8 京 K computer

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―

応 専

関口芳弘　庄司文由　塚本俊之

理化学研究所

施設の概要

スーパーコンピュータ「京」の施設は，神戸市のポートアイランドの南に位置する計算科学研究機構に設置されている．この施設は図 -1 に示す 4 つの建物で構成されており，「京」の性能を常時保障できる構造，安定運用に必要な設備を持つとともに，研究交流や多様な知識の融合を促進できる研究・教育環境の拠点として整備されている．

「京」が設置されている計算機棟と，研究者が常駐する研究棟は免震構造になっており，震度 6 強レベルの地震でも激しい揺れを減免できるように地震対策が施されている．

「京」の計算機本体は，筐体数で 864 台，筐体間を接続する通信ケーブルが合計 20 万本（総延長 1,000km）にもなる超大規模システムであるため，計算機棟の構造は以下を満足する設計となっている．

- 床の耐荷重：各筐体の重量は約 1.5t
- フロア全体の耐荷重：1,300t 超
- 筐体設置レイアウトの自由度の確保
- 通信ケーブル長の短縮化・均一化

これらを満たした結果として，「京」の計算機筐体は計算機棟の 3 階に，グローバルファイルシステムは計算機棟の 1 階に設置されており，それぞれの下の階には冷却のため空調機が設置される構造となっている（図 -2）．

「京」は大規模システムの中でトップクラスの低消費電力（2011 年 6 月版 GREEN500 の省エネスパコンランキング第 6 位）を誇るが，全体では 10,000kW を超える電力を必要とする．そのため，

図 -1　計算科学研究機構

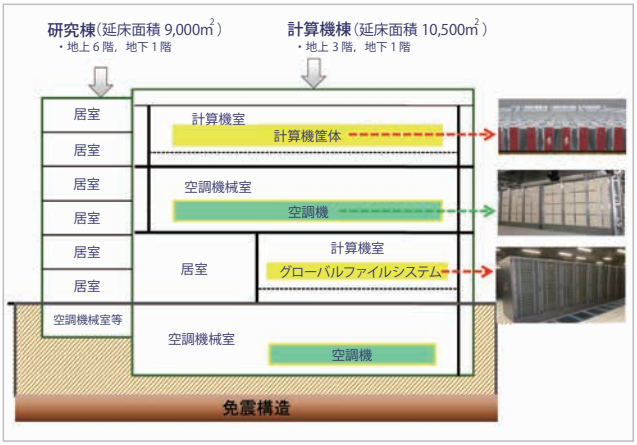

図 -2　計算機棟・研究棟

「京」を安定運用するためには確実な電力供給および冷却システムが重要な要素となっている．

まず，確実な電力供給だが，施設全体で必要となる電力は以下の 3 つの供給源で賄われている．

- 関西電力からの受電電力：11,000kW
- CGS☆1 発電機：5,000kW × 2（平均）
- 太陽光発電パネル：50kW

☆1 CGS：CoGeneration System（電熱併給システム）．

CGS は瞬時電圧低下（以下，瞬低）対策に利用されているだけでなく，排熱を冷暖房に有効活用しており，太陽光発電パネルと合わせて，環境に配慮した省エネシステムとなっている（図 -3）．

「京」は 88,128 個の膨大な数の CPU で構成されており，発生する熱量も多大である．そのため，CPU，通信用 LSI（Inter Connect Controller，以下 ICC）の冷却には水冷方式を，メモリや電源等の冷却には空冷方式を採用したハイブリッド形式（図 -4）で設計されている．この冷却方式の採用により「京」の故障率を大幅に減らすことができ，結果として安定運用に大いに貢献している．

