Vol.2017-HPC-158 No.4
2017/3/8

温度センサーを用いた「京」のジョブ消費電力推定精度向上の検討

石井 雅俊[†1,a)]　中尾 宏[†1]　中島 善康[†1]
山本 啓二[†2]　塚本 俊之[†2]　末安 史親[†3]

概要：「京」をはじめとする大規模クラスタシステムは，運用コストに占める電力料金の割合が非常に大きく，ジョブ毎の消費電力を考慮してジョブを実行することで，規定電力を超過しない運用が求められている．そのため，実行されたジョブ毎の電力情報のデータベースの構築を進めているが，「京」システムでは全てのノードに電力計が備わっていないため，簡単に作成することができない．これまで，各ノードに取り付けられた既存の温度センサー情報を用いたジョブ電力の推定について検討を行っているが，その電力推定の平均誤差が 5%程度あるためさらなる推定精度の向上が望まれている．「京」では水冷と空冷が混在した複雑な冷却システムのため，温度センサーを用いた電力の推定精度を向上させるには，実機を反映した温度-電力モデルが必要である．本検討ではシステムボード上の冷却機構や，冷却水温と吸気温度の温度依存性を考慮した温度-電力モデルおよび電力推定式を新たに構築し，その電力推定精度を検証した．その結果推定を誤差 2%まで向上できることを確認した．これにより，電力制約下でのシステムの稼働率をさらに向上させることが可能になると考えられる．

キーワード：スーパーコンピュータ，「京」，電力超過対策，電力推定，熱モデル

1. はじめに

「京」等のスーパーコンピュータや大規模クラスタシステムでは，システム全体の消費電力が数 MW を超えるものもあり，運用コストに対する電力コストの占める割合が高い．このため，システム全体の消費電力を考慮した運用が求められており，電力制限下でシステム全体の性能やエネルギー効率を最適化するための方法が提案されている[1]-[3].

一般的に計算機の消費電力は CPU やメモリアクセス等負荷に応じて消費電力が変動する．「京」では，共用開始から 1 年が経過した頃からソフトウェアの最適化が進んだ結果，システム全体の消費電力が大きく変動し，契約電力の上限を超える状況が時折発生するようになった．頻繁な契約電力の超過は電力契約の見直しにつながるため，運用コストへの影響は非常に大きい．そのため，システム全体の消費電力を適切にコントロールすることが運用上の課題となってきた[1].

この課題に対して，ジョブスケジューリングの段階で今後の電力がどのように推移するかを予測して，電力超過に備えることが検討されている[2]．これは，実行実績のあるジョブについて，ユーザ ID，グループ ID およびノード数，指定経過時間，実行時間，実行開始/終了時間，ジョブの形状，ジョブ名，ジョブスクリプト，電力等のジョブの実行に関するデータをジョブ実行実績として蓄積することで次回の実行時の電力を予測する手法である．残念ながら，「京」においては，個々のノードに電力計が設置されておらず，直接ジョブ電力を計測することはできない．そこで，既存のラック温度センサーを用いて温度変化を電力に換算することでジョブ毎の電力を推定する方法が検討されてきた[1]．しかし，その電力推定の平均誤差が最大で 5%程度あるためさらなる推定精度の向上が望まれている．「京」では水冷と空冷が混在した複雑な冷却システムのため，温度センサーを用いた電力の推定精度を向上させるには，実機を反映した温度-電力モデルが必要である．本検討ではシステムボード上の冷却機構や，冷却水温と吸気温度の温度依存性を考慮した温度-電力モデルおよび電力推定式を新たに構築し，その電力推定精度を検証した．

2. 「京」の概要

「京」は 82,944 台の計算ノードと 1.27 PiB のメモリ，11 PB のローカルファイルシステム，30 PB のグローバルファイルシステムなどから構成されている．図 1 にそのシステム構成の概要を示す．

「京」の計算ノードは，864 台のラックで構成されており，1 ラックあたり 24 枚のシステムボードが収められている．さらに 1 システムボードに 4 個の CPU が搭載されている構成となっている．システムボード上には CPU 以外にメモリと Tofu インターコネクトのコントローラーである ICC が搭載されている．計算ノードの CPU と ICC および電源パワーデバイス用の素子は水冷による冷却方法が適用されている．また，それ以外のメモリ等は空冷されている[5].

「京」の運用に必要な電力は商用電力と自家発電により供給されている．自家発電の設備には定格出力 5 MW 強のガスタービンによるコジェネレーションシステムを 2 台備え，通常運用時は 1 台ずつ交互に運転を行っている．商用

†1 株式会社富士通研究所
Fujitsu Laboratories Ltd.
†2 国立研究開発法人理化学研究所
Riken
†3 富士通株式会社
Fujitsu Limited
a) ishii.masatoshi@jp.fujitsu.com

電力と合わせ，通常運用時には約 18 MW が「京」の上限電力となる[6].

