2005－HPC－103（23）
2005／8／5

計算科学のための超並列クラスタPACS-CSの概要

朴　泰祐[†,†††]　佐藤　三久[†,†††]　宇川　彰[††,†††]

筑波大学計算科学研究センターで現在開発中の超並列クラスタ PACS-CS (Parallel Array Computer System for Computational Sciences) は、計算物理学、計算物質科学、計算生命科学等、広範囲な計算科学アプリケーションを対象とした新しい発想の超並列クラスタである。計算ノードに関してはメモリへのアクセスバンド幅を最重要ポイントと考え、通常の高性能クラスタとは異なり、ノード当り単一プロセッサという構成を取っている。並列処理用ネットワークは、ノード台数増加に伴うスイッチコストを削減しつつ、実空間モデルに基づく大規模科学技術計算に適するよう、Gigabit Ethernet のトランキングに基づく多次元ハイパクロスバ網を構築する。

これらのコンセプトの下で実装密度を従来の 2-way SMP ノードと同一に保つために、合計 8 ポートの Gigabit Ethernet を持つ単一 CPU ノードを 19inch ラックマウント型の 1U サイズに 2 台搭載可能とする、新型マザーボードを開発する。PACS-CS は 2006 年 6 月に稼動開始予定であり、最終的なシステム規模は、総 CPU 数 2560 台、総ピーク性能 14.3 Tflops となる。

PACS-CS: A massively parallel cluster for computational sciences

Taisuke Boku,[†,†††] Mitsuhisa Sato[†,†††] and Akira Ukawa[††,†††]

We have been developing a large scale PC cluster named PACS-CS (Parallel Array Computer System for Computational Sciences) at Center for Computational Sciences, University of Tsukuba, for wide variety of computational science applications such as computational physics, computational material science, computational biology, etc. We consider the most important issue on the computation node is the memory access bandwidth, then a node is equipped with a single CPU which is different from ordinary high-end PC clusters. The interconnection network for parallel processing is configured as a multi-dimensional Hyper-Crossbar Network based on trunking of Gigabit Ethernet to support large scale scientific computation with physical space modeling.

Based on the above concept, we will develop an original motherboard to configure a single CPU node with 8 ports of Gigabit Ethernet, which can be implemented in the half size of 19 inch rack-mountable 1U size platform. PACS-CS will start its operation on June 2006 with 2560 CPUs and 14.3 Tflops of peak performance.

1. はじめに

計算科学における大規模・高性能計算機の需要は近年増加の一途をたどっている。人間生活に直接結びつく、物性科学・バイオインフォマティクス・生命科学・工学応用は言うに及ばす、素粒子・宇宙等の基礎科学においても、次世代の大規模シミュレーションのために Pflops 級までの要求が既に出ている。これらの要求に計算機科学が応えるためには、計算科学分野との密接な協力が不可欠である。

筑波大学計算科学研究センター[1)]では、その前進である計算物理学研究センター、さらにそれ以前の時代から伝統的に、実アプリケーションに即した超並列計算機システムの構築とその上でのアプリケーション実行という形で、計算科学と計算機科学の研究者が共同研究を行うという形態を取り続けてきた[2)~4)]。つまり、大規模計算機システムの利用者と開発者という、両極に立つ者が互いに問題を共有し合いこれを解決することにより、極めて理想的な研究環境を提供してきたと言える。CP-PACS[4)]は計算物理学研究センターにおいて産学連携の研究体制の下で開発され、ピーク性能 614 Gflops、Linpack 性能 368 Gflops を達成し、1996 年 11 月の TOP500 リスト[5)]の第 1 位となった。CP-PACS は素粒子物理学・物性物理学等における大

† 筑波大学大学院システム情報工学研究科コンピュータサイエンス専攻
Department of Computer Science, Graduate School of Systems and Information Engineering, University of Tsukuba
†† 筑波大学大学院数理物質科学研究科物理学専攻
Department of Physics, Graduate School of Pure and Applied Sciences, University of Tsukuba
††† 筑波大学計算科学研究センター
Center for Computational Sciences, University of Tsukuba

規模計算の他、計算宇宙物理学における複合系計算のプロトタイプ構築[6] 等、延べ 10 年間に渡り、数々の大規模計算を行ってきた。しかし、その後の HPC プラットフォームの性能向上は目覚しく、最新の TOP500 リストでは 500 位のエントリマシンですら 1 Tflops 以上の性能を持つに至っている。

