

スーパーコンピュータ「京」の概要

Overview of the K computer System

● 宮崎博行
● 渡邊 貞

● 草野義博

● 新庄直樹

● 庄司文由

● 横川三津夫

あらまし

スーパーコンピュータ「京」は、文部科学省が推進する「革新的ハイパフォーマンス・コンピューティング・インフラ(HPCI)の構築」計画のもと、富士通が独立行政法人理化学研究所と共同で開発した、超大規模並列スーパーコンピュータである。これまで培ってきた先端技術を結集し、CPU、インターコネクトなどのハードウェア、およびOS、ファイルシステム、運用管理ソフトウェアなどを開発した。システムの規模としてLINPACK性能10 PFLOPSが実現できるレベルを目標とし、CPU数にして8万個以上を結合するアーキテクチャを採用している。このような大規模システムの運用には、高性能であるとともに省電力および高い信頼性を実現することが必要である。

本稿では、システムのハードウェア構成の概要を提示するとともに、性能、電力、信頼性という観点でどのような施策を講じたかを中心に概説する。

Abstract

RIKEN and Fujitsu have been working together to develop the K computer, with the aim of beginning shared use by November 2012, as a part of the High-Performance Computing Infrastructure (HPCI) initiative led by Japan's Ministry of Education, Culture, Sports, Science and Technology (MEXT). Since the K computer involves over 80 000 compute nodes, building it with lower power consumption and high reliability was important from the availability point of view. This paper describes the K computer system and the measures taken for reducing power consumption and achieving high reliability and high availability. It also presents the results of implementing those measures.

まえがき

富士通は、1977年に日本初のスーパーコンピュータ「FACOM230-75 APU」を開発して以来30年以上にわたり最先端のスーパーコンピュータの開発・提供を続けてきた（図-1）。その中で、プロセッサをはじめとするハードウェアやソフトウェアなどを自社開発し、技術力を培ってきた。

その技術力を結集し、世界最高の性能を持つ超大規模並列計算機システムとして開発したのが、スーパーコンピュータ「京」^(注)である。

「京」は、文部科学省による「京速計算機システム」開発計画に応じて富士通と理化学研究所が共同で開発したもので、その名のとおりに1秒間に1京回の計算（浮動小数点演算、10 PFLOPS）性能を持つことを目標とした。また、開発に当たっては、ベンチマークテストにおけるピーク性能だけでなく、実際の研究で用いられるアプリケーションでの高い実効性能を出すことも目標とした。システム全

（注） 理化学研究所が2010年7月に決定したスーパーコンピュータの愛称。⁽¹⁾

体の設置と運用を1か所で可能にするために消費電力を抑えること、および、大規模システム全体の実運用に耐え得るだけの信頼性を備えることも必要であった。

そのため、次に述べる4点を開発目標として掲げた。

- ・科学技術計算向け高性能CPUの開発
- ・超大規模並列計算機向け新インターコネクトアーキテクチャの開発
- ・低消費電力
- ・高信頼性・高可用性

CPUおよびインターコネクトアーキテクチャについては本特集掲載の別稿^{(2), (3)}で詳説することとし、本稿では「京」システムの概要を説明し、システムレベルにおける低消費電力、高信頼性・高可用性に対する施策と成果について述べる。

