Vol.2014-ARC-213 No.20
Vol.2014-HPC-147 No.20
2014/12/10

「京」におけるジョブ単位の消費電力推定の検討

宇野 篤也[1,a)]　肥田 元[2]　池田 直樹[2]　井上 文雄[1]　塚本 俊之[1]　末安 史親[3]　庄司 文由[1]

概要：近年，東日本大震災後の電力事情の悪化等から，システムの消費電力を考慮した運用を行う必要性が増してきている．最近の HPC アーキテクチャは CPU やメモリアクセスの負荷に応じて電力が大きく変動する傾向にあり，消費電力が実行されるジョブの効率等に依存することが知られている．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ためには，実行されるジョブの消費電力を予測することが非常に重要となる．しかし，「京」のような大規模システムでは大小さまざまなジョブが常に同時に実行されているため，システム全体の消費電力の変動から個々のジョブの消費電力を把握することは難しい．本研究では「京」の計算ラックに設置された温度センサ等の情報を元に，ジョブ毎の消費電力を推定できないか検討を行った．本稿ではその手法と結果について報告する．

1. はじめに

スーパーコンピュータ「京」は，理化学研究所と富士通株式会社が共同開発した汎用並列スーパーコンピュータである．運用は理化学研究所 計算科学研究機構（AICS）が行っており，2012 年 9 月から共用を開始している．「京」は低消費電力 CPU の採用など消費電力を抑えるように設計されているが，その消費電力は無負荷時で約 10MW，高負荷時には 14MW を超える場合もあり，電力設備への負荷や電力契約上の制約等の観点から，システムの全体の消費電力を考慮した運用が求められている．

「京」に限らず，近年のアーキテクチャは CPU やメモリアクセスの負荷に応じて電力が大きく変動する傾向にあり，システムの消費電力は実行されるジョブの効率等に依存することが知られている．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ためには，実行される個々のジョブの特性と消費電力を把握し，システム全体の消費電力を予測することが重要となる．しかし，「京」のような大規模システムでは大小様々なジョブが常に同時に実行されており，計算ノード毎に電力計が設置されていない限り個々のジョブの消費電力を把握することは難しい．実行されたジョブのプロファイル情報から消費電力を推定する方法もあるが，プロファイル情報はジョブの実行後に取得できる場合がほとんどで，システムの消費電力をリアルタイムに管理する場合には，この方法は使えない．

1　独立行政法人理化学研究所 計算科学研究機構
2　株式会社富士通ソーシアルサイエンスラボラトリ
3　富士通株式会社
a)　uno@riken.jp

「京」の場合，一部の計算ラックを除いてラック単位で電力を測定することはできないが，各計算ラックにはいくつかの温度センサが取り付けられている．そこで我々は，これら温度センサの情報とシステム全体の消費電力の情報を組み合わせることにより，個々のジョブの消費電力を推定することができないか検討を行った．

本稿では，温度センサの情報とシステム全体の消費電力から個々のジョブの消費電力を推定する方法と，その推定結果について報告する．

2. 「京」の概要

「京」は，82,944 台の計算ノードと 1.27PiB のメモリ，11PB のローカルファイルシステム (LFS) と 30PB のグローバルファイルシステム (GFS) から構成される汎用並列スーパーコンピュータである．**図 1** に京のシステム構成概要を示す [1].

