Vol.2013-ARC-207 No.29
Vol.2013-HPC-142 No.29
2013/12/17

TSUBAME-KFC: 液浸冷却を用いたウルトラグリーンスパコン研究設備

遠藤 敏夫[1,a)]　額田 彰[1,b)]　松岡 聡[1,c)]

概要：高性能計算システムの将来の性能向上に向けた最大の制約条件は消費電力であり，計算機器の電力のみならず冷却に係る電力も合わせて議論される必要がある．著者らはスパコン冷却電力の大幅削減を目的として，200TFlops 級の高性能高密度スパコンを，液浸冷却および大気冷却方式により冷却するコンテナ型研究設備である，TSUBAME-KFC を構築してきた．電力性能評価の結果，PUE=1.15，4.5GFlops/Watt という，世界最高クラスの性能を達成した．

1. はじめに

スーパーコンピュータの今後のさらなる高性能化のための最大の課題は消費電力の問題である．スパコンの最も大きな意義はその高い性能であるが，それに加えて以下に少ない電力・エネルギー消費で高性能を実現可能かに，大きく注目が集まっている．エクサスケールスパコンの 2020 年前後の実現に向けて，各国の研究機関により研究が推進されており [1], [2]，現実的に構築可能な 20MWatt 級以下にて実現するために，50GFlops/Watt という電力性能比を実現することが不可欠とされている．

その現れとして，2007 年に始まったスパコン世界ランキングである Green500?では，電力性能比 (Flops/Watt) によるランキングを年二回発表している．東京工業大学学術国際情報センターで運用している TSUBAME2.5 スーパーコンピュータの前身である TSUBAME2.0 は，2010 年秋の本ランキングにおいて世界二位，および Greenest Production Supercomputer 賞を獲得した [4]．2013 年 9 月にアップグレードを行った TSUBAME2.5 も．2013 年秋の同ランキングにおいて上位を獲得することが期待されている．

同センターでは，現状の TSUBAME スパコンの省エネ設計・運用だけにとどまらず，次世代のスパコンのための更なる省エネ技術の研究を行っている．その一環として，スパコン運用時において無視できない電力を消費する冷却部分に注目し，先進的な冷却機能を取り入れたコンテナ型研究設備 TSUBAME-KFC (Kepler Fluid Cooling) について報告する．この設備においては，スパコン計算ノードをまるごと液体に浸す，液浸冷却と呼ばれる手法と，冷媒液そのものの冷却のための大気冷却を併用する．この方式により，冷却方式の効率を示す PUE(power usage effectiveness) を，理想値である 1 に近づけることをめざす．また，一台の計算ノードに 2CPU(ソケット) と 4GPU と搭載する高密度実装および，液浸冷却によるノード内ファン除去の影響により，Green500 の指標である Flops/Watt においても世界トップクラスの 4GFlops/Watt を目指す．本稿ではこの TSUBAME-KFC の電力性能の予備評価結果について報告する．

2. 電力性能の指標

本節では，電力性能の指標として，Green500 で利用される Flops/Watt と，PUE(power usage effectiveness) について述べる．いずれもスパコン電力性能のあくまで一面のみを表す指標であることに注意する必要がある．

2.1 Green500 の指標

Green500 においては，Linpack ベンチマーク [5] 実行時の測定値に基づき「*Linpack* 速度性能 *(Flops)*/実行時電力 *(Watt)*」を求め，その値の高さによってランキングが定められる．なお Green500 にランクされるのは，同時に発表される速度性能ランキング Top500 において 500 位の性能以上となったシステムである．

ここで分母となる電力は，以下のような条件で測定されたものとなる．

(1) 計算ノードおよびネットワーク機器の消費電力を測定

[1] 東京工業大学学術国際情報センター
[a)] endo@is.titech.ac.jp
[b)] nukada@matsulab.is.titech.ac.jp
[c)] matsu@is.titech.ac.jp

する．

(2) ストレージ機器の電力については含めなくてよい．

(3) 冷却のための電力については含めなくてよい．

(4) Linpack のカーネル実行時間のうち，全実行時間でなく，下記の範囲の電力について平均する．全実行時間のうち，連続する 20%以上の時間を測定する．その時間が 1 分未満である場合には 1 分以上を測定する．ただしカーネル実行時間のうち，最初の 10%と最後の 10%は測定に含めないこと．

条件 (3) に示したように，スパコンの運用の効率に大きく影響を与える冷却電力が含まれていないことに注意する必要がある．つまり冷却電力の進歩は，この指標には原則的に反映されない．

