解説 2

Blue Gene/L システム
―スーパーコンピューティングへのグランドチャレンジ―

IBMコーポレーションは、1999年12月、当時の世界最速スーパーコンピューターの500倍の演算処理能力を持つペタ・フロップス・マシンの研究計画（Blue Gene計画）を発表しました。これは、生命科学のグランドチャレンジと同時に、それを実行するペタスケールの計算機を構築するという計算機工学的なグランドチャレンジの双方にアプローチすることを目的としたものでした。

2005年11月に開催されたSC|05（スーパーコンピューター、ネットワーク、ストレージに関する国際学会）において、ローレンス・リバモア研究所に設置された64ラックのBlue Gene/Lは、標準ベンチマークで280.6Tflop/sの性能を達成し、再び世界一の座を獲得しました。

Blue Gene/Lは、従来の並列計算機の限界を大きく引き上げる革新的なスーパーコンピューターといえます。本稿では、そのようなブレークスルーを達成するための本質的な技術的チャレンジと、Blue Gene/Lによる実際の取り組み、さらにそのアーキテクチャー的な特徴について解説します。

日本アイ・ビー・エム株式会社
Senior Technical Staff Member
ディープコンピューティング開発研究所
技術開発担当部長
清水 茂則 Shigenori Shimizu

［プロフィール］
1983年に日本IBMに入社以来、基礎研究部門で、コンピューターアーキテクチャー・並列処理・回路設計・テクノロジーなどの研究・管理に従事。2005年より、ディープコンピューティング開発研究所にて、技術開発を担当。

日本アイ・ビー・エム株式会社
東京基礎研究所
専任研究員
寒川 光 Hikaru Samukawa

［プロフィール］
1984年、日本IBMに入社。1995年より東京基礎研究所にて数値計算・数値解析・コンピューターアーキテクチャー・High Performance Computingの研究に従事。

日本アイ・ビー・エム株式会社
東京基礎研究所
副主任研究員
土井 淳 Jun Doi

［プロフィール］
1999年に日本IBMに入社以来、インダストリー分野向けのCADやモデリングに関する研究開発に携わり、2004年より、ディープコンピューティング向けのアプリケーションのプログラム最適化を担当。

Article 2

Blue Gene/L System
– A Grand Challenge to Supercomputing –

IBM Corporation announced in December 1999 the research plan named “Blue Gene Plan”, which aimed to create “Peta-Flop/s Machine” that was capable of processing capabilities of 500 times the world’s best supercomputer at that time. This plan aimed to make two grand challenges: One to life sciences, and the other to computer engineering through creation of a petascale supercomputer.

At SC|05 (an international workshop on supercomputers, networks and storage) held in November 2005, the Blue Gene/L system with 64 racks installed at Lawrence Livermore National Laboratory once again was ranked as the No.1 supercomputer in the world, having achieved 280.6Tflop/s Linpack performance under the standard benchmark.

The Blue Gene/L system is an innovative supercomputer that significantly raises the limitations of conventional parallel computers. This article explains essential technical challenges fought to achieve such a breakthrough, the actual usage of the Blue Gene/L and its architectural features.

❶ スーパーコンピューティングのブレークスルー

スーパーコンピューターの最も基本的な構成要素は、プロセッサーチップであり、これは半導体の継続的進歩にドライブされ、「ムーアの法則」と呼ばれる比率（18カ月で2倍程度）で性能が向上しています。

実際に、Power Architecture™の最初のプロセッサーであるPOWER™1は1990年に発表され、50 Mflop/s程度の性能でしたが、最新のPOWER5™は7.6 Gflop/sの性能を提供し、15年間で約150倍の性能向上を果たしました。その結果、シミュレーション技術の進歩と相まって、スーパーコンピューターによるシミュレーションは、理論や実験を補足する位置付けから、むしろリードする位置付けに変わりつつあります。