上記の内容について，次章以降にその特徴を紹介する．

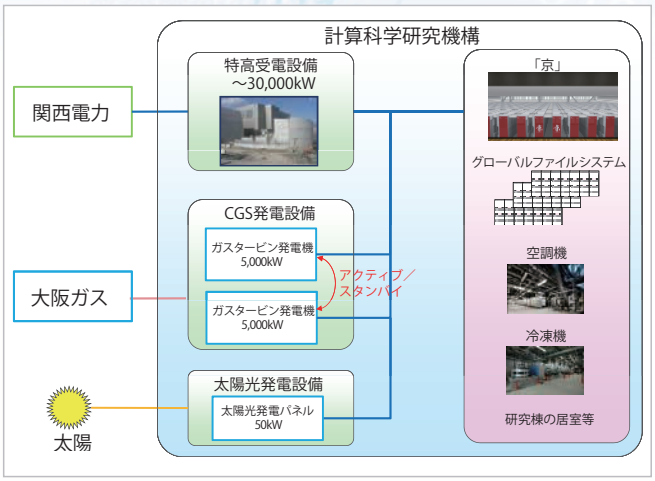

図 -3　3 系統の電源供給

構造

「京」本体が置かれる計算機室は 60m × 50m= 3,000m^2 に及ぶ広大な面積を有している．これは一般的な体育館の約 2 倍の広さである．この広大な計算機室を「柱なし」で構成した（図 -5）．柱があると，筐体の設置間隔を均一にできず，筐体と筐体を接続する通信ケーブルの長さに不均一をもたらし，結果として「京」全体の性能を落とす恐れがあった．そこで，柱をなくすことで筐体配置上の制限をなくし，筐体を自由にレイアウトできるようにした．

無柱とするために，2 つの技術を用いた．1 つは「免震構造」である．地震が来ると，一般の建物は地震によってひねられる．ひねられて建物が壊れること

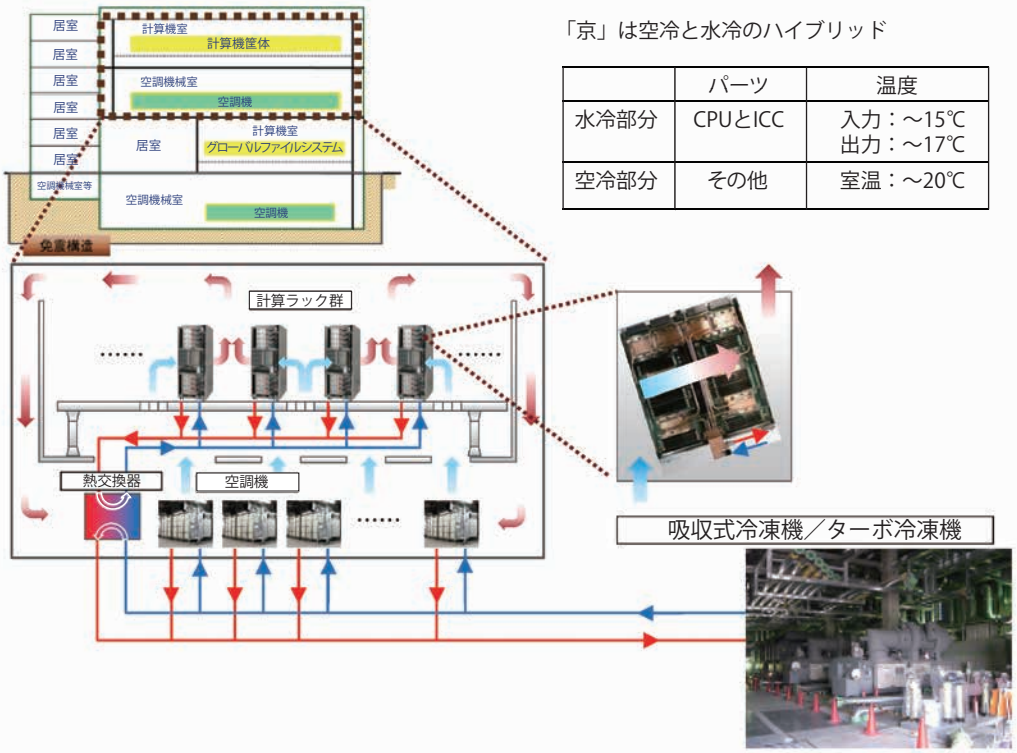

図 -4　冷却システム

図 -5　無柱の 3 階計算機室

図 -6　積層ゴム免震装置

を防ぐためには，耐震構造か免震構造を採用する必要がある．耐震構造では，柱や梁を太くしたり，柱の本数を多くしたり，筋交い＝ブレースをたくさん入れたりしなければならない．ところが免震構造にすると，地震動に対してほぼ水平に揺れるだけになるため，ひねりに抗しなくてもよくなり，柱やブレースを減らすことができるのである．