また，条件 (4) に示したように，実行時間の一部の計測でよいことになっているが，Linpack ベンチマークの特性として，消費電力が一定ではないことに注意が必要である．実行前半に計算密度および消費電力が高く，後半になるにつれ消費電力が下がる傾向にある．実際に，我々が 2013 年秋の Green500 に提出した TSUBAME2.5 の性能は下記のようなものであった：速度性能は 2.831PFlops であり，その際のカーネル実行時間全体の平均消費電力が 1125kWatt であった．一方，Green500 の規則に則りカーネル実行時間のうち 70%から 90%の部分の平均消費電力は 922.5kWatt となり，後者の値に基づき 3.069GFlops/Watt という値を提出した．

2.2 PUE

PUE(power usage effectiveness) はデータセンター等で広く電力使用効率を表す指標として用いられ，以下で求められる．

PUE= (IT 機器の消費電力*+*付帯設備の消費電力*) /IT* 機器の消費電力

本稿では，付帯設備の消費電力として冷却設備の消費電力を用いる．定義から，PUE の値は必ず 1 以上であり，1 に近いほうが電力使用効率が良いとされる．PUE=2 は，IT 機器と同等の電力を，冷却などの付帯設備が消費していることを示す．現在のスパコンセンターやデータセンターにおける PUE が 1.5 ～ 2 程度とされる．それに対し TSUBAME2 (2.0 時代) の年間平均 PUE は約 1.29 であり，より 1 に近く効率的であると言える．

PUE についての注意事項は，分母が IT 機器の消費電力であることに起因する．IT 機器 (スパコンにおいては計算ノードなど) の省電力化が進み，その一方で冷却電力などがそれほど下がらない場合には，計算上の PUE は悪化するという現象が起きる．本来のスパコン全体の電力効率を把握するためには，PUE だけでなく前節のような仕事と電力の比も考慮する必要があると考えられる．

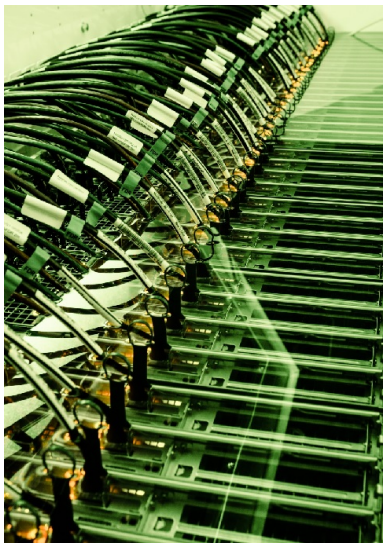

図 2 GRC Carnot Jet システムの油槽の中に計算サーバが浸された様子

図 3 TSUBAME-KFC の外観．20 フィートコンテナの隣に冷却塔が設置されている

3. 液浸冷却を導入する TSUBAME-KFC

3.1 概要および冷却方式

TSUBAME2.0/2.5(以下総称して TSUBAME2) では，省エネルギー化のための設計・運用により世界トップクラスの省エネ性を実現しているが，次世代のスパコンを更に省エネとするための研究開発を行っている．その一環として，冷却機器電力に注目する．TSUBAME2 においては 2.2 節に述べたように，PUE 1.29 と，すでに通常のデータセンターよりはるかに電力効率のよい冷却方式を実現している．しかしそれでも冷却電力だけで電気代は年間約二千万円に上り，さらなる削減が求められる．現状の TSUBAME2 の冷却方式において，消費電力を増大させている箇所は，約 9 ℃の冷媒水を生成するチラーとなる．冬期を除き，外気温より低温に冷却するには，コンプレッサーなどを必要とするためである．

そこで我々は，近年注目されている温液冷却方式に注目し，テストベッド TSUBAME-KFC の構築および評価を，NEC, NVIDIA, Green Revolution Cooling 社 (以下 GRC) などと協働で行っている．本システムの概要を図 1 に，外

図 1 TSUBAME-KFC の冷却方式の概要．計算サーバが発した熱は，冷媒油から水へ，そして自然大気中へ放出される．

観を図 3 に示す．温液冷却方式はドイツ SuperMUC[7] などでも導入されているが，SuperMUC では冷媒液パイプを計算ノード内に通しており，このためには専用計算ノード設計が必要となる．それに対し我々は，温液冷却のなかでも，デファクトスタンダードなラックマウントサーバを利用することができる，液浸冷却方式を採用する．このために，GRC の Carnot Jet システムを用いる．このシステムでは，1000 リットル強の絶縁性の冷媒液 (油) が満たされた油槽型ラックに，計算サーバをまるごと浸すことによって冷却を行う．計算サーバ群が油槽型ラックに格納されている様子を 図 2) に示す．