「京」の計算ノード構成

まず、「京」システムの中心である、計算ノードの概要を示す。計算ノードは、CPU、メモリおよびインターコネクトで構成されている。

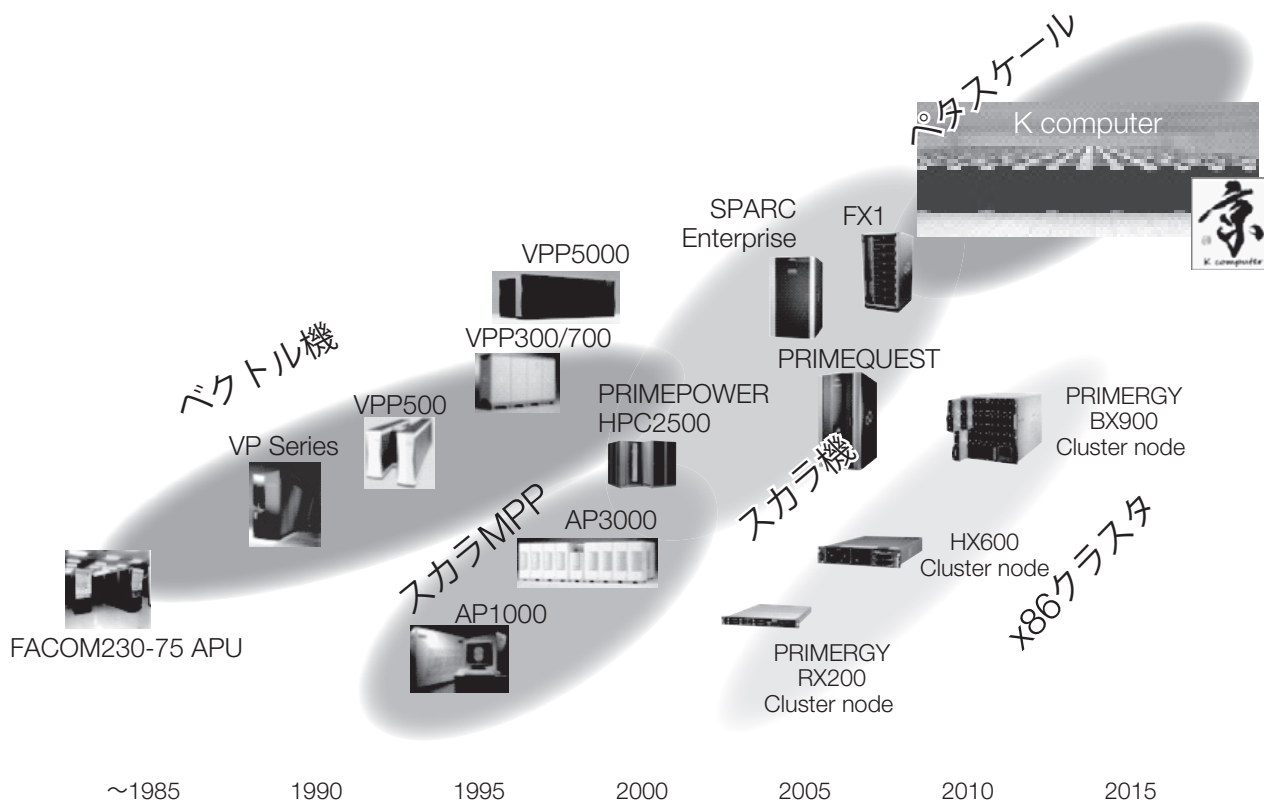


図-1 スーパーコンピュータ開発の歴史

(1) CPU

「京」のCPUとしては、理論性能128 GFLOPSの8コアCPUである「SPARC64 VIIIfx」を開発した(図-2)。(4) SPARC64 VIIIfxは富士通が開発した最先端45 nm半導体プロセス技術を利用し、プロセス技術と設計の両面から低消費電力化を図ったことで、電力性能で世界トップレベルの2.2 GFLOPS/Wを実現した。また、科学技術計算・解析に適する大規模な命令拡張HPC-ACEを適用した上、L2キャッシュとして6 Mバイト/12wayのセクタキャッシュ、およびハイエンドテクニカルコンピューティングサーバFX1で有効性を実証したVISIMPACT(ハイブリッド並列を高効率に実現するアーキテクチャ技術)を発展的に適用することで、HPC(High Performance Computing)分野での高い実行性能を実現した。さらに、従来別LSIで実現されていたシステムコントローラ・メモリアクセスコントローラ(MAC)の機能をCPUに統合することで、強力なメモリスループットと低レイテンシを実現した。

(2) メモリ

メインメモリとしては、汎用のDDR3-SDRAM-DIMM(Dual Inline Memory Module)を採用した。このDIMMはサーバやPCクラスタなどでも用いられているコモディティなモジュールであるため、マルチソースを可能とし、約1年の製造期間で合計約70万枚というボリュームで安定した供給と

質を保つことが可能であり、電力などでも最善のものを選択できるというメリットがあった。その上で、1CPUあたり8チャンネルのメモリインタフェースを備えることで、メモリバンド幅はピーク64 Gバイト/秒と競合を凌駕^{りようが}し、科学技術計算で必要とされる高いメモリスループットを確保した。

(3) インターコネクト

計算用ネットワーク(インターコネクト)としては、8万ノード超の超大規模並列計算機に向けインターコネクトアーキテクチャ Tofuを開発・適用した(5) Tofuインターコネクトは、超大規模のCPU群をスケラブルに接続する低レイテンシ、高スループットの直接網インターコネクトであり、6次元メッシュ/トラス構成を採ること高い可用性、運用性を実現している(図-3)。Tofuでは、拡張次元オーダによるルーティングアルゴリズムを適用することで、システム内に故障ノードを内包する状況でもユーザに対しては故障ノードが存在しないように1~3次元トラス結合したCPUグループを提供できる。また、ユーザに割り当てるCPUグループとして、システムの一部を切り出した場合でも、ユーザには完全な1~3次元トラス結合したCPUグループを提供できる。

「京」の計算ノードは、上記CPU/DIMMおよびインターコネクトを実装したICC-LSI(Inter-Connect Controller-LSI)から構成される。Tofuインターコネクトは直接網であることから、ICCには他ノード間のパケットを中継する機能が存在する。この部分をルータ部と呼び、残りの部分をノード部と呼んでいる(図-4)。ノード部が故障すると該当ノードを利用していたジョブだけが影響を受けるが、一方ルータ部故障の場合は該当ノードを

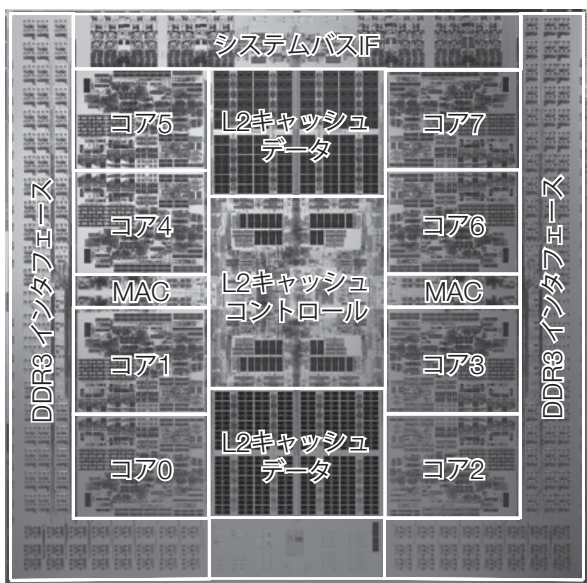
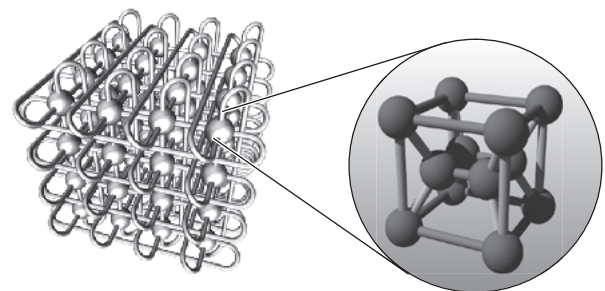


