2 京 K computer

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―

応 専

黒川原佳　庄司文由
理化学研究所

2011 年 6 月と 11 月の 2 期連続で TOP500 リストの No.1 を獲得した「京」の特徴は，LINPACK の性能が高いという以外にも，実アプリケーションでも高い実効性能を出せること，さまざまなユーザニーズに対応するための柔軟な運用が可能であること，省電力性能が高いこと，障害に強いことなど，共同利用施設として，多くの研究者や技術者に利用されることを想定して設計されていることである．

本稿では，「京」のシステム概要と，使いやすさを実現するためのさまざまな機能を紹介する．

最近のスーパーコンピュータの技術トレンド

図 -1 は過去 20 年間の TOP500 リスト No.1 の性能値の推移である．平均すると年率で約 1.9 倍速くなっており，ムーアの法則（18 カ月で性能が 2 倍）を超えるスピードで性能向上が続いていることが分かる．一方で，消費電力や設置面積などの制約条件は大きく変わっていないため，性能向上のスピードとバランスするように，低消費電力化と省スペース化が求められている．

そのような背景から，トップエンドのスーパーコンピュータの最近の技術トレンドには，大きく 2 つの流れがある．1 つは超並列化，もう 1 つはアクセラレータ（GPGPU（General-purpose computing on graphics processing units ; グラフィクス用プロセッサによる汎目的計算）など）の採用である．

図 -1 が示していることは，単体性能の向上よりもシステムの全体性能が向上するスピードの方が速いということであり，トップエンドのスーパーコンピュータは，ノード数が増える方向，つまり超並列化の方向