計算機棟は，49 台の積層ゴム免震装置で支えられている（図 -6）．この免震装置が中心から最大 70cm 変位することによって，地震動を建物に伝えないようにする．免震装置により計算機棟内での最大加速度は 200gal 以下に抑えられ，設置された筐体が転倒する，ずれるなどの問題を引き起こさない．震度 5 程度の中地震では無被害，震度 6 強の大地震でも小破程度に抑えられ，建物の主要な機能は確保される．

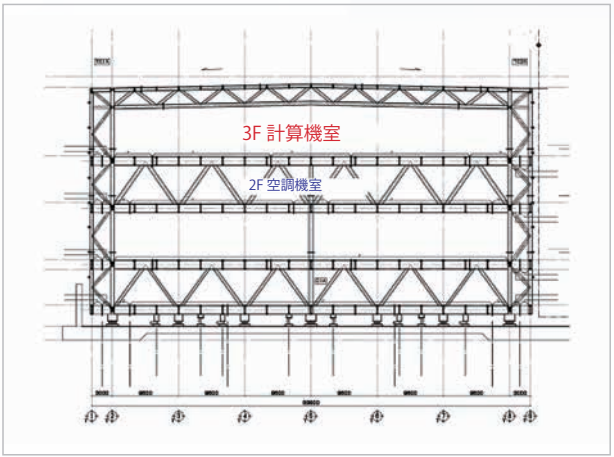

図 -7　橋梁技術によるトラス構造

もう 1 つは「橋梁技術」を用いたことである．「京」本体を収容する計算機室は 3 階となる．高い位置に作られる大きな計算機室を無柱化するために，橋梁に使われる技術を用いた．図 -7 に計算機棟の断面を示す．3 階計算機室と 2 階空調機室の部分に着目していただきたい．2 階空調機室に三角形のトラス構造を見ることができるであろう．計算機室はこのトラスによって支えられている．つまり，計算機室は幅 60m 長さ 50m の「トラス橋」なのである．トラスの上部を無柱の計算機室とし，トラス階を空調機械室としても使用している．

計算機室の床下は，深さ 1.5m のフリーアクセスとしている．この内部に，電力ケーブル，通信ケーブル，CPU 冷却水配管が収容されている．また，フリーアクセス内部は，空調冷気のサプライチャンバ☆2 としても機能している．

電力設備

電力設備における最大の特徴は，データセンターや電算機センターに必須の設備と思われる無停電電源装置（以下，UPS）がないことである．近年電算機センターでは，超並列スーパーコンピュータが設置され，計算能力の向上とともに消費電力も指数関数的に増加している．商用電源の停電ならびに瞬低

