

1 スーパーコンピュータ「京」の運用状況

山本啓二 宇野篤也 塚本俊之 菅田勝文 庄司文由

理化学研究所 計算科学研究機構

スーパーコンピュータ「京」

スーパーコンピュータ「京（けい）」（以下、「京」）は、理化学研究所と富士通株式会社が共同開発し、2011年6月と2011年11月のTOP500リスト¹⁾において世界最速と認定されたスーパーコンピュータである^{2)~4)}。2010年9月より兵庫県神戸市のポートアイランドにある理化学研究所計算科学研究機構（AICS）への設置が始まり、2012年6月末に完成した。

「京」は、2012年9月28日に共用を開始してから1年以上が経過したが、2011年、2012年のゴードン・ベル賞の受賞をはじめとして、生命科学、物質材料、防災・減災、ものづくり、基礎物理等の幅広い分野から、「京」でなければできないような画期的な成果が続々と出てきつつある。本稿では、いくつかの統計情報を交えながら、共用開始後の「京」の運用状況を紹介する。

システム概要

「京」の名前は、1秒間に1京回（1兆の1万倍）の計算ができることに由来する。ここでは、「京」のシステム概要を紹介する。図-1に示すように、「京」は82,944台の計算ノードと5,184台のI/Oノード、2種類のファイルシステム、管理用／制御用サーバ、Pre/Postサーバおよびフロントエンドサーバから構成される^{5), 6)}。

1つの計算ノードには、CPU、16GBのメモリモジュール、CPU間のデータ通信を担うインターコネクトコントローラ（ICC）が搭載されている。つ

まり、1秒間に1京回もの計算速度は8万個以上という膨大な数のCPUにより実現できている。CPUは富士通製のSPARC64TM VIIIfxで、8個のプロセッサコアとコア間で共有する6MBの2次キャッシュを持つ。動作周波数は2GHzでピーク性能の128GFLOPSを58Wで実現する。これは汎用CPUとしては世界最高クラスの電力あたり性能である。

I/Oノードは計算ノードとファイルシステム間のI/O処理を担う。ファイルシステムは30PBのグローバルファイルシステム（GFS）と11PBのローカルファイルシステム（LFS）の2種類を持つ。利用者のファイルはGFSに置かれる。利用者が「京」にジョブを投入する際、そのジョブで必要とされるGFS上のファイルは、ジョブ実行前にLFSにステージング機能（ステージイン）によりコピーされ、ジョブが終了すると、ジョブの実行中に作成されたLFS上のファイルはGFSへとステージング機能（ステージアウト）により移動される。これらのステージングはジョブマネージャにより自動的に行われる。フロントエンドサーバには利用者がインターネットから接続し、「京」へのジョブ投入や、プログラミング等を行う。Pre/Postサーバでは「京」にジョブを投入するための前処理やジョブ投入後のデータ整理等の後処理を行うことができ、1TBのメモリと8つのCPUを搭載している。管理用／制御用サーバではジョブスケジューラ等が動いており、利用者管理やジョブ管理が行われる。

計算ノード間はICCで接続され、Tofuと呼ばれる6次元メッシュ／トーラスの直接網型のネットワークを構成している⁷⁾。本来、直接網型のネットワークは故障に弱いという欠点があるが、Tofuで

は、冗長な通信経路を活用することで故障部分を迂回し、システム全体の稼働を続けられるようになっている。プログラミングの観点からは利用者は、1次元、2次元、3次元のトラスネットワークとして利用することができる。

「京」のオペレーティングシステムにはLinuxが採用されており、プログラミング言語

としてはFortran, C, C++言語が利用可能である。また並列化の標準ライブラリであるOpenMPやMPIなどをサポートし、汎用性と高いスケーラビリティを提供している^{6), 8)}。「京」では、ノード内はOpenMPや自動並列機能を用いたスレッド並列を、ノード間はMPIを用いたプロセス並列を行うハイブリッド並列のプログラミングモデルを推奨している。

「京」には、信頼性、可用性、保守性に関してもさまざまな工夫が盛り込まれている。「京」のような超大規模システムでは、故障が発生することをあらかじめ想定した対策がきわめて重要である。たとえば故障部分を迅速に交換できるように作られている。また重要なコンポーネントである電源ユニットや管理サーバなどは2重化されており、ホットスワップ機能により故障の際の運用への影響を最小限にとどめることができる。