図-2 SPARC64 VIIIfx



6次元メッシュ/トラス

ノードグループ

図-3 Tofuインターコネクト

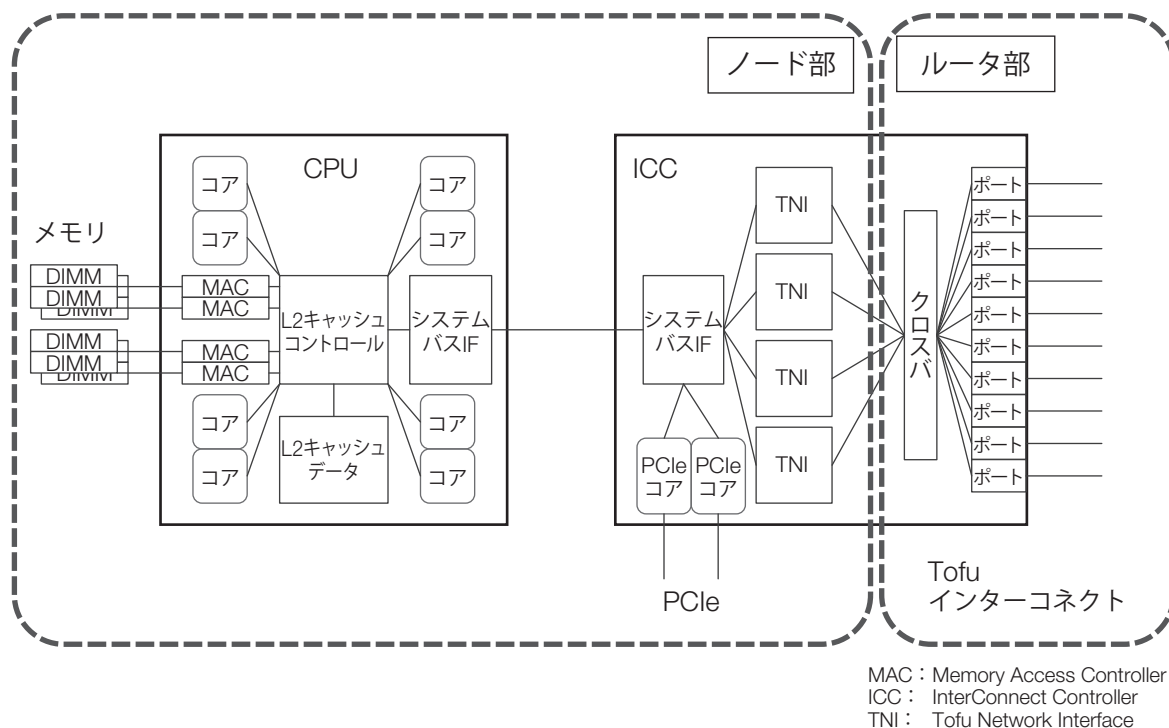


図-4 ノード部/ルータ部の概念

パケットの中継点として利用していた全てのジョブへ影響がある。ルータ部の故障率は回路物量にほぼ比例するため、ノード部の故障率より圧倒的に低い。このことから、影響の差異に応じてICC-LSIの故障を分類し、かつ故障発生時のリアクションを変えることで、ノード部故障の場合にはシステムの動作を継続する仕組みを実現している。

「京」のラック構成

「京」の計算ノードは、専用に開発されたラックに搭載される。ラックには、計算ノードを四つ実装するシステムボード（SB）（図-5）を24枚、システムブートやファイルシステムアクセスのためのIO用ノードを一つ実装するIOシステムボード（IOSB）を6枚搭載でき、合計で102ノードが1ラックに搭載される。

計算ノードおよびIOノードに使用されるCPU/ICCおよび電源パワーデバイス用の素子には水冷を適用し、ラック側に搭載した水冷配管および水冷ホースから各SB/IOSBへ冷却水を供給している。水冷を適用した理由は、システムの信頼度確保と実装の高密度化、低消費電力化（リーク電流減）のためである。SB/IOSBに搭載されるDIMM

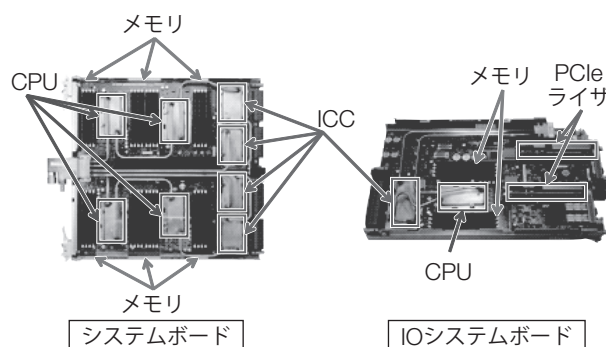


図-5 システムボード・IOシステムボード

は汎用品を使用するため空冷されるが、ボードを横向きに実装したため風向も横向きにしなければならない。しかし、ラック同士も高密度に配置させるため、SBをラック内で斜めに実装させることで、角度のついた部分の前面から吸気し、SBを通しDIMMを冷却した後、後面から排気する構造とした（図-6）。

