温水冷却の効率に関する定量的評価の試み

庄司　文由、野中　丈士

理化学研究所計算科学研究センター

塙　敏博

東京大学情報基盤センター/最先端共同 HPC 基盤施設

１．はじめに

循環冷却水を用いた直接・間接水冷技術は HPC システムの冷却方式として広く浸透しており、得に世界最高性能の HPC システムではデファクトスタンダードとなりつつある。また、近年の技術革新により、高性能 CPU はより高温での動作に耐えられるようになり、エネルギー効率改善の目的としての温水冷却（Warm Water Cooling）は国内外の多くの HPC センターおよびデータセンターで採用されている。中にはアメリカ暖房冷凍空調学会（ASHRAE）が定める高温直接水冷（High-Temperature Direct Liquid Cooling）ガイドラインの W3-W5 クラス（>30℃）の冷却水供給温度を超える HPC システムも報告されている［1]。しかしながら、温水冷却の効果を正しく評価するためには、冷却の電力を節約できるというポジティブな点に加え、演算性能の低下や消費電力の増加等のネガティブな点も等しく考慮に入れる必要がある。我々はこれまで大規模 HPC チャレンジの一環として、最先端共同 HPC 基盤施設（JCAHPC）の関係者及びに施設運用主体の富士通株式会社のご協力の元で、2016 年より運用している Oakforest-PACS（OFP）を用いて循環冷水温度の変更による定量的評価を進めてきた（2019 年 10 月および 2021 年 5 月）[2]。冷水温度を高く設定することで、外気による自然な冷却を活用し、冷凍機等を駆動するための電力が節約できることを定量的に評価 [3]した。また、OFP 側にインストールされている Intel Parallel Studio XE 内の単体ノードベンチマーク向け LINPACK（Intel Optimized LINPACK Benchmark for Linux）を用いて、単体ノードでの冷水温度による演算性能への影響について定量的評価も行った [4]。本稿では複数ノード（CPU）を利用するジョブへの影響を評価するために、並列有限要素法コード（GeoFEM）を用いて、冷水温度を上げた場合の消費電力の増加および演算性能の低下について定量的な分析を試みた。その際に単体ノードでの実行性能を基に性能の近い CPU 同士をグループ化し、その影響の大きさを比較する調査も行った。特定のシミュレーションコードを用いた調査結果であるが、この様に単体および複数の CPU を使うジョブを、異なる冷水温度で実行し、分析結果に基づいた効率的な施設運用の実現向けの運用手順確立の手助けになれば幸いである。

２．実験方法

温水冷却の定量的評価という事で、通常運用時の冷却水設定温度（12℃）とそれよりも高い温度（18℃）への変更を施設運用側にお願いした。また、実行時の詳細データを得る目的で CPU のクロック周波数や消費電力データを一般ユーザー権限で取得できるよう turbostat のコマンド実行を可能とする設定も行って頂いた。本大規模 HPC チャレンジでは 4200 ノード（CPU）提供されたため、12℃と 18℃の冷却水設定温度で全ノードを用いた単体 GeoFEM を 3 回実行し、実行性

能を基に 16 ノード毎にグループ化し、並列 GeoFEM を 3 回実行し性能のバラつきを調査した。

温水冷却法では外気による自然な冷却を活用するため、冷却能力は外気温に大きく依存する。図 1 の左上は気象庁が公開している我孫子市の気温データ（10 分毎）を基に大規模 HPC チャレンジ当日の気温変化を示す。冷却水温度が 18℃に設定されている時間帯では気温が 18℃以下であったが、冷却水温度設定が 12℃の時間帯では外気温が 12℃を上回っているのが確認できる。図 1 の右上は提供された 4200 ノードのラック間での分布を示す。その際に 1 ラック（120 ノード）を全面（FRONT）、背面（REAR）に分けて表示している。このノード分布から 1 番から 18 番目の縦ラック群（R01-R18）と 19 番から 36 番目の縦ラック群（R19-R36）が主に利用されていることがわかる。図 1 の下段には両ラック群（R01-R18、R19-R36）への冷却水供給温度の推移を示す。これらの図からラックの全面と背面での冷却水供給温度の差のみならず、ラック群によっても冷却水供給温度に差があるのが確認できる。