「京」のジョブ処理状況

ここでは共用開始以降の「京」でのジョブの処理状況について述べる。

我々は、利用者が「京」を使う際にストレスなくジョブを実行できる環境の提供を目指している。たとえば、ジョブが実行されるまでの待ち時間をできるだけ短縮したり、システムの稼働率およびスケジューリング効率をできるだけ向上させることである。

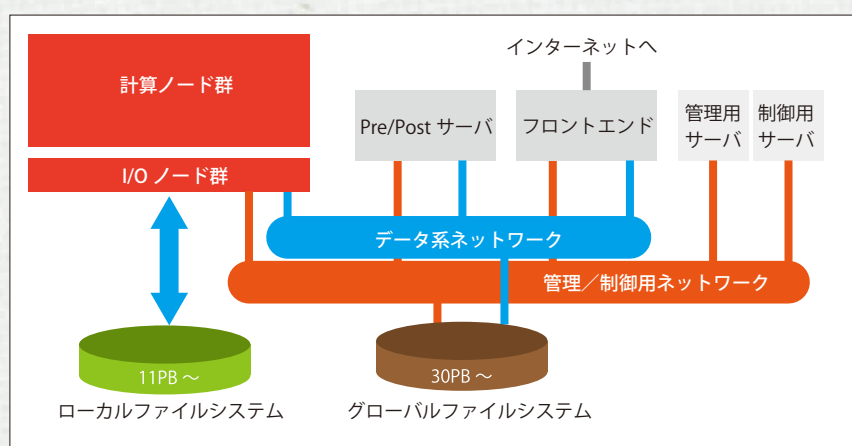


図-1 「京」システム概要

スケジューリングアルゴリズムにはFCFS^{☆1}とバックフィル^{☆2}を採用している。また、スケジューリング効率を上げるため、3つのグループ (Small, Large, Huge) のジョブキュー (「京」ではリソースグループと呼ぶ) を用意している。Smallは1～384ノードのジョブを、Largeは385～36,864ノードのジョブを、Hugeは36,865～82,944ノードのジョブをそれぞれ投入できるようになっている。SmallとLargeについては投入できるノード群を分割している。通常運用では、SmallとLargeが利用可能で、Hugeは1カ月のうち連続した3日間 (以下、大規模実行期間) のみ利用できる。

「京」にはジョブの種類として、通常ジョブ、ステップジョブ、バルクジョブ、会話型ジョブがある。通常ジョブは普通のバッチジョブである。ステップジョブはジョブの実行順序を利用者が定義できるジョブで、前のジョブの実行が成功すれば次のジョブを実行する等の一連の処理が実現できる。バルクジョブは複数のジョブを同時に投入することのできるジョブで、同じアプリケーションをパラメータを変えて何度も実行するようなアンサンブル計算等に利用できる。