しかし、より現実的かつ大規模な問題にアプローチするには、まだまだ計算性能が足りないのも事実です。例えば、そこそこの大きさのタンパク質（周囲の水分子と合わせて3万原子程度）の折り畳み構造（protein folding）のシミュレーションでさえも、1Pflop/sの計算機で1年間ほどかかると試算されています。

この問題にアプローチするために、1999年にBlue Gene計画がスタートしました。当時の世界最速スーパーコンピューターの数百倍の計算性能を持つ計算機を5年計画で実現しようという取り組みです。

数百倍もの性能向上となると、従来技術の単純な延長で達成するのは不可能であり、明らかなブレークスルーが求められました。最大の問題は、熱対策と設置面積、信頼性でした。

汎用のプロセッサーチップを大量に並べて数百Tflop/s規模の計算機を作るとすれば、その発熱量は数十Mワット規模になると試算されます。これは、平均的な米国家庭の1万所帯以上の電力消費に匹敵します。限られた体積から発生するこれだけの熱量を冷却するのは極めて困難であり、結局、システムを稼働できないということになります。設置面積も数千平方メートル規模が予想され、物理的な設置スペースの問題にとどまらず、信号伝達の電気的な問題をも引き起こします。また、PCクラスターで既に経験している信頼性の問題は、システム規模の増大とともに、桁違いに深刻な問題となります。

性能対熱効率・性能対面積効率・信頼性の指標において、特定のアプリケーションの実行に特化した専用マシンとスーパーコンピューターの間には、大きな開きがあります。つまり、専用マシンの方がこれらの指標において桁違いに良い値を実現します。従って、Blue Geneの設計アプローチの基本は、専用マシンの効率を維持しつつ、比較的広範なアプリケーションに適用可能な大規模並列システムを実現することでした。

性能対熱効率・性能対面席効率は、組み込み用のPowerPC®プロセッサーコアに、大規模並列マシンの実現に必要になる周辺回路を付加する、いわゆるSoC（System on Chip）アプローチで達成されました。比較的低周波数で動作する組み込み用のPowerPCは、高速サーバーに使用されているハイエンドのPowerチップに比較して、2〜10倍程度の性能対電力効率を実現します。

比較的低周波数で動作する組み込み用のプロセッサーを基本的な構成要素として用いるということは、裏を返せば、高性能のハイエンド・プロセッサー・チップを用いる場合と比較して、圧倒的に多くのプロセッサーを並列動作させることを意味します。プロセッサー数に比例して容易にスケールすることが可能なメモリーシステム、およびプロセッサー間を接続する相互接続網が必要になります。そこで、Blue Gene/Lでは、分散メモリーとMPI（Message Passing Interface）と呼ばれるメッセージ交換に基づくプログラミングモデルを用いることによって、このスケーラビリティーを実現しようとしています。そして、3次元トーラス状の相互接続網とグローバルツリーと呼ばれる2進木状の相互接続網によって、65,536個の計算ノードから構成される最大規模のシステムの場合でさえも、プロセッサーの数に比例したバンド幅を低遅延時間で実現するスケーラビリティーが実現されています。

そして、これらの相互接続網を実現するための回路を、SoCのアプローチで、プロセッサーコアと同じ半導体チップ上に実装することによって、性能とともに、高い性能対面積効率が実現されます。実際、計算ノードに必要なコンポーネントは、組み込み用PowerPCプロセッサーを用いたSoCチップとDRAM（Dynamic Random Access Memory）チップだけであり、非常

にコンパクトな計算ノードが実現されます。

これは同時に、システムの信頼性にも大きく寄与します。最大構成では65,536個の計算ノードとなるわけですが、それぞれノードが高い信頼性を持つことが非常に重要になります。均質なノードから構成されるシステム全体のMTBF（Mean Time between Failure）は、各ノードのMTBFをノード数で割った値になりますから、各ノードの信頼性が低いと、システム全体は頻繁に故障することになってしまいます。これがクラスターなどの大規模化を妨げる重大な要因の一つになっています。Blue Gene/Lでは、各計算ノードは、おおむねプロセッサーSoCチップとDRAMチップの2種類の部品だけであり、非常にシンプルな構成です。また、詳しくは述べませんが、SoCのデザインアプローチを用いることによって、信頼性向上のための回路を加えることが可能となり、メモリーアレイや信号伝送経路に対して、誤り検出・誤り訂正・再試行などの機能が付加されています。

