Vol.2014-HPC-146 No.4
2014/10/2

電力消費量の上限を考慮した「京」の運用

井上文雄[†1]　宇野篤也[†1]　塚本俊之[†1]　松下聡[†1]
末安史親[†2]　池田直樹[†3]　肥田元[†3]　庄司文由[†1]

「京」では現在，通常は小中規模のジョブ（36,864 ノード以下）を実行し，大規模ジョブ（36,865～82,944 ノード）は特定の期間（以下，大規模ジョブ実行期間）に実行するという運用を行っている．通常の運用では「京」の消費電力は契約電力内に収まっているが，大規模ジョブ実行期間において契約電力を超過する事例が発生した．頻繁な契約電力の超過は電力契約の見直し等につながり，運用に及ぼす影響は無視できないものである．そこで，これを回避するために，投入予定の大規模ジョブを消費電力の観点で事前に審査することにした．すなわち，過去の動作実績等から推測した大規模ジョブ実行時の消費電力が運用上の上限を超えないことが確認されたジョブのみ投入を許可することにした．加えて，消費電力を 24 時間監視できる体制の構築，及び最大電力量を超過した際のジョブ停止プロセスの整備など運用方法の変更を実施した．本稿では，これらの対策と今後の取り組みについて報告する．

1. はじめに

スーパーコンピュータ「京」（以下，「京」）は， 理化学研究所と富士通株式会社が共同開発した，生命科学・医療，エネルギー，防災・減災，次世代ものづくり，物質と宇宙といった様々な分野のプログラムを高速に処理できる汎用性の高いスーパーコンピュータで，2012 年 9 月に共用を開始して以来，概ね安定して運用している[1][2]．しかし，低消費電力の CPU の採用などにより消費電力を抑えているものの，規模が大きいためにシステム全体の消費電力は 10MW を超えており，運用コストに占める電力料金の割合は非常に大きい．

一般的に計算機の消費電力は実行されるジョブにより変動するが，特に「京」では規模が大きいためにジョブによる消費電力の変動が非常に大きい．2012 年 9 月の共用開始当初は，ほとんどのジョブについて，チューニングがさほど進んでいなかったことや，規模が大きくなかったことなどから，大規模ベンチマーク等の特殊なケースを除き，消費電力が問題となることはなかった．しかし，共用開始から 1 年が経過した頃から，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様になった．これは，実行されるジョブのチューニングが進み，かつ，大規模な計算が実行されることが多くなってきたことによると思われる．このような電力超過は運用への影響が大きいため，いかに電力の消費をコントロールするかが運用上の課題となってきた．そこで，最初の取り組みとして消費電力が契約電力を超えないようにするための対策を実施することにした．

本稿では今回実施した緊急対策である，ジョブの緊急停止と事前審査制度の取り組みについて述べる．

†1 理化学研究所　計算科学研究機構
†2 富士通株式会社
†3 株式会社富士通ソーシアルサイエンスラボラトリ

2. 「京」の運用

2.1 「京」と運用設備

「京」は，計算ノード 82,944 ノード，メモリ容量 1.27PB，ローカルファイルシステム（以下，LFS）11PB，グローバルファイルシステム（以下，GFS）30PB，およびフロントエンドサーバなどの周辺機器から構成されている．図 1 に「京」のシステム構成概要を示す．

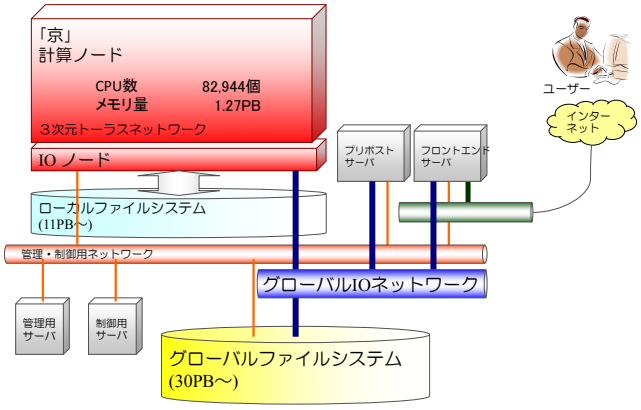

図 1　「京」のシステム構成概要

「京」の運用で消費される電力は全て熱となるため，排熱に必要な大規模冷却設備（冷凍機および空調機，CPU 水冷システム）が常時稼働している．

これらの周辺設備を含めた「京」の運用全体に必要な電力は，商用電力（関西電力）と自家発電により供給されている．自家発電用の設備として，ガスタービン発電による定格出力 5MW のコジェネレーションシステム（以下，CGS）を 2 台備えているが，通常時は 1 台ずつ交互に運転している[3]. 図 2 に理化学研究所計算科学研究機構（以下，AICS）の電源設備を示す．