「京」の消費電力は無負荷時で約 10MW，高負荷時で

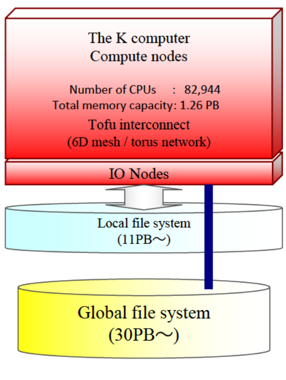

図 1　「京」のシステム構成

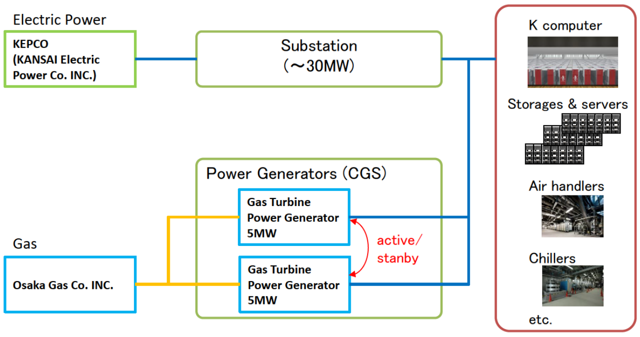

図 2 AICS の電源設備

14MW を超える．「京」の運用に必要な電力は，商用電力（関西電力）と自家発電により供給されている．**図 2** に AICS の電源設備を示す．自家発電設備として，ガスタービンによるコジェネレーションシステム（CGS）を 2 台備えている．CGS 1 台の定格出力は約 5MW で，通常運用時は 1 台づつ交互に運転を行い，商用電力からの電力と合わせて全体の消費電力を賄っている．

2.1 「京」の運用

「京」の運用形態は，36,864 ノード以下の規模のジョブの実行が可能な通常運用と，36,865 ノード以上の規模のジョブを実行できる大規模ジョブ実行運用の 2 つに大きく分けることができる．毎月第二火曜から 3 日間を大規模ジョブ実行期間として設定している [2][3].

前述のとおり，「京」の運用で消費される電力は，CGS と商用電力で賄われていて，電力会社とは供給電力の上限を決めて契約を行っている．この上限を超え，電力超過が発生した場合[*1]，電力会社に対して違約金の支払いが発生する．この電力超過が頻繁に発生するようであれば，供給元より契約電力自体の見直しを迫られることになり，結果として運用経費の増大につながる．そのため，電力超過を防ぐことが重要となる [4].

運用面から電力超過を防ぐ手段として簡単かつ効果的な方法は，電力超過が発生すると判断された時点で実行中のジョブを強制的に停止することである．大規模ジョブ実行期間内では，ほぼ 1 ジョブ単位でジョブが実行されるめ，電力超過が発生した場合には速やかに該当ジョブを停止することができる．一方，通常運用期間では大小様々なジョブが同時に多数実行されている．そのため，電力超過が発生した場合に超過の原因となったジョブを特定することは難しく，そのままでは手当たり次第にジョブを停止するしかない．

そこで，通常運用時においても，電力超過が発生した場合に適切にジョブを停止できるよう，ジョブ単位での消費電力の推定方法を検討した．

*1 毎時ごとの 0～30 分または，30～60 分の 30 分間における平均使用電力が契約電力を超えた場合

表 1 温度センサの仕様

計測可能範囲	0～100 度
計測誤差	±約 2 度
計測解像度	0.5 度 (補正前) (測定範囲によっては結果補正あり)

3. ジョブ単位の消費電力の推定

ジョブ単位での消費電力の推定方法として，以下の方法を検討した．

- ジョブが使用するノード数を用いた推定
- ラックに取り付けられた温度センサを用いた推定

3.1 使用ノード数による消費電力の推定

「京」で実行されている個々のジョブが使用しているノード数とシステム全体の消費電力から，ノード単位の平均消費電力を求めることができる．この場合，ジョブ単位の推定消費電力はノード数に単純に比例するので，電力超過時には削減すべき電力量をもとにジョブを順次停止すればよい．しかし，実際にはジョブ毎の消費電力は異なっているため，ジョブを停止しても予測した電力を削減できるとは限らない．また，規模の大きなジョブが複数同時に実行されているような場合には，どのジョブを停止すればよいか判断することは難しい．そのため，単純にノード数から消費電力を推定してジョブを停止するのは効率のよい方法とは言えない．特に大規模のジョブを途中で停止する場合，中断により失われる計算資源量は無視できない量となる．効率の面からいえば，電力超過分だけ消費電力を減らせるようにジョブを停止するのが理想的である．このように，停止すべきジョブを適切に選択することは非常に重要である．

3.2 温度センサ情報を利用した消費電力の推定

「京」の場合，ジョブの実行時の電力の大部分は，CPU とメモリ，Tofu インターコネクトのコントローラである ICC によって消費される．そこで，ジョブ実行時の各種温度変化から消費電力を推定する方法について検討を行った．ファイル I/O 時の消費電力の変動についても調査を行ったが，計算ノードおよびディスクラックの消費電力にはほとんど変化がみられなかった．