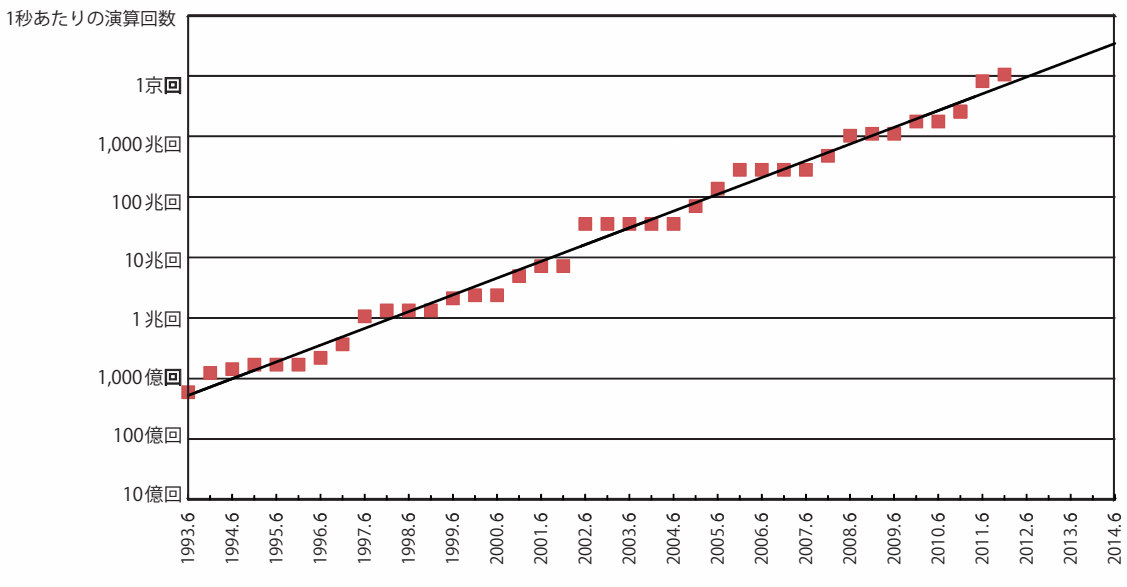

図 -1
TOP500 リスト No.1 の演算性能の推移

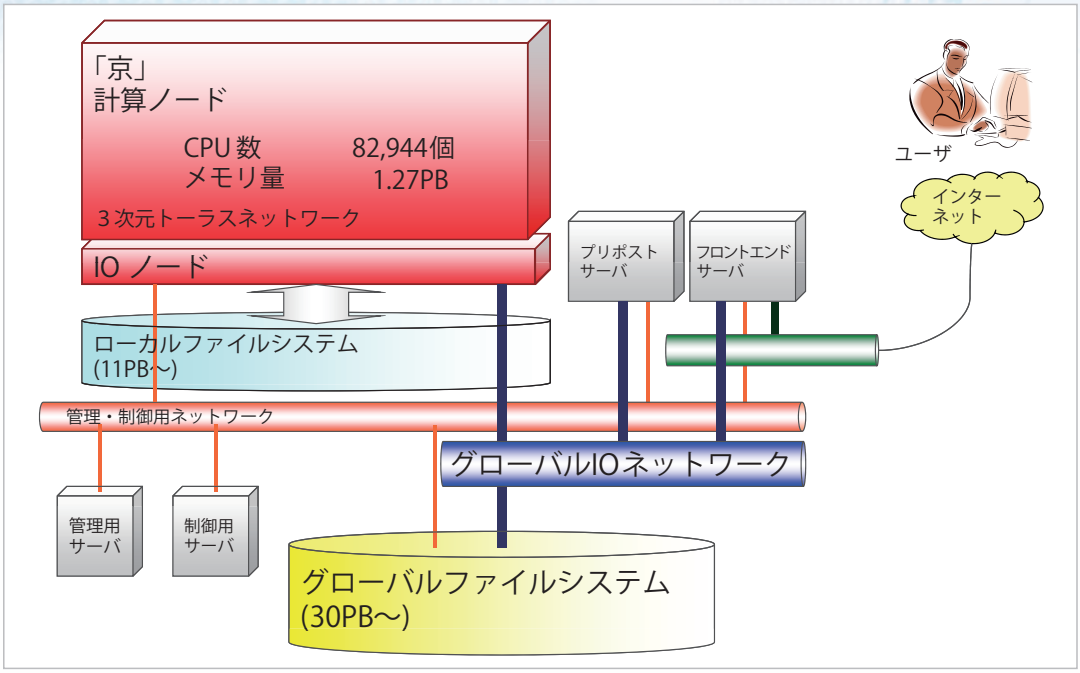

図-2
「京」のシステム構成概要

に進んでいることを意味している．超並列型アーキテクチャの典型は，IBM 社の BlueGene シリーズである．CPU のクロック周波数を抑えて消費電力を削減する一方で，大量のノードを高密度実装することで，省スペースを実現している．しかし，超並列化はアプリケーション開発者に，既存コードに対してアルゴリズムを含む大規模な改変を強いるという難点がある．

もう 1 つの流れはアクセラレータである．もともとグラフィクス用のデバイスだった GPU（Graphics Processing Unit）を汎用的な計算に活用する試みは以前からあったが，NVIDIA 社が CUDA（Compute Unified Device Architecture）と呼ばれる開発環境を整えたことにより急速に普及した．GPU は CPU に比べ構造がシンプルで大量の演算器を搭載するため，演算密度が低くなる複雑な計算には向かないが，科学技術計算で頻繁に現れる同じ計算を大量に反復するような場合に高い効果を発揮する．また，シンプルな分，消費電力も少なく，実装密度を上げることができる．ただ，応用範囲が一部のアプリケーション領域に限定されているのが現状である．

「京」のシステム構成概要

図-2 に「京」のシステム構成の概要を示す．

システムは，大きく 4 つのパートから構成される．計算機の心臓部である計算ノード群，ローカルファイルシステム，グローバルファイルシステム，そしてフロントエンドサーバなどの周辺機器群である．

「京」は 8 万個以上の計算ノードを持ち，システム全体では 1PB 以上のメモリ容量を有する．各計算ノード間は，「Tofu（Torus fusion）インターコネクト」と呼ばれる 6 次元メッシュ／トーラスネットワークで物理的に接続される．

計算ノード群の傍らには，ジョブ実行時のディスク IO のための一次領域としてのローカルファイルシステムが配置されている．さらに，計算ノード群とグローバル IO ネットワークで接続されたグローバルファイルシステムがあり，ここにユーザのホーム領域や保存するデータ用領域が置かれる．

