

「京」の運用のノウハウを活かした 大規模HPCシステム向けソリューション

Solutions for Large-scale HPC Systems Leveraging Know-how of Operating the K Computer

● 末安史親 ● 井上俊介 ● 関澤龍一

あらまし

スーパーコンピュータやPCクラスタといったHPC(High Performance Computing)システムの運用では、単位時間あたりにより多くの計算を実行できることが重要である。従来のHPCシステムでは利用者数がある程度限られていたため、運用者はシステム上で実行されるアプリケーションの種類や性能を把握することが可能であった。しかし、近年のシステムの大規模化に伴い、利用者数や実行されるアプリケーションの数や種類が増加し、運用者はシステムが効率的に利用されているかを把握することが難しくなってきた。スーパーコンピュータ「京」の運用においても、システム全体での資源(CPU、メモリ、I/O、通信、消費電力)の効率的な利用が課題となっている。この課題に対し、「京」を運用している理化学研究所計算科学研究機構(AICS)は、アプリケーションによる各資源の利用状況の採取と、消費電力の利用効率化への取り組みを開始した。富士通はAICSの取り組みを支援し、アプリケーション情報の自動採取環境、および消費電力予測システムを構築した。

本稿では、「京」の資源利用効率化を支援する大規模HPCシステム向けソリューションについて紹介する。

Abstract

When operating high-performance computing (HPC) machines such as supercomputers and PC clusters, faster computing in a unit time is crucial. With previous HPC systems, administrators could manage the types and features of applications executed on the system, as the number of system users was limited. However, system capabilities have improved in recent years, catering to more users and increasing the number and variety of applications that can be run simultaneously. Consequently, administrators are finding it harder to ascertain whether or not a system is operating efficiently. Ensuring that resources (CPUs, memories, I/O, communication, and power consumption) are efficiently deployed on a system-wide basis also presents a challenge when operating the K computer. To address this challenge, RIKEN Advanced Institute for Computational Science (RIKEN AICS), Japan, an operator of the K computer, has embarked on a project to extract data on how application programs use each resource, and to enhance the power use efficiency. In support of these initiatives at the AICS, Fujitsu has developed a platform to automatically collect application data and a system to forecast power consumption. This paper presents a solution designed for the operation of large-scale HPC systems, based on the resource use enhancement developed for the K computer.

まえがき

計算機センターなどで運用されているスーパーコンピュータやPCクラスタといったHPC（High Performance Computing）システムにおいては、単位時間あたりにより多くの計算を実行できることが重要視され、ハードウェアの高性能化やシステムの大規模化が進められてきた。こうした状況において、運用者にとっては、システムが有する大量の資源を効率的に利用できているかを把握することが課題となる。

スーパーコンピュータ「京」^(注)を運用する理化学研究所計算科学研究機構（以下、AICS）は、資源利用効率の改善のために、各アプリケーションの資源利用状況を運用者が採取する取り組みを開始した⁽¹⁾。その中で富士通は、「CPU」「メモリ」「I/O」「通信」という従来の一般的な資源に、新たに重要性が高まっている「消費電力」を加えた五つを対象とし、アプリケーションごとの情報を自動的に採取する環境を構築した。またAICSは、得られた情報を用いて、消費電力を考慮したシステム運用の検討を進めている。富士通はその取り組みを支援するとともに、消費電力の予測システムを開発した。

本稿では、大規模HPCシステムの効率的な資源利用における課題と、「京」での対応としてAICSの取り組みと富士通が開発したソリューションについて述べる。

HPCシステムの現状と大規模化に伴う課題

HPCシステムが利用されている計算分野は、生命・医療、エネルギー、材料、ものづくり、防災、宇宙など多岐にわたる。これらの計算は、演算量（CPU使用量）とデータ量（メモリ使用量）の観点から大きく二つに分類できる。演算量に比べてデータ量が多い計算と、データ量に比べて演算量が多い計算である。同じ分野のアプリケーションは比較的似た性能特性を持つ場合が多く、浮動小数点演算性能やメモリスループット性能といった性能指標も似ている場合が多い（図-1）。こうした理由から、従来大学や研究機関などでは自身の計算分