システムを効率的に運用するにあたり、「京」で実行されているジョブの状況を分析することは非常に重要である。特に、OSやファイルシステム、ス

☆1 First-Come and First-Served : ジョブの到着順にスケジューリングを実施する方法。

☆2 ノード利用効率を改善するため、空きノードがある場合にはジョブの実行順序を変更してスケジューリングする方法。

スーパーコンピュータ「京」の利用

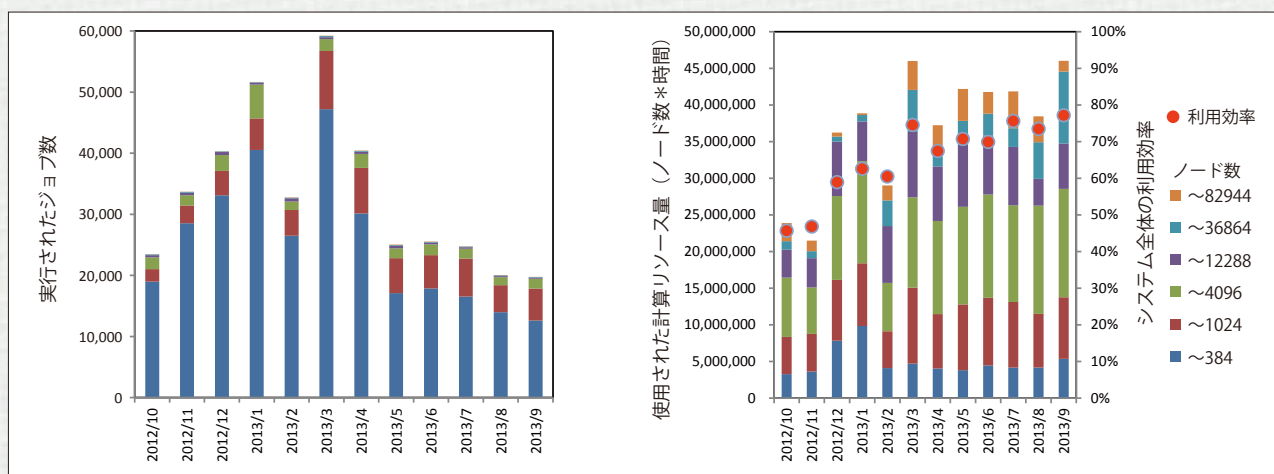


図-2 実行されたジョブの数（左）と使用された計算リソース（右）

ケジューラ等の動作にかかる各種パラメータを決定する上で、ジョブの規模や数、I/O量、ジョブ実行までの待ち時間などのジョブの特性を把握することは重要な役割を持つ。

● ジョブ実行状況

図-2は2012年10月から2013年9月までの1年間に「京」上で実行されたジョブの数と使用された計算リソース^{☆3}の状況を月ごとにグラフ化したものである。2012年11月と2013年2月の使用された計算リソース量は他の月と比べると少ないが、これはこれらの月にシステムのメンテナンスをそれぞれ1週間程度行い、その間システムが停止していたためである。

グラフから2013年2月以降ジョブの傾向が大きく変わっていることが分かる。これは、システム構成を変更したことによる。共用開始時にはシステムを複数のリソースグループに分割せず、1つのリソースグループですべてのジョブを処理する構成としていた。これは、大きなジョブの間で小さなサイズのジョブが実行されることで、システム全体の利用効率^{☆4}の向上を狙ったものであった。しかし実際には経過時間の長い小さなジョブが大きなジョブの実行を阻害し、システム全体の利用効率を低下さ

せる要因となっていた。そのため、2013年2月に計算ノードをSmallとLargeのリソースグループに分割し、システム全体の利用効率の改善を図った。この変更以降、システムの利用効率が改善されただけでなく、比較的規模の大きいジョブが適切に実行されるようになったことで、ジョブの平均サイズが大きくなっている。また、ファイルステージングのタイミングの変更やジョブの同時実行数を変更するなどのシステムパラメータの最適化を続けており、2013年9月の時点でシステム全体の利用効率は約80%を達成している。これは、直接網のネットワーク持つシステムとしては非常に高い値^{☆5}である。

「京」の共用開始から1年間で、約275,000個のジョブが実行され、計算リソースとして420,000,000ノード時間が消費された。図-3は実行時間別にジョブ数と使用された計算リソースの割合を示したものである。実行されたジョブの約60%が実行時間30分未満のジョブで、2時間未満のジョブは約80%であった。また、実行時間が2時間未満のジョブが計算リソースの約20%を消費していることが分かる。一方、実行時間が12時間を超えるジョブの数は全体の約5%で、計算リソースの約40%を消費している。

図-4は実行されたジョブの実行効率^{☆6}の分布

☆3 ジョブの実行経過時間と使用されたノード数の積。

☆4 利用可能な計算リソース量に対する実際に利用された計算リソース量の割合。

☆5 「京」ではノード間通信を有効に活用するため、ジョブに対して隣接した計算ノード群の塊を割り当てる。そのため、一般的なシステムと比べてシステムの利用効率を上げることが難しい。