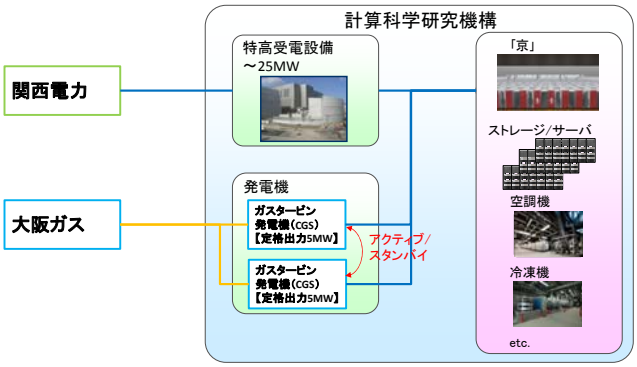

図 2　AICS の電源設備

共用開始当初の AICS 全体の最大消費電力の見込を表 1 に示す.

表 1　AICS 全体の消費電力見込（共用開始時）

内訳	消費電力
「京」本体（含むLFS）	10MW
ジョブ実行による増分	～ 4MW
その他施設（含むGFS）	～ 3MW
合計	～17MW

「京」本体の消費電力とその他施設の消費電力は，共用開始前の試験利用期間の実績値から算出した値であり，ジョブの最大消費電力は実行効率が最も高い LINPACK を実行した際の消費電力を参考に算出した．17MW を電力供給の上限とし, この電力を賄えるように CGS の 1 台での発電電力 5MW を除いた 12MW 分を関西電力と契約した.

2.2 通常運用と大規模ジョブ実行

「京」では，計算資源をノード時間積という単位で管理している．システム全体のノード時間積を，全計算ノード数×時間とし，それを予め各グループに対して配分している．ユーザはグループに配分された計算資源の範囲内で，使用する計算ノード数と実行時間の上限を指定してジョブを実行する．通常の運用では，36,864 計算ノード以下の小中規模ジョブが実行可能である．36,865 計算ノード以上のジョブについては，毎月第 2 火曜日から 3 日間の大規模ジ実行期間を設けている．通常運用時と大規模ジョブ実行期間の計算ノード数と実行時間の制限値を表 2 に示す.

表 2　計算ノード数と実行時間の制限値

	使用計算ノード数		実行時間
	（下限）	（上限）	（上限）
通常運用	1	36,864	24時間
大規模ジョブ実行期間	36,865	82,944	8時間

3. 電力需給状況

図 3 は電力の需給状況のグラフ（以下, 電力トレンド図）である．図中の青色が AICS 全体の消費電力，赤色が関西電力からの受電電力, 茶色が CGS の発電電力をそれぞれ示している．この図は，通常運用時における電力トレンドを示しており，緑の破線で示す電力供給の上限を超えていないことが確認できる. しかし, 大規模ジョブ実行期間では，2013 年度は 4 月，11 月，および 2 月に電力超過[a]が発生した．原因は，LINPACK 並に電力を消費するジョブが複数回実行されたためであった．図 4 に 11 月の大規模ジョブ実行期間で電力超過が発生した時の電力トレンド図を示す.