「京」では、ラックに搭載される6IOノードを3種類の用途に分けて定義している。

- Boot-IOノード（BIO）：2ノード
- Local-IOノード（LIO）：3ノード
- Global-IOノード（GIO）：1ノード

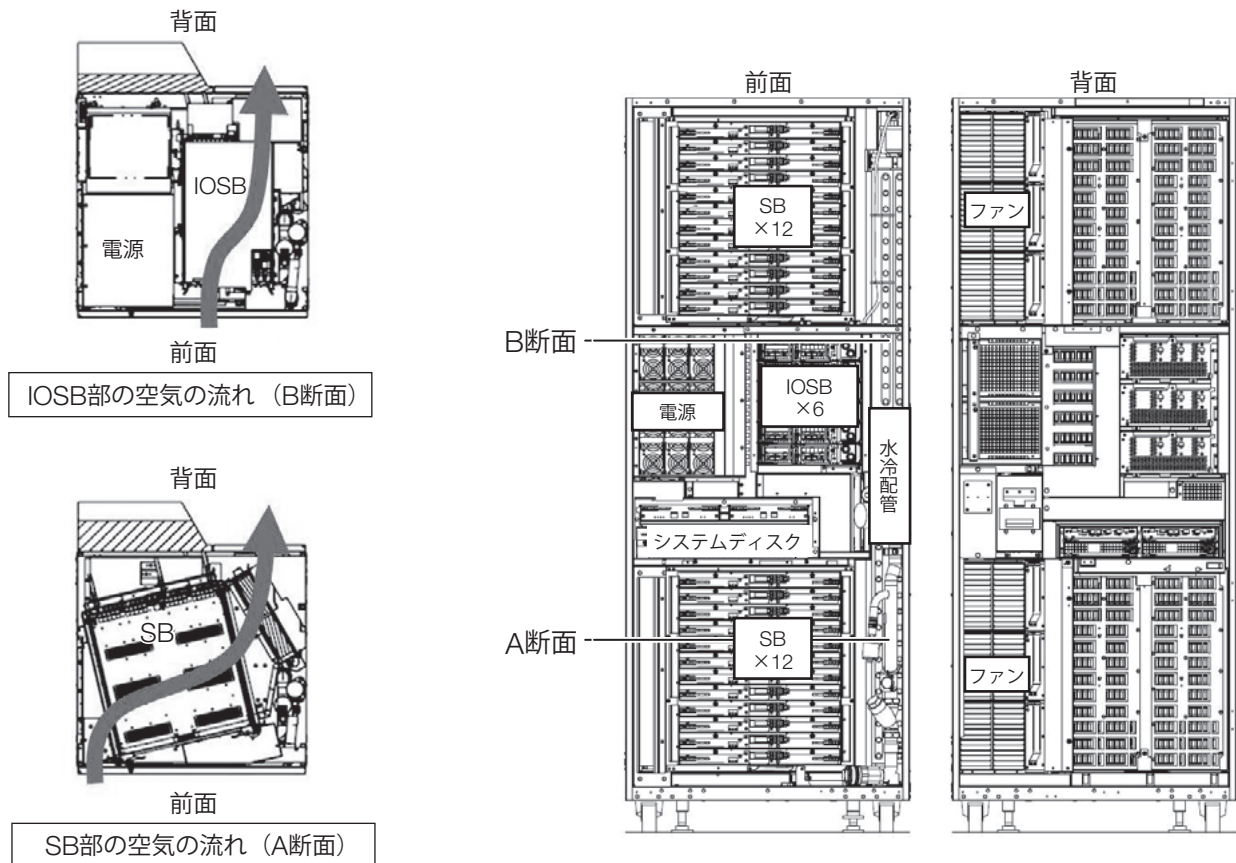


図-6 ラック実装および冷却風向

BIOノードは、「京」ラック内に実装されるシステムブート用ディスク（ETERNUS DX80）と8G-FCインタフェースで接続されている。BIO自身はディスクブートにより起動し、ラック内の残り100ノードに対するTofu経由のブートサーバとして動作する⁽⁶⁾。ラック内の二つのBIOノードは運用系・待機系としての役割を持ち、運用系故障時には待機系BIOがブートサーバの役割を担う。

LIOノードは、「京」ラックと隣接して設置されるディスクラックに搭載されるローカルディスク（ETERNUS DX80）と接続される。

GIOノードは、InfiniBand-QDRインタフェースを1本持ち、外部グローバルストレージシステムに接続される。これらストレージシステムは、大規模かつ高性能、高信頼なファイルシステムとしてFEFS（Fujitsu Exabyte File System）⁽⁷⁾を採用し、LIOまたはGIOがダウンした場合は隣接する別ラックに搭載されるLIO/GIOのパスを経由してファイルシステムへのアクセスを継続可能な構成を採っている。

また、ラックには電源ユニット（PSU）、空調ファン、およびシステム制御機構（SP）を搭載しており、それぞれが冗長構成を採ることで単一故障によるラックのダウンを防いでいる。