「京」の計算ラックにはいくつかの温度センサが搭載されている．今回の消費電力の推定に使用した温度センサの情報は，ラック吸気温度，System Board(SB) 排気温度，冷却水入力温度，CPU 温度である．**表 1** にこれらの温度センサの仕様を示す．これらの温度データは，現状では 10 分毎に取得することができる．ここでは，ジョブの実行による CPU 温度変化と SB 排気温度変化を以下のように定義した．

- CPU 温度変化 = CPU 温度 - 冷却水入力温度
- SB 排気温度変化 = SB 排気温度 - ラック吸気温度

「京」の一部の計算ラックに搭載された電力センサを使用し，これらの温度変化とジョブの消費電力の関係について調査を行った．

3.2.1 CPU

CPU 温度変化と消費電力の関係について調査した．

図 3 に CPU 部分の冷却機構を示す．本来ならば，CPU 温度と冷却水の上昇温度から発熱量を算出し消費電力を推定するのが望ましいが，「京」の計算ラックには冷却水の排出温度を測定するセンサは取り付けられていない．そこで，CPU 温度と冷却水入力温度の差 (CPU 温度変化) を利用することにした．CPU で生じた熱は冷却水により冷やされ，冷やしきれなかった熱が CPU 温度の変化量として測定される．冷却水により一定の熱量が運ばれ，残った熱により CPU 温度が上昇すると考えると，CPU 温度変化から消費電力をある程度推定することは可能と判断した．

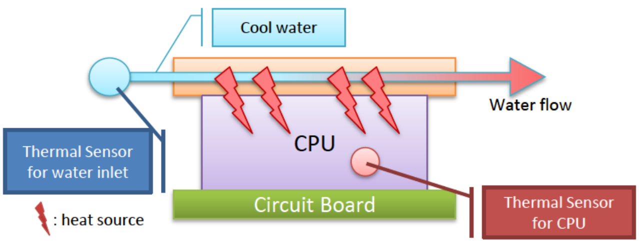

図 3　CPU の冷却機構

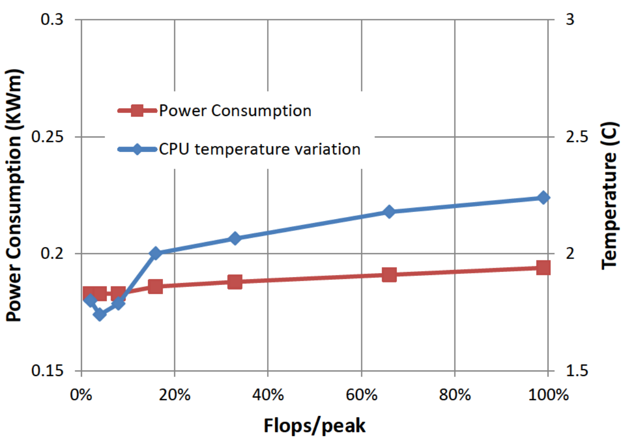

図 4　CPU 負荷と CPU 温度変化，消費電力の関係

図 4 に CPU の負荷を変化させた場合の，CPU 温度変化と消費電力変化の関係を示す．ここでは，浮動小数点演算数と固定小数点演算の割合を変えることで flops 値を変化させながら消費電力を測定した．1 ラックの 96CPU 全てで同じプログラムを実行し，その平均値を求めている．縦軸は消費電力と温度変化を，横軸は CPU の理論性能に対する flops 値の割合をそれぞれ表している．flops 値の低い領域では，CPU 温度変化が比例していない部分もあるが，全体的には CPU 温度変化も消費電力も flops 値に比例しており，CPU 温度変化から消費電力が推定可能であるとわかる．

3.2.2 メモリ

メモリ負荷と消費電力の関係について調査した．

図 5 に「京」の System Board の構成を示す．1 枚の System Board には，計算ノード (CPU1 台と ICC1 台，メモリ 16GiB で構成) が 4 台載っている．CPU と ICC は主に冷却水で冷やされるので，SB 排気温度変化は，主にメモリの発熱によって生じると考えられる．