アプリケーションの実行性能を高めるための工夫

CPU

「京」で採用した CPU である SPARC64™ VIIIfx は，8 個のプロセッサコアを備えたマルチコア構成となっている．

さらに各コア内に SIMD（Single Instruction

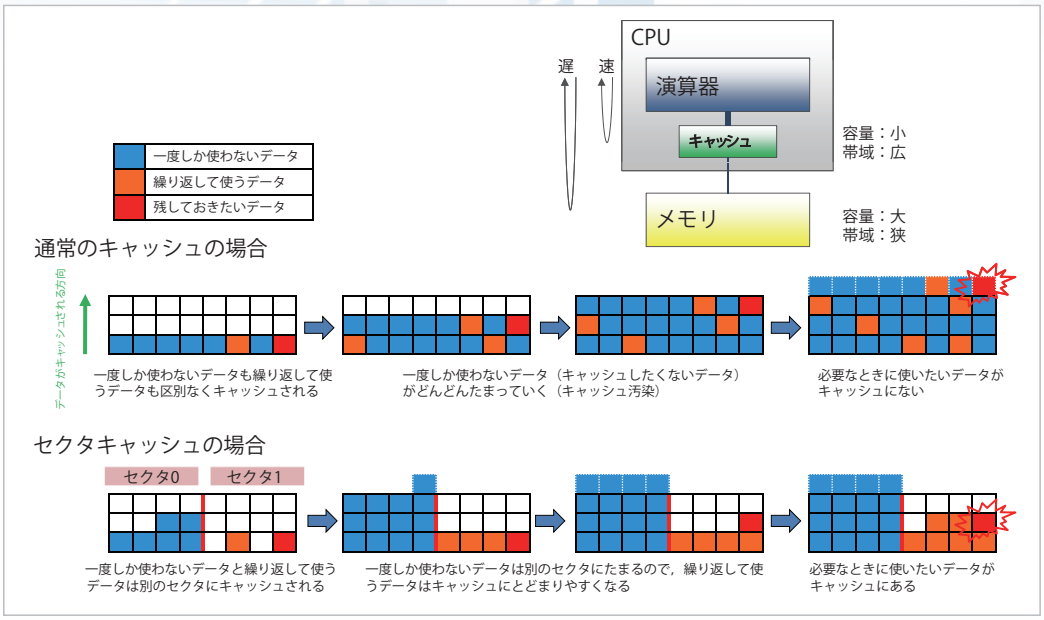

図 -3
セクタキャッシュの動作イメージ

Multiple Data）演算機構，コアあたり 256 本の浮動小数点レジスタ，6MB の共有キャッシュ，プロセッサコア間の同期を高速に実行するためのハードウェアバリア機構，セクタキャッシュ機構など，さまざまな新機能を盛り込んだ．その結果，倍精度浮動小数点演算でコアあたり 16GFLOPS，CPU チップあたり 128GFLOPS という高性能を達成している．特にレジスタについては，x86 アーキテクチャなどの CPU と比較しても倍以上の本数を備えており，コンパイラによる最適化の自由度が高く，さまざまな演算パターンにおいて高い実効性能を得やすい構成といえる．

セクタキャッシュは，本 CPU で初めて採用された機能で，ユーザがキャッシュメモリをソフトウェアによって制御できるようにしたものである．従来，キャッシュメモリの動作をユーザが直接コントロールすることはできず，ハードウェアで自動的に制御される．そのためユーザには，キャッシュの存在を特に意識しなくても，その恩恵を受けられるというメリットがある反面，比較的長期にキャッシュしたいデータとキャッシュが不要なデータを分けて扱うなどの細かな制御ができないという問題があった．そこで SPARC64^{TM} VIIIfx では，図 -3 のようにキャッシュメモリを 2 つの領域（セクタ）に分け，どちらの領域にデータをキャッシュするのか，ユーザが指定できるようにした．

たとえば，片方の領域（セクタ 0）を一度しか使わないデータ用，もう片方（セクタ 1）を繰り返し使うデータ用というように使い分けることで，繰り返し使うデータをキャッシュ上にとどまりやすくすることができるため，キャッシュの利用効率を高める効果が期待できる．