「京」のシステム全体構成

「京」の計算ラックは、システム全体で864ラックによって構成される。図-7に示すように計算ラックは隣接2ラックでZ軸接続（IOノードを含め17ノード）を構成する。計算ラック4ラックに対してディスクラックを1ラックの割合で、Y軸方向（横方向）に45ラック（計算ラック：36ラック、ディスクラック：9ラック）を配置する。また、X軸方向には24ラックを配置し、フロア全体では24×45の長方形に配置している。ディスクラックにはローカルディスクを12台搭載し、各ローカルディスクにはX軸方向に隣接するLIOの2ノードが接続される。

「京」システムは、計算ラックやディスクラック、並列ファイルシステムだけで構成されるので

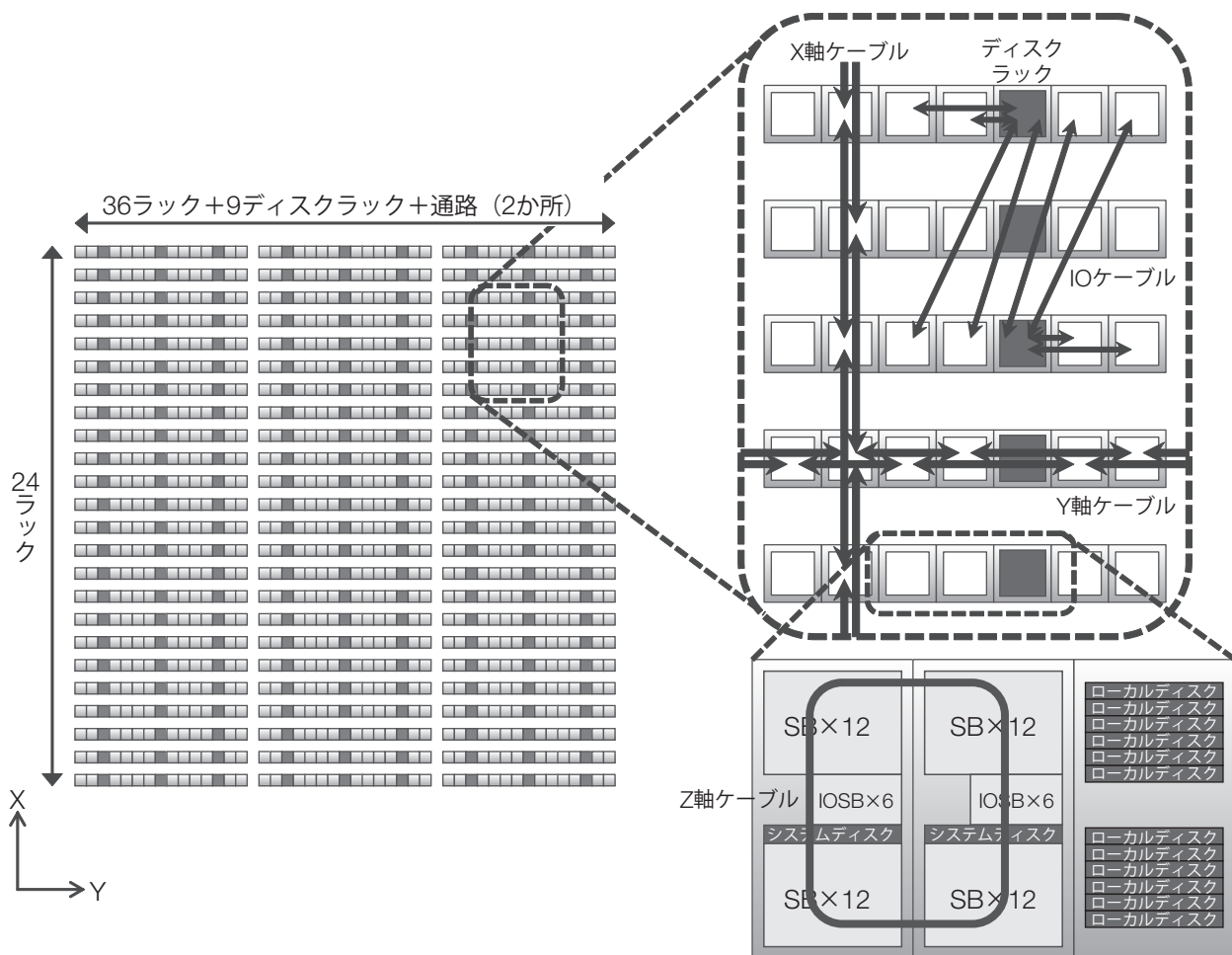


図-7 計算ラック構成

はなく、システム管理、ジョブ管理のための周辺サーバハードウェアおよび運用管理ソフトウェアによる制御が必要不可欠である⁽⁸⁾。また、故障時のハードウェア交換に際し、リモート通報システムおよび保守サーバによる保守オペレーションが行われる。理化学研究所に設置されている「京」システム全体の概略を図-8に示す。

システムレベルの省電力施策

「京」システム全体での低消費電力化のため、まずCPU/ICCの開発において徹底的な省電力化を行った。クロックゲーティングによるダイナミック電流の削減、およびHigh-Vthトランジスタの採用で、動作周波数は2.0 GHzに抑えリーク電流を大幅に低減した⁽⁹⁾。さらにリーク電流はジャンクション温度が高くなると急激に増加する特性があることから、水冷を採用しジャンクション温度を通常の85℃から30℃にまで落とすことで、リーク電流