計算科学研究センターでは、今後のより拡大する大規模科学計算のための、CP-PACS に代わるより強力なプラットフォームの姿を模索してきたが、新しいコンセプトに基づく超並列 PC クラスタシステムの構築を 2005 年度より開始することになった。このシステムは PACS-CS と名付けられ、ピーク性能としては CP-PACS の 20 倍以上を目指している。本稿では、PACS-CS のコンセプト、実装方針、予備性能評価、今後の計画について述べる。

2. PACS-CS の開発コンセプト

CP-PACS[4] が開発された 1990 年代前半から中頃は、大規模並列処理技術が開花した時代であった。各大型計算機メーカーは競って超並列計算機（MPP: Massively Parallel Processor）を構築し、それらは各地の大型計算機センターや国立研究所に導入された。当時はマイクロプロセッサの動作周波数が数百 MHz、専用ネットワークのノード当りバンド幅も数百 MB/s という、超並列処理アルゴリズムにとってほぼ理想的な性能バランスが保たれていた。

その後 MPP の開発は、単なる MPP 構築だけのための技術開発では成り立ち難くなり、over Tflops の時代に入ってからは PC クラスタの台頭が目覚しい。近年の TOP500 リストを見ても、高性能 PC クラスタは上位に食い込むだけでなく、あらゆる規模・階層に満遍なく普及している。

このような現状に対し、我々はここ数年間の中期的目標に立つプラットフォーム開発を検討してきた。性能レンジとしては 10 Tflops 級であるが、重要なことはこれまでの我々の超並列システム開発と利用技術の延長上に、いかにして次世代計算機に繋がりかつ現状で十分利用できるシステムを構築するかというコンセプトである。現在の MPP の状況を見れば、全ての要素をオーダーメイドで構築することはもはや不可能であり、従って開発のポイントは**いかにして対価格性能比の高いコモディティ技術を効率的に取り込み、アプリケーションの実効性能の高いシステムを実現するか**ということになる。ここで重要なことは、単にコモディティ製品をそのままの形で購入し、これらを組み合わせてシステムを構築するだけでは、我々のニーズには不十分であるということである。

現在、日本国内においても 10 Tflops 級のクラスタがいくつか導入されている[7),8)]。これらに共通する特徴は、

- 市販の 2-way から 4-way SMP PC サーバ（IA-32 または IA-64）を使用
- Myrinet、Infiniband 等の SAN を fat-tree あるいは clos 網のような多段スイッチ構成で利用

ということである。世界的に見れば、ネットワークにさらに強力な Quadrix 等の MPP 向けのものを使用している例もあるが、ノード構成に関しては概ねこの方針が取られている。

これらは、PC サーバノード、NIC、スイッチ等について、既に HPC 向けに用意されている「売れ筋製品」をアセンブルしクラスタとして構築した結果であると言える。従って、システム全体の対価格性能比や保守性という点で、大学や国立研究所の大型計算機センターでの運用に即している。なぜならば、それらのセンターではアプリケーションユーザが広範に渡り、多数の単一プロセッサをグリッド的に利用する例から数千プロセッサによる並列処理まで、広いスペクトルの利用に応える必要があるからである。

これらの動向に対し、我々筑波大学計算科学研究センターでは、基本的に異なる姿勢で大規模科学計算のプラットフォーム作りを考えている。

- ある程度絞られた応用分野と利用方法を想定し、できる限りアプリケーションの実効性能を高める
- 実効性能に直接関係する属性、すなわちバンド幅・レイテンシ・容量・台数といったファクタを最優先する
- 利用方法を限定することにより、システム構成上の無駄を省く
- できる限りコモディティ技術を利用しコストを抑える
- 市販プラットフォームでこれが満足できない場合、最低限のコストでシステムの一部を開発する

これらの結論として得られる方向性を一言で言えば**コモディティ部品（チップ等を含む）を要素とした超並列計算機を開発する**ということである。このコンセプトの下に、我々は PACS-CS の設計を行った。