☆2　サプライチャンバ：冷却用空気の供給空間．

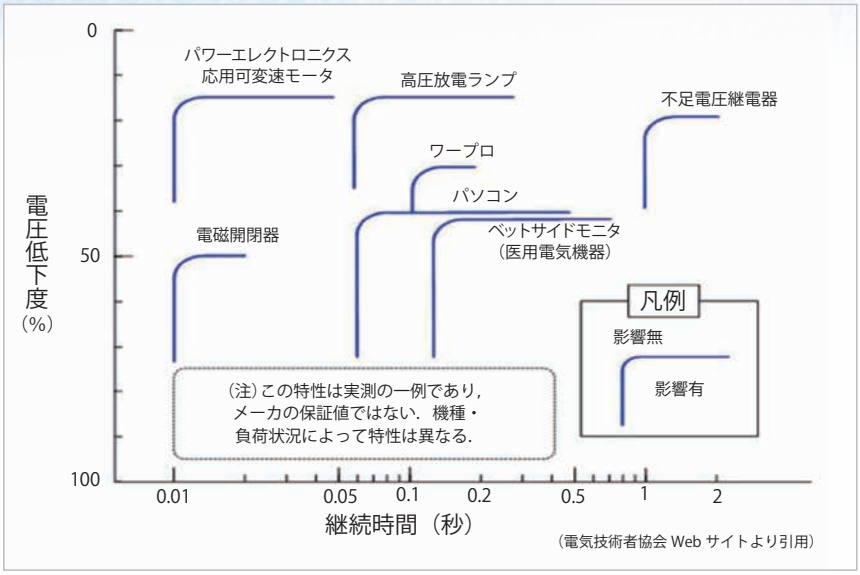

図 -8　機器瞬低耐量

対策に対してこれまで大多数の施設において UPS で保護してきたが，スパコンの消費電力の増加とともに必要 UPS 容量も増加してしまい，そのイニシャルコストの負担や設置場所の確保が困難となっている．また，UPS の電力変換効率は 95% 程度であり，大規模システムでの電力損失も無視できないほどの電力になる．さらに，UPS の蓄電池は 5 年ないし 7 年ごとに交換が必要であり，大容量となれば運用費を圧迫する．鉛蓄電池であれば環境への悪影響も避けられない．よって計算機システムだけで 10,000kW 以上の大電力を消費する「京」において，UPS の採用は困難であった．

しかし瞬低はたびたび発生するものであり，それは計算機を停止させてしまう恐れがある．いったん計算機システムが停止してしまうと，再度立ち上げるのに数時間から 1 日程度を要する．その間，計算機の利用ができなくなり，稼働率を下げる．よって，瞬低対策は計算機の稼働率を上げるために必須であるが，UPS を用いない方法で行う必要がある．

瞬低の時間ならびに電圧低下によって影響を受けるかどうかは，機器によって異なる．一般的には図 -8 で示されるようになっている．

図によればパソコンの場合，電圧低下度 60%，継続時間 0.06 秒程度を超えると瞬低により停止してしまう．逆に言えば，その程度までの瞬低であれば影響を受けないのである．「京」の瞬低耐量についてのデータは取られていないが，計算機システムにおいても規模が大きいほど瞬低耐量も大きくなると考えられる．これは，計算機はコンデンサの塊であり，稼働中に計算機自身に充電される電荷量が多いため，それによって自身の電圧をしばらく保持できるからである．したがって計算機自体は軽微な瞬低では停止しない．

しかし，計算機を冷却する機器の瞬低耐量はあまり大きくない．これは，冷却系機器へ電力を供給する動力盤内の電磁開閉器の励磁を保持できず，開放してしまうためである．そのため，計算機は瞬低に耐えても冷却機器が耐えられず，結果として計算機が停止してしまう可能性もある．

以上の状況から，「京」の電源設備においての瞬低対策は 2 つのレベルに分けて検討した．第 1 のレベルは絶対に停電させてはいけない部分である．ここは CGS 発電設備を UPS として使うことにより保護することにした．具体的には，データを保存するグローバルファイルユニット，研究者が研究を行う研究棟は CGS と高速限流遮断機を組み合わせることによって UPS と同等の無停電保護を行っている．

第 2 のレベルは，短時間の瞬低だけを保護する部分である．ここは「京」自身が持つ電荷によって保持可能な時間だけ保護することにした．具体的には，冷却系の電磁開閉器に，1 秒間閉路状態を保持できる遅延開放式電磁開閉器を採用し，瞬低耐量の強化を行っている．