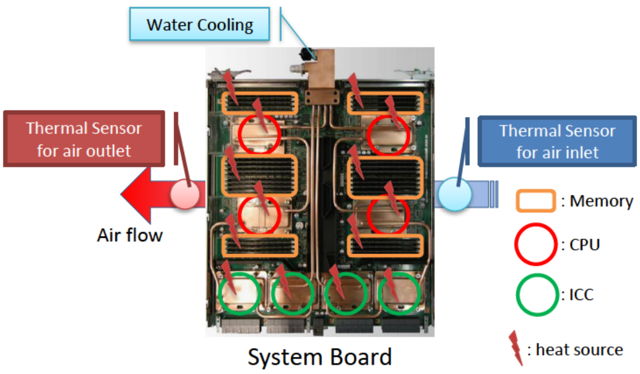

図 5　System Board の構成

図 6 にメモリ負荷 (メモリスループット) を変化させた場合の SB 排気温度変化と消費電力変化の関係を示す．1 ラックの 96CPU 全てで同じプログラムを実行し，24 枚の System Board の平均値を求めている．縦軸は消費電力と温度変化を，横軸はメモリスループットをそれぞれ表している．グラフからメモリスループットに比例して消費電力が変化していることがわかる．SB 排気温度変化も同様にメモリスループットに比例して変化しており，SB 排気温度変化から消費電力を推定することが可能であることがわかる．

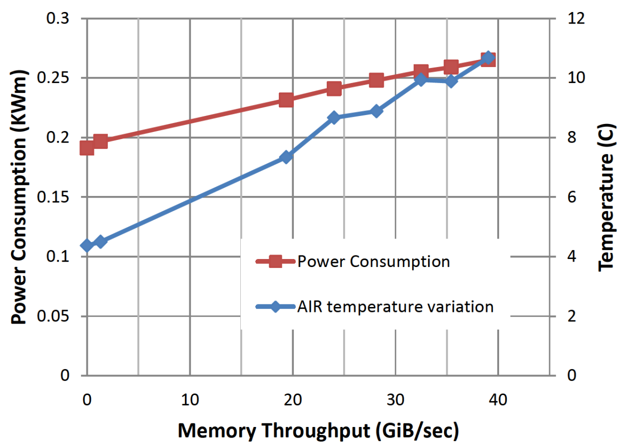

図 6　メモリスループットと消費電力，SB 排気温度変化の関係

3.2.3 ICC

ICC は設計上，消費電力は一定となっている．実際に ICC の消費電力が一定であるかどうか，ICC の負荷を変えて消費電力の変化を調査した．

図 7 に ICC の負荷を変化させた場合の消費電力の関係を示す．ICC には 4 つの TNI(Tofu Network Interface) が繋がっている．リンクあたりの性能は 5GiB/sec × 2（双方向）である．今回の測定では，通信に使用する TNI の数を変えて ICC の負荷 (通信量) を変化させた．縦軸は消費電力を，横軸は使用する TNI の数に応じたネットワークスループットを表している．図 7 から TNI の数に応じて消費電力が変化していることがわかる．TNI を通して通信されるデータはメモリに対しても同様に負荷をかけるため，通信のデータ量に応じてメモリも電力を消費することになる．よって，図 7 のネットワークスループットはメモリスループットとみなすことができる．図 7 の TNI の数を変化させた場合の消費電力の変化と，図 6 のメモリ負荷を変化させた場合の消費電力の変化を比較すると，その傾向が一致していることがわかる．これは，使用する TNI の数を変化させた時の消費電力の変動はメモリによるものであって，ICC の負荷が変わっても ICC 自体の消費電力は変動していないことを示している．この結果は，ICC の消費電力は一定ということと一致している．