3. PACS-CS の設計方針

我々のこれまでのプラットフォームであった CP-PACS は、MPP として以下の特徴を持っていた。

- 300 Mflops のピーク性能を持つ 2048 台のノードによる超並列システム
- 単一 CPU のノード構成と擬似ベクトル処理機構を持ち、高バンド幅メモリに支えられた高い CPU 実効性能（CPU 性能 1 Gflops 当り 4GB/s）
- ３次元ハイパクロスバ網（以下、3D-HXB と略）による高いノード当り通信バンド幅（CPU 性能 1 Gflops 当り 1GB/s）とシステム全体のバイセクションバンド幅（644GB/s）
- 専用ネットワークと支援ソフトウェアによる低レ

イテンシ通信
- 分散された I/O ノードと RAID-5 構成の高バンド幅ディスク装置
- 超並列向け専用 OS による高速なジョブ起動

これらのうち、特にメモリバンド幅とネットワークバンド幅に関する数値的特性は MPP ならではのものであり、現在の PC クラスタとは大きな差がある。これらは主に、この数年間で CPU 性能（動作周波数）が飛躍的な伸びを示しているのに対し、メモリとネットワークの性能が追いついていないという現状から来ている。しかし、我々は現在のコモディティ技術を利用することにより、できるだけこの姿に近いシステムを構築する方法を提案する。以下にその設計方針を示す。

全体構成　コモディティ技術に支えられた対価格性能比の良いプロセッサとネットワークを利用し、数千プロセッサ規模の計算科学のためのインフラストラクチャを構築する。単に既成の PC サーバをネットワークで結合するというだけではなく、必要に応じてボード設計等を行う。

プロセッサ　IA-32 互換機のような高性能コモディティプロセッサをベースに考える。CPU 周波数はある程度高い必要があるが、メモリバンド幅とのバランスを考えたリーズナブルな速度と、消費電力にも配慮した選択を行う。

計算ノード構成　メモリバンド幅、ネットワークバンド幅の両面から考え、SMP 構成は取らない。両バンド幅を実効性能の基本的要件と考え、この点で極力妥協せずにシステムを構築する。

ネットワーク　システム全体のバイセクションバンド幅を無闇に追求せず、実空間モデル等で基本的になる隣接・放送・縮退通信を高速に処理するネットワークを安価に実現する。

ディスク装置　数千ノードのシステム上で、計算途中での一時利用ディスクを高いバンド幅で提供するために、各計算ノードにはある程度の容量のハードディスク装置を個別に搭載する。

実装密度　できる限り高いメモリバンド幅・ネットワークバンド幅を提供しつつ、ノードの実装密度に関しては従来の SMP 型 PC サーバと同等のものを目指す。

以上の方針に従った結果、PACS-CS の実現のためには、専用マザーボードの開発が不可欠であるという結論に至った。CPU 性能当りのメモリとネットワークのバンド幅を追求しつつ、実装密度を通常の SMP ノードと同等に保つためには、コンパクトな単一 CPU 用マザーボードと、通常の I/O バス構成に囚われない計算ノードの実装が必要である。しかし、このために個々のパーツを LSI レベルから開発するのではなく、通常の PC マザーボードを設計・開発するのと同様に、パーツレベルではコモディティ製品を応用する。

4. PACS-CS の実装

前節で述べように、我々は PACS-CS を従来の高性能クラスタの一種というよりも、コモディティ部品で構成された超並列計算機という位置づけで考える。これまでに述べたコンセプトと設計方針に基づき、以下のように PACS-CS を実装する。

4.1 ノード構成

コモディティプロセッサの情勢に鑑み、CPU として IA-32 または IA-64 を検討した。現在のこれらのプロセッサの 2005 年前半時点での最大動作周波数は IA-32 が 3.6GHz、IA-64 が 1.6GHz である。いずれも、消費電力等の観点から動作周波数が頭打ちになっており、Intel を始めとする各メーカーは dual core 構成のプロセッサに向かっている。

我々の目標はノード当りの実効性能の向上であり、CPU 性能とメモリバンド幅を少しでもバランスさせることである。従って、dual core や闇雲に周波数だけが速いプロセッサの投入は効果がないだけでなく、消費電力の点ではむしろ中間的な周波数のプロセッサに高バンド幅メモリを搭載するのが望ましい。また、IA-64 の現状を見ると需要が当初予想ほど伸びておらず、チップセットを含む足回りの充実度から見ると IA-32 との格差が大きい。