この対策により，「京」に 1 秒間以内の軽微な瞬低が発生しても冷却系は停止せず，計算機システムの継続稼働が可能となっている（図 -9）．

施設運用開始後，現在までに商用側停電事故はまったくないが，瞬低は数回起こっている．これまでのところ一度だけ電圧低下率 53.2%，継続時間

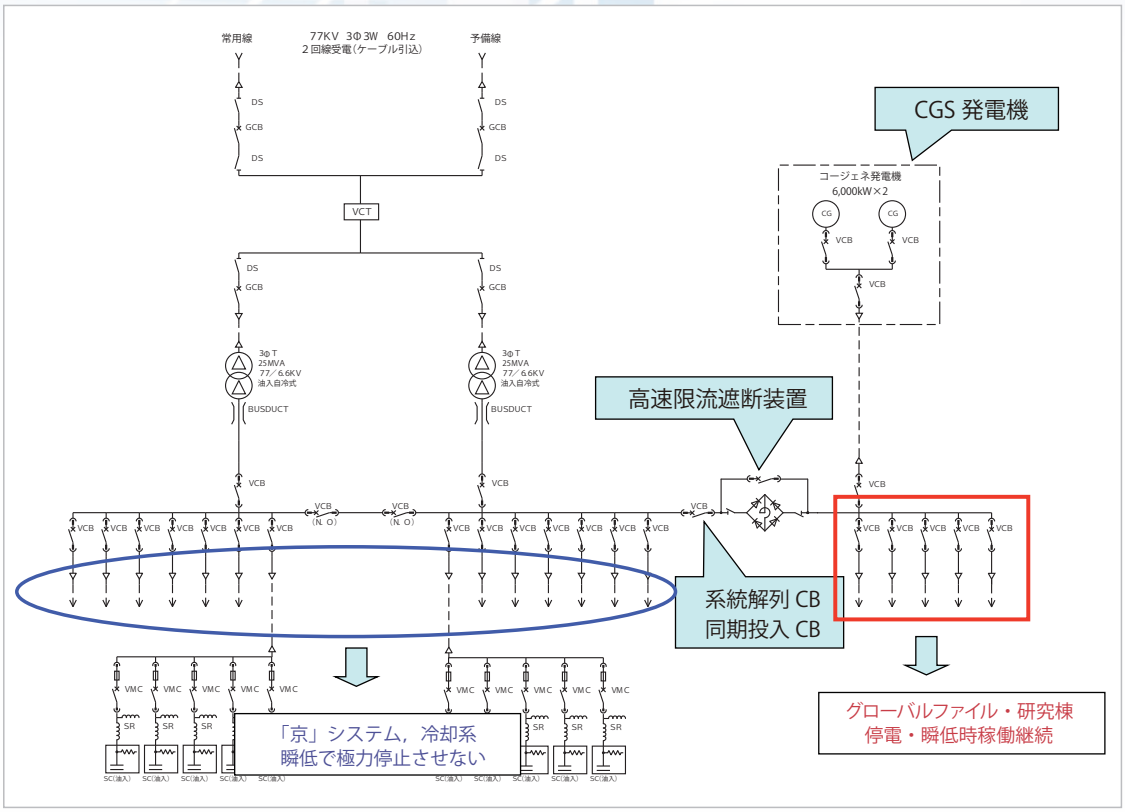

図 -9
CGS 発電設備による瞬低対策

199ms=11.9 サイクルという大きな電圧降下率，長い瞬低時間を持つ瞬低が起こった．この時は計算機にも影響があり 1/3 程度のノードが停止してしまった．しかし，冷凍機など冷却系機器の運転は継続していたため，計算機の再立ち上げも短時間に行うことができた．

本件以外では，計算機を利用しているユーザにまったく影響することなく運用を継続することができている．

空調／冷却設備

通常の空調機に用いられるベルト式シロッコファンの効率は 30％程度にとどまる．今回採用したダイレクトドライブプラグファンの全圧効率は 70％にもなる．これによりファンモータの容量＝消費電力を，シロッコファン方式に比べ半分以下にすることができた．

高効率ファンを採用したことにより，すべての機器を運転しても空調機械室内は 75dB 程度の騒音値であり，会話も可能なレベルである．騒音値が低いということは，エネルギー変換効率が高く，効率的に空気を搬送できているということである．これにより冷却空気で連続的につながっている計算機室の騒音環境の悪化を防ぐことにも貢献できた．

また，「京」では CPU と ICC を直接水で冷やす方式を採用している．水は空気の 4 倍の熱容量があり，非圧縮性流体でもあるので冷却媒体の搬送動力を空冷の 1/4 以下にすることができ，施設の消費エネルギーも低減できる．一次冷水は空調用冷水と共用化し，2 階空調機室に設置した熱交換器を介して各筐体へ送水している．送水温度は 15℃ ± 1℃に制御されている．理研では加速器施設でのマシン冷却水設備での実績があり，本 CPU 冷却設備において，加速器施設の技術を転用することにより，安定的な設備とすることができた．

環境への配慮

近年，電算センターやデータセンターでは，計算機の高速化，大容量化によりその消費電力は年々指数関数的に増大している．CPU 単体や計算機システムとしての省電力性はもちろん，システムを収容