セクタキャッシュ機構は，コンパイラ判断による利用と，ソースコード中の指示行による指定の両方が可能である．

●インターコネクト

本システムのもう 1 つの重要な構成要素である CPU（ノード）間を接続するネットワークについては，「Tofu インターコネクト」と呼ばれる高性能・高信頼性の独自ネットワークを開発した．Tofu インターコネクトの構成は，ノード間を直接接続する直接結合網で，物理的には 6 次元メッシュ／トーラスの接続である．

ユーザはこのネットワークを論理的に 3 次元トーラスのネットワークとして利用できる．図 -4 に示すように，各ノードは 3 次元の各方向に対して，それぞれ 5GB/s（双方向）の帯域幅を持つリンクで接続されている．

ICC（Inter Connect Controller）は，Tofu インタ

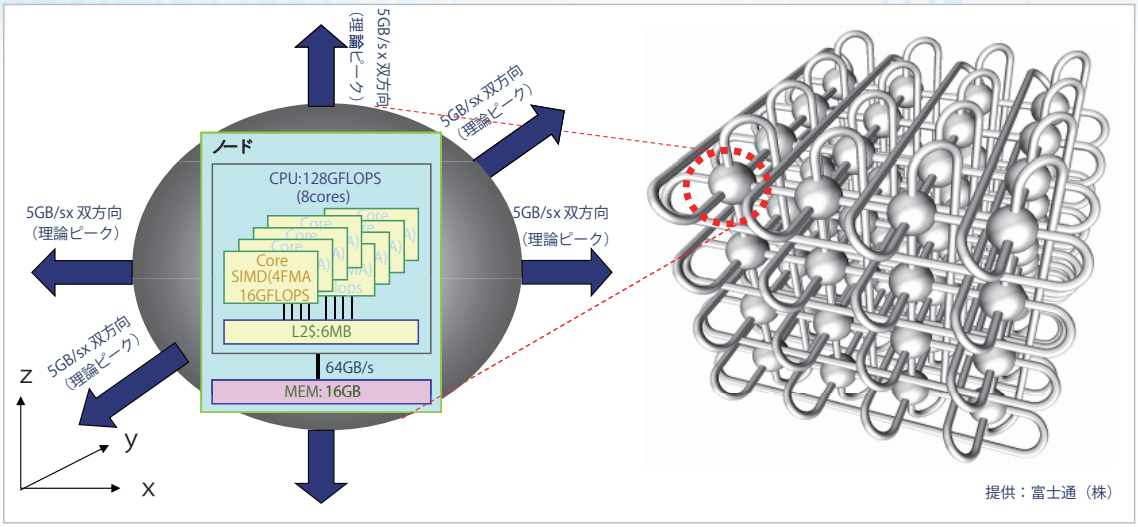

図-4
3次元トーラスのイメージ

ーコネクトを構成するためのLSIチップで，各ノードに1つずつ実装されている．ICCは4つのDMA（Direct Memory Access）エンジンを有し，4方向の同時通信およびRDMA（Remote Direct Memory Access）通信を行うことができる．Tofuインターコネクトの6次元メッシュ／トーラスは，6次元のうち，3つの次元がトーラス，残りの3つの次元がメッシュとなっている．各ノードから見るとトーラスの次元に対して正方向と負方向の2リンク，メッシュも端のノードを除けば次元に対して2リンク必要となるが，2つの次元では2ノードだけでのメッシュとなり，それらは1リンクずつで結合されているため，合計10本のリンクが必要となる．ICCは全部で10本の外部リンクを有している．これに加え，4本のCPU接続リンク，そして，PCI-Expressポートを有する．ICCの全スイッチング容量は100GB/sで，バーチャルカットスルーによるパケット転送が可能となっていて，低遅延で広帯域なインターコネクトネットワークを構成することができる．

消費電力を削減するための工夫

「京」のシステム開発においては，高性能と低消費電力の両立が重要なポイントであった．そのため，既存の計算センターの電力設備の調査結果や，将来の技術動向等を考慮して，冷却設備や周辺装置などを含むすべての装置の消費電力の合計を「演算性能10PFLOPSで30MW以下」という目標を設定した．