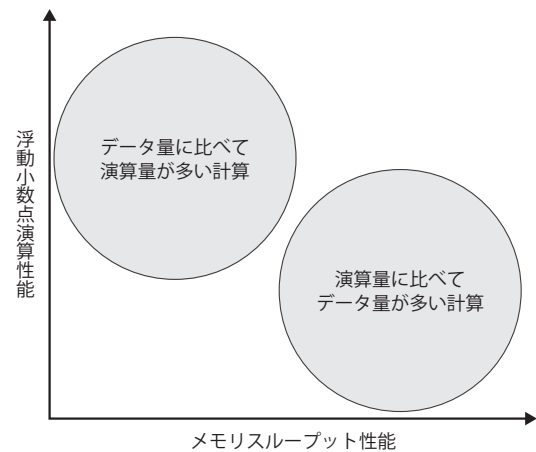


図-1 HPC向けアプリケーションの性能特性イメージ

野に適したシステムを個別に導入することが一般的であった。運用者は新たなシステムに必要なスペックを把握する必要があるが、利用者がある程度限られていたため、システム上で実行されているアプリケーションに関する性能情報の把握が可能であった。

しかし、近年はシステム性能の向上に伴い、導入コストや設備の削減が必要となってきた。そこで、大規模なHPCシステムを様々な組織が共用するようになり、計算分野や資源利用効率の異なる多種多様なアプリケーションが実行されるようになった。これにより、運用においてどのようなアプリケーションがどの程度効率的に実行されているかを把握できない利用者が多くなった。その結果、システム全体がどの程度効率良く利用されているかを運用者が把握することは難しくなってきた。

従来のHPCシステムの資源は、CPU、メモリ、ネットワークなどのハードウェアが中心であった。しかし、システムの大規模化に伴って、消費電力およびその変動幅が非常に大きくなってきた。そのため、消費電力を資源の一つと捉えてシステムを運用する重要性が増しており、アプリケーション単位とシステム全体の両面から効率的に利用するための取り組みがなされている^{(2), (3)}。

そこでAICSは、利用者の各アプリケーションの資源（CPU、メモリ、I/O、通信、消費電力）利用状況を採取する取り組みを開始し、富士通はそれらの情報を自動採取する環境を構築した。

(注) 理化学研究所と富士通が共同開発したスーパーコンピュータ。「京」は理化学研究所の登録商標。

アプリケーション情報の自動採取環境

本章では、富士通が構築した各アプリケーション情報の自動採取環境について述べる。環境構築に当たっては、各資源の性能特性指標として採取すべき情報を検討した。HPCシステムで実行されるアプリケーションは演算量とデータ量の観点から大きく二つに分類できることから、CPUに関する性能情報として浮動小数点演算性能を、メモリに関する性能情報としてメモリスループット性能を採取することとした。I/Oおよび通信については、アプリケーションが実行された各計算機に蓄積された情報を採取する。消費電力の採取については、次々章で詳述する。ほかにも、利用者がアプリケーション実行方法をシステムに指示するための実行スクリプトなどを採取する。

これらの情報はアプリケーション実行後にシステムで自動的に収集され、全ての情報をデータベースで一元的に蓄積・管理する。また、採取された情報を集計するための様々な集計ツールを整備し、アプリケーション情報を多面的に分析可能とした。

アプリケーション情報自動採取の仕組みを図-2に示す。

この仕組みによって、各アプリケーションにおける五つの資源の利用状況を運用者が容易に把握できるようになった。

アプリケーション情報の活用による効率化

HPCシステムでは、資源利用状況に偏りや許容上限超過が生じた場合、性能劣化やシステムトラブルなどの問題が発生する可能性がある。例えば、複数のアプリケーションが同一ディスク装置へ大量にアクセスすると、読み書き性能が低下する可能性がある。これを防ぎ、かつシステム全体を効率的に利用するためには、前章までで得られた情報を基に、アプリケーションの実行を適切に制御することが重要である。