以上の点から、我々は Intel Low Voltage Xeon EM64T 2.8GHz を計算ノード用 CPU として採用する。このプロセッサに DDR2 400MHz の SDRAM を 2-way interleaved 構成で装着し、6.4GB/s の理論ピークバンド幅を提供する。EM64T であるため、SSE3 までの SIMD 命令が利用可能で、理論ピーク性能は 5.6Gflops になる。従って、相対メモリバンド幅は CPU 性能 1Gflops 当り 1.14GB/s となる。CP-PACS 等に比べると決して満足できる値ではないが、2-way SMP 構成の Xeon や、4-way SMP 構成の Itanium2 のような標準的な PC クラスタに比べ、かなり高い値を維持している。

この他、個々の計算ノード上には RAID-1 仕様のローカルなハードディスクを設け、スタンドアロンの PC として運用できるようにする。この上で Linux オペレーティングシステムを実行可能とする。

システム全体で 10 Tflops 級のピーク性能を達成するには、ノード数は 2000 台規模になる。最終的に 2560 台のノード数（= CPU 数）を持つシステムを構築する。

4.2 ネットワーク

高性能 PC クラスタ向けネットワークとしては、Infiniband や MyrinetXP のような SAN (System Area Network) が主流となっており、最近では 10 Gigabit Ethernet も候補になりつつある。これらのネットワークは各ノードに対して太いリンク（500MB/s

～1GB/s）を提供し、階層化されたスイッチ網によってある程度のシステムワイドなバイセクションバンド幅を確保している。

PACS-CS で想定されるアプリケーションの基本的な並列化手法は、実空間離散化に基づく超並列処理である。素粒子物理学における QCD 計算、物性物理学における実空間密度汎関数法、宇宙物理学における輻射流体計算等はいずれもこの範疇に属する。これらの手法は、例えば流体力学におけるルジャンドル変換や、問題を FFT に帰着させる方法等に比べ、絶対的な総計算量が増加する傾向にある。しかしながら、問題を多次元メッシュ化されたノードに直接マッピングすることにより、実空間での相互作用は隣接通信に帰着され、ネットワークへの負荷が大幅に削減される。

このように MPP のスタイルを踏襲する計算手法では、広範囲な通信を適度なバンド幅で支援する一般的なクラスタ向け SAN は適さず、単純なメッシュ結合、あるいは CP-PACS で採用された 3D-HXB[9] のようなネットワークが望ましい。3D-HXB は単一の NIC ではなく 3 次元方向に対応した 3 つの NIC によって外部スイッチに接続される。これは、ノード当りに必要な総ネットワークバンド幅を 3 つの NIC に分散させることに相当し、NIC を結合するバスと NIC そのものに要求されるバンド幅を低減するという効果を持つ。例えば、3D-HXB を適度なバンド幅で実装するとなると、1 次元方向当り 200～300MB/s 程度のバンド幅が確保できれば、3 次元同時転送（実空間隣接通信では必要十分）を行った際の総バンド幅を 1GB/s 程度にまで高めることができる。

以上の考え方から、我々は PACS-CS のネットワークを、Gigabit Ethernet（以後、GbE と略）のトランク技術に基づく 3D-HXB 網とすることに決めた。GbE のトランク利用は従来から研究されており、数本程度のトランクであれば高い効率で通信が可能であることが知られている[10]。これに加え、各ノードでソフトウェアによる最大 2 ホップのルーティングを行い、3D-HXB 網を実現する。ただし、実際には想定されるほとんどのプログラムでは隣接通信が基本であり、ルーティングを行う局面は少ないと予想される。

ネットワークを束ねるスイッチに関しても、GbE は非常に高い対価格性能比を実現可能である。高性能の SAN に比べ、GbE の NIC はボードレベルでも 1 万円単位、ネットワークチップ単価ではさらに安くなる。また、SAN のスイッチは元々大規模化が容易なようにバックプレーンの性能や拡張ポート数を大きく取るためにイニシャルコストが高い。しかし、GbE に基づく 3D-HXB であれば、単一リンク当りのスイッチポート数は非常に低い。例えば、4096 ノード構成でも、1 次元当りに必要なのは僅か 16 ポートのスイッチである。無論、全体で 768 台（16 × 16 × 3）のスイッチが必要になるが、この程度のポート数の L2 スイッチの単価は極めて安く、スイッチ側でも大幅な対価格性能比の向上が可能である。