この目標を達成するためのキーとなるのは，主要な構成要素であるCPUの高性能化と低消費電力化である．CPUを製造する上で基本となる半導体プロセスには，当時最先端の45nm CMOSプロセスを採用することで，高性能化と低消費電力化の両立を目指すとともに，消費電力に直接影響する動作周波数をやや低めの2GHzに設定した．さらに，命令実行に必要のない回路を動的にストップできる機能など，設計上の工夫により動作時の消費電力の削減を目指した．また，冷却効率のよい水冷方式を採用し，CPUのジャンクション温度を30℃にまで下げることで，さらなる消費電力と故障率の低減を実現した．

これらの低消費電力化技術や前述の高速化技術によって，CPUあたりの理論演算性能128GFLOPSに対して，消費電力58Wを達成している．ワットあたりの演算性能は2.21GFLOPSで，これは汎用CPUとしては現在でもトップクラスの数値である．

システム全体の消費電力については，システム評価中に実施した測定によれば14～15MW程度であり，当初の目標である10PFLOPSで30MW以下を十分にクリアしている．

高信頼性を実現するための工夫

「京」の開発プロジェクトの目標は，LINPACKで10PFLOPSを達成することだけではない．完成

後に，共用施設として全国の研究者がいつでも必要なコンピュータ資源を安定して利用できるシステムとして，必要十分な機能を具備していなければならない．その一方で，本システムは，主要部品のCPU チップだけでも 8 万個以上，システム全体では，100 万個以上もの部品から構成される超大規模システムである．個々の部品の信頼性を高めることはもちろん重要であるが，これほどまでに部品点数が膨大になると，それだけでは対応しきれない．そこで，システムとしての稼働率を高めるために，「壊れない」，「壊れてもすべてが止まらない」，「壊れてもすぐ直せる」システムでなければならず，そのための機能を備える必要がある．

システムの信頼性を向上させる上で最も重要なことは，CPU 自身の信頼性を高めることである．SPARC64™VIIIfx のジャンクション温度の 30℃は，パソコンやサーバ等で使用されている他の CPU と比べて格段に低く，85℃で駆動した場合と比較して，数十倍寿命が伸びるという試算もあり，故障率低減に大きな効果があると期待される．また CPU 内の回路には徹底した「エラー検出機能」を備えている．万一エラーを検出した場合には，「エラー訂正機能」やエラーを検出した命令を自動的に再実行する「命令再実行機能」により，一般的な CPU であればシステムのダウンにつながりかねない一時的なエラー（間欠故障）が起こっても，自動的に再実行され，システムの動作には影響が出ない．さらに，「命令再実行機能」でも救えない永久的な故障の場合は，故障した CPU をシステムから切り離した上で，システムの残りの部分の運用を継続することができる．

また，CPU 間を接続するネットワークにも信頼性を高めるための機能を実装した．一般的な 3 次元トーラスネットワークのような直接網ネットワークでは，ある計算ノードが故障すると，その影響が故障部分だけにとどまらずに周辺に及びやすく，広範囲の計算ノードが利用できなくなるケースが多い．そこで今回開発した Tofu インターコネクトでは，冗長なリンクを代替経路として活用することで，このような事態を避けることができるようになっている．そのため，故障が発生してもその影響は最小限の範囲に限定され，故障個所以外のシステムは運用を継続することができる．結果として，システムの稼働時間を増やすことができる．

これらの機能に加えて，ログインサーバ，管理サーバなどの各種サーバの二重化，データパスの二重化などによって，システム全体の信頼性，可用性の向上を図っている．

運用性と利便性を高めるための工夫

本システムは，多くの研究者・技術者がストレスなく使えなければならない．そこで，大規模システムとして運用中の地球シミュレータの実績や経験も参考にして，運用性に優れ，使いやすい共用施設のシステムを目指して，先端的な技術を採り入れたシステム開発を行った．

これまで述べたようなハードウェアが持つ高い機能を有効に活用し，その性能を最大限に発揮させるために，CPU 内の自動並列化をサポートした最適化コンパイラ，デバッガ，性能チューニングツール，数値計算ライブラリなどを開発した．このシステムでは，アプリケーションの流通や既存システムとの互換性などを考慮し，科学技術計算の分野で広く使われているプログラミング言語である Fortran，C/C++，さらに並列プログラミングの標準ライブラリである MPI（Message Passing Interface）とデータ並列言語の XPFortran をサポートしている．