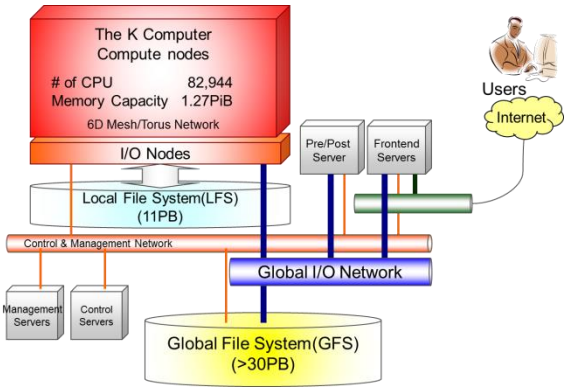


図 1 「京」のシステム構成

3. 温度センサーを用いたジョブ電力の推定

「京」には計算ノード毎の電力計が備わっていないため，ジョブ毎の電力を求めることが難しい．さらに，「京」は既に運用を開始しているため，現在の運用を大きく変えるような手段を導入することが難しい．よってジョブ毎の電力を求めるため，「京」システムの全ラックに既に設置されている温度センサー情報から電力を推定する方法が検討されてきた[1-2]．図 2 に温度センサーによるジョブ電力推定のための電力-温度熱回路モデルを示す．熱回路では熱源Qは電流源として表され，消費電力が全て熱に変換されると仮定すると熱源Qは消費電力Pに等しい．吸気と排気の温度をそれぞれT_{in}，T_{out}とし，システムボード内部の熱抵抗をRとする．「京」では，水冷部の冷却水量や，空冷部の風量は変動しないため，熱抵抗Rは一定値であるとすると，消費電力は(1)式のように推定することができる．

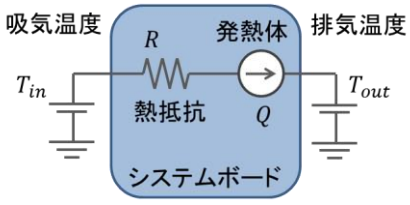


図 2 電力-温度熱回路モデル

$$P = Q = \frac{\Delta T}{R} = \frac{T_{out} - T_{in}}{R} \quad (1)$$

「京」では，ジョブの実行時の消費電力の大部分は CPU とメモリ，Tofu インターコネクトのコントローラーである ICC によって消費される．ICC の消費電力は一定で，ジョブによって消費電力が変動をするコンポーネントは CPU とメモリのみとなることが報告されている[1]．したがって，計算ノードの電力を推定するには，CPU，メモリと固定電力が分かればよい．

「京」では，全(864)計算ラックに対して 5 分毎にラック吸気温度，システムボードの排気温度，水冷入力温度，各 CPU 温度の情報を採取している．CPU は水冷されているため，CPU 温度と水冷入力温度との差をT_{cpu}とし，メモリは空冷されているため，システムボード排気温度とラック吸気温度との差T_{air}の情報から電力を以下の推定式で推定する方法が検討されている．

$$P = a \cdot T_{cpu} + b \cdot T_{air} + c \quad (2)$$

計算ラック 864 台のうちラックの約 3 割には電力計が取り付けられており，このラック電力値と温度センサー情報から係数a，b，cを求め，「京」で実行されたジョブ単位での推定値と測定値を比較した結果が報告され，その結果最も精度の高いジョブでは二乗平均平方根(RMS)誤差が 0.68%と非常に高精度で推定が可能であることが示される一方，ジョブよっては誤差が 5.45%と大きく，さらなる推定精度の向上が望まれている．この誤差がジョブによって異なることの要因として，(2)式に反映されていない要素があることが推測される．「京」では水冷と空冷が混在した複雑な冷却機構であることから，より実機を反映した熱モデルを構築することで電力推定精度の向上が期待される．

4. 実機を反映した熱モデルの構築

各温度センサーの温度変化は，CPU やメモリ等のコンポーネントでの電力消費により発生した熱量が熱伝導されることで発生する．計算ラック当たりの温度センサーは，ラック吸気温度，水冷入力温度が各 1 個，システムボードの排気温度が 24 個(システムボードに 1 個)，CPU 温度が 96 個(各 CPU に 1 個)実装されている．これらの温度センサーを用いて実機を反映した詳細な電力-温度熱モデルを構築するためには，CPU 毎やシステムボード毎のメモリ電力の測定が必要となる．しかし現状の「京」では電力を測定する手段がシステム監視用の全ラックの合計電力と一部のラックに取り付けられたラック電力計のみで，電力-温度熱モデルの構築に制限があった．そこで，各 CPU 電力とシステムボードごとのメモリ電力の各コンポーネント電力の測定系を構築した．この電力情報と温度センサー情報から実機を反映した詳細な熱モデルを構築した．

システムボードの供給される冷却水温度は，通常稼働時は 15 °C であるが，負荷変動等により 15 °C から 18 °C の範囲で変動している．また，ラック吸気温度はラックの配置位置により 18 °C から 27 °C の範囲でバラツキがあることが分かっている．これらの温度の変動により熱抵抗等のモデルのパラメータが変化することが考えられるためこれを考慮した電力-温度熱モデルのパラメータの温度依存性について検討した．

4.1 各コンポーネント電力の測定

上述のように「京」は既に運用を開始しているシステムであり，システムボード上に新たに電力計を設置することは難しい．「京」のシステムボード上には各 CPU，システムボード上のメモリ，ICC への電力供給に Point of Load(POL)電源が使われており，この POL 電源の出力値を読み出すことができるツールを導入した．本ツールは 10 分間の平均電力を取得することができる．このツールはマネージメント用 CPU を経由して取得しているため，「京」システム全体の電力を取得しようとすると，保守用のネットワーク負荷が過負荷となるため，あくまで限定したラック毎の電力しか測定することができない．このため，電力-温度熱モデルの構築のみに使うこととする．