冷媒油の冷却のために，一度冷媒水との熱交換を行い，その冷媒水は屋外に設置された冷却塔により蒸散熱により冷却される．我々は，計算サーバ群の格納された油槽ラックと熱交換器を，屋外に設置された 20 フィートコンテナに格納した．そしてその隣に冷却塔を設置した．冷却塔においては，冷媒水を塔の上部からゆっくり降下させ，その過程で外気による冷却を行う．この方式において，冷却のために電力を消費する主な部分は，冷媒油を循環させるポンプ，冷媒水を循環させるポンプ，冷却塔に設置されたファンとなる．空冷方式において電力を大きく消費するコンプレッサーなどは存在しない．

なおこの方式においては，冷媒水を外気の露点温度よりも低温に冷却することができない．それであっても，液浸方式の採用により，年間のほぼ全ての季節においてスパコンの冷却が可能と期待されている．

3.2 計算サーバ群

TSUBAME-KFC で用いられている計算サーバは，NEC/SMC 104Re-1G 改 40 台であり，1U 筐体それぞれに下記のように 2CPU と 4GPU を搭載する非常に高密度なものである．

- Intel Xeon E5-2620 v2 (IvyBridge 世代) 6 cores 2.1GHz × 2
- NVIDIA Tesla K20X GPU × 4
- DDR3 メモリ 64GB
- SATA3 SSD 120GB
- 4x FDR InfiniBand インタフェース × 1

また主要ソフトウェア環境は以下の通りである．

- CentOS Linux (x86_64) 6.4
- GCC 4.4.7, Intel compiler 2013.1.039
- CUDA 5.5
- OpenMPI 1.7.2

40 台合計の理論性能は，倍精度で 210TFlops 程度，単精度で 630TFlops 程度となる．

40 台の計算サーバ群が 1 基の 42U 油槽型ラック (上図に見られるように横置きとなる) に格納されている．計算サーバ群の合計消費電力は事実上の最大で 40kWatt 強となる．一般的なデータセンターにおいてはラックあたり 5~15kWatt 程度，高密度実装された TSUBAME2 においてもラックあたり 30kWatt 程度となっており，KFC においてはさらに高くなっている．それであっても，Carnot Jet 自身の冷却可能容量は約 100kWatt となっており，十分対応可能となっている．

3.3 計測機器

本システムの評価のためには，細粒度な電力計測が必要であるため，下記のような計測機器を設けている (図 4)．計算サーバおよび InfiniBand スイッチごとの電力計測を行うため，それぞれについて電力センサおよび電力メータ Panasonic KW2G を設置した．40 基以上の電力メータのデータはロガー AKL1000 に集約される．これらにより，

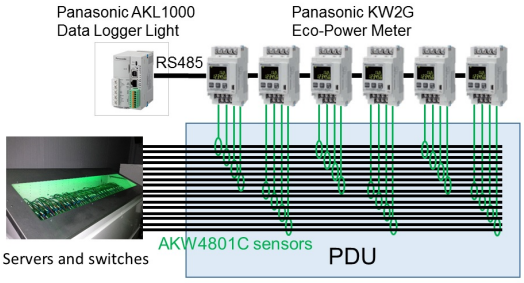

図 4 TSUBAME-KFC の電力計測機器

各計算サーバおよびスイッチの消費電力を，1 秒間隔で取得することができる．なお計算ノード内部には，各 CPU，各 GPU などについて温度センサを備えている．

その他，冷媒油ポンプ，冷媒水ポンプ，冷却塔の電力をそれぞれ取得することができる．

3.4 液浸冷却および本システム独自の対応

本システムの実現においては，先進的な冷却手法であるが故に，下記のような種々の対応を GRC および NEC と協働で行った．これらの一部は我が国独自のものである．