を大幅に低減した。

LSIの消費電力削減のためにクロックゲーティング設計を行うと負荷変動がより顕著となるため、電源の過渡応答特性を改善することが重要な課題となった。過渡応答特性改善と大規模ラックへの効率的な電源供給、そして高密度実装を両立するため、中間バスコンバータ方式を採用した。過渡応答特性改善のために負荷の近傍へ非絶縁タイプの小型POL（point of load）電源を配置し、一方筐体全体へ効率良く電源供給するために筐体電源から各SBへは48 V給電としている。SB上には絶縁型バスコンバータを搭載することで48 Vを降圧してPOL電源へ電力を供給している。このような方式にすることで、絶縁トランスの個数を減らすことができ、高効率でかつ高密度な電源供給系を実現した。

システムレベルの高信頼性・高可用性施策

SOC (System-on-a-Chip) とすることで、LSI 上の回路量は増大する一方で、LSI以外の部品点数は減少する。よって大規模LSIの故障率がシステムのそれに占める割合は年々大きくなっており、大規模LSIの故障率を低減することは「京」システムの安定運用において大変重要な意味を持つ。

● 高信頼性施策

高信頼に対しては個々の部品の信頼度を上げること、部品の利用方法により信頼度を上げる取組みを行っている。システムの信頼性を向上させる取組みについては、

- ・ジョブ運用の継続
- ・システム可用性確保

という二つの枠組みで捉えている。表-1では、ジョ

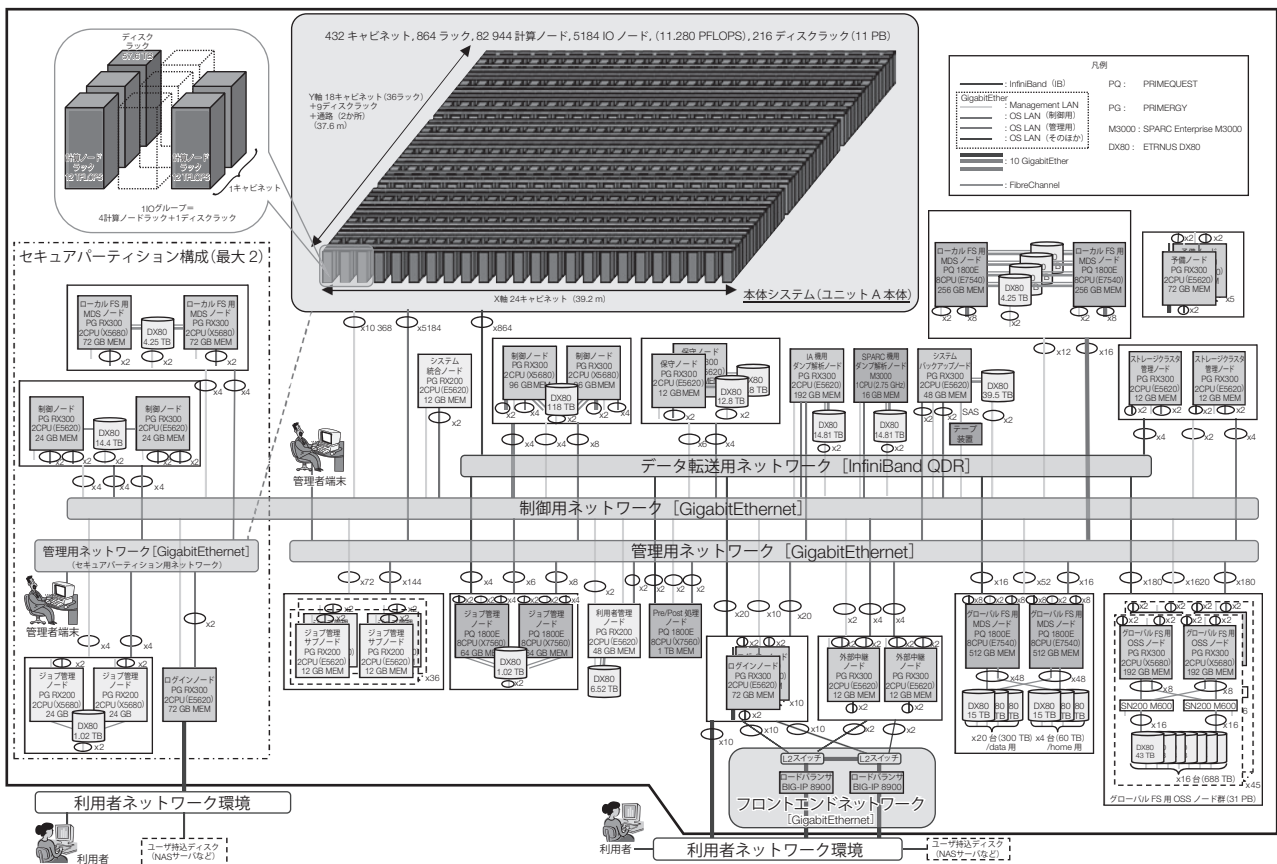


図-8 システム全体の概略

表-1 冗長構成/活性交換による高信頼化

主要部品	冗長構成	活性交換
筐体電源	Yes	Yes
冷却用ファン	Yes	Yes
サービスプロセッサ (SP)	Yes (二重化切替え)	Yes
システムボード (SB/IOSB)	No ⇒ B軸で故障回避	Yes
CPU/ICC	No ⇒ 水冷, リトライ, ECCで改善	(SB活性交換)
POL電源	No ⇒ 水冷で信頼度確保	(SB活性交換)
中間コンバータ	Yes	(SB活性交換)
DIMM	No ⇒ 拡張ECCで救済	(SB活性交換)
システムディスク	Yes ⇒ コントローラ, 電源は冗長 HDDはRAID5+ホットスペア	Yes (モジュール)