「京」の計算ラックは 3 相 200 V で受電して，電源ユニット(PSU)で 48 V に変換し，中間バスコンバータ(IBC)で 12 V に変換し，POL 電源に供給する電源構成となっている[5]．ジョブ電力を把握するためには，PSU, IBC の変換損失を含む電力値が必要であるが，ツールで得られた電力は POL 電源の出力値であるため，上記の電力変換損失は含まれていない．そのため，一部の計算ラックに取り付けられた電力計の測定値から変換損失を求める必要がある．

CPU とメモリの POL 電力測定値をそれぞれP'_{cpu}，P'_{mem}とし，それぞれの変換効率を考慮した補正係数をα_{cpu}，α_{mem}とし，さらに電力変動しない固定電力をP_{const}とすると，ラック電力P_{rack}は(3)式で表すことができる．PSU, IBC の変換損失も含む CPU 電力P_{cpu}，メモリ電力P_{mem}はそれぞれ(4), (5)式から求めることができる．

$$P_{rack} = \alpha_{cpu}P'_{cpu} + \alpha_{mem}P'_{mem} + P_{const} \quad (3)$$

$$P_{cpu} = \alpha_{cpu}P'_{cpu} \quad (4)$$

$$P_{mem} = \alpha_{mem}P'_{mem} \quad (5)$$

CPU 補正係数α_{cpu}を求めるため，1 ラック内の全ての CPU を同時に 4 段階にステップ的に負荷を変化させて，その時の POL 電力とラック電力を測定した．

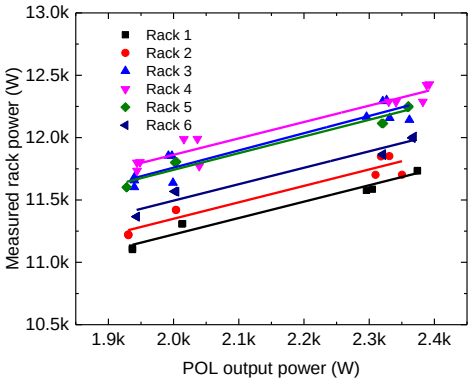


図 3 CPU の POL 電力とラック電力の関係

図 3 に CPU の POL 電力とラック電力の関係を示す．ここで，POL 電力は 1 ラックに搭載されている全 96CPU の合計電力である．この結果から POL 電力とラック電力はこの電力変動範囲において線形の関係を示しており，その傾きから CPU の補正係数α_{cpu}は 1.331 と求めた．

同様にメモリ補正係数α_{mem}を求めるため 1 ラック内の全てのメモリを同時に 4 段階にステップ的に負荷を変化させてその時の POL 電力とラック電力を測定した結果を図 4 に示す．メモリの負荷変動においては CPU の電力変動が見られたため，縦軸はラック電力計の測定値から式(4)から求めた CPU 電力を引いた電力を示している．また，POL 電力は 24 システムボードの全メモリ電力の合計値である．その結果，メモリにおいても CPU 同様にこの電力変化の範囲においては線形の関係を示しており，メモリの補正係数α_{mem}は 1.174 であること分かった．

この方法により，CPU 毎の電力，システムボード毎のメモリ電力を電源の変換効率も含めて正確に測定することが可能となった．

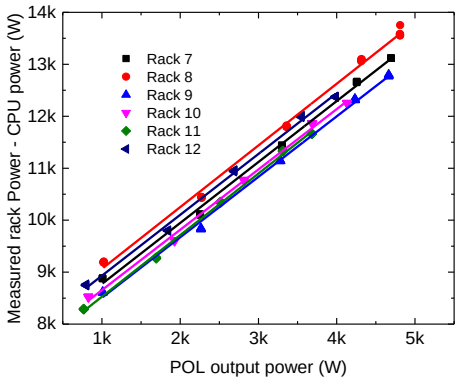


図 4 メモリの POL 電力とラック電力の関係

4.2 実機を反映した熱モデルの検討

上記より測定した CPU 毎，システムボード毎のメモリ電力値から，より詳細な電力温度モデルについて検討した．

4.2.1 CPU

1 計算ラック中の全 96CPU の各電力P_{cpu}と CPU 温度T_{cpu}と冷水入力温度T_{water}との差ΔT_{cpu}の関係を図 5 に示す．この結果，1 システムボードには CPU0~3 の 4 個の CPU が搭載されているが，CPU1,2 は CPU0,3 に比べ同じ電力値でも温度上昇が高いことが分かった．この CPU による温度の違いは，CPU の冷却機構によるものと推測される．図 6 にシステムボードに搭載された水冷ユニット概略を示す．システムボード上に実装された 4 個の CPU は水冷方式で冷却されており，システムボード内に供給された冷却水は 2 方に分岐され，それぞれ 2 個ずつ CPU を順番に冷却する構造となっている．このため，下流にあたる CPU1,2 の冷却水は上流の CPU0,3 の発熱により温度が上昇していると考えられる．よって，CPU1,2 と CPU0,3 は異なる熱モデルの構築