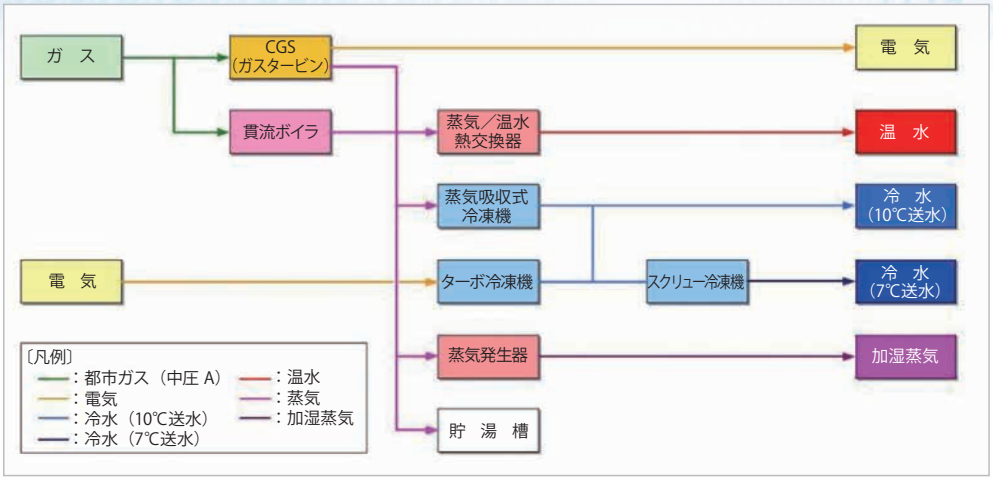

図 -10　エネルギーフロー

する施設全体の省エネ性も課題となってきている．

建屋も含めた電算センター全体の省電力性を評価する指標が PUE（Power Usage Effectiveness）である．PUE は，計算機と空調・冷却機器，照明などを含めた計算機施設全体の消費エネルギーを，計算機システムのみの消費エネルギーで割り算して求める．定義により，PUE は 1 以上の数値になるが，1 に近づくほど省エネ性に優れた施設ということになる．すなわち，空調・冷却機器，照明などの消費エネルギーが少ないほど，評価が高くなる．

建家設備での消費エネルギーの大部分は，空調・冷却設備で消費されている．いかにこの部分の消費を少なく抑えるかが技術的課題である．米国の環境保護庁（EPA）が公開している「2011 年に達成すべき PUE の目標値」は，標準の「Improved Operations」で 1.7，実現したい最良値「Best Practice」で 1.3，画期的な新技術が開発された場合の指標「State of the art」が 1.2 である．既存のデータセンターの PUE は 2.3 ～ 2.5 程度が一般的である．国内の最新のデータセンターにおいても 1.6 ～ 1.8 程度のようである．

「京」がいくら高速で高性能であっても，莫大なエネルギーを消費し，環境負荷を増大させるものであってはならない．システムを設置する建屋設備においては徹底的に省エネ性にこだわった設計を行った．都市ガスを燃料とした CGS は，発電するとともにその排ガスから排熱を回収し，蒸気を発生する．発電と排熱回収を合計した熱効率は 75％以上となっている．また，熱源機として高 COP（Coefficient Of Performance）を実現するインバータターボ冷凍機や CGS からの排熱回収による蒸気焚き冷温水発生機を設置しており，2 種類の冷凍機を状況に応じて最適に組み合わせて運転している．さらに，計算機センターにおける電力ロスの大きな部分を占める UPS を排し，CGS と高速限流遮断機を組み合わせることによって，UPS と同等の機能を持たせている．それらの技術の組合せにより，本施設の PUE は設計上 1.3 ～ 1.5 を実現することを目標とした．図 -10 に本施設のエネルギーフローを示す．

大阪ガスより中圧ガスが供給され，CGS と貫流ボイラで消費される．ただし，貫流ボイラは CGS 停止時に温熱源とするための小規模なものであり，通常時は運転されていない．CGS は最大出力 6,120kW のものが 2 台設置されている．高負荷計算を行うときは 2 台の CGS を運転し，中低負荷計算時は CGS1 台のみの運用としている．CGS は発電するとともに，排気からの熱を回収して蒸気を発生する．その蒸気を熱源として蒸気焚き吸収式冷凍機にて冷水を製造し，ベースロード冷却負荷を担わせる．蒸気は温水製造，加湿用蒸気の発生にも利用される．関西電力より 77kV で受電した電力は，降圧された後，計算機本体や空調機，冷凍機へ配電される．冷熱源として電動式インバータターボ冷凍機が設置してあり，高 COP とするため部分負荷用として運転している．