☆6 ピーク性能に対するジョブの実効性能の割合。

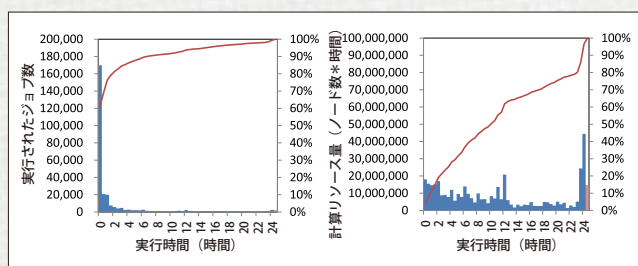


図-3 実行時間別のジョブ数（左）および計算リソース使用状況（右）（実線は累積値を示す）

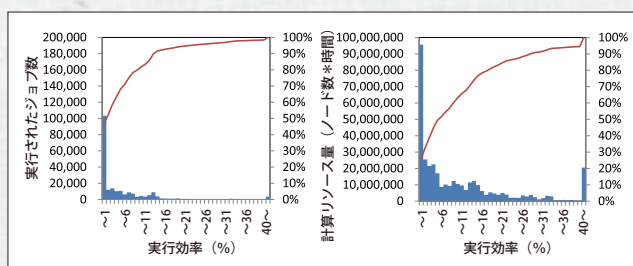


図-4 実行効率別のジョブ数（左）および計算リソース使用状況（右）（実線は累積値を示す）

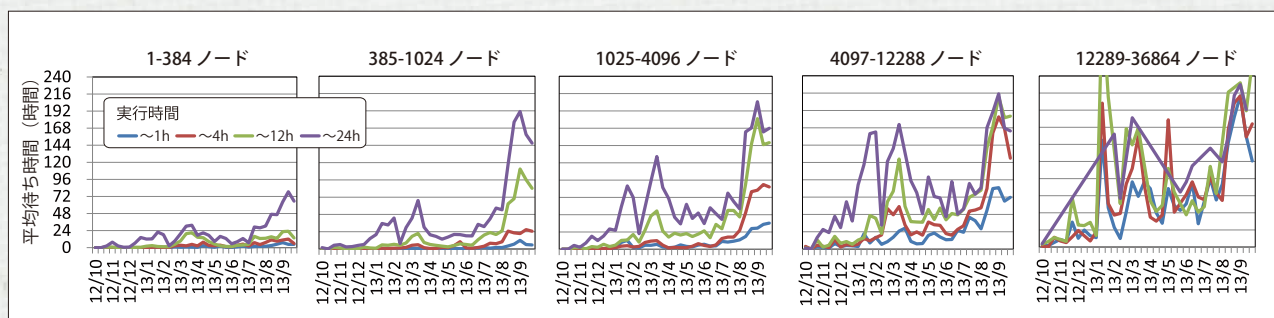


図-5 計算リソース別のジョブ実行までの待ち時間

を表している。全体の約90%のジョブが実行効率13%以下で、計算リソースの約90%が実行効率28%未満のジョブで消費されている。

ジョブ実行待ち時間

図-5はジョブが必要とする計算リソース（ノード数、実行時間）別に、ジョブがシステムに投入されてから実行されるまでの待ち時間を示したものである。グラフから分かるように、必要な計算リソースの大きさが大きくなるほどジョブの待ち時間が増加している。

「京」では、利用者の計算リソースを前期（4月～9月）、後期（10月～3月）の2期に分けて管理している。そのため、各期末が近づくと投入されるジョブが増加し、この時期はジョブ実行待ち時間が長くなる傾向がみられる。実際に、3月と9月の待ち時間が他の月と比べて非常に長いことが図-5から分かる。

安定運用の対策と運用実績

「京」は、膨大な数のCPU、ICC、メモリモジュール等から構成される複雑なシステムである。一般

的にはこのような部品点数が多いシステムは、信頼性を確保することが困難であると言える。たとえば、1つのCPUの故障頻度が年に1回だったとすると、PC単体の故障率だと考えれば十分な信頼性と言えるが、「京」では82,944個から構成されるので、確率的には約6分に1個の割合でいずれかのCPUが故障することになり、システムとしてとても使いものにならない。「京」のような超大規模構成では、システムの信頼性を上げるために、部品単体の信頼性を上げることに加え、故障が起きた時にその影響が広範囲に及ばないようにすること、故障箇所を迅速かつ可能な限り稼働を継続したまま交換できることなどの観点が重要になる。

ここでは、「京」における信頼性向上のための対策と、共用開始後1年間の故障発生状況を紹介する。

● 「京」における信頼性向上のための対策

「京」の設計においては、まず主要なパーツであるCPUとICCについては駆動温度を30度以下に設定した。これは通常のPCに使われるCPUと比較するとかなり低い温度である。この温度は、温度が低いほど故障率が低下するという経験則（アレニ

スーパーコンピュータ「京」の利用

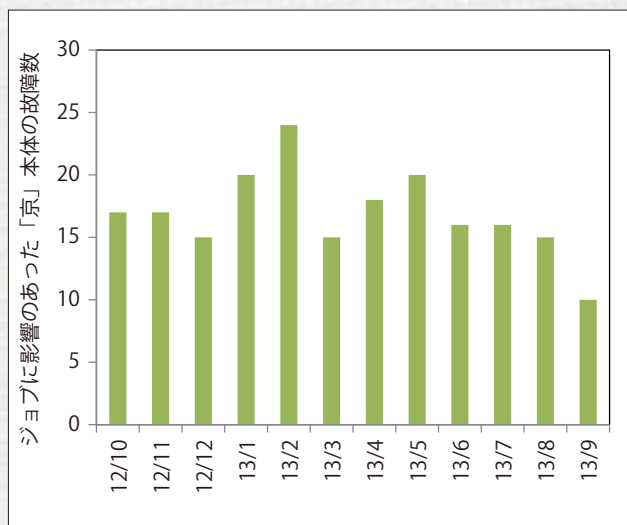


図-6 2012年10月から2013年9月までの「京」本体の故障数

ウスの法則)に基づいて設定されたもので、駆動温度を通常のPCと同程度の85度に設定した場合と比較して、製品寿命が約60倍伸びることが我々の試算から明らかになっている。