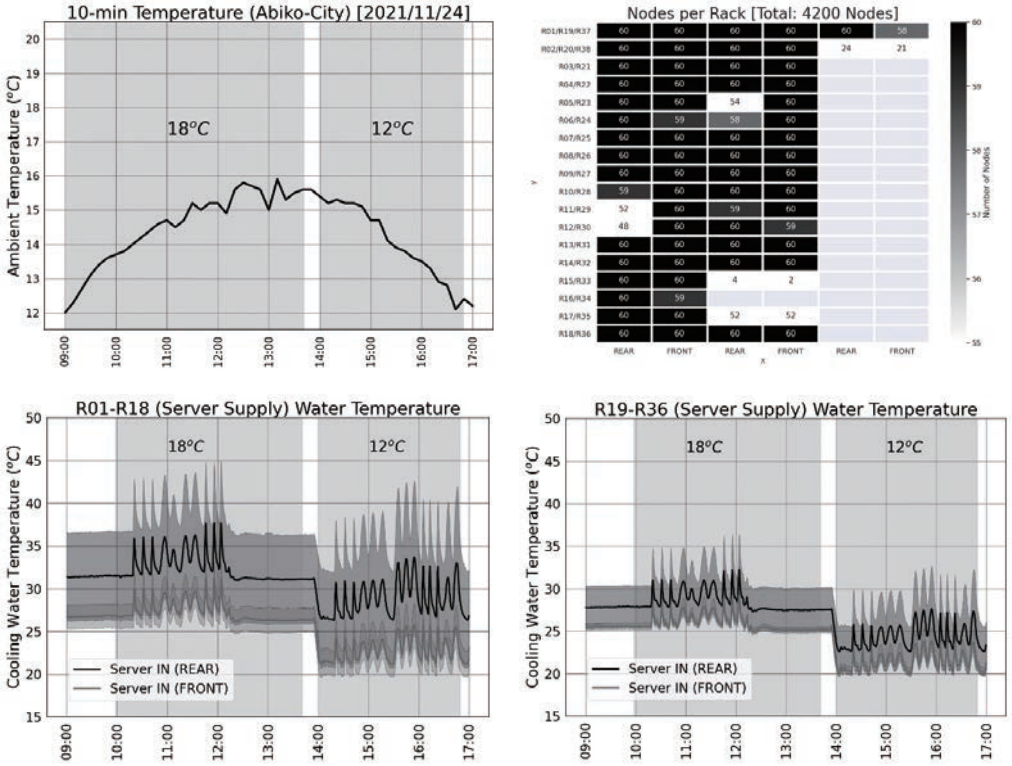

図 1　当日の我孫子市の温度変化（左上）、ラック内の利用ノード数（右上）、R01-R18 ラック群への冷却水供給温度の推移（左下）、R19-R36 ラック群への冷却水供給温度の推移（右下）

3．単体ノード性能

表 1 に単体ノード性能を得るために利用した GeoFEM に実行パラメータと問題サイズを示す。これはウィークスケーリングのベンチマーク向けに用意された設定ファイルであり、単体ノード（CPU）であっても 8 プロセス（各 8 スレッド）で実行を行う設定である。図 2 の左上は単体ノードでの GeoFEM 実行時間（3 回）の最も早かった時間（BEST TIME）の分布を表している。ピークが 2 つあることが確認できるが、実行時間の分布幅は数%と影響が少ないことから、今回はこの点についての詳細な分析は行っていない。重要なのは、冷却水設定温度を変更しても GeoFEM