ブ運用継続観点から主要部品の冗長構成による運用継続性を担保し、システム可用性観点では活性交換によるシステム全体停止を伴わない故障部品交換により可用性を担保することを示している。

ジョブ運用継続観点では、ハードウェア冗長構成、1ノード1CPU構成などによる部品点数の削減、ECC（Error Correcting Code）やリトライによるエラー訂正、およびLSIやPOL電源を水冷し稼働温度を低減させることによる半導体故障発生率低減の施策を行い、故障発生によるノードダウン・ジョブダウンが起きる確率を下げている。Tofuケーブルとして電気ケーブルを採用したこともジョブ運用継続性を高めるための施策の一つである。

● 高可用性施策

システム可用性確保の観点では、6次元メッシュ/トラスによる故障ノード回避の仕組みをSB故障・活性保守時にも拡張して適用できるように、SB1枚4ノードの縮退でも迂回ルーティング可能になるようなSB内Tofu座標配置を持たせた。具体的には、図-9のようにSB上の4ノードが同一Y、B座

標に閉じるよう、SB上にはA軸・C軸接続を持たせ、B軸接続がSBの外に出るよう実装した。

また、前述したIOノード3種については、IOノード故障で配下計算ノードのIOパスが使えなくなならないよう、IOパス冗長として共通の接続先を持つ交替IOノードを用意しておき、計算ノード群はTofu経由で交替IOノードへアクセスを継続できるよう構成されている（図-10）。BIOはラックに2ノ

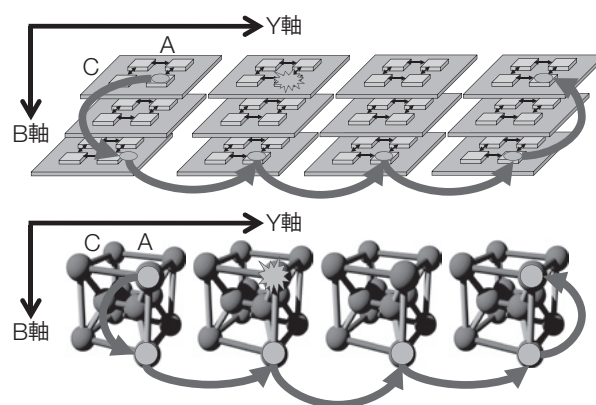


図-9 Tofu迂回概念

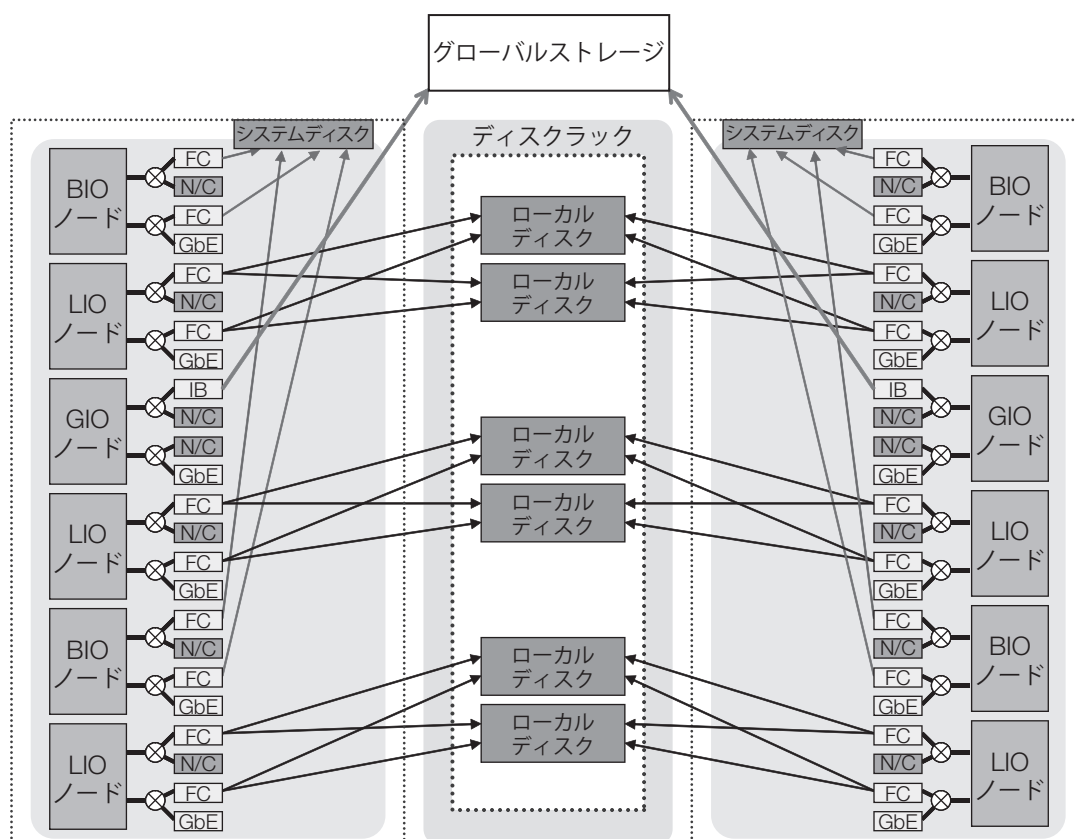


図-10 IO冗長パス

表-2 TOP500/Green500結果

	2011/6	2011/11	単位
TOP500 ランキング	1	1	位
コア数	548 352	705 024	個
性能 (Rmax)	8162.00	10 510.00	TFLOPS
電力 (Power)	9898.56	12 659.9	kW
Green500 ランキング	6	32	位
電力性能	824.56	830.18	MFLOPS/W

ド実装され、運用/待機の冗長構成を採る。LIOはラックに3ノード実装され、ディスクラックに実装されるローカルディスクを共有する他ラックのLIOとの冗長構成を採る。GIOはラックに1ノード実装され、LIOと同様に他ラックのGIOとの冗長構成を採る。

ベンチマーク結果

前述の開発目標に沿って開発された「京」は、まだ導入作業の過程であった2011年6月の段階で、システムの一部(672ラック)を利用してHPC-LINPACKベンチマークテストを行った。その結果、2011年6月発表のTOP500リストで1位となる8.162 PFLOPSの性能を達成した。⁽¹⁰⁾ LINPACKベンチマークの走行時間は28時間超であり、この期間68 544ノードが故障なく動作し続けたことは、システムの高信頼性、ジョブ運用継続性の高さを示すものである。詳細を2011年11月発表の結果と合わせて表-2に示す。

スパコンの総合的な性能を評価するHPCチャレンジベンチマークにおいても、2011年11月発表のHPCチャレンジ賞(クラス1)の4部門 {Global HPL, Global RandomAccess, EP STREAM (Triad) per system, Global FFT} 全てで1位を獲得した。また、「京」を用いた実アプリケーションによる成果が認められ、2011年11月発表のゴードン・ベル賞の最高性能賞に選出された。これらは、LINPACKベンチマークに特化したマシンではなく、幅広い範囲のアプリケーションに対応できる汎用性を備えていることを示すものである。

む す び

「京」システムは、その開発目的から大規模HPCシステムであり、高信頼性・高可用性を重視して

設計段階から取り組んだ。富士通と理化学研究所は、目標とするピーク性能、アプリケーション実効性能の達成、および高信頼性・高可用性システムの構築を推進してきた。2011年6月のTOP500ベンチマークで1位、Green500ベンチマークで6位(汎用スカラCPUでは1位)を獲得し、続く2011年11月にはTOP500ベンチマークで引き続き1位を獲得するとともにLINPACK性能10 PFLOPSを達成しており、計画どおり着実に目標を達成している。今後は導入が終わり運用フェーズに入るため、実アプリケーションでの性能およびシステム実運用での評価を行っていく。

参考文献

- (1) M. Yokokawa et al.: The K computer: Japanese next-generation supercomputer development project. ISLPED, p.371-372 (2011).
- (2) 吉田利雄ほか: スーパーコンピュータ「京」のCPU SPARC64 VIIIfx. *FUJITSU*, Vol.63, No.3, p.254-259 (2012).
- (3) 安島雄一郎ほか: スーパーコンピュータ「京」のインターコネクT Tofu. *FUJITSU*, Vol.63, No.3, p.260-264 (2012).
- (4) T. Maruyama et al.: SPARC64 VIIIfx: A New-Generation Octocore Processor For PETASCALE Computing. *IEEE Micro*, Vol.30 Issue2, p.30-40 (2010).
- (5) Y. Ajima et al.: Tofu: A 6D Mesh/Torus Interconnect for Exascale Computers. *Computer*, Vol.42, No.11, p.36-40 (2009).
- (6) 師尾 潤ほか: スーパーコンピュータ「京」のオペレーティングシステム. *FUJITSU*, Vol.63, No.3, p.273-279 (2012).
- (7) 酒井憲一郎ほか: スーパーコンピュータ「京」の高性能・高信頼ファイルシステム. *FUJITSU*, Vol.63, No.3, p.280-286 (2012).
- (8) 平井浩一ほか: スーパーコンピュータ「京」の運用管理ソフトウェア. *FUJITSU*, Vol.63, No.3, p.287-292 (2012).
- (9) H. Okano et al.: Fine Grained Power Analysis and Low-Power Techniques of a 128GFLOPS/58W SPARC64™ VIIIfx Processor for Peta-scale Computing. *Proc. VLSI Circuits Symposium'00*,

p.167-168 (2010).

(10) TOP500 Supercomputing Sites.

<http://www.top500.org>

著者紹介



宮崎博行 (みやざき ひろゆき)

次世代テクニカルコンピューティング
開発本部システム開発統括部 所属
現在、次世代HPCの仕様策定および
LSI開発業務に従事。



庄司文由 (しょうじ ふみよし)

独立行政法人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グ
ループ 所属
現在、理研においてシステム開発作業
に従事。



草野義博 (くさの よしひろ)

次世代テクニカルコンピューティング
開発本部システム開発統括部 所属
現在、次世代HPCのシステム開発に
従事。



横川三津夫 (よこかわ みつお)

独立行政法人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グ
ループ 所属
現在、理研においてシステム開発全般
のとりまとめに従事。



新庄直樹 (しんじょう なおき)

次世代テクニカルコンピューティング
開発本部システム開発統括部 所属
現在、次世代HPCの開発統括業務に
従事。



渡邊 貞 (わたなべ ただし)

独立行政法人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 所属
現在、理研においてプロジェクト全体
を統括。