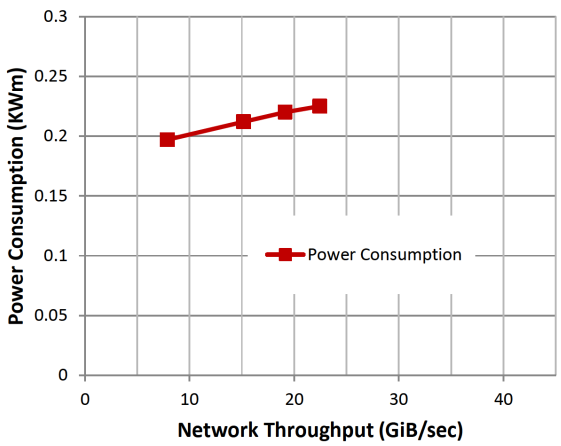

図 7 ICC の負荷と消費電力の関係

以上の結果から，CPU 温度変化および SB 排気温度変化からジョブ単位での消費電力推定は可能と判断した．

4. 消費電力の推定

温度センサ情報をもとにジョブの消費電力の推定を行った．

ジョブの実行時に消費された電力は全て熱となると仮定し，CPU 温度変化と SB 排気温度変化から消費電力を推定する．CPU 温度変化と消費電力変化 (図 4)，SB 排気温度変化と消費電力変化 (図 6) の関係から，消費電力の推定式を次のように定めた．

$$P = a \cdot T_{cpu} \times b \cdot T_{air} + c$$

P はシステム全体の消費電力を，T_{cpu} は平均 CPU 温度変化を，T_{air} は平均 SB 排気温度変化をそれぞれ表す．

図 8 に 2014 年 4 月から 7 月までの平均 CPU 温度変化，平均 SB 排気温度変化とシステム全体の消費電力変化のグラフを示す．このデータをもとに a, b, c を求めた．各値は，以下のようになった．

$$a = 1.01966119707989$$

$$b = 0.254083571677359$$

$$c = 8.35685250443171$$

「京」の全計算ノードを使用したジョブの消費電力の計測データと，温度センサの情報を基に推定した結果の比較を行った．現状では，システム全体の消費電力は 1 分毎に取得できるが，温度センサの情報は 10 分毎にしか取得できないため，全ノードを使用して 1 時間以上実行されたジョブについて調査を行った．4 月から 7 月の 4 か月間で対象となるジョブは 6 本あった．

図 9 に，ジョブ実行時の消費電力の実測値と推定値を示す．実測値は 1 分毎に，推定値は 10 分毎にプロットしている．推定値が実測値と比べて若干高めになっているが，その変動の傾向は一致していることがわかる．

次に，実際の運用で電力超過が発生した状況を想定し，複数ジョブが実行されている状況での消費電力の推定を行った．**図 10** に，実際に「京」上で実行されたジョブ毎のノード数と推定消費電力のグラフを示す．ここでは，1,000 ノード以上を使用したジョブを対象とし，消費電力はジョブ実行による変動値を表している．図 10 の上のグラフがジョブ毎のノード数を，中央のグラフが温度センサ情報を基にしたジョブ毎の推定消費電力を，下のグラフがノード数に比例して求めたジョブ毎の推定消費電力をそれぞれ表している．同じ時間帯の同じ色は同一ジョブ示していて，グラフ中の青の折れ線はシステム全体で使用されたノード数を，赤の折れ線は計測されたシステム全体の消費電力をそれぞれ表している．図 9 の場合と同じく，温度センサ情報による消費電力推定の場合，全体的に高めの推定値となっている．ノード数比例で消費電力を推定した場合と比べて，温度センサ情報に基づく推定は，ジョブ毎の消費電力の違いを見て取ることができる．

現状では，全ノードを使用する場合以外のジョブでは，温度センサ情報に基づく推定消費電力の検証が十分ではない．「京」の一部の計算ラックにはラック単位で測定可能な電力計が設置されている．今後は，これらを活用し温度センサ情報に基づく推定消費電力の検証を行っていきたいと考えている．