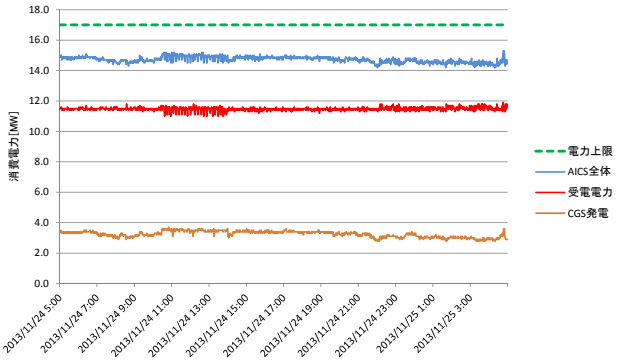

図 3　電力トレンド図（通常運用時）

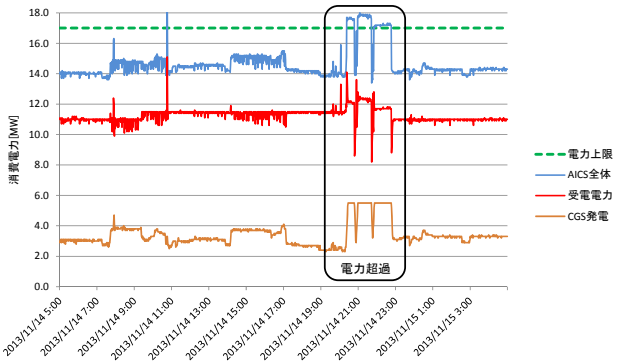

図 4　電力トレンド図（大規模ジョブ実行期間）

4. 電力超過対策

電力超過が発生すると電力会社に対して違約金の支払いが発生する．それだけでなく，頻繁に超過するようだと契約電力を上げることを電力会社から求められ，運用経費の増大に直結することになり，運用への影響は非常に大きい．実際 2013 年度の電力超過の影響により，2014 年度は契約電力が 0.75MW 増の 12.75MW となった.

電力超過を抑止する対策として，自家発電量を増やす方法，システムの一部を停止する方法，ジョブの実行をコン

a）電力超過とは，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ることである.

トロールすることで消費電力を制御する方法について検討した．

4.1 自家発電量を増やす方法

自家発電量を増やすことで電力上限を上げることができる．そのためには，CGS を 2 台運転する必要があるが，当然のことながら超過した電力を賄うためのガス料金の分だけ運用経費が増大する．また，CGS は休止状態から発電可能な状態になるまでに 2 時間程度を要するため，電力超過が発生してから起動していては間に合わない．そのため，あらかじめ 2 台の CGS を稼働させておく必要があり，運用経費的には望ましくない．これらの理由から，自家発電量を増やす方法は採用しなかった．

4.2 システムの一部を停止する方法

システムの一部を停止し，消費電力を抑えることで，電力超過の可能性を減らすことができる．しかし，この方法ではシステム全体の計算資源が少なくなり，全てのグループに対して事前に割り当てた計算資源を提供できなくなる．そのため，「京」では採用できないと判断した．

4.3 ジョブの実行をコントロールする方法

消費電力を考慮してジョブの実行をコントロールできれば電力超過を防ぐことができる．ジョブの実行をコントロールする方法として以下の 2 つを検討した．

(1) 電力超過を発生させたジョブの緊急停止
(2) 電力超過を発生させるジョブの除外

(1)により，消費電力を確実に上限以下に下げることが出来る．しかし，ユーザの観点からはジョブの緊急停止は可能な限り避けるべきで，この点では，事前に電力超過を発生させるジョブを除外する方法が有効である．そこで，まずは(2)を実施し，万一電力上限を超過する場合は(1)を実施する方法を採用することにした．

4.3.1 ジョブの緊急停止

ジョブの緊急停止とは，大規模ジョブ実行時に電力が許容範囲を超えた場合に，実行中のジョブを強制的にキャンセルする仕組みである．例えば，2 つの大規模ジョブが実行されている場合，どちらのジョブの影響で許容範囲を超えたか判断することは難しい．この場合，実行中の全てのジョブを強制的にキャンセルすることになる．ジョブの緊急停止フローの概略を図 5 に示す．

ジョブの緊急停止は，「京」システム，施設監視，システム監視の独立した 3 つのシステムを連携させることで実現している． 施設監視担当は常時電力トレンドを監視し，消費電力が契約電力を超える可能性が高まった場合，システム監視担当へジョブの緊急停止を電話で依頼する．依頼を受けたシステム監視担当は，ジョブ停止スクリプトを用いて実行中の大規模ジョブを停止させる．このスクリプトは実行中の大規模ジョブを特定し，該当するジョブに対して停止コマンドを実行する．次に，対象ジョブの停止を確認後，関係者にメールで連絡する．

