特集

PCクラスタ

Technologies for PC Cluster Systems

あらまし

パーソナルコンピュータ（PC）のハードウェア性能の劇的な性能向上とともに，複数のPCを高性能インタコネクトで結合し，高い処理性能を実現するPCクラスタシステムが注目を浴び構築されている。しかし，PCクラスタ上で，高速で安定したアプリケーション性能を実現するにはいくつかの機能と技術が必要である。

本稿では，PCクラスタについての一般知識とPCクラスタ上で安定した高いアプリケーション実行性能を実現するために重要な技術について述べた後，RWCプロジェクトで開発されたSCoreクラスタシステムソフトウェアの概要，そして富士通のPCクラスタに対する取組みおよび将来の展望について述べる。

Abstract

Based on the recent major upgrades of the hardware performance of personal computers (PCs), PC cluster systems, which link multiple PCs together via high-performance interconnections to achieve high throughput capacity, have been constructed and are coming to prominence. However, consistent, high-speed execution of applications on a PC cluster still requires some new functions and technological improvements. This paper gives a general description of PC cluster systems and the key technologies for achieving stable and high-application performance on a PC cluster. It also gives an overview of the SCore cluster system software developed by the RWC project and Fujitsu's present activities and future prospects for PC cluster systems.

住元真司（すみもと　しんじ）

ITコア研究所グリッド&バイオ研究部 所属
RWCPに出向中はSCoreの研究開発に従事。富士通帰任後，現在，PCクラスタシステムの研究開発と普及活動に従事。

まえがき

近年のパーソナルコンピュータ（PC）の普及による劇的なマイクロプロセッサの高速化と低コスト化により，複数台のPCとネットワークを用いたシステム構築が広まっている。このシステムをPCクラスタと呼ぶ。現在，PCクラスタにはその冗長性を利用して高信頼システムの実現をねらったものと，複数のコンピュータ資源を用いて大規模システムの実現をねらったものがある。

近年，とくに後者のPCクラスタの発展が目覚しく，2002年11月に発表された世界中のハイパフォーマンスコンピュータのLinpack性能を競うTOP500リスト（http://www.top500.org）においては，PCを2,304台用いたPCクラスタが5.6 TFLOPSの計算性能を実現し世界で第5位にランクインされた。これ以外にも上位100位以内に15台のPCクラスタがランクインされ，その勢いはとどまることを知らない。TOP500リストにランクされているPCクラスタでは，一つを除きLinuxが採用されている。

PCクラスタがベクタ型コンピュータなどほかの商用スーパーコンピュータと異なる点は，最低2台のPCがあればだれでも構築可能なことである。それも，LinuxとPCクラスタ用のフリーソフトウェアを利用すれば，ソフトウェアのコストを掛けずに構築できるのである。しかし，安定した性能を実現するためには，様々な知識とノウハウが必要であるのも事実である。

本稿では，複数のPCを用いて大規模システムの実現をねらったPCクラスタについて，その基本的な知識と技術を解説する。そして，PCクラスタ上で商用スーパーコンピュータに匹敵する機能と性能を実現するSCoreクラスタシステムソフトウェア（以下，SCore）を紹介し，富士通のPCクラスタへの取組み，今後の展望について述べる[1]。

PCクラスタの特徴と重要な機能

ここ数年のPCの性能向上と低コスト化には目を見張るものがある。PC向けのマイクロプロセッサの動作周波数は既に3 GHzを超えている。また，メモリ，ディスクについても高速化，大容量化が進んでいる。

コンピュータネットワークについてもインターネットの爆発的普及により，ギガビットクラスのネットワークの普及が急速に進んでいる。Ethernetといったコモディティネットワークのほか，高性能なクラスタ専用ネットワーク（インタコネクト）の開発も盛んである。

このような中，PCを高性能インタコネクトで接続したPCクラスタが注目を浴びている。たくさんのプロセッサ，メモリ，ディスクを搭載しているPCクラスタは次のような処理に向いている。

・台数分のプロセッサ処理能力による高速な計算処理
・台数分のメモリ容量（帯域）による巨大データ処理
・ディスク（I/O）容量（帯域）による巨大データ検索

これらの処理を，PCクラスタ上で実行するためには，複数のPCに処理を割り当てる機能が最低限必要である。しかし，さらにPCクラスタを活用しようとする場合，以下に挙げる機能が必要になる。

（1） ノードPC（PCクラスタを構成するPC）間通信を処理ボトルネックとしない高性能通信機能
（2） PCクラスタ上で効率よく処理を実行するジョブスケジューリング機能
（3） 長時間実行ジョブに対する耐故障機構
（4） クラスタファイルシステム
（5） 計算結果の可視化システム

以下では，ここで挙げた機能のうち特に重要である，高性能インタコネクト，ジョブスケジュール機能，耐故障性のための機能について述べる。

高性能インタコネクトと高性能通信

PCクラスタが注目され始めた1990年代半ばは，Ethernetが普及の途中にあった。しかし，通信速度は10 Mbpsと遅くPCクラスタには不十分であった。このため，PC間の通信性能を高めるために，ギガビットクラスのMyrinet，SCIなどのクラスタ向け高性能インタコネクトが開発され使われるようになった。クラスタ向けインタコネクトの性能は時代とともに高速化し，現状ではMyrinet2000，SCI，QsNETといった2〜3 Gbpsのものが主流となっている。一方，Ethernetを代表とするコモディティネットワークも高速になり，現在では1 GbpsのGigabit Ethernetが普及期を迎えている。さらに

ネットワークの高速化はとどまらず，10 G Ethernet, Infinibandといった10 Gbpsクラスのネットワークの開発が進んでいる。

このような高性能インタコネクトがPCクラスタ上で威力を発揮できるかどうかは，実行するアプリケーションに依存する。アプリケーションによっては100 MbpsのEthernetで十分なものがある一方，数Gbpsの高性能インタコネクトを必要とするものもある。一般