご意見をいただいた。ここに感謝の意を表する。

参　考　文　献

池淵大輔ら，『細粒度パワーゲーティングを適用した汎用マイクロプロセッサ Geyser-1』，情報処理学会研究報告 2010-ARC-187/2010-EMB-15, No. 1, pp. 1-6, 2010

B. Rountree, et al., “Beyond DVFS: A First Look at Performance Under a Hardware-Enforced Power Bound,” IPDPS Workshops (HPPAC), pp. 947-953, 2012

D. Hackenberg, “Node Power Consumption Variability,” Energy-Efficient HPC WG Workshop, 2014

Y. Yoshida, et al., “Analyzing Performance and Power-Efficiency Variations among NVIDIA GPUs,” International Conference on Parallel Processing (ICPP), No. 65, pp. 1-12, 2022

P. Sinha, et al., “Not All GPUs Are Created Equal: Characterizing Variability in Large-Scale, Accelerator-Rich Systems,” International Conference on High Performance Computing, Networking, Storage and Analysis (SC22), No. 65, 15 pages, 2022

Y. Inadomi, et al., “Analyzing and Mitigating the Impact of Manufacturing Variability in Power-Constrained Supercomputing,” SC15, No. 78, pp. 1-12, 2015

E. Totoni, et al., “Scheduling for HPC Systems with Process Variation Heterogeneity,” PPL Technical Report, 10 pages, 2015

B. Acun, et al., “Variation Among Processors Under Turbo Boost in HPC Systems,” ICS, No. 6, pp. 1-13, 2016

R. Sakamoto, et al., “Production Hardware Overprovisioning: Real-world Performance Optimization using an Extensible Power-aware Resource Management Framework,” IPDPS, pp. 957-966, 2017

D. Chasapis, et al., “Power Efficient Job Scheduling by Predicting the Impact of

Processor Manufacturing Variability,” International Conference on Supercomputing (ICS), pp. 296-307, 2019
草場智也ら,『A64FX プロセッサにおける電力・性能ばらつきの評価・分析』, 情報処理学会研究報告 2023-HPC-188, No. 21, pp. 1-6, 2023（情報処理学会コンピュータサイエンス領域奨励賞受賞）
Sandia National Laboratories, “Power API,” http://powerapi.sandia.gov/ （2023 年 10 月 3 日アクセス）