が必要であると考えられる．

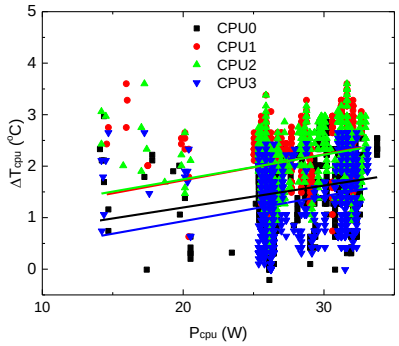


図 5 各 CPU 電力と CPU 温度と水冷入力温度との温度差の関係

この冷却順序を考慮した熱モデルを図 7 に示す．図 7(a)は，上流側の CPU0，3 の熱モデルを示しており，CPU0 で発生した熱P_{cpu0}は熱抵抗a_{cpu0}を通り CPU 温度センサーT_{cpu0}に到達するモデルで，(6)式のように表すことができる．よって，CPU0 電力P_{cpu0}は，CPU 温度T_{cpu0}と水冷入力温度T_{water}から(7)式により求めることができる．

$$T_{cpu0} - T_{water} = \Delta T_{cpu0} = a_{cpu0} P_{cpu0} \quad (6)$$

$$P_{cpu0} = \frac{\Delta T_{cpu0}}{a_{cpu0}} = \frac{T_{cpu0} - T_{water}}{a_{cpu0}} \quad (7)$$

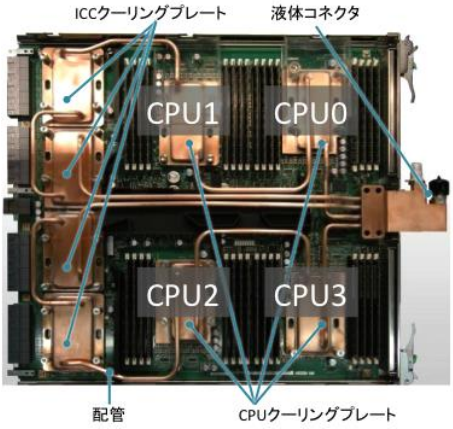
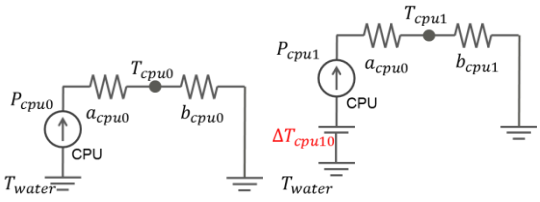


図 6 「京」の水冷ユニット概略



(a) CPU 0,3 のモデル　　(b) CPU 1,2 モデル

図 7 CPU 電力-温度熱回路モデル

一方 CPU1，2 では，上流の CPU0,3 による冷却水の温度上昇ΔT_{cpu10}を考慮し，その熱回路モデルを図 7(b)に示す．ここで，ΔT_{cpu10}は，上流の CPU0 電力P_{cpu0}による冷却水温の温度上昇であるため，ΔT_{cpu10}は CPU0 電力P_{cpu0}に比例すると考えられ，その比例係数をa_{cpu10}とする．図 5 の CPU1 電力P_{cpu1}と CPU1 温度T_{cpu1}と水冷入力温度T_{water}の差ΔT_{cpu1}との関係を(8)式で表す．ここで観測されるΔT_{cpu1}とP_{cpu1}の傾きをa_{cpu1}と定義する．CPU1,2 と CPU0,3 のクーリングプレートの熱的な構造が同じであるため，熱源から温度センサーまでの熱抵抗は CPU0 と同じa_{cpu0}とすると，a_{cpu1}とa_{cpu0}の差が CPU0 による温度上昇によるものであると考えられる．よってa_{cpu10}は(10)式のように求められる．

$$\Delta T_{cpu10} = a_{cpu10} P_{cpu0} \quad (8)$$

$$T_{cpu1} - T_{water} = \Delta T_{cpu1} = a_{cpu1} P_{cpu1} \quad (9)$$

$$a_{cpu10} = a_{cpu1} - a_{cpu0} \quad (10)$$

以上から図 7(b)の CPU1,2 の電力-温度熱回路モデルは(11)式のように表すことができる．CPU1 電力P_{cpu1}は，CPU0, CPU1 温度T_{cpu0}, T_{cpu1}と水冷入力T_{water}から(12)式により求めることができる．

$$\begin{aligned} T_{cpu1} - T_{water} &= \Delta T_{cpu1} = a_{cpu0} P_{cpu1} + \Delta T_{cpu10} \\ &= a_{cpu0} P_{cpu1} + a_{cpu10} P_{cpu0} \\ &= a_{cpu0} P_{cpu1} + a_{cpu10} \frac{\Delta T_{cpu0}}{a_{cpu0}} \end{aligned} \quad (11)$$

$$\begin{aligned} P_{cpu1} &= \frac{\Delta T_{cpu1}}{a_{cpu0}} - a_{cpu10} \frac{\Delta T_{cpu0}}{{a_{cpu0}}^2} \\ &= \frac{T_{cpu1} - T_{water}}{a_{cpu0}} - a_{cpu10} \frac{T_{cpu0} - T_{water}}{{a_{cpu0}}^2} \end{aligned} \quad (12)$$