また「京」は、4ノード(CPU)が搭載されたシステムボード単位で保守を行い、システムボードの交換作業中も、それ以外の部分は運用を継続できるため、単一の故障による運用への影響は4ノード(CPU)分しかない。加えて、メモリモジュールについては、1bitエラー^{☆7}を検出した場合は、実行中のジョブが終了した時点で交換する運用を行っており、ジョブが途中で打ち切りになるケースはない。

これらの工夫により、故障が起きたとしても、運用への影響を最小限に抑えることができている。

● 共用開始からの1年間の運用実績

共用開始後の約1年間に発生した、ジョブ実行に影響のあった「京」本体の故障数を図-6に示す。本故障には計算ノードとI/OノードのCPU、ICCおよび電源の故障が含まれ、メモリモジュールの故障は含まない。CPU、ICCおよび電源を合わせて約20万個の部品に対し、月あたりの故障数は12～

^{☆7} たとえば、倍精度浮動小数点数は64bitで表現されるが、このうちの1bitが変化すること。「京」で採用しているメモリにはECC機能があり、1bitの誤りを訂正し、2bitの誤りを検出できる。

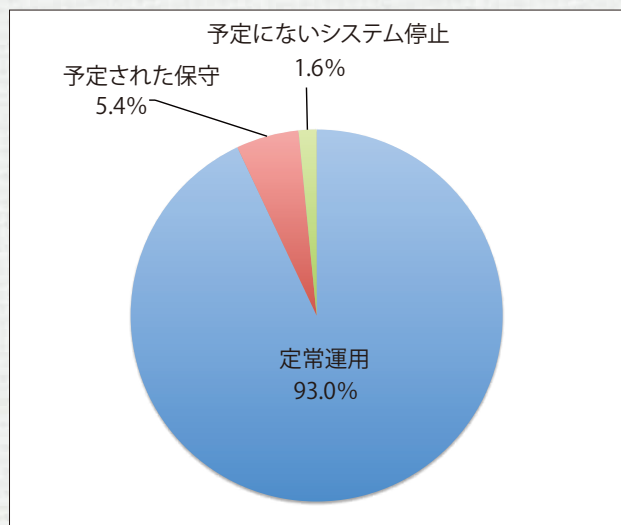


図-7 2012年10月から2013年9月までの運転時間の内訳

24件のため故障率は0.006～0.012%程度ときわめて低い。

また、共用開始後の約1年間の運転時間の内訳を図-7に示す。共用を開始してしばらくはGFS等のトラブルで全系を停止することが数回あったが、2012年10月から2013年9月までの実績では、定常運用と予定された保守を除いた、予定にないシステム停止はわずか1.6%(6日弱)ときわめて安定して稼働していることが分かる。今後は、予定にないシステム停止の原因究明と解決および予定された保守時間の短縮に努め、システムとしてさらに安定した稼働を目指したい。

消費電力

「京」は世界トップレベルの計算性能を持つ一方で、稼働させるために大量の電力を消費する。ここでは、「京」の安定的な稼働を支えるための電力供給設備と、「京」の電力の状況を紹介する。

● 電力消費の内訳

共用を開始して以来、利用の増加に伴い消費電力も微増の傾向が続いている。図-8は2012年10月から2013年9月までの各月当たりの平均電力の内訳を示しているが、AICS全体では平均15,200kW