また，運用性を高めるには，大量のデータを効率よく処理し，CPU やメモリを効率的に管理できることが重要である．そのため，グローバル／ローカルファイルシステムで用いている並列分散ファイルシステムや，ジョブが利用する CPU やメモリ，Tofu インターコネクト，階層型ファイルシステムなどの資源を効率良く配分し，アプリケーションを円滑に実行するジョブスケジューラを開発した．

本システムは，汎用のスカラプロセッサを用いたシステムであることに加え，オペレーティングシステムとして広く普及した Linux を採用している．ユ

ランキング	国	システム名	演算性能（PFLOPS）	実行効率（%）	1ワットあたりの演算性能	実行時間（時間）
1	Japan	K computer	10.510	93	824.56	29.47
2	China	Tianhe-1A	2.566	55	635.15	3.37
3	US	Jaguar	1.759	75	253.07	17.27
4	China	Nebulae	1.271	43	492.64	1.91
5	Japan	TSUBAME2.0	1.192	52	852.27	2.40

表 -1
第 38 回 TOP500 リストの上位 5 位

ーザに標準的な利用環境を提供することで，共用施設として幅広い計算科学の分野で利用しやすいシステムになると考えている．これにより，すでに開発されたアプリケーションの「京」へのポーティングが容易になるなど，より多くの応用分野でのシステム利用が促進されると期待できる．

「京」の性能実証

●TOP500

冒頭にも記したとおり，「京」は，2011 年 6 月と 11 月の 2 期連続で TOP500 リストの No.1 を獲得した．その結果を詳しく考察し，設計時の目標がどの程度達成できたかを紹介する．

表 -1 に 2011 年 11 月に発表された第 38 回 TOP500 リストの上位 5 位までを示す．

演算性能についてはすでに述べたように，「京」は，10PFLOPS を達成した初めてのスーパーコンピュータとなった．次に実行効率であるが，これは，設計上の演算性能に対して，実際にどの程度の性能が達成できたかを表す指標である．表 -1 にあるように，「京」の実行効率は 93% と，他のマシンと比較しても突出して高く，きわめて優秀といえる．これは，レジスタが強化されたことや，セクタキャッシュの寄与が大きいと考えている．

また，低消費電力という点でも，上位 5 位の中では第 5 位の TSUBAME2.0 にわずかに及ばないもののトップクラスであることが分かる．一般に規模が大きくなると通信のための電力が余分に必要となることから，電力性能比は悪化することを勘案すると，「京」の電力性能はきわめて高いといえよう．

さらに，実行時間も特筆すべき点である．LINPACK は計算機にきわめて高い負荷をかけるベンチマークであり，耐久試験を行っているとも言える．その意味で，「京」の約 30 時間という実行時間は，他のマシンが軒並み数時間，長いものでも 17 時間程度であることと，規模の違いを考慮すると，「京」の信頼性は群を抜いていることが分かる．

このように，LINPACK 性能 10PFLOPS を達成したほかにも，他のマシンよりも圧倒的に高い実行効率，低消費電力，高信頼性を同時に実証することができた．

●HPC チャレンジ賞

スーパーコンピュータの性能をより多角的に評価するためのベンチマークとして，近年 HPC チャレンジベンチマークが注目されている．HPC チャレンジベンチマークは，28 項目にわたりスーパーコンピュータの性能を評価することができるが，その中の特に重要な 4 項目の No.1 マシンを表彰するのが HPC チャレンジ賞である．

表 -2 に示すとおり，「京」は，2011 年の HPC チャレンジ賞において，4 部門すべてで 1 位を獲得することができた．

なおこの性能値は，「京」の約 2 割の資源を用いて計測されたものであり，全資源を使った場合は，さらに性能値が向上すると見込まれる．

この結果により，「京」が LINPACK だけではなく，より幅広い分野のアプリケーションに柔軟に対応できるポテンシャルを持っていることを示すことができた．

参考までに，過去に 4 部門すべてで 1 位を獲得した事例は，2005 年と 2006 年の BlueGene/L 以外にはなく，いかに「京」が他のスーパーコンピュータに対して突出した性能を有しているかがお分かりいただけると思う．