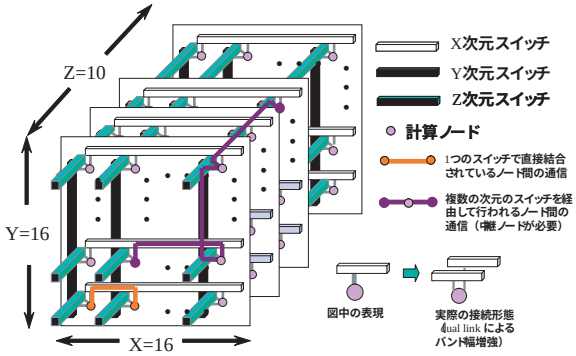

図 1　PACS-CS のネットワーク構成

以上の考え方に基づき、PACS-CS のネットワークを図 1 に示すような構成に設計した。計画では、2560 ノードを結合するため、16 × 16 × 10 構成の 3D-HXB 網となる。

1 次元方向当りの GbE リンク数は 2 とした。この結果、1 次元方向の単方向通信バンド幅は 250MB/s（125MB/s×2）となり、3 次元全ての同時転送を実行する場合の単方向バンド幅は 750MB/s にもなる。CPU 当りのネットワークバンド幅から見れば、これは 2-way SMP ノードに Infiniband を 1 本接続した場合の 1.5 倍もの性能を、極めて安価に実現できることになる。具体的なネットワークバンド幅（3 次元）は単方向で 1 Gflops 当り 134MB/s となる。CP-PACS での値（換算すると 1GB/s）には大きく劣るものの、一般のクラスタが提供する 25～80MB/s 程度に比べ優れたバンド幅を提供可能である。さらに、このバンド幅を単一 PCI バスではなく次元方向別の複数の PCI バスで支えているため、高倍率の PCI-Express を導入することなく足回りを支えることが可能である。

また、上記の並列処理データ転送用ネットワークとは別に、システム全体に一般的なネットワークサービス（NFS, NIS, DNS 等）を提供するための通常のツリー構造ネットワーク（これを運用系ネットワークと呼ぶ）と、システムコンソール機能を集約し、PACS-CS システム全体を統合的に管理するための独立なツリー構造ネットワーク（これを管理系ネットワークと呼ぶ）もそれぞれ用意する。これらに供される GbE ポートはデータ通信用とは別途用意する。

そして、さらにこれら全てのスイッチ（システム全体で数百台）を SNMP で監視・管理するための監視系ネットワークを設ける。特にデータ転送用ネットワークのスイッチはツリー構造を持たないため、個々のスイッチに管理系リンクを張り、集約的に管理する。

4.3　マザーボード開発と全体仕様

以上のノード構成とネットワーク構成を実現する

マザーボードは市販品では存在しない。不要な 2nd CPU ソケットがなく、最低限のメモリスロットと多数の GbE ポートの高密度実装という条件を満たすため、我々は PACS-CS の専用ボードを開発する。

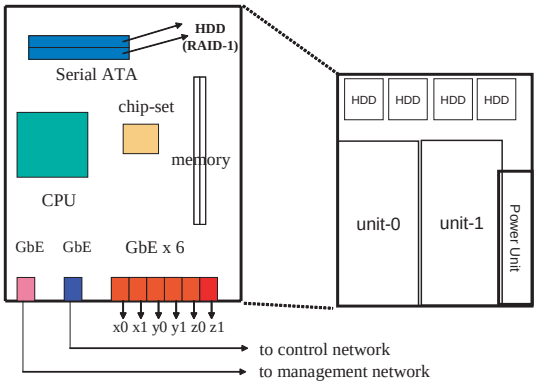

図 2 マザーボードとシャーシへの組み込みイメージ

図 2 にマザーボードの概略とラックシャーシに組み込んだイメージを示す。CPU、メモリ、チップセット、ローカルディスク（RAID-1 ミラーリング）等については通常の PC サーバと変わりはない。特徴的なのは合計 8 本の GbE ポートである。我々は、多数の GbE ポートを実装するため、PCI バスのような拡張スロットを用いることなく、NIC chip を直接オンボード実装する。これらのポートは標準的な RJ-45 インタフェースで外部スイッチと接続される。