TOP500 に登録するための全ノードによる

LINPACK ベンチマーク試験は 2011 年 10 月 7 日～8 日にかけて行われた．LINPACK は CPU 負荷率の高い計算であり，実運用に入った後の最大消費電力と同等であると考えられる．LINPACK 時，CGS は 2 台運転しており，それぞれの負荷率は 78%，発電出力は 3,800kW，合計発電電力は 7,600kW であった．CGS の負荷率が 100% ではなかったため，CGS の総合効率は最大時より劣る．LINPACK ベンチマーク試験は約 30 時間にわたって行われたが，高い電力需要が安定して続いていたため，PUE は 1.34 となった．

LINPACK ベンチマーク試験後，「京」はソフトウェアの調整運転に入った．2011 年 11 月から 2012 年 4 月にかけて，「京」の消費電力は通常時 10,000kW ～ 11,000kW の範囲で推移している．この程度の消費電力であれば CGS1 台での運用で，電力，回収熱量とも十分である．受電電力を主とし，消費電力の変動に合わせ，不足する分を発電する運用を実施している．CGS の発電出力は 2,800kW ～ 3,800kW となっている．この間，月平均 PUE は 1.36 ～ 1.38 で運用が続いている．このことから，「京」の消費エネルギーのほぼ 1/3 のエネルギーで冷却が可能であることが示された．

この PUE 値は，日本データセンター協会環境・基準 WG「PUE/DCiE 計測方法に関するガイドライン（Ver2.1）」によって計算したものである．PUE=1.3 台での施設運用は，計算機システムが大規模，高密度，高集積された計算機センターとして，大変優れたものとなっていると自負している．ただし，PUE は本来，他施設と比較しその優劣を競うものではない．本ガイドラインにおいても「自社内（グループ企業内）のデータセンターで比較する際に用いる指標として活用することを推奨する」と述べられており，「PUE/DCiE 値を ILF（IT 機器の消費エネルギー），CLF（冷却の消費エネルギー），PLF（電力ロス等）に分類し，（中略）3 つに分類することにより，現状の消費エネルギーの比率を把握することができ，改善対象の発見が容易になる」ことが目的であることに注意したい．今後の運用においても，PUE 値を指標にして改善点を発見し，より効率的な施設運用を心がけていきたいと考える．

まとめ

「京」は，幅広い分野のアプリケーションに対応できる汎用性と，多くの利用者のさまざまなニーズに対応するための柔軟性を兼ね備えたスパコンである．しかし，いくら「京」の基本性能が高くても，それを運転するための電気設備，空調・冷却設備，免震設備等の周辺設備が機能しなければ，安定的な運用は望むべくもない．そのために，計算科学研究機構の施設は，本稿で紹介したような「京」のポテンシャルを最大限に引き出すためのさまざまな工夫を盛り込んだことがお分かりいただけたかと思う．

特に「京」は，これまでの大学等の計算センターに設置されているスパコンと比べ，桁違いに規模が大きく，消費する電力も大きい．そのため，それを支える設備もこれまでにない規模となったが，2010 年 6 月と 11 月に世界一を獲得したときの LINPACK ベンチマークの実行や，アプリケーションによる大規模実行などの，システムに高い負荷がかかる場合でも，電気設備および冷却設備が長時間にわたりきわめて安定的に機能し，世界一獲得の重要なアシスト役を果たした．

「京」は 2012 年 6 月に完成し，9 月末には供用が始まる．「京」によってさまざまな分野の研究が加速され，多くのブレークスルーがもたらされると期待されている．それらを実現するためにも，施設の安定的な運用と改善のための努力を今後も続けていく必要があろう．

（2012 年 4 月 27 日受付）

関口芳弘　sekiguchi@riken.jp
運用技術部調査役．施設の電源，冷却設備の運転・管理を総括している．建屋建設工事時は総括監督員として従事．

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．

塚本俊之　tsukamoto_yuki@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．