五つの資源に関する情報はそれぞれアプリケーション実行制御に有益な情報となり得る。その中でも、過去に「京」において消費電力の上限超過が発生した⁽⁴⁾ことから、AICSは消費電力を対象とした資源利用効率化に力を入れて取り組んでいる。

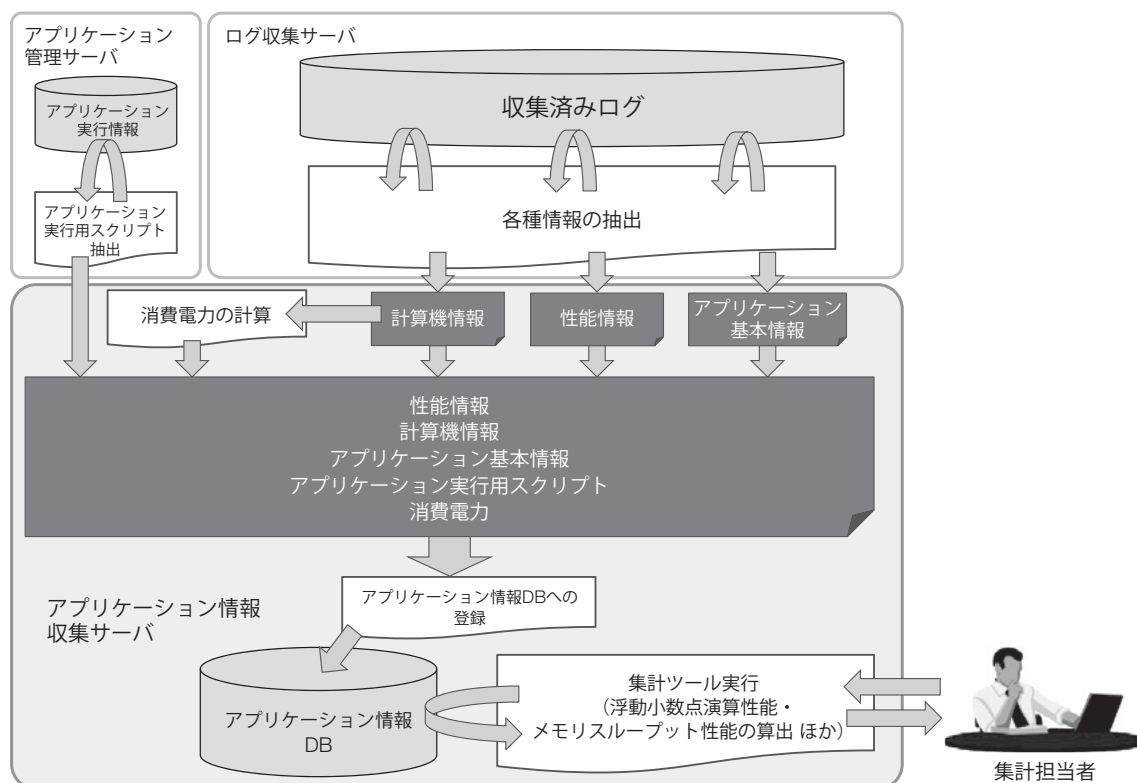


図-2 アプリケーション情報の自動採取環境

そしてこれを実現するため、アプリケーションの消費電力の履歴を基に、今後実行されるアプリケーションの消費電力を実行前に予測する手法を開発した。⁽⁵⁾ 富士通は、本予測手法の開発を支援するとともに、「京」で実行されるアプリケーションの消費電力予測システムを開発した。