の実行性能にはあまり影響がない点である。図 2 の右上は GeoFEM 実行時のノード単位の消費電力を表している。実行性能とは異なり冷却水設定温度が高くなると消費電力がわずかに増えるということも確認できる。図 2 の下段は各冷却水設定温度域での GeoFEM 実行時間を基にグループ化（16 ノード）した 256 グループの箱ひげ図である。この箱ひげ図の箱の部分（色が濃い部分）はデータの中央 50%の分布を表しており、ひげの先端は最小値と最大値を表している。また、箱ひげから離れた部分に点として表示されているものは外れ値である。箱ひげ図の下部分は最も早かった時間（BEST TIME）を表しており、この数値を元に右から左にかけてソートしてプロットしている。この図からは、最も遅かった時間つまり、箱の上部分と、最も速かった時間との間に相関が見られず、その傾向は冷水温度を変えても変わらないことが分かる。

GeoFEM	8 Process per Node 8 Threads per Process Problem Sizes: 256 x 128 x 128 (1 Node) 1024 x 256 x 256 (16 Nodes) 1024 x 512 x 256 (32 Nodes)

表 1　GeoFEM 向け実行パラメータ（単体ノード、並列処理）

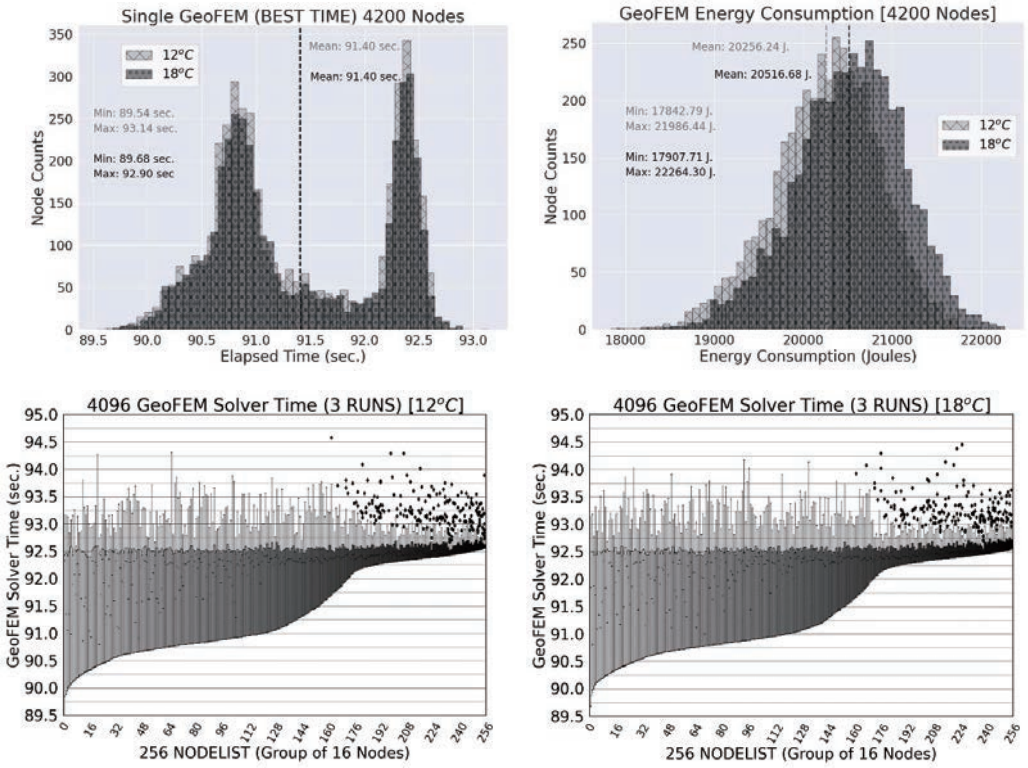

図 2　両設定温度域での単体ノード GeoFEM 実行時間（左上）と消費電力（右上）の分布、GeoFEM 実行時間を元にグループ化（16 ノード）した 256 グループの箱ひげ図（下段）