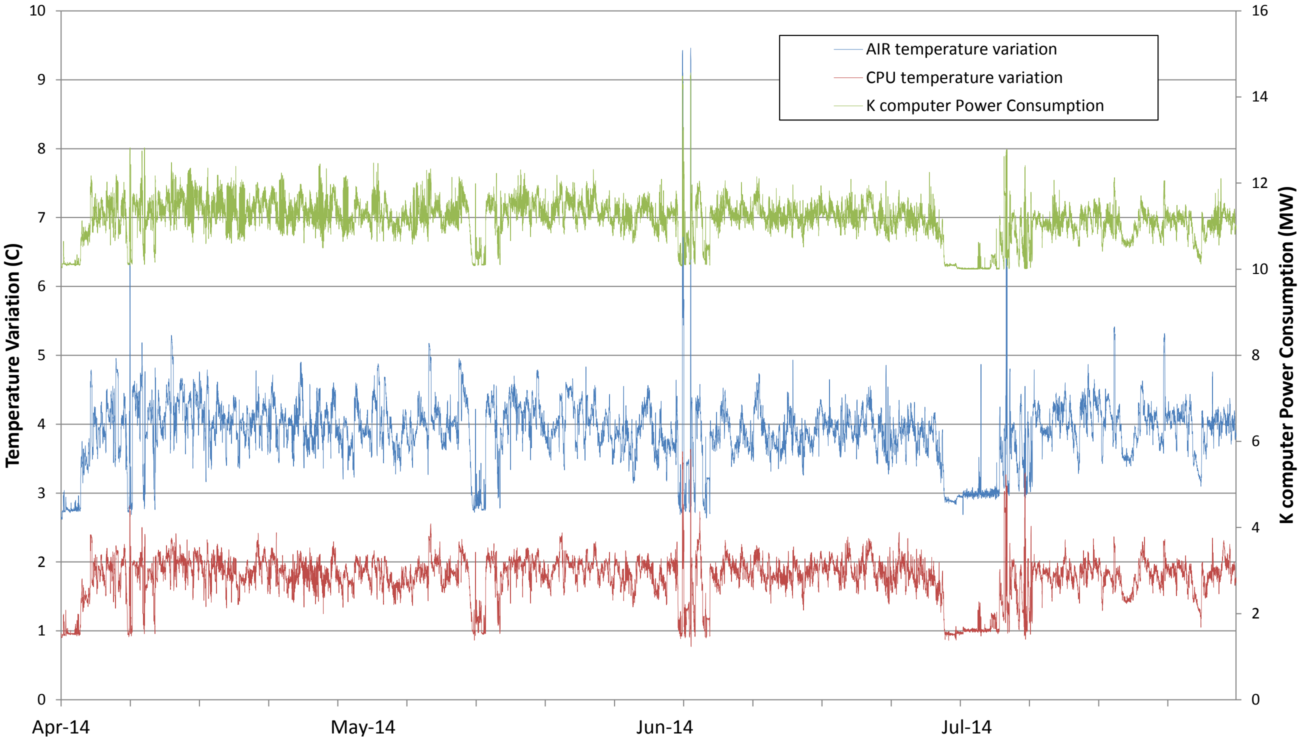

図 8 平均 CPU 温度変化，平均 SB 排気温度変化とシステム全体の消費電力変化の関係

5. おわりに

本研究では，各計算ラックに取り付けられた温度センサ情報とシステム全体の消費電力から個々のジョブの消費電力を推定する方法について検討を行った．

CPU の温度変化と System Board の排気温度変化のデータを使用した消費電力の推定式を求め，ジョブ単位の消費電力の推定を行った．この推定式により，ジョブ単位の大まかな消費電力を推定することができたが，現状では温度センサの精度やサンプリング間隔の問題等から正確な消費電力の推定は難しい．ジョブ実行時のプロファイル情報を利用することができれば，より正確な消費電力の推定が可能であると思われるが，プロファイル情報はジョブ実行が終了した後でなくては取得できない．そのため，電力超過発生時に速やかにジョブを停止することはできない．本手法では，随時取得できる温度センサから消費電力をリアルタイムに推定することができ，消費電力の大きなジョブから順次停止させることができる．

今後は，さらなる消費電力の推定精度の向上と，実運用への応用について検討を続けていきたいと考えている．

参考文献

[1] 黒川原佳，庄司文由：スーパーコンピュータ「京」システム概要，情報処理，Vol.53，Vol.53，No.8，pp.759-766 (2012).

[2] 山本啓二，宇野篤也，塚本俊之，菅田勝文，庄司文由：スーパーコンピュータ「京」の運用状況，情報処理，Vol.55，No.8，pp.786-793 (2014).