このほかにも多数のブレークスルーが結集され、2004年秋に発表されたTOP500（世界中のスーパーコンピューターの性能ランキング）から、3期にわたって1位の座を維持しています。2005年11月に発表された最新のTOP500リストでは、ローレンス・リバモア研究所に設置された64ラックの最大システム（最大性能：360Tflop/s、Linpack性能：280.6Tflop/s）が堂々1位の座を獲得しています。

❷ Blue Gene/Lの概要

Blue Gene/LのプロセッサーSoCチップのブロック図を**図1**に示します。

Blue Geneのプロセッサーチップは、二つのPPC440（PowerPC 440）プロセッサーコアを中心として構成されています。それぞれのコアには、ダブルFPU（Floating point number Processing Unit：浮動小数点演算装置）と呼ばれる文字通り二つのFPUが付加されています。それぞれは既存のFPUでPowerPCの浮動小数点演算命令を持ち、二つのFPUを用いて並列に計算ができるアーキテクチャーになっています。さらにL2キャッシュ、L3キャッシュのディレクトリー、eDRAM（Enhanced Dynamic Random Access Memory）による4MバイトのL3キャッシュなどを持ち、トーラス、グローバルツリー、グローバルバリア、Gビットイーサネットなどの通信機能もプロセッサーチップ上に実装されています。