図 3 には 12℃の冷却水温度での単体ノード GeoFEM を実行した際の CPU 周波数の挙動を表して

いる。LINPACK での CPU 周波数の挙動[4]と違い、実行時では AVX512 利用による CPU 周波数の低下（周波数スロットリング）[6]が発動されないため、コンスタントに最高周波数（1.5GHz）で実行されるのが確認できる。図 3 の右側には色が濃い時間帯（15 秒－85 秒）で最高周波数を下回る 57 ノードを表している。この 57 ノードの周波数が低下した原因については別途分析が必要である。

全般的に見ると、LINPACK では周波数スロットリングが発動される際に CPU によって異なる周波数で推移していたが、GeoFEM はメモリバンド律速型のため、演算器の稼働率が低く、周波数スロットリングが発動される機会が（ほとんど）ないため、実行性能にも影響が少なく、冷水温度を変えてもその傾向は変わらないと考察できる。

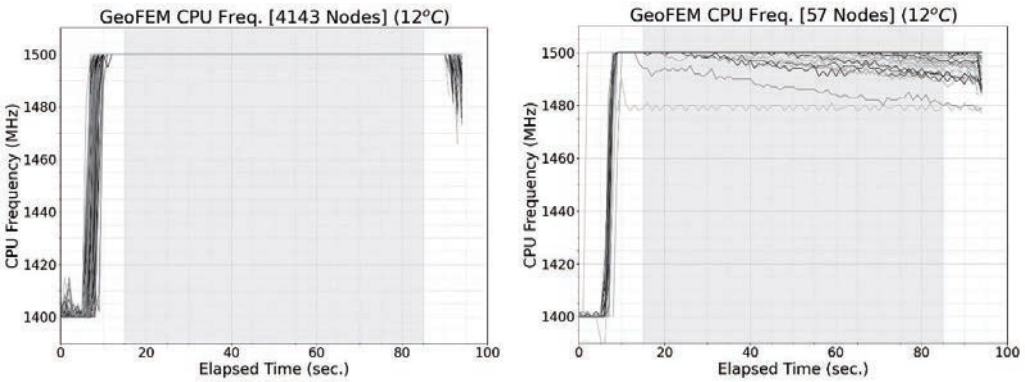

図 3 単体ノード GeoFEM 実行時の CPU 周波数の挙動

４．並列 GeoFEM 性能

図 4 は 12℃と 18℃の冷却水温度域での並列 GeoFEM を実行した際の実行時間を表している。上段は各 16 ノードの 256 グループでの最も早い実行時間（BEST TIME）をプロットしている。グループ番号は最も早かったノードを順番にグルーピング化しているため、性能順のグループ番号である。左寄りに最も早かったノードが集まっているグループであり、右寄りに最も遅かったノードが集まっているグループであるが、256 のグループ間での性能差がほとんど見られなかった。図 4 の上段は冷却水温度を変更した際の実行時間を表しているが、グループ間での明確な差異が見られなかった。また、下段の左側は 12℃の冷却水温度設定で 32 ノードの並列 GeoFEM の実行時間をプロットしたものであるが、同じようにグループ間（128）での性能差は数%に留まった。右側の図は真上の図から CPU 周波数が異なる挙動を見せていた 57 ノード（図 3）が入っているグループを排除したプロットである。図 5 は既に[2]で報告している冷却水温度設定の変更による単体ノード LINPACK 性能と消費電力の変化であるが、図 2 の単体ノード GeoFEM との比較の為にあえて表している。これらの図を比較すると冷却水温度設定を上げた場合は AVX512 処理も利用する CPU インテンシブな LINPACK では消費電力の上昇幅は少なく、AVX512 を利用しない GeoFEM の方が明確に確認できる。しかしながら、施設の消費電力削減[3]のメリットの方が上回る可能性が高いとも考えられる。また、GeoFEM の様な CPU インテンシブでないジョブを実行する際にはジョブスケジュラーが割当てるノードの影響が少ないとも考えられる。しかしながら、AVX512 を利用する CPU インテンシブのジョブは割当てられるノードによっては性能のバラつきが出てくる可能性が否定できないため、更なる調査が必要であると考える。