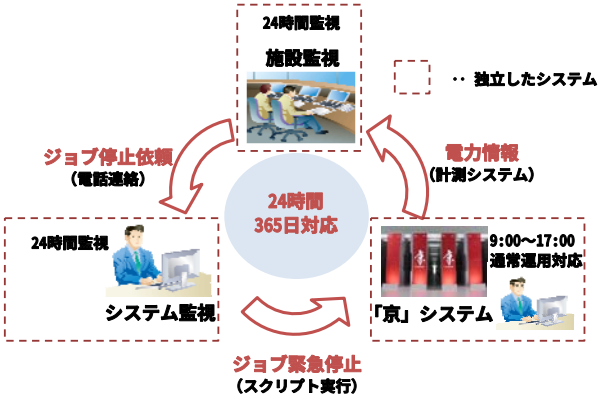

図 5　ジョブの緊急停止フローの概略

電力超過には 30 分間という時間的制限があるため，ジョブの緊急停止は，消費電力が消費電力の上限を超過する数分前に完了させる必要がある．時間的ロスを最小限にするため，施設監視からシステム監視への連絡は電話連絡とし，余分な手続きや判断を極力排除した．

4.3.2 事前審査制度

電力超過を発生させるジョブを除外するため，事前審査制度を導入した．事前審査により，投入予定の大規模ジョブの消費電力を推測し，「京」全体の消費電力が供給電力の上限に収まるようにコントロールすることを目指す．事前審査制度のフローを図 6 に示す．

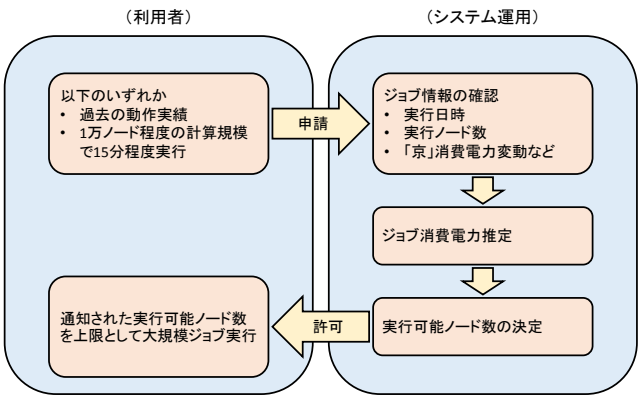

図 6　事前審査制度のフロー

大規模ジョブ実行時の消費電力は，当該ジョブの過去の実行実績から推測することとした．そのために，過去に実行実績がある場合はそのジョブを，実績がない場合は投入予定のジョブを 1 万ノード程度の計算規模で 15 分程度実行

してもらい，その消費電力から大規模実行時の消費電力を推測する．具体的な方法は以下の通りである．まず，申請のあったジョブの実行日時，使用された計算ノード数などを確認し，該当する電力トレンド図から「京」の消費電力の変動を読み取り,その値をジョブの消費電力と見なす(図7)．その消費電力を使用された計算ノード数で割り，1 ノード当たりの消費電力とする（式 1)．許容電力（4MW）を 1 ノードあたりの消費電力で割った値を，実行可能ノード数とする（式 2)．そしてユーザには，この実行可能ノード数を上限とした大規模ジョブの実行を許可する.ただし，実行可能ノード数のジョブであっても，同時に複数実行された場合には，電力超過が発生する可能性がある．そのため，ジョブの同時実行数を制限することでこれを防いでいる．