Global HPL	性能値 (TFLOPS)	システム名	設置機関
1 位	2,118	K computer	理研 AICS/ 日本
2 位	1,533	Cray XT5	オークリッジ研究所 / 米国
3 位	736	Cray XT5	テネシー大学 / 米国

Global Random Access	性能値 (GUPS)	システム名	設置機関
1 位	121	K computer	理研 AICS/ 日本
2 位	117	IBM BG/ P	ローレンスリバモア研究所 / 米国
3 位	103	IBM BG/ P	アルゴンヌ研究所 / 米国

EP STREAM (Triad) per system	性能値 (TB/ s)	システム名	設置機関
1 位	812	K computer	理研 AICS/ 日本
2 位	398	Cray XT5	オークリッジ研究所 / 日本
3 位	267	IBM BG/ P	ローレンスリバモア研究所 / 米国

Global FFT	性能値 (TFLOPS)	システム名	設置機関
1 位	34.7	K computer	理研 AICS/ 日本
2 位	11.9	NEC SX- 9	海洋研究開発機構 / 日本
3 位	10.7	Cray XT5	オークリッジ研究所 / 米国

Global HPL：大規模連立一次方程式の解を求めるベンチマークプログラム．主にシステムの演算性能を評価．
Global Random Access：任意のノード間でランダムな通信を行うベンチマークプログラム．主にシステム全体の通信性能を評価．
EP STREAM（Triad）per system：乗加算演算の反復計算を行うベンチマークプログラム．主にメモリに対するリードライト性能を評価．
Global FFT：高速フーリエ変換を行うベンチマークプログラム．主にシステム全体の通信性能を評価．

表 -2　2011 年 HPC チャレンジ賞の 4 部門の上位 3 位まで

●ゴードン・ベル賞

ゴードン・ベル賞は，実際のアプリケーションの実効性能と計算科学の成果に対してアメリカ計算機学会が授与する賞である．

筑波大と理研および富士通の研究チームは，「「京」による 100,000 原子シリコン・ナノワイヤの電子状態の第一原理計算」というテーマで 2011 年のゴードン・ベル賞の最高性能（Peak Performance）賞を受賞した．これにより，「京」がベンチマークだけでなく，実際のアプリケーションでも高い性能を発揮することを示すことができた．

このように，「京」は，当初の性能に対する設計目標をすべて達成したと同時に，幅広い計算科学分野のさまざまなアプリケーションに対応できることが実証された．

現在は，試験利用という形で，40 本以上のアプリケーションの最適化とチューニングが先行して進んでいる．そのうちの半数近くのアプリケーションが，数万ノード，数 PFLOPS 規模で十分な並列化効率を達成している．これは，「京」の使いやすさ，チューニングのしやすさが寄与していると考えている．また，一部のアプリケーションからは，科学的な成果も出始めており，今後さらに多くのアプリケーションでさまざまな成果が出てくることが期待される．

「京」の周辺装置および外部接続

「京」の周辺装置には大きく分けて 3 つの要素がある．1 つ目はローカル／グローバルのファイルシステムを構成するサーバやストレージ装置，2 つ目にはさまざまな管理や制御系のサーバ群，最後にそれらを繋ぐネットワークである．

「京」ではローカル／グローバルともにファイルシステムに FEFS（Fujitsu Exabyte File System）を用いている．並列ファイルシステムのメタデータを処理するための MDS（Meta Data Server）には，ローカル／グローバルともに高信頼性サーバである PRIMEQUEST E1800 を 2 台で冗長構成として，ストレージ装置へのパスも冗長化されている．また，OSS（Object Storage Server）は，ローカルでは「京」の IO ノード（2,592 台）を用い，グローバルでは IA サーバである PRIMERGY RX300S6（90 台）を用いている．各 OSS から ストレージ装置（ローカルは 2,592 台，グローバルは 720 台）へのパスは