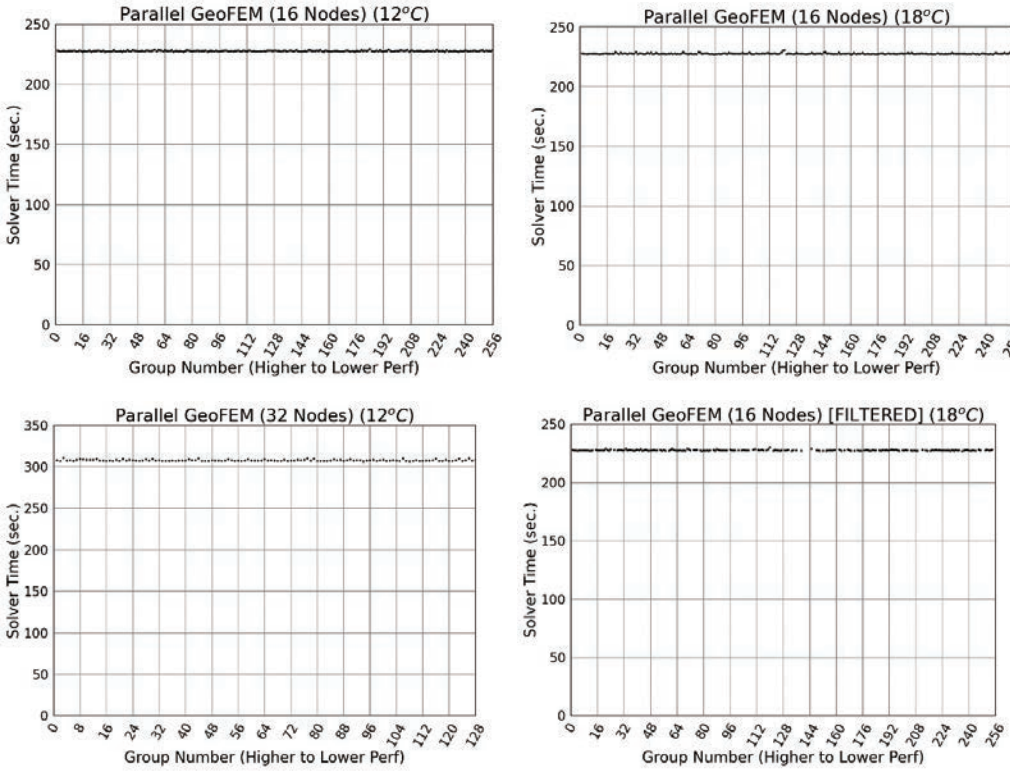

図 4　ソーティングされたグループ毎の並列 GeoFEM（16、32 ノード）の実行時間

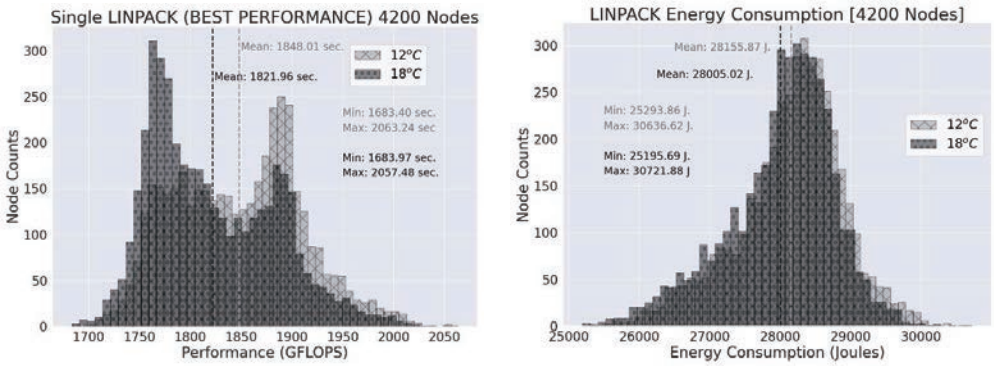

図 5　両設定温度域での単体ノード LINPACK 性能（左）と消費電力（右）の分布

5．まとめ

今回の実験は、冷水温度を高くしていった場合の性能に対する影響を、複数ノードジョブに対して定量的に評価するという目的で実施した。ノード数が増えた場合の影響を正しく評価するためにウィークスケーリングの評価が可能な GeoFEM を用いた。制限時間の問題等で主に 16 ノードのグループでの測定のみとなったが、単体での GeoFEM 性能の順に CPU をソーティングし、グループ化して、グループ間の並列性能の比較を行った結果、GeoFEM の様なメモリバンド律速で AVX512 処理を利用しないコードでは、冷却水の温度を上げて運用しても、クロック周波数の低下は起きにくく、結果としてジョブの実行性能への影響は小さいことが観測された。しかしなが

ら、HPL のような演算律速なコードでは、冷却水の温度が上がるとクロック周波数の低下とそれによる性能低下が発生しやすくなり、並列ジョブの場合はその影響がより顕著になると予想される。冷却水温度の上昇に対する反応には CPU 毎に差があり、実行性能に加えて、その再現性にも影響が出る可能性がある。実行のたびにパフォーマンスが変わるとなると、課金制度にも影響する。運用的には、より特性が近いノードを割り当てることで、並列ジョブの実行性能への影響を減らすことができるかも知れない。決定的な結論に至るまでには更なるデータの採取と定量的評価が必要であるが、本稿が分析結果に基づく効率的な施設運用の実現向けの運用手順確立の手助けになれば幸いである。

６．謝辞

本研究は、大規模 HPC チャレンジという制度がなければ到底実現できなかった。このような機会を与えていただいたことに深く感謝致します。また、今回の実証実験を行う上で、JCAHPC の関係者、運用主体の富士通株式会社の皆様に大変なご苦労をおかけしました。お詫びと共に深く感謝申し上げます。GeoFEM のコードの提供および OFP 上での実行に関するサポートについて、東京大学情報基盤センター　中島教授および Intel 株式会社　堀越氏に多大なご協力をいただきました。

参　考　文　献

[1] H. Shoukourian et al. “SuperMUC - the first high-temperature direct liquid cooled petascale supercomputer operated by LRZ,” in Contemporary High Performance Computing: From Petascale toward Exascale, Volume 3, J. S. Vetter, Ed. CRC Press, 2019, ch. 10.