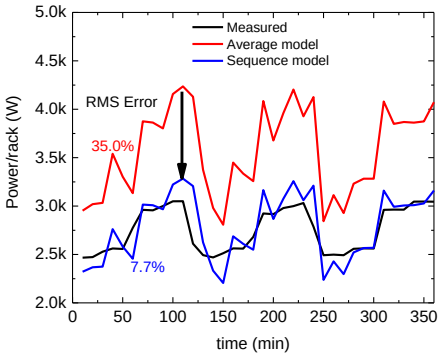


図 8 新たなモデルによる電力推定精度の検証結果

このモデルパラメータa_{cpu0}，a_{cpu1}を図 5 の CPU 電力と CPU 温度と水冷入力温度との差温度の関係から求めた．このモデルの効果を検証するため，全ての CPU のモデルパラメータが同じ図 7(a)の熱回路モデルで電力を推定した場合と本熱回路モデルで電力を推定した場合および実際の CPU 電力測定値と比較した結果を図 8 に示す．全て同じ CPU パラメータの場合には CPU が高負荷になったときに冷却の下流の CPU 温度が上流の CPU 発熱により上昇するため，電力が大きく見積もられているが，本モデルを導入

することで，その影響が低減され，CPU 電力の推定精度の RMS 誤差が 35%から 8%まで低減できることが分かった．

4.2.2 メモリ

1 計算ラック中の各メモリ電力P_{mem}とラック吸気温度T_{inlet}とシステムボードの排気温度T_{outlet}との差ΔT_{mem}の関係を図 9 に示す．

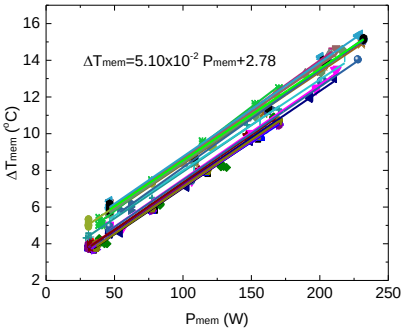


図 9 メモリ電力と吸排気温度差の関係

図 9 からメモリ電力はラック吸気温度T_{inlet}，システムボードの排気温度T_{outlet}との温度差ΔT_{mem}に比例することが分かる．また，メモリ電力が 0 の場合でも温度差ΔT_{mem}が発生しており，これは，メモリ以外の熱源がシステムボード上にあり,その温度上昇によるものであると考えられる．よって，メモリ電力P_{mem}による吸排気温度差ΔT_{mem}は(13)式のように定義する．ここから，メモリ電力P_{mem}はラック吸気温度T_{inlet}とシステムボードの排気温度T_{outlet}から(14)式で求めることができる．

$$T_{outlet} - T_{inlet} = \Delta T_{mem} = a_{mem} P_{mem} + b_{mem} \quad (13)$$

$$P_{mem} = \frac{\Delta T_{mem} - b_{mem}}{a_{mem}} = \frac{T_{outlet} - T_{inlet} - b_{mem}}{a_{mem}} \quad (14)$$

4.3 熱モデルパラメータの温度依存性調査

4.3.1 CPU

システムボードに供給される水冷入力温度T_{water}は，通常稼働時には 15 °C であるが，負荷変動等により 15 °C から 18 °C の範囲で変動している．CPU は水冷されているため空冷に比べると熱容量が高く，外気温の影響を受けにくいと考えられる．そのため，水温のみを変化させてその温度依存性について調査した．冷却水温度の変化は施設側の供給水温を変化させることにより行った．図 10 に(6), (9)式で定義されるa_{cpu0}とa_{cpu1}に関して，水冷入力温度T_{water}を 15 °C から 18 °C に変化させた時の温度依存性について測定した結果を示す．図 10 から水冷入力温度T_{water}を 15 °C から 18 °C まで変化させた時のa_{cpu0}とa_{cpu1}のT_{water}による変化は約 1%であった．本解析結果から CPU 熱モデルパラメータの温度依存性を(15)式で定義する．

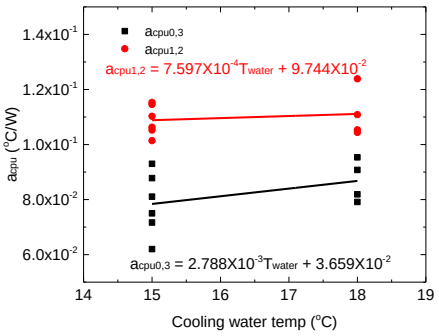
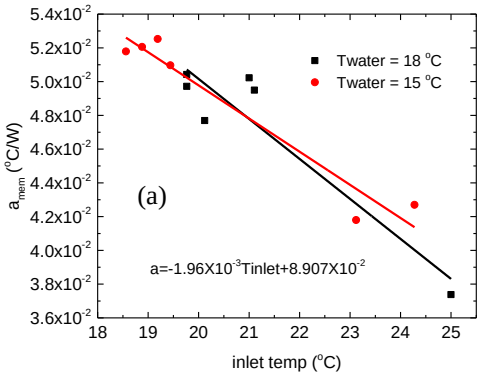


図 10 CPU 電力-温度熱モデルパラメータの温度依存性

$$a_{cpu1} = 7.597 \times 10^{-4} T_{water} + 9.744 \times 10^{-2}$$
$$a_{cpu0} = 2.788 \times 10^{-3} T_{water} + 3.659 \times 10^{-2} \quad (15)$$