一つのプロセッサーチップには、二つのプロセッサーの組が含まれていますが、二つのプロセッサーの使い方を次のうちから選択することができます。一

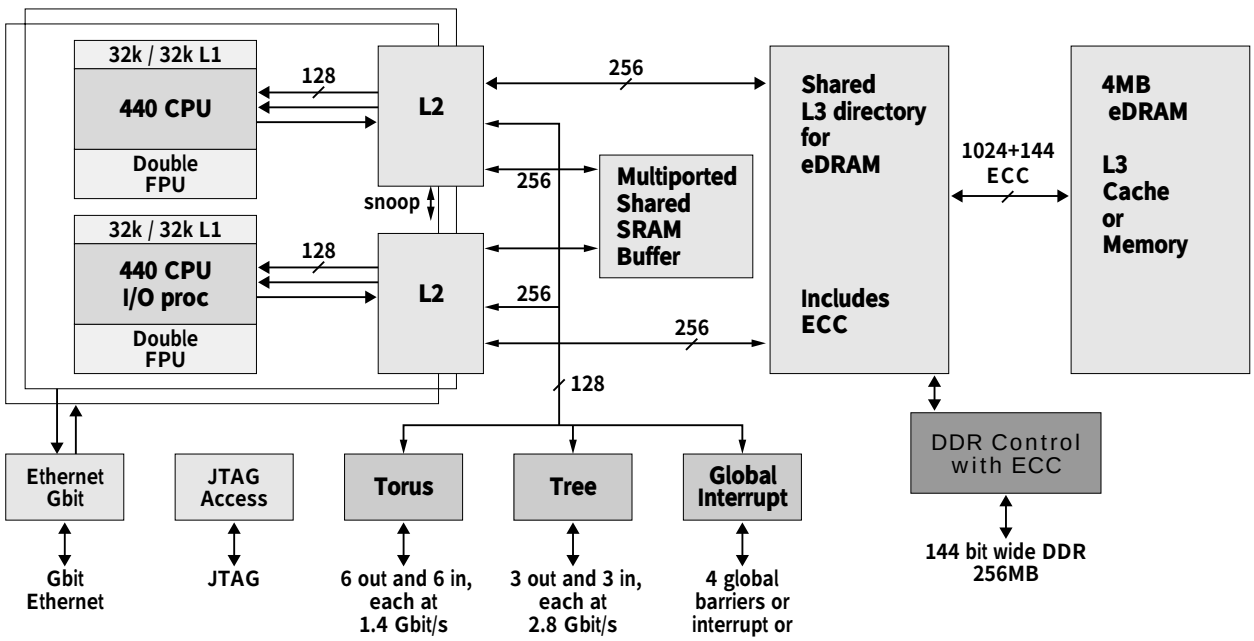

図1. プロセッサーチップ

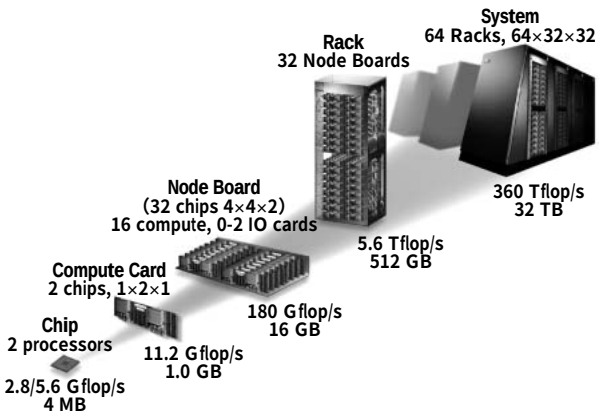

図2. Blue Gene/Lの構成

つ目はコプロセッサーモードと呼ばれるモードで、一つのプロセッサーは、計算専用としてユーザープログラムを実行し、もう一方のプロセッサーは、MPIの通信処理専用に稼働します。二つ目は、仮想ノードモードと呼ばれるモードで、二つのプロセッサーがそれぞれ独立のプロセッサーとしてユーザープログラムを実行します。

それぞれのプロセッサーチップには、DDR1（Double Data Rate 1）DRAMから構成される512Mバイトのローカルメモリーが付加され、計算ノードを構成します。2組の計算ノードが1枚のプロセッサーカードに実装され、16枚のプロセッサーカードがノードボードに搭載されます。ノードボードには、このほかに1～2枚のプロセッサーカードが搭載され、こちらは計算ノードとしてではなく、I/O（Input/Output：入出力）ノードとして用いられます。32ノードボードを1筐体に搭載します。つまり、1,024計算ノードが1筐体に実装されます（**図2**）。

並列計算機としての基本構成単位は1筐体の半分の512ノードのミッドプレーンと呼ばれる構成です。ミッドプレーン内ではネットワークは相互接続され、8×8×8の3次元トーラスがネットワークの基本となります。さらに複数のミッドプレーンをリンクASICを用いて接続することで、大規模な並列計算機を構成することができます。

理論最大性能は1プロセッサーが2.8Gflop/sなので、1ラック当たり5.6Tflop/sになります。また消費電力は、1ラック当たり25Kワット程度で、1ワット当たり200Mflop/sと高い性能対電力効率が実現されています。

2.1. プロセッサーSoCチップ

ベースとなるPPC440コアは32ビットアドレッシングの組み込み用のプロセッサーで、700MHzで動作します。このPPC440コア自身は、浮動小数点ユニットは持ちませんが、APU（Auxiliary Processor Unit）ポートに付加プロセッサーを密結合できる仕様になっています。Blue Gene/LではこのAPUポートに二つのFPUからなるダブルFPUを付加しています。L1は32Kバイト／32Kバイトの命令／データキャッシュで、ダブルFPUに対しては128ビットのインターフェースを持ちます。二つのプロセッサーのL1間にはコヒーレンシーの機能がサポートされていません。これを補う目的で、ロックボックスや共有SRAM（Static Random Access Memory）などの同期の手段を備えています。共有SRAMは16Kバイトで二つのプロセッサーコアから共有されるので、二つのプロセッサー間の高速なデータ交換などに使用されます。L2は容量が2Kバイトと小さいものの、フル連想性を備え、128バイトのキャッシュラインを16個保持できます。また二つのL2間はスヌープによるコヒーレンシーが実現されています。L3はeDRAMにより実現されており、4Mバイトの容量で、二つのプロセッサーで共有されます。メモリーはDDRコントローラーがチップ上にあり、プロセッサーチップごとに512MバイトのDDR1 DRAMが付加されます。