GbE ポートは全部で 8 本用意されるが、うち 6 本（ボード中央に位置する x0, x1, y0, y1, z0, z1）は 2 本ずつ束ねられ、3D-HXB 網の 3 つの次元方向のスイッチにそれぞれ接続される。この他に運用系と管理系の 2 系統のネットワーク用の個別ポートがある。以上、合計 8 本の GbE は 2 本ずつ 1 つの NIC chip (Intel 製を予定) で管理され、それら 4 つの NIC chip は各々独立な 133MHz/64bit PCI-X バスに接続される。従って、これらを賄う総 I/O バスバンド幅は 4GB/s となり、全 GbE ポートの双方向総バンド幅のさらに倍という、余裕を見た設計になっている。

図 2 の右側に示すように、システムはこのボード 2 枚を並べて標準的な 19 インチラックの 1U の厚さに収められる。スペース削減のため、2 台のマザーボードで電源ユニットが共有される。

計算ノード群の他に、3D-HXB を構築するための大量の GbE スイッチが必要となる。システム構成が $16 \times 16 \times 10$ であるため、16 ポートのスイッチを単位として用いるのが最適であるが、実装密度を考慮して 48 ポートのスイッチ（2005 年度後半には現在の 24 ポートスイッチ並みのポート単価で出回ると予測）を VLAN 構成によって区切って用いる。16 ポートスイッチ単位で計算すると、2560 ノードのためには X, Y 各次元用に 320 台、Z 次元用に 512 台のスイッチが必要になり、これを 48 ポートスイッチで構成すると 351 台となる。さらに、各ノードの並列処理用 GbE ポートは 6 本ずつであるから、全ノードとスイッチ間を結ぶ Ethernet ケーブルは 15,360 本になる。これらのスイッチとケーブルを効率的にラッキングする設計を検討中である。

以上をまとめた現在の設計仕様を表 1 に示す。

表 1 PACS-CS の 2005 年 7 月現在における設計仕様

ノード台数	2560 ($16 \times 16 \times 10$)
理論ピーク性能	14.3 Tflops
ノード構成	単一 CPU/ノード
CPU	Intel LV Xeon EM64T 2.8 GHz, 1MB cache
メモリ容量	2GB/ノード
メモリバンド幅	6.4GB/s (1Gflops 当り 1.14GB/s)
並列処理ネットワーク	3 次元ハイパクロスバ
リンクバンド幅	単方向 250MB/s/次元 単方向 750MB/s (3 次元同時)
バイセクションバンド幅	640 GB/s
ローカルディスク容量	160 GB/ノード (RAID-1)
ファイルサーバディスク容量	10TB (RAID-5)
オペレーティングシステム	Linux (FedoraCore3)
システム管理ソフトウェア	SCore
プログラミング言語	Fortran90, C, C++, MPI
システム規模	総ラック数：59 総消費電力：570kW

4.4 システムソフトウェア

運用系ネットワークを利用することにより、システム全体はフラットな IP アドレス空間を持つ一般的な Linux クラスタを構成する。システム全体の管理は大規模クラスタ管理ミドルウェアである SCore を利用する[11]。この上で、適当なキューイングとバッチ管理を行い、ユーザジョブの管理を行う。また、3D-HXB を構成するネットワークハードウェアを有効利用し、GbE のトランク利用と HXB 上のルーティング制御を行う特殊ドライバを、ハードウェアに先行して開発している[12]。並列プログラミングはこのデバイスをベースとした通常の MPI で行う。

運用系ネットワーク上のファイルサーバについては、数千ノードからの集中アクセスが発生しないよう、基本的なファイルアクセスは、一旦各ノードのローカルディスクへのファイルコピーを行った上で行い、このファイルコピーを適当にスケジューリングすることによってネットワークトラフィックを制御する。このための支援ソフトウェアを準備し、ジョブ実行のバッチシステムと連動させる予定である。また、SCore を利用するため、バイナリ実行時のファイルシステムへの負荷はそれほど高くならないと考えている。

5. 予備性能評価

現在、PACS-CS の構築に合わせ、各種アプリケーションプログラムの作成・改良を行なっている。特に主要なアプリケーションに関しては、設計仕様作成段階で予備的な評価が行われた。具体的には、PACS-CS で想定される仕様と等価な単体ノードを市販 PC サーバ上で構成し、並列化されたアプリケーションの単体プロセッサ上での実行速度を評価した。一例として、QCD（量子色力学）計算において我々が標準的に用いているベンチマークプログラムを、PACS-CS で搭載予定の CPU とメモリに合わせてチューニングしたバージョン[13] における、単体ノードでの実行結果を表 2 に示す。