4.3.2 メモリ

ラックの配置位置により，ラック吸気温度は 18 °C から 27 °C の範囲でバラツキがある．この温度の変動により熱抵抗等の電力-温度熱モデルのパラメータが変化することが考えられるため，その温度依存性について調査した．(12)式で定義されるメモリ熱モデルのモデルパラメータa_{mem}，b_{mem}について，ラック吸気温度T_{inlet}の異なるラックで水冷入力温度T_{water}を変化させた時の変化について測定した．その結果を図 11(a), (b)に示す．図 11(a)からメモリ電力P_{mem}と吸排気温度差ΔT_{mem}特性の傾きa_{mem}は，ラック吸気温度T_{inlet}が 18 °C から 25 °C のラックではa_{mem}は 5.2×10^{-2}から 3.8×10^{-2} まで変化し，その傾きは 1.96×10^{-3} であった．水冷入力温度T_{water}の依存性は見られない．また，b_{mem}もラック吸気温度T_{inlet}により変化する水冷入力温度T_{water}が 15 °C の時では 3~-3 °C まで変化し，その傾きは 8.48×10^{-1} であった．さらに水冷入力温度T_{water}の変化によりラック吸気温度T_{inlet}とb_{mem}特性の傾きはほぼ変わらずに水冷入力温度T_{water}の変化量と同じ 3 °C ほど高くなることが分かる．ここから，システムボード内部では，システムボードの排気温度T_{outlet}はシステムボード内を流れる冷却水との熱交換により冷却されており，ラック吸気温度T_{inlet}と水冷入力温度T_{water}による温度依存性はその相互作用によるものであると考えられる．実測データからこれらの(13)式のメモリ推定式の温度依存性を(16), (17)式で定義する．今後は，この温度依存性モデルの妥当性を評価するため，水冷と空冷の相互作用についての詳細な物理モデルの構築が必要である．



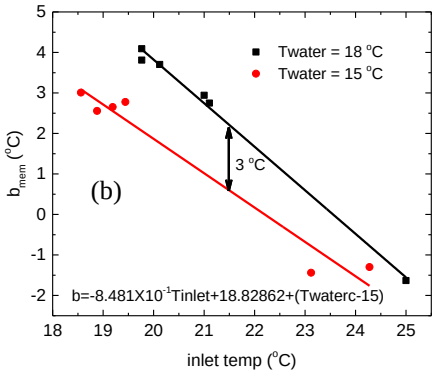


図 11 メモリ電力-温度熱モデルパラメータの温度依存性

$$P_{mem} = \frac{\Delta T_{mem} - b_{mem}}{a_{mem}} = \frac{T_{outlet} - T_{inlet} - b_{mem} + (T_{water} - 15)}{a_{mem}} \quad (15)$$

$$a_{mem} = -1.96 \times 10^{-3} T_{inlet} + 8.91 \times 10^{-2}$$
$$b_{mem} = -8.48 \times 10^{-1} T_{inlet} + 18.86 + (T_{water} - 15) \quad (16)$$

4.3.3 固定電力

固定電力の吸気温度や冷却水温度の依存性について測定した. 固定電力の算出にはラック電力計の測定値から(4), (5)式から求めた CPU とメモリ電力を差し引いた電力で評価した. 図 12 にその結果を示す. 固定電力のラック吸気温度T_{inlet}，水冷入力温度T_{water}の温度依存性は見られなかった. 固定電力P_{fixed}は測定電力の平均値の 7.70 kW とし，ラックごとのバラツキは最大で 608 W であった.

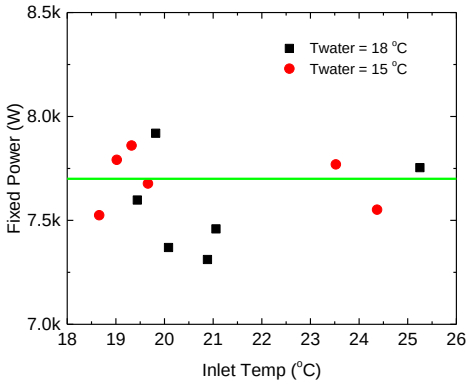


図 12 固定電力の温度依存性

$$P_{fixed} = const. = 7700\ kW \quad (17)$$

5. 推定動作の検証

これまで検討した電力-温度熱モデルから導いた推定式について，7 ラックに関して「京」の 2016 年 9 月 4 日から 10 日までの 1 週間の実運用中の電力について，温度センサーによるラック電力推定値とラック電力計の測定値との比較により電力推定精度の検証を行った．温度センサーによるラック推定電力P_{rack}は, CPU とメモリのラック合計電力P_{rack_cpu}，P_{rack_mem}およびラックの固定電力P_{fixed}の合計で(18)式で示す．CPU のラック合計電力P_{rack_cpu}は全 96 個の CPU 電力の合計値で(19)式より求める. 各 CPU 電力は(7), (12), (14)式から求めた．メモリのラック合計電力P_{rack_mem}は全 24 システムボードの合計値で(20)式から求めた．各ボードのメモリ電力は式(15), (16)から求めた．温度センサーによる電力推定値とラック電力測定値と比較した結果を図 12 に示す．ここで，ラック電力の測定値は温度の取得間隔と同様に 5 分間の平均値とした．また，ラックごとの推定誤差の二乗平均平方根(RMS)を表 1 に示す.