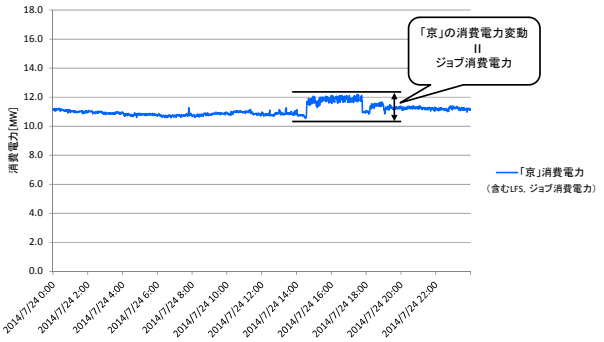

図 7　ジョブの消費電力

$$1ノード当りの消費電力=\frac{ジョブ消費電力}{計算ノード数（実行時）} \quad (1)$$

$$実行可能ノード数=\frac{許容電力（4MW）}{1ノード当たりの消費電力} \quad (2)$$

表 3 に事前審査結果の事例として，8 月大規模ジョブ実行期間の審査結果の一部を示す．

表 3 事前審査結果サンプル（8 月大規模ジョブ実行期間）

大規模ジョブ実行時		ジョブ消費電力推測値（MW）	消費電力差異（実測値－推測値）
計算ノード数	ジョブ消費電力実測値（MW）		
37,544	0.59	1.17	-0.58
80,199	3.74	0.42	3.32
82,944	2.12	2.16	-0.04
37,544	0.87	1.17	-0.30
65,536	1.21	0.78	0.43
80,000	0.94	0.94	0.00
82,944	0.38	1.60	-1.22
82,944	3.44	0.99	2.45

この表は，大規模ジョブ実行時の計算ノード数，ジョブ消費電力の実測値と推測値，およびこれら消費電力の差異を示している．計算ノード数は，ジョブ実行時に使用された計算ノード数である．ジョブ消費電力の実測値は，事前審査時と同様に該当する電力トレンド図から読み取った「京」の消費電力変動値としている．ジョブ消費電力の推測値は，事前審査時にもとめた 1 ノード当たりの消費電力に使用された計算ノード数をかけたものである．この表からわかるように，実測値と推測値が異なる事例が発生している．事前審査開始からこれまでに 24 件の審査を実施し，83 件の大規模ジョブが実行されたが，このうち実測値と推測値が 1MW 以上異なる事例が 25 件発生している．原因としては，推測にもちいたジョブの実行時間が短いなど事前審査時のジョブ情報の精度に問題があることや，事前審査時と大規模ジョブ実行時でジョブ実行時のパラメータが異なることなどが考えられる．詳細な原因については分析中である．

5. おわりに

「京」の供用開始から 2 年が経過し，実行されるジョブのチューニングが進むとともに規模も大きくなり，契約電力の上限を超える状況が発生するようになった．対策としてジョブの緊急停止方法の整備と事前審査制度を実施した．ジョブの緊急停止では，実行中の大規模ジョブを停止するためのスクリプトやジョブ停止プロセスを整備した.また，ジョブの事前審査制度では，ジョブの消費電力を過去の動作実績から推測し，その値を基に実行許可ノード数を決定するプロセスを策定した．これらの対策は 6 月の大規模ジョブ実行期間より運用を開始している．

これらの対策により,事前審査制度を導入した 6 月以降，大規模ジョブ実行期間中に電力超過は発生していない．参考までに 8 月の大規模ジョブ実行期間の電力トレンドを図 8 に示す．

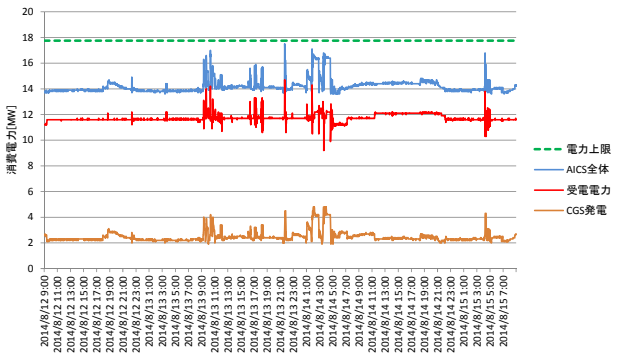

図 8　電力トレンド図（8 月大規模ジョブ実行期間）

これらの対策にはまだ改良の余地がある．例えば，ジョ

ブの緊急停止に関しては，人手を介さずに自動的にジョブを停止する環境の構築，事前審査制度に関しては，ジョブの消費電力の予測精度の向上などである．さらに，将来的には消費電力を計算資源と同様に管理できないか，その可能性を検討していきたいと考えている．

参考文献

[1] 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.
[2] 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)
[3] 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）.