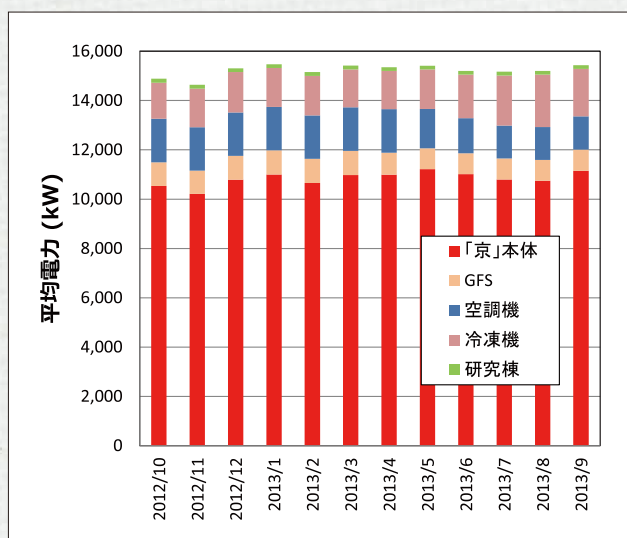


図-8 2012年10月から2013年9月までの電力内訳

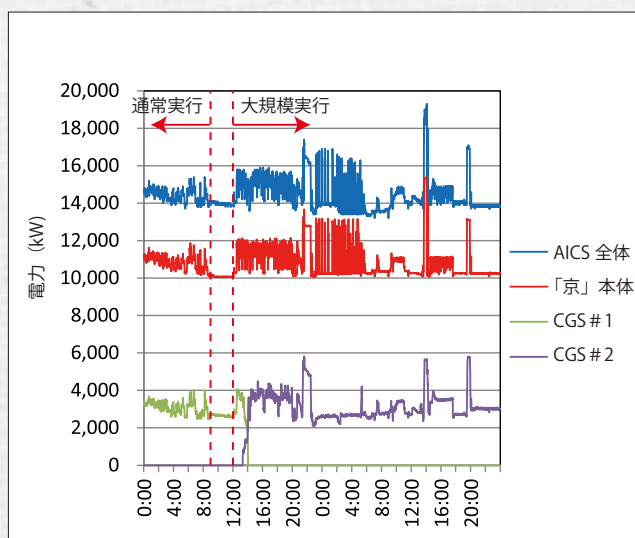


図-9 電力消費推移 (2013/11/12 ~ 13)

程度の電力となっている。その内訳は、最大負荷である「京」本体（LFSを含む）の電力が全体の72%を占めており、冷却系設備（冷凍機と空調機）の電力が21%、利用者のデータを格納するGFSの電力が6%、研究棟の電力が1%となっている。「京」本体とGFSの消費電力を合計したPUE値^{☆8}は月平均で1.3を下回っており、計算機センターとして非常に高いエネルギー効率を達成している。なお、PUEの計算式は以下のとおりである。

$$PUE = \frac{\text{「京」本体の消費電力} + \text{GFSの消費電力} + \text{空調機の消費電力} + \text{冷凍機の消費電力}}{\text{「京」本体の消費電力} + \text{GFSの消費電力}}$$

● 電力供給

電力供給は、関西電力からの受電電力（契約電力12,000kW）と定格出力5,000kWのガスタービン発電機（CGS）による自家発電で構成されている。AICS全体の消費電力に対する割合としては、大まかに関西電力からの受電電力分が2/3、CGSによる自家発電分が1/3である。またCGSは、関西電力からの電力供給が停止した場合に、GFSと研究棟への電力供給を担保する無停電電源（UPS）の

役割も担っている^{9), ☆9}。CGSは2台設置されており、1台が常時稼働し、もう1台が待機系として定期的に切り替える運用を実施している。

図-9は2013年11月12日からの2日間の電力消費推移を示している。「京」本体の電力消費がAICS全体の電力消費の推移を支配しているのが明らかに見てとれる。最初の9時間は通常運用の期間であるが、「京」本体の消費電力は、100個以上の中小規模のジョブ群により平均化され11,000 ± 500kWの範囲に納まっている。しかしその後、11月12日の12時頃から大規模実行期間が始まり、「京」の全ノードを使うような大規模ジョブが実行されることにより、電力消費に大きな変動が発生している。消費電力の変動幅は、実行されるジョブの特性に大きく依存している。ほぼ1つのジョブしか実行されない大規模実行期間は、複数のジョブによる平均化が起きないため極端に消費電力が変動する現象が発生しやすい。実際に図-9の「京」本体の消費電力の推移を見ると、システムが待機状態にある時の10,000kWから高負荷ジョブが実行されている時の15,300kWまで極端に変化していることが分

☆8 Power Usage Effectiveness: データセンタ全体の消費電力を、サーバなどのIT機器の消費電力で割った値で、エネルギー利用効率を示す指標の1つ。

☆9 関西電力からの電力供給が停止した場合は、「京」本体は停止するが、データが保存されているGFSはCGSからの電力により運用を継続できる。停電によって実行中のジョブは異常終了となるが、元となるデータがあれば再実行が可能であるという意味でGFSの運用を優先している。