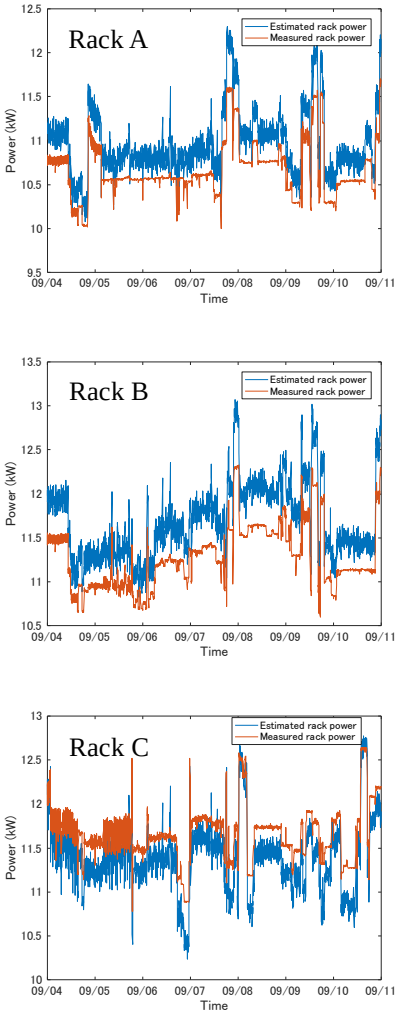
$$P_{rack} = P_{rack\_cpu} + P_{rack\_mem} + P_{fixed} \quad (18)$$

$$P_{rack\_cpu} = \sum_{j=0}^{23} \sum_{i=0}^{3} P_{cpu\ ij} \quad (19)$$

$$P_{rack\_mem} = \sum_{j=0}^{23} P_{mem\ j} \quad (20)$$

表 1 から RMS 誤差はラックにより異なり，168.0 ~ 632.6 W であった．ラック電力の平均値と比較すると推定誤差は 1.54 ~ 5.47%であった．この推定誤差の要因を考える上で，図 13 から推定値と測定値のオフセットに着目した．図 13 から，推定値と測定値にはオフセットが見られ，その差はラック F を除く 6 ラックではほぼ一定であることが分かる. これを確認するため，推定誤差の度数分布をプロットした

結果を図 14 に示す．その結果，ラック F を除く 6 ラックにおいて，ラックごとの中心値のバラツキは見られるが，分散カーブはほぼ正規分布を示しており，その半値幅はほぼ同じであることがわかる．ラックごとの標準偏差を計算した結果を表 1 に示す．ここから，ラック F を除き推定誤差の標準偏差は 113.0W から 180.4W であり，RMS 誤差の主要因はラックごとのオフセットのバラツキによるものであると考えられる．この原因は、固定電力の算出の際，ラックにより約 680 W バラツキがあったため，このバラツキによるものであると考えられる．電力を考慮したジョブスケジューリングに適用するには，ジョブによって変動する電力の推定が重要である．固定電力はラック間でバラツキがあっても合計電力が分かればよいため，今後は解析ラック数を増やしてその固定電力の全体の平均値を求める必要がある．



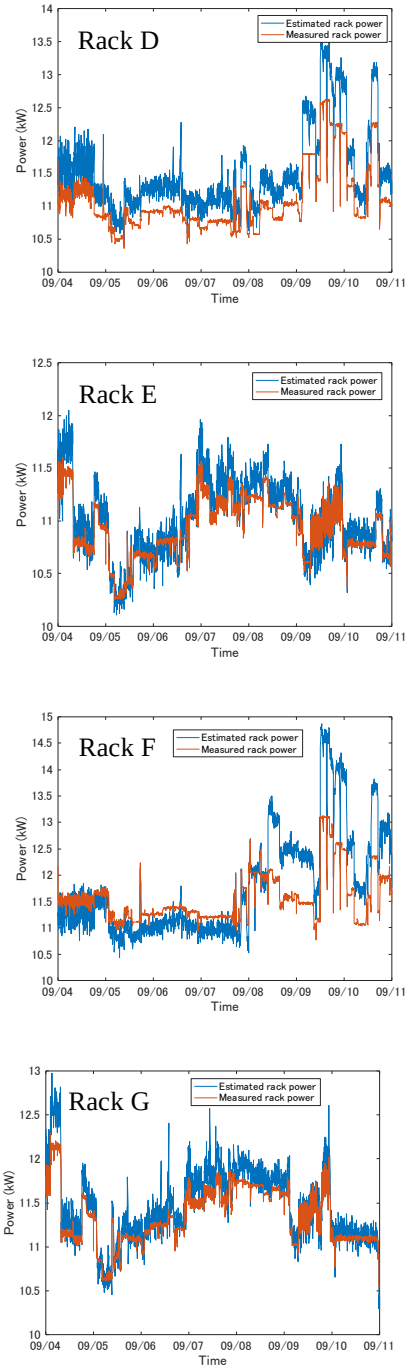


図 13 ラックごとの消費電力の推定値と測定値の比較

また，ラック F の推定誤差が大きくなる原因について調査した結果, ラック吸気温度が 9 月 8 日 10 時を境に 23.3 ℃から 24.7 ℃ と約 1.5 ℃ 上昇していることが分かった. このため，今回構築した温度依存性のモデルに反映されていない要素があると考えられる．今後この影響を考慮したモデルを導入することでさらなる推定誤差の向上が可能であると考えられる．