以降の章では、本稿では消費電力に関する取り組みについて述べる。

温度情報を用いた消費電力推定環境

「京」は、システム全体と一部のラックの消費電力をそれぞれ測定している。しかし、システム全体やラックを専有して実行されるアプリケーションは少ない。そのため、ほとんどのアプリケーションは何らかの方法で消費電力を推定する必要がある。本章では、AICSが開発した温度情報を用いたアプリケーション単位の消費電力推定手法⁽⁴⁾の概要と、富士通が構築した消費電力推定環境について述べる。

「京」では、一つのシステムボード（SB）に四つのCPUと32枚のメモリを搭載している。CPUは水冷、メモリは空冷であり、温度情報としてはラック吸気温度、SB排気温度、SB水冷温度、およびCPU温度などが採取可能である。これらの情報を用いて、CPU温度変化とメモリの発熱による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－SB水冷温度

SB排気温度変化＝SB排気温度－ラック吸気温度

こうして得られた値を用いて、次の式により消費電力を求める。

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

P はシステム全体の消費電力推定値を、 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。係数 a 、 b 、および c は、実測値と推定値の標準誤差が最小となるように求める。この式において、アプリケーションが実行されたラックやSBの T_{cpu} および T_{air} を用いることで、当該アプリケーションの消費電力を推定できる。

富士通では、図-2に示す計算機情報の一つとして温度情報を収集し、アプリケーションごとの消費電力を推定している。そして、各アプリケーションに関するほかの情報とひも付けられてDBに蓄積される。

消費電力予測システム

大規模HPCシステムの利用者は、特定分野に特化した計算を実行する。そのため、利用者単位で見るとアプリケーションの種類はさほど多くなく、結果として性能や消費電力の特性も似ていることが多い。したがって、過去に実行されたアプリケーションの消費電力の実績を分析することで、今後実行されるアプリケーションの消費電力を予測することが可能である。実際にアプリケーション実行回数の多い利用者の実績を数人抽出し分析したところ、消費電力と使用する計算機台数には相関がみられる場合が多いことが分かった。そこで、計算機台数に応じた消費電力の予測手法として以下の三つを検討した。

(1) 計算機台数の差異による消費電力の変化を予測式で予測する手法

アプリケーション実行に必要な計算機台数と消費電力に比例関係があると仮定し、実績から利用者ごとに作成した予測式（1次式）を用いて、実行前アプリケーションの計算機台数から消費電力を予測する。利用者ごとの予測式と計算機台数の情報のみを用いて予測可能なため、計算に必要なデータ量を減らした高速な予測が可能である。ただし実際には、消費電力と計算機台数が必ずしも比例関係ではないため、本手法では予測誤差が大きくなる傾向がある。

(2) 計算機台数ごとの最大消費電力を用いる手法

実行前アプリケーションと同じ利用者かつ同じ計算機台数の実績のうち、最大の消費電力を当該アプリケーションの予測値とする。本手法では、個々のアプリケーションの消費電力が大きめに予測されるため、システム全体での消費電力の最大値を予測する際に有効である。

(3) 計算機台数ごとの平均消費電力を用いる手法

(2) とほぼ同様であるが、最大値ではなく、平均値で予測する点が異なる。計算機台数ごとに平均消費電力を算出するため、各利用者の実行するアプリケーションが複数の電力特性を持つ場合や、消費電力と計算機台数が比例関係でない場合に、(1) よりも個々のアプリケーションの消費電力を高精度に予測できる。三つの手法のうち、システム全体の予測誤差はこの手法が最も小さい。