スーパーコンピュータ「京」の利用

かる。

このような短時間かつ大きな電力変動の問題は、「京」のような規模になって初めて顕在化したことであり、今後運用方法を含めて対応を検討していく必要がある。

利用者対応について

「京」の登録利用者数は1,700名を超えた。「京」は常時多数の利用者がジョブを投入・実行しており、稼働率はきわめて高い。「京」の利用者はさまざまな分野にまたがっており、使用する言語環境やジョブの規模など、使い方やニーズは利用者ごとに大きく異なる。一方、「京」は大規模かつ複雑なシステムであるため、各利用者のニーズに合った最適な利用方法が分かりにくい場合もある。「京」を安定的かつ効率的に利用してもらうためには、このような利用者からの問合せ対応がきわめて重要になってくる。ここでは利用者からの問合せ状況とサポート体制について紹介する。

● サポート体制

利用者からの問合せは、「京」に関する利用者選定業務と利用者支援業務を行う登録施設利用促進機関である一般財団法人高度情報科学技術研究機構（以下、RIST）が最初に対応する¹⁰⁾。問合せは、メール、電話に加え、AICSに設置されているRISTの問合せ窓口でも受け付けている。RISTへの問合せのうち、「京」の運用全体にかかわるものや、非公開情報を含むもの、不具合等は、その対応がRISTから「京」の運用機関であるAICSへと引き継がれる。すなわち、「京」の利用申請時から継続して利用者を支援するRISTと、「京」の開発から運用まで豊富な経験と知識を持つAICSの2つの組織が連携し利用者をサポートしている。

● 問合せ状況

ここでは、問合せのうち、RISTからAICSに引き継がれたものについて見てみる。図-10は、共用

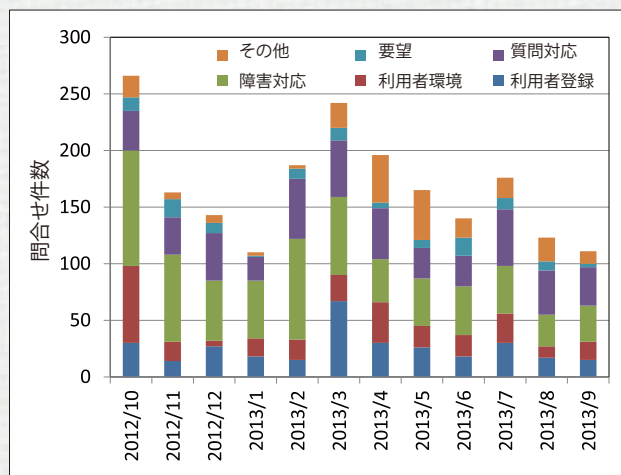


図-10 問合せ状況

開始からの問合せ件数の推移である。問合せ件数の総数は約2,000件に上るが、その内訳は、「京」で実行するプログラムの高速化や「京」独自機能の利用法等に関する質問対応、またはシステムの不具合等の障害対応に分類されるものが多い。これまでに問合せ件数が多かったのは、共用開始直後の2012年10月と、利用期間の期末にあたる2013年3月である。共用開始直後ではシステムの不具合もあり、利用者環境と障害対応に分類される問合せが多い。利用期間の期末では次年度の利用者登録に関する問合せが多い。2013年度に入ってシステムが安定するにつれて、障害対応の件数は落ち着いてきていることが分かる。

「京」はこれまでにさまざまな機能の追加や改善を行ってきているが、その中には利用者からの問合せが発端になったものが少なくない。コンパイラやジョブスケジューラなどのシステムソフトウェアの改良が例である。バグが見つかることは、運用する立場からすると決して望ましいことではないが、利用者からの指摘によるバグの顕在化や、機能の追加等を通じて、「京」はより使いやすく、より安定に運用できるよう改善されているという側面もある。今後も、さまざまな利用者からの問合せに対応しながら、システムのさらなる改善につなげていきたい。

まとめ

「京」は共用を開始して以降、幅広い分野の多くの利用者に使われてきており、画期的な成果も数多く報告されている。LINPACK ベンチマーク^{☆10}の性能でこそ、世界 No.1 を譲ったものの、実際のサイエンスのアプリケーションにおける実行効率という観点では、現在でも世界最高のポテンシャルがあると我々は思っている。今後、世界トップレベルの成果創出のペースをさらに加速していくためにも、「京」の安定的な稼働と、高度化を継続して進め、利用者により良い利用環境を提供できるように努力していきたい。