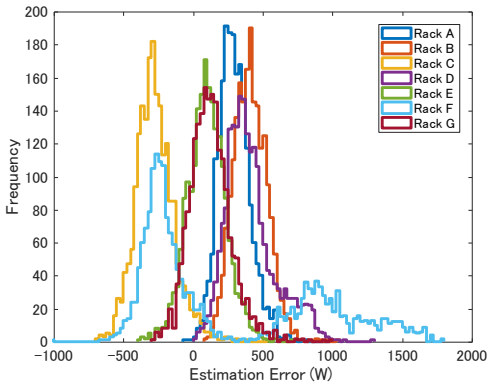


図 14 推定誤差の度数分布

現状のモデルで今回解析した全 7 ラックの負荷変動による電力変化のみに着目すると平均で 1.85%の精度で電力の推定が可能であることが分かった．本手法による電力予測精度の向上は，電力制約下においてジョブスケジューリングする際の電力マージンの削減につながり，その結果，システム稼働率をさらに向上させることが可能になると考えられる．

表 1 電力推定値と測定値のラックごとの誤差

ラック名	二乗平均平方根誤差		誤差の標準偏差	
	(W)	(%)	(W)	(%)
ラック A	317.1	2.97	113.0	1.06
ラック B	426.3	3.78	118.2	1.05
ラック C	308.1	2.64	139.2	1.19
ラック D	448.9	4.05	180.4	1.63
ラック E	168.0	1.54	137.5	1.26
ラック F	632.6	5.47	631.6	5.46
ラック G	193.5	1.71	148.8	1.31
平均	356.4	3.17	209.8	1.85

6. まとめ

既存のラック温度センサーを用いてジョブ毎の電力推定の推定精度の向上について検討した．「京」では水冷と空冷が混在した複雑な冷却システムのため，温度センサーを用いた電力の推定精度を向上させるには，実機を反映した温度-電力モデルが必要であり，CPU の冷却機構や，冷却水温と吸気温度の温度依存性を考慮した温度-電力モデルおよび電力推定式を新たに構築し，その電力推定精度を検証した．その結果，7 ラックの電力推定誤差は二乗平均平方根誤差が 1.54~5.47%であることが分かった．最も精度が悪い 5.74%の推定誤差のラックはラックの吸気温度の変化により推定精度が悪化しており，今後この影響を考慮したモデルの構築が必要であると考えられる．また，ラック電力推定誤差の度数分布を解析した結果，大きなラック電力の推定誤差の発生要因として固定電力のバラツキによるものであることが分かった．固定電力のバラツキを除く変動電力は，1.85%の推定精度で推定できることが分かった．

電力予測精度が向上することで，電力制約下でジョブスケジューリングする際の電力マージンを削減することができ，その結果，システム稼働率をさらに向上させることが可能になると考えられる．また，「京」の複雑な冷却系においても電力と温度の関係を高精度でモデル化することが可能であることから，大規模クラスタシステムおよびデーターセンターの電力と温度の関係をモデル化することにより，冷却系のモデル予測制御等によるシステム全体電力の最適化への応用が期待される．

参考文献

[1] 宇野 篤也, 肥田 元, 井上 文雄, 池田 直樹, 塚本 俊之, 末安 史親, 松下 聡, 庄司 文由, “消費電力を考慮した「京」の運用方法の検討”, 情報処理学会論文誌コンピューティングシステム, 2015, vol. 8, no. 4, p. 13-25.
[2] 山本 啓二, 末安 史親, 宇野 篤也, 塚本 俊之, 肥田 元, 池田 直樹,庄司 文由, “過去の実行実績を利用したジョブの消費電力予測”, 情報処理学会研究報告ハイパフォーマンスコンピューティング,2015, vol. 2015-HPC-151, no. 2, p. 1-7.
[3] 黄 巍, 岩澤 直弘, カオ タン, 和 遠, 近藤 正章, 中村 宏, “エネルギー効率を考慮した電力制約下でのスループット指向ジョブスケジューリング”, 情報処理学会研究報告ハイパフォーマンスコンピューティングと計算科学シンポジウム HPCS2015, 2015, HPCS2015, p. 150-158.
[4] 宮崎博行, 草野義博, 新庄直樹, 庄司文由, 横川三津夫, 渡邊貞, “スーパーコンピュータ「京」の概要”, 雑誌 FUJITSU, 2012, vol. 63, no. 3, p. 237-246.
[5] 前田秀樹, 久保秀雄, 島森 浩, 田村 亮, 魏 杰, “スーパーコンピュータ「京」のシステム実装技術”, 雑誌 FUJITSU, 2012, vol. 63, no. 3, p. 265-272.
[6] 井上 文雄, 宇野 篤也, 塚本 俊之, 末安 史親, 池田 直樹, 肥田 元, 庄司 文由, “電力を考慮した「京」の運用改善への取組み”, 情報処理学会研究報告ハイパフォーマンスコンピューティング,2016, vol. 2016-HPC-153, no. 36, p. 1-5.