- 計算サーバ群を液体に浸す上で，物理的稼働部分の除去を行う必要がある．サーバ内にあるファン (12 個) が該当するため，その除去を行った．またハードディスクも該当するが，今回は物理的稼働部分を持たない SSD を採用した．
- 計算サーバ群を液体に浸す上では，プロセッサとヒートシンク間のグリスなども代替する必要がある．今回は金属シートを用いた．
- Carnot Jet システムが標準で用いる冷媒油は引火点が約 177 ℃であり，日本の消防法においては第四類危険物第三石油類に該当することが分かった．これを大量に用いることは今後の展開がより困難になると考え，代替の冷媒油として Exxon Mobil Spectrasyn8 PAO を採用した．これは引火点 260 ℃であり，危険物該当外となる．それでも，設計時および設備完成時に，消防署による確認が行われている．
- IT 機器をコンテナに格納し，屋外に設置したものはコンテナ型データセンターと見なされ，原則的に建築確認対象外となる．しかしそのためには稼働時には無人であるなどの必要条件がある (国土交通省の通知による)． その条件を満たすことについて，設計時に都庁に確認を行った．

以上のような TSUBAME-KFC システムは 2013 年 10 月に完成し，その電力性能，PUE などについては予備評価中であり，その結果を次節以降に示す．

表 1 計算サーバの温度および電力の比較．空冷および液浸冷却の場合を比較する．

冷却方式	空冷 (26 ℃)	液浸 (油温 29 ℃)	液浸 (油温 19 ℃)
温度 (℃)			
CPU1	46	42	33
CPU2	50	40	31
GPU1	52	47	42
GPU2	59	46	43
GPU3	57	40	33
GPU4	48	49	42
消費電力 (Watt)	749	693	691

4. 評価

4.1 冷却方式の計算サーバへの影響

まず，TSUBAME-KFC で採用する液浸冷却方式が，計算サーバの電力および温度にどう影響するか評価した．そのために，3.2 節に述べた 1 ノードを液浸冷却および通常の空冷で冷却した場合を比較した (表 1)．空冷の場合には，冷却ファンは装着されており，液浸の場合には除去されている．

表には，空冷 (吸気温度 26 ℃) の場合と，液浸冷却において油温を 29 ℃と 19 ℃とした場合を示す．いずれの場合も，計算サーバの 4 つの GPU 全てで倍精度行列積演算 (CUBLAS ライブラリ) を実行し続けた場合を示す．

空冷 (26 ℃) の場合と液浸 (29 ℃) のケースを比較すると，前者の気温のほうが低いにも関わらず，CPU や GPU の温度については，後者のほうが低く，冷却効率が高いことがわかる．これは空気と冷媒油のうち，後者のほうが比熱が大きく，熱を効率的に奪うためと推測される．液浸 (19 ℃) の場合はさらに温度が低くなっている．

計算サーバの電力を比較すると，空冷より液浸のほうが 7.8%程度低くなっていると分かった．この理由としては，(1) 液浸の実験においては計算サーバ内部のファンが除去されている，(2)CPU/GPU 温度の低下，が推測される．液浸 (29 ℃) と液浸 (19 ℃) の電力の差が小さいことから，上記のうち (1) が支配的と推測している．

4.2 PUE

TSUBAME-KFC の PUE の評価を行う．前節と同様に，各計算サーバの全 GPU で行列積演算を行う場合について評価し，結果を図 5 に示す．本図の”TSUBAME-KFC”については，液浸された計算サーバ，ネットワーク機器，液浸冷却機器の電力測定結果から得られた値を示す．比較対象の空冷の場合 (air cooling) については，測定の困難さのために，下記のような推定値を含む．計算サーバ電力については，1 台の電力の外挿値である．また，冷却電力に

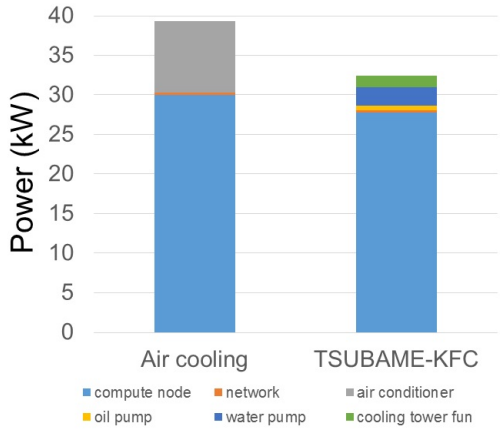

図 5 液浸冷却による TSUBAME-KFC の PUE の評価 (PUE=1.15). 比較対象の空冷 (air cooling) は，推定値を含む.

ついては，PUE を 1.3 と仮定した場合 (冷却電力=IT 機器 ×0.3) を図に示している.