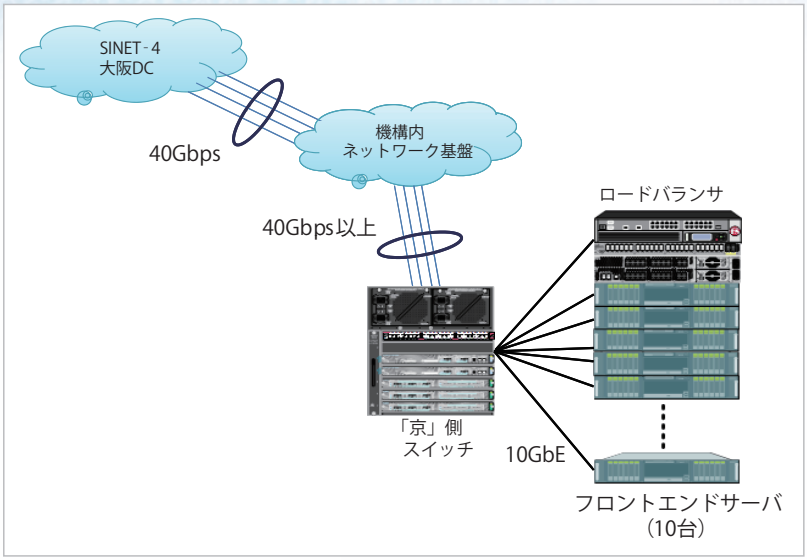

図-5 「京」へのアクセスのイメージ

冗長化しており，どれか1台のMDSあるはOSSがダウンしても，パスが切れても，24時間365日のサービスが提供可能なハードウェア構成となっている．

「京」はさまざまな役割を持ったサーバ群を用いて，システムの制御や管理を行っている．特に「京」は，管理対象である計算ノードやIOノードが8万台を超えるため，ノード管理やジョブ管理のサーバ群は，「京」をクラスタと呼ばれる分割された単位で管理しており，各クラスタに二重化した管理サブサーバを配置し，それらを管理する上位サーバにも二重化した主管理サーバを配置する階層化構成となっている．管理サブサーバはノードのさまざまな監視や効率的なジョブ操作（プログラムの並列起動や統計情報の収集など）を行っている．

「京」では制御・管理系のネットワークにEthernet，およびファイルシステムのネットワークにInfiniBand（QDR）を用いている．Ethernet系は「京」の管理に用いる系と，サービスプロセッサや各種サーバ，またストレージ装置のハードウェア情報管理に用いる制御系のネットワークがそれぞれあり，GbEの総ポート数は2万ポートを超える．また，EthernetとInfiniBandのスイッチの配置は，各クラスタで区分けされている．各装置からスイッチのパス，スイッチ自体もすべて冗長構成となっており，どこか1つのパスが切れてもすべての経路は確保される構成となっている．

図-5に「京」へのアクセスのイメージを示す．

「京」を外部から利用するためのサーバは，フロントエンド（IAサーバ10台）で構成されており，利用者が意識することなく，自動的にロードバランスされる．フロントエンドから外部（Internet）へは，理研計算科学研究機構内ネットワーク基盤を通し，学術ネットワーク基盤であるSINET-4堂島DCまで40Gbpsの帯域により接続されている．

まとめ

冒頭でも触れたとおり，「京」の特徴は圧倒的な演算スピードだけではない．実アプリケーションでも高い実効性能を出せること，さまざまなユーザニーズに対応するための柔軟な運用が可能であること，省電力性能が高いこと，障害に強いことも大きな特徴であり，共用施設であることを考えれば，むしろこちらの方が重要である．今年の秋には共用が開始されるが，本稿で紹介した「京」の特徴が有効に機能し，新たな知見やブレークスルーがもたらされると期待される．

参考文献

1) 横川三津夫，庄司文由：京速コンピュータ「京（けい）」とは何か？世界最速レベルの計算性能を目指して，原子力学会誌，Vol.52，No.12（2010）．
2) 庄司文由：京速コンピュータ「京（けい）」とその利用，応用物理学会誌，Vol.80，No.7（2011）．

（2012年4月27日受付）

黒川原佳（正会員） motoyosi@riken.jp

次世代スパコン開発実施本部　開発グループシステム開発チームの開発研究員．

庄司文由 shoji@riken.jp

次世代スパコン開発実施本部　開発グループシステム開発チームのチームリーダ．