参考文献

- 1) TOP500 : TOP500 Supercomputer Sites, Top500.org (online), available from (<http://www.top500.org>) (accessed 2014-2-26).
- 2) 横川三津夫, 庄司文由: 京速コンピュータ「京 (けい)」とは何か, 日本原子力学会誌, Vol.52, No.12, pp. 14-18 (2010).
- 3) 庄司文由: 京速コンピュータ「京 (けい)」とその利用, 応用物理, Vol.80, No.07, pp.574-578 (2011).
- 4) Yokokawa, M., Shoji, F., Uno, A., Kurokawa, M. and Watanabe, T.: The K Computer: Japanese Next-generation Supercomputer Development Project, in Proceedings of the 17th IEEE/ACM International Symposium on Low-power Electronics and Design, ISLPED '11, pp.371-372.
- 5) 黒川原佳, 庄司文由: システム概要—世界トップクラスの演算性能と使いやすさを両立—, 情報処理, Vol.53, No.8, pp.759-766 (2012).
- 6) 清水俊幸, 安島雄一郎, 吉田利雄, 安里 彰, 志田直之, 三浦健一, 住元真司, 長屋忠男, 三吉郁夫, 青木正樹, 原口正寿, 山中栄次, 宮崎博行, 草野義博, 新庄直樹, 追永勇次, 宇野篤也, 黒川原佳, 塚本俊之, 村井 均, 庄司文由, 井上俊介, 黒田明義, 寺井優晃, 長谷川幸弘, 南 一生, 横川三津夫: スーパーコンピュータ「京」の構成と評価, 電子情報通信学会論文誌 D, Vol.J96-D, No.10, pp.2118-2129 (2013).

^{☆10} 連立一次方程式の求解を行うベンチマークプログラムで、測定値はスーパーコンピュータの性能ランキングに使用されている。

- 7) Ajima, Y., Sumimoto, S. and Shimizu, T.: Tofu: A 6D Mesh/Torus Interconnect for Exascale Computers, Computer, Vol.42, No.11, pp.36-40 (2009).
- 8) Adachi, T., Shida, N., Miura, K., Sumimoto, S., Uno, A., Kurokawa, M., Shoji, F. and Yokokawa, M.: The Design of Ultra Scalable MPI Collective Communication on the K Computer, Comput. Sci., Vol.28, No.2-3, pp.147-155 (2013).
- 9) 関口芳弘, 庄司文由, 塚本俊之: 計算科学研究機構の施設と設備—「京」の安定運用を支える基盤—, 情報処理, Vol.53, No.8, pp.801-807 (2012).
- 10) 高度情報科学技術研究機構: HPCI ポータル, 高度情報科学技術研究機構 (オンライン), 入手先 (<https://www.hpci-office.jp/>) (参照 2014-2-26).

(2014 年 5 月 30 日受付)

山本啓二 (正会員) keiji.yamamoto@riken.jp

東京大学大学院情報理工学系研究科コンピュータ科学専攻修了。博士 (情報理工学)。東京大学情報基盤センター特任助教を経て、理化学研究所計算科学研究機構に入所。2013 年より運用技術部門システム運転技術チーム特別研究員。「京」の運用およびシステムの高度化に従事。

宇野篤也 (正会員) uno@riken.jp

2000 年筑波大学大学院工学研究科博士課程修了。博士 (工学)。現在、(独) 理化学研究所 計算科学研究機構運用技術部門システム運転技術チーム チームヘッド。京の運用およびシステムの高度化に従事。

塚本俊之 tsukamoto_yuki@riken.jp

1986 年名古屋大学大学院博士課程理学研究科満了。同年富士通 (株) 入社。宇宙分野の業務システム開発に従事。2010 年理化学研究所次世代スーパーコンピュータ開発実施本部開発研究員。2014 年理化学研究所計算科学研究機構運用技術部門施設運転技術チームヘッド。施設の運転および高度化に従事。

菅田勝文 k_sugeta@riken.jp

1997 年岡山理科大学工学部情報工学科卒業。1997 年日本タイムシェアにて移動体通信用交換機の開発に従事。2012 年 (独) 理化学研究所計算科学研究機構運用技術部門所属。京コンピュータの運用および高度化に従事。

庄司文由 shoji@riken.jp

1998 年金沢大学大学院自然科学研究科単位取得退学。1998 年広島大学情報教育研究センター助手。2006 年理化学研究所次世代スーパーコンピュータ開発実施本部開発研究員。2010 年同チームリーダー。2014 年理化学研究所計算科学研究機構運用技術部門長。京コンピュータの運用および高度化に従事。博士 (理学)。