本実験においては，TSUBAME-KFC の IT 機器 (計算サーバおよびネットワーク機器) の電力は約 28.1kWatt，冷却機器の電力は約 4.3kWatt であり，PUE は 1.15 と算出される．これは TSUBAME2 の 1.29(年間平均) と比較すると大幅に改善されている.

現状において，冷却電力のうち最大の電力を消費しているのは冷媒水ポンプ (2.4kWatt) である．また，計算サーバの負荷を下げても，このポンプの電力はこれ以上下がらないことが分かった．ポンプがオーバースペックであり，また低負荷時の出力制御の柔軟性に欠けると考えられ，この点の改良は今後の研究に含まれる.

4.3 Green500 指標による電力性能比

本節では，Green500 指標による Flops/Watt 値の評価を行う．2.1 節に述べたように，Green500 指標には冷却電力の影響は含まれないが，節に述べたように，サーバ内ファンの除去，チップ温度の低下により空冷の場合より有利と期待される.

評価結果を表 2 に示す．比較対象として，TSUBAME2.0(2010 年秋時点)，TSUBAME2.5 および，2013 年春の Green500 において世界一となったイタリア CINECA システムを示す．Linpack ソフトウェアとしては，TSUBAME-KFC と TSUBAME2.5 においては，NVIDIA 社から提供された GPU 利用 In-core 版を，TSUBAME2.0 においては，我々が開発した hybrid 版 [4] を用いた場合の性能である.

表のうち，”(Green)”は Green500 の規則に基づきカーネル計算時間の一部の区間を平均した値を，”(All)”はカーネル計算時間全体を平均した値を示す.

TSUBAME-KFC の Flops/Watt 値は，CINECA の記録である 3.209GFlops/Watt を大幅に更新した，4.503GFlops/Watt を達成した [*1].

5. おわりに

次世代のスーパーコンピュータのさらなる電力性能比の向上のために，ウルトラグリーン・スパコン研究設備 TSUBAME-KFC を構築し，予備評価を行った．液浸冷却および大気冷却を用いることにより，計算サーバ電力と冷却電力の双方を低減することができる．予備評価の結果, PUE においては 1.15, 電力性能比においては 4.503GFlops/Watt と，世界トップクラスの省電力性を達成することができた.

今後は，油槽ラック中の冷媒の挙動の解析や，ポンプの詳細な制御による更なる冷却効率の向上および，夏季における冷却性能の評価などを行う予定である．このような解析結果を，次世代スパコンの設計にフィードバックしていく予定である.

謝辞 本研究は文部科学省概算要求「スパコン・クラウド情報基盤におけるウルトラグリーン化技術の研究推進」の援助によります．システム構築・評価にあたって多大なご協力をいただいた，NEC, NVIDIA, Green Revolution Cooling, SUPERMICRO, Mellanox, 東京工業大学学術国際情報センターをはじめとする皆様に深く感謝します.

参考文献

[1] International Exascale Software Project: http://www.exascale.org/iesp
[2] 科学技術振興機構: 戦略的創造研究推進事業 CREST ポストペタスケール高性能計算に資するシステムソフトウェア技術の創出. http://www.postpeta.jst.go.jp
[3] The Green500 List: http://www.green500.org
[4] 遠藤 敏夫, 額田 彰, 松岡 聡: スーパーコンピュータ TSUBAME 2.0 における Linpack 性能 1 ペタフロップス超の達成． 情報処理学会論文誌コンピューティングシステム, Vol. 4, No.4 (ACS 35), pp.169–179 (2011).
[5] A. Petitet, R. C. Whaley, J. Dongarra, A. Cleary: HPL - A Portable Implementation of the High-Performance Linpack Benchmark for Distributed-Memory Computers. http://www.netlib.org/benchmark/hpl/
[6] Top500 Supercomputer Sites: http://www.top500.org
[7] Leibniz Supercomputing Centre: SuperMUC Petascale System. http://www.lrz.de/services/compute/supermuc

[*1] 本稿執筆時点では 2013 秋の Green500 順位は未発表であるが，口頭発表時点では発表に含める

表 2 Green500 指標による電量性能比の比較.

システム 時期	TSUBAME2.0 2010 秋	CINECA 2013 春	TSUBAME2.5 2013 秋	TSUBAME-KFC 2013 秋
Linpack 速度 (TFlops)	1192	98.51	2831	125.1
電力 (KWatt)				
(Green)	1244	30.70	922.5	27.78
(All)	1440	-	1125	31.18
電力性能比 (GFlops/Watt)				
(Green)	0.958	3.209	3.069	4.503
(All)	0.828	-	2.517	4.012