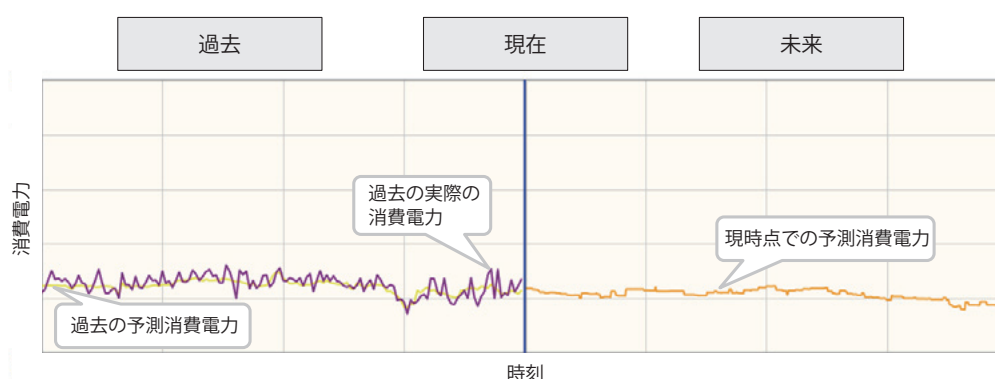


図-3 システム全体の消費電力の予測・可視化結果

富士通では(3)の手法を用いて、近い将来のシステム全体の消費電力をリアルタイムに可視化するシステムを開発した。システム全体の消費電力を予測・可視化した結果の例を図-3に示す。このグラフでは、アプリケーションごとの予測消費電力をそれらの実行予定時間帯で積み上げており、中心が現時刻、それより左側が過去、右側が未来を表す。グラフ上に表示されるデータは過去の予測消費電力と実際の消費電力、および現時点での予測消費電力である。本システムにより、アプリケーション実行による消費電力の大きな変動がいつ頃発生するかを、運用者が把握できるようになった。

む す び

本稿では、大規模HPCシステムにおける課題として資源の効率的な利用を挙げ、その解決に向けたAICSの取り組みと、富士通が開発したソリューションについて述べた。

AICSは今後、本稿で紹介したソリューションで採取した様々な情報を基に、アプリケーションの資源利用効率改善を支援し、システムの運用効率改善や消費電力の削減を目指す。富士通は、今後もAICSの取り組みを支援するとともに、現在開発を進めているソリューションを更に発展させ、ほかのHPCシステムへの展開や次世代HPCシステムの開発・運用に活かしていく。特に消費電力については、これまでに開発した消費電力予測システムを活用し、将来的に個々のアプリケーションの電力特性に応じた実行制御の実現を目指す。これにより、消費電力の効率的な利用が期待できる。

最後に、本ソリューション開発への取り組みに際してご協力いただきました理化学研究所計算科学研究機構運用技術部門の皆様には、この場をお借りして厚く御礼申し上げます。

参考文献

- (1) 黒田明義ほか：「京」での運用効率改善・電力の効率化に向けたジョブ解析システムの開発. 情報処理学会研究会報告, Vol.2016-HPC-156, No.7 (2016).
- (2) 黒田明義ほか：スーパーコンピュータ「京」を用いたアプリケーション性能特性と使用電力の相関解析. 情報処理学会論文誌, ACS, Vol.8, No.4, p.1-12 (2015).
- (3) 秋元秀行ほか：システム消費電力の上限を意識したポスト「京」向けジョブ運用ソフトウェアの実現に向けて. 情報処理学会研究会報告, Vol.2015-HPC-152, No.1 (2015).
- (4) 宇野篤也ほか：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, ACS, Vol.8, No.4, p.13-25 (2015).
- (5) 山本啓二ほか：過去の実行実績を利用したジョブの消費電力予測. 情報処理学会研究会報告, Vol.2015-HPC-151, No.2 (2015).

著者紹介



末安史親 (すえやす ふみちか)

テクニカルコンピューティング・ソリューション事業本部
計算科学ソリューション統括部
スーパーコンピュータ「京」の運用保守業務に従事。



井上俊介 (いのうえ しゅんすけ)

(株) 富士通システムズ・イースト
第三産業ソリューション本部
デジタルエンジニアリング第二事業部
HPC向けアプリケーションの高度化業務に従事。



関澤龍一 (せきざわ りゅういち)

テクニカルコンピューティング・ソリューション事業本部
計算科学ソリューション統括部
スーパーコンピュータ「京」の運用保守業務に従事。