2.2. ダブルFPUアーキテクチャー

できるだけ少ないトランジスター数・電力消費で、できるだけ高い実効性能を実現するという観点から、Blue Gene/LのダブルFPUが設計されました。ダブルFPUは、単に一つの命令で同時に二つの浮動小数点演算を行うような、通常のSIMD（Single Instruction/Multiple Data）よりも広い範囲の演算ができるように設計されていて、例えば複素数演算のように，多くの数値演算プログラムで効率よく利用することができます。ダブルFPUを用いると最大で1クロックごとに四つの浮動小数点演算が実行されるので、700MHzという低い動作周波数にもかかわらず、プロセッサー当たり2.8Gflop/s、プロセッサーチップ当たり5.6Gflop/sという高い演算性能が実現されます。

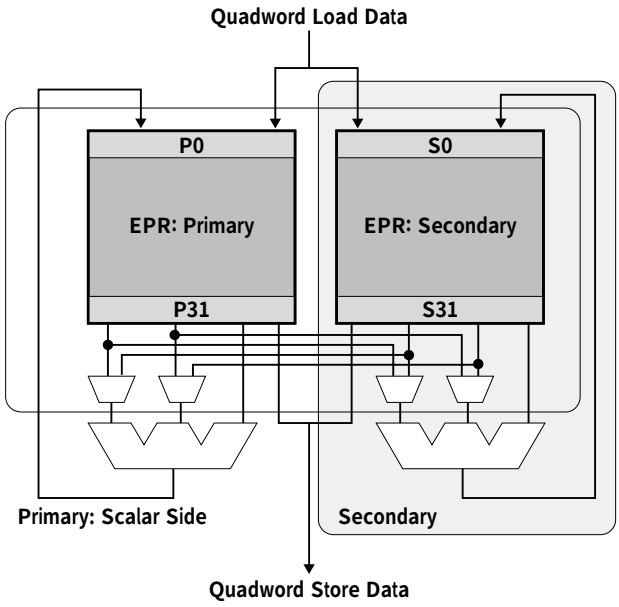

図3. ダブルFPU

図3に示すように、ダブルFPUは、32要素（各要素は64ビット長）のレジスターファイルを2セット持ちます。既存のPowerPC浮動小数点演算命令は1次側（primary side）のレジスターファイルに作用し、拡張された命令は2次側のみか、1次側と2次側の両方に同時に作用します。通常のSIMDのように、1次側と2次側の同じ位置のレジスターに対して同じ演算を同時に実行するパラレル演算のほかにも、オペランドのどちらかを1次側と2次側で入れ替えて使用するクロス演算や、1次側か2次側のデータを共通に使うクロスコピー演算などがサポートされていて、SIMDよりも広いクラスのデータ並列性が提供されるので、SIMOMD（Single Instruction Multiple Operation Multiple Data）と呼ばれます。

2.3. 相互結合網

MPI通信を、広帯域、低遅延時間、高効率で行うために3次元トーラス、グローバルツリー、グローバルバリアの三つの相互結合網が用いられます。そのほかに外部通信用のGイーサネットと、システム制御、モニタリング、ブートに用いられるJTAG（Joint Test Action Group）制御ネットワークがありますが、ここでは説明を省きます。

3Dトーラスは、主にMPIのポイント・ツー・ポイント通信に使われます。各計算ノードからは、3軸方向の正負、合わせて6方向に対して、それぞれ双方向にリンクが張られ、隣接する計算ノードと接続されます。リンク1本当たり1.4Gbpsの帯域で、ノード当たり合計2.1GB/sの帯域が提供されています。また、1ホップ（隣の計算ノードへの転送）の遅延は、100ns程度と高速です。複数ホップ（複数ノードを経由した転送）の経路制御は、ハードウェアによって自動的に行われ、プロセッサーに負荷を掛けないため高速です。