[2] 庄司文由, 野中丈士, 塙敏博, ″温水冷却の効率に関する定量的評価の試み,″ スーパーコンピューティングニュース Vol.22 No.6 (2020 年 11 月) https://www.cc.u-tokyo.ac.jp/public/VOL22/No6/09_2020shoji.pdf

[3] J. Nonaka, T. Hanawa and F. Shoji, “Analysis on the Impact of the Cooling Water Temperature on the HPC System and Facility - Case Study: Oakforest-PACS (OFP) System and Facility,” ISC2020, Frankfurt, Germany, 2020. Available: https://2020.isc-program.com/presentation/?id=post131&sess=sess325.

[4] J. Nonaka, T. Hanawa and F. Shoji, “Analysis of Cooling Water Temperature Impact on Computing Performance and Energy Consumption,” 2020 IEEE International Conference on Cluster Computing (CLUSTER), Kobe, Japan, 2020, pp. 169-175, doi: 10.1109/CLUSTER49012.2020.00027.

[5] K. D. Stroup and P. Peltz, “Measuring and mitigating processor performance inconsistencies,” in Cray User Group Conference (CUG 2019), 2019. [Online]. Available: https://cug.org/proceedings/cug2019proceedings/includes/files/pap124s2-file1.pdf

[6] Mathias Gottschlag, Frank Bellosa, ″Mechanism to Mitigate AVX512-Induced Frequency Reduction,″ Technical Report, Karlsruhe Institute of Technology, 2018. https://arxiv.org/abs/1901.04982