表 2 QCD ベンチマークにおける単体ノードの予備性能評価 (on LV Xeon 2.8GHz, EM64T)

プログラミング	性能 (Gflops)
Fortran90 (SSE3 vector on)	1.451
SSE3 組み込み関数利用	1.910
SSE3 アセンブラ記述	1.873

このベンチマークのコア部分は、キャッシュが有効利用できない複素数ベクトル処理であるが、このような状況でもプログラミング次第でピーク性能の 34%もの実効性能が得られている。この効率をさらに高めるため、プログラム及びアルゴリズムの改良を進めている。

これ以外にも、実空間密度汎関数法による物性第一原理計算、アンサンブルモデルによる気象予測、生物系統樹の構築等、様々な分野のアプリケーションの PACS-CS 上での実行に向けての開発・性能評価を行っている。特に QCD 及び物性計算は PACS-CS における主要アプリケーションと位置づけられており、単体プロセッサ性能だけでなく、ネットワークでの通信も加味した仮想評価を行なった。現在想定している 3D-HXB ドライバが予想通りの性能を発揮した場合、想定している典型的な問題サイズにおいて、通信時間が全実行時間に占める割合は、QCD の場合で 23%程度、物性計算の場合で 10%程度と予測している。

6. おわりに

PACS-CS は 2004 年度後半から製造請負に関する政府調達作業を開始し、2005 年 7 月に株式会社日立製作所がこれを落札した。現在、システム実装に向けた実質的な検討を進めている。ハードウェア調達とは別に、3 次元ハイパクロスバ網用のネットワークドライバの開発も進めている。現在のプロトタイプでの性能評価に基づき、実機向けの開発を行うための調達も進めている。

PACS-CS は 2006 年 6 月に稼動開始予定であり、その理論ピーク性能は 14.3 Tflops で完成時には日本国内での最高性能クラスタとなる見通しである。さらに、単純なピーク性能だけでなく、本稿で述べたようなバンド幅重視設計に基づく実効性能の高さも期待される。

筑波大学計算科学研究センターでは、PACS-CS を中心とした各種大規模計算科学アプリケーションを実行する予定である。これまでの HPC クラスタでは不可能であった各種問題にチャレンジすると共に、本プロジェクトによって開発された技術が他の計算科学向けクラスタ構築に役立つことを期待する。

謝辞 本プロジェクトを進めるに当り、システム基本設計及びアプリケーション性能予測等多くの面で協力を頂いた、筑波大学計算科学計算センター関係者諸氏に感謝する。

参 考 文 献

1) http://www.ccs.tsukuba.ac.jp/
2) http://www.rccp.tsukuba.ac.jp/people/shirakaw/PAX/
3) T. Shirakawa, et al., “QCDPAX - an MIMD array of vector processors for the numerical simulation of quantum chromodynamics”, Proc. ACM/IEEE Conference on Supercomputing, 1989.
4) T. Boku, et al., “CP-PACS: A massively parallel processor for large scale scientific calculations”, Proc. ICS’97, 1997.
5) http://www.top500.org/
6) T. Boku, et.al., “Heterogeneous Multi-Computer System: A new platform for Multi-Paradigm Scientific Simulation”, Proc. of International Conference on Supercomputing, 2002.
7) 工藤 知宏 他, “AIST スーパークラスタ構想”, 情処研報 2002-HPC-91 (SWoPP 松山 2002, 2002.
8) R. Himeno, “PC Cluster as Main HPC resource at Supercomputing Center”, invited talk, CLUSTER2004, 2004.
9) 朴 泰祐 他, ”ハイパクロスバ・ネットワークにおける転送性能向上のための手法とその評価”, 情報処理学会論文誌 Vol.36, No.7, pp. 1610-1618, 1995.
10) 住元 真司 他, “複数の Ethernet を束ねる Network Trunking 機構の提案と 1024 プロセッサ PC クラスタ上での性能評価”, HPCS2004 論文集, 2002.
11) http://www.pccluster.org/
12) 住元 真司 他, “PACS-CS のための Ethernet を用いた高性能通信機構の設計”, 情処研報 2005-HPC-103 (SWoPP 武雄 2005), 2005.
13) 石川 健一, “QCD 性能評価ベンチマーク Mult-Bench_v2.62_sse3_64”, 2005.