[3] Keiji Yamamoto, Atsuya Uno, Hitoshi Murai, Toshiyuki Tsukamoto, Fumiyoshi Shoji, Shuji Matsui, Ryuichi Sekizawa, Fumichika Sueyasu, Hiroshi Uchiyama, Mitsuo Okamoto, Nobuo Ohgushi, Katsutoshi Takashina, Daisuke Wakabayashi, Yuki Taguchi, Mitsuo Yokokawa: The K computer Operations: Experiences and Statistics, Proceedings of International Conference on Computational Science (ICCS), (2014)

[4] 井上文雄，宇野篤也，塚本俊之，松下聡，末安史親，池田直樹，肥田元，庄司文由：電力消費量の上限を考慮した「京」の運用，情報処理学会研究会報告 Vol.2014-HPC-146 No.4 (2014).

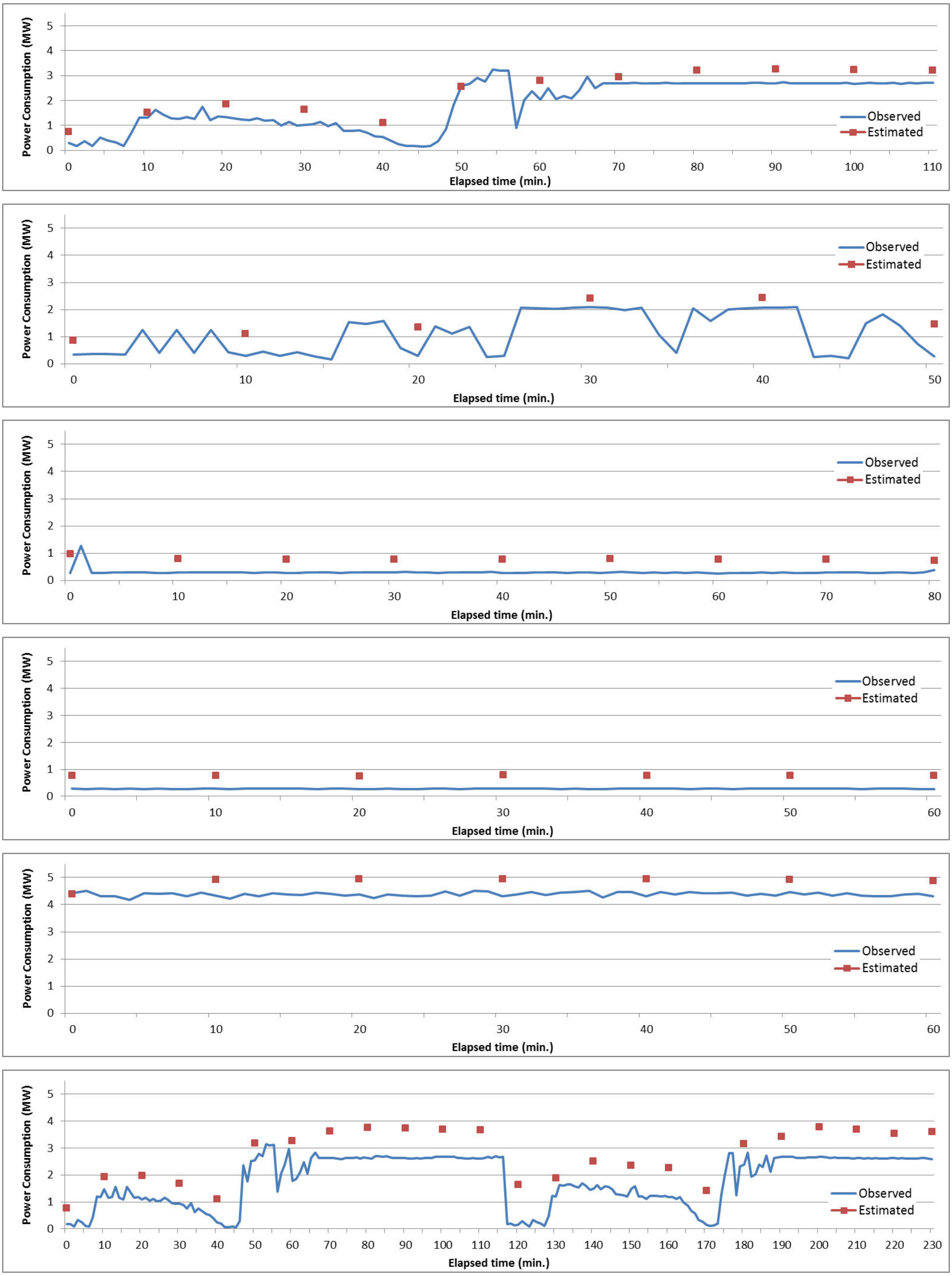

図 9　ジョブ実行時の消費電力の実測値と推定値

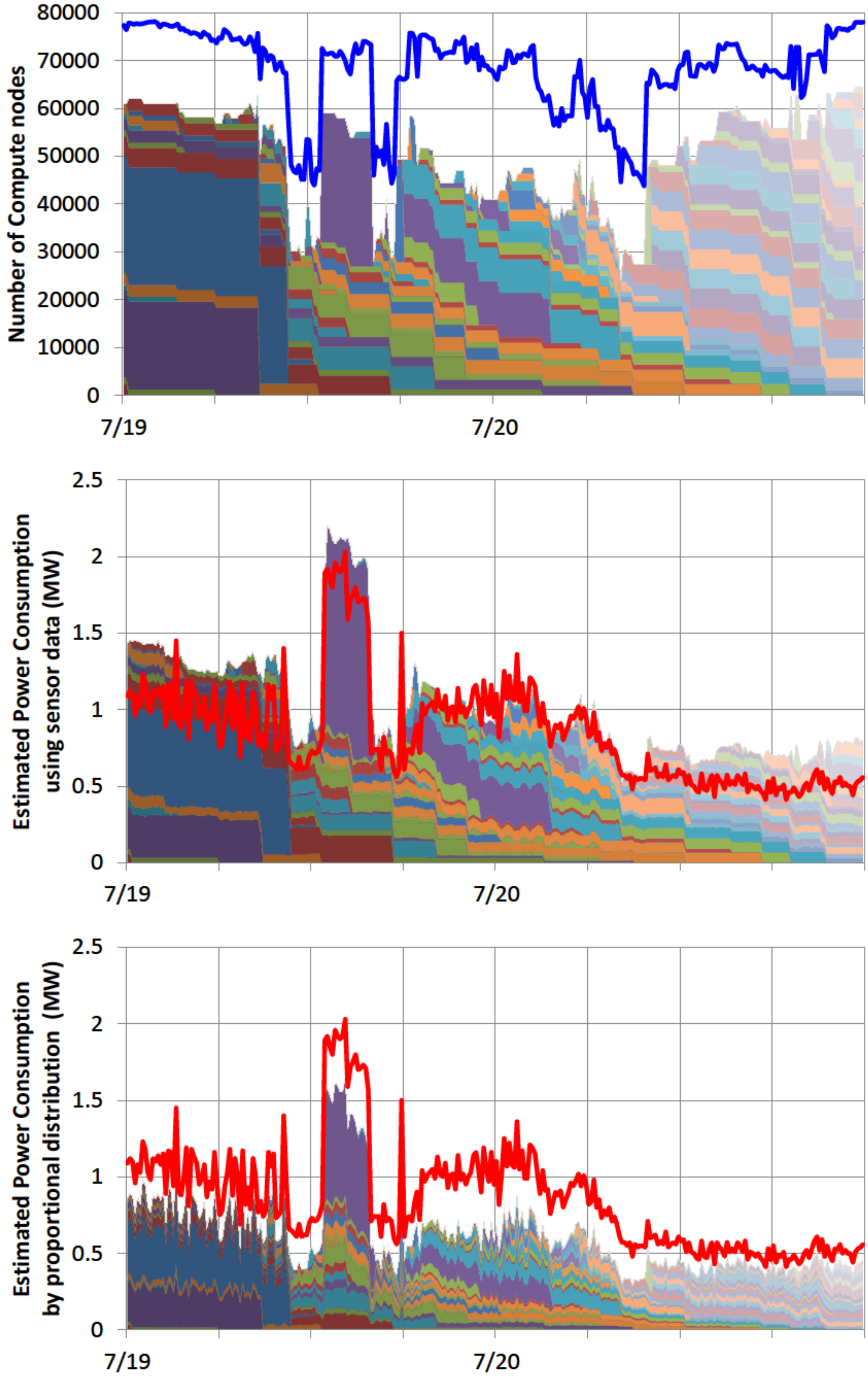

図 10 ジョブ単位のノード数（上）と推定消費電力（中：温度センサ情報使用，下：ノード数比例）