グローバルツリーは、主に、MPIの縮約オペレーションや同報通信に使われます。ツリーの3方向にそれぞれ2.8Gbpsのリンクが双方向で張られ、ノード当たり合計2.1GB/sの帯域が提供されます。各プロセッサーチップ上のグローバルツリーの制御回路には、算術／論理回路が付加されていて、縮約オペレーションの総和演算などは、プロセッサーに負荷を掛けずに、ハードウェアによって自動的に高速に行われます。

また，バリア処理（MPI_barrier）は，グローバルバリア網により，各プロセッサーチップ中の専用回路で実行され、プロセッサーの負荷なしに自動的に次々とノード間を高速に伝達されます。

❸ システムソフトウェア

I/Oノードでは、組み込み用Linux（MCP：Mini Control Program）が動き、複数の計算ノードを管理します。例えば、最大構成の65,536計算ノードのシステムでは、1,024のI/Oノードが付加され、それぞれ、64の計算ノードを管理します。システム全体は、外部から見ると1,024ノードのクラスターのように見ることもできます。計算ノードでは、CNK（Compute Node Kernel）と呼ばれる専用のシングルプロセスのカーネルが稼働し、アプリケーションは、この上で排他的に実行されます。CNKはPOSIX（Portable Operating System Interface for UNIX™）インターフェースの多くをサポートしています。アプリケーションが発行するファイルI/Oは、glibcからの要求として発行され、計算ノードからは、clibcを経由して、I/Oノードへのポイント・ツー・ポイント通信要求として発行され、グローバルツリーを経由してI/Oノードに送られます。I/Oノードは、これを受け取り、Linuxデーモンが、Gビットイーサ

ネットを経由してファイルサーバーへ要求を出します。

また、3次元トーラスとグローバルツリーを用いてメッセージ交換ライブラリーが実装されています。通信ソフトウェアは、上位にMPI層（レイヤー）、下位にネットワークをアクセスするパケット層を備え、両者を短遅延時間でつなぐために中間には1層のメッセージ層を持つだけです。またコレクティブ通信用には、最適化されたパスを実装しています。このメッセージ交換ライブラリーを用いて、アプリケーションプログラムは、メッセージ・パッシング・モデルで並列処理記述されます。

❹ おわりに

より多くのノード数に対してスケールするようなアルゴリズム、アプリケーションの開発は、今後もますます加速されていくものと思われます。実際、Blue Gene/Lにも多くのアプリケーションが移植されつつあります。その一方で、このことは、多数のノードをより強力なネットワークで接続することを要求します。ノード当たりの問題のサイズが小さくなるか、あるいはノードの性能が高くなる度合いに比例して、ノード間をより低遅延時間のネットワークで接続することが必要になります。そうしないと、多数のノードに並列化したことによるオーバーヘッドが並列化による効果を打ち消してしまうからです。半導体は、性能・電力消費に関して比例縮小則が維持できない状況になっています。実際、最大性能のプロセッサーの性能対電力効率と、最大の性能対電力効率を与えるプロセッサーのそれとの開きは、どんどん大きくなっています。このような流れの中で、Blue Gene/Lの設計思想・アーキテクチャーは、今後のスーパーコンピューターに対する一つの明確な方向性を与えるものと思われます。

[参考文献]

[1] A. Gara, et. al., 'Overview of the Blue Gene/L system architecture', IBM Journal of Research and Development, vol. 49, no. 2/3, 2005

[2] J. E. Moreira, et. al., 'Blue Gene/L programming and operating environment', IBM Journal of Research and Development, vol. 49, no. 2/3, 2005

[3] 'Blue Gene: A vision for protein science using a petaflops supercomputer', IBM Systems J., vol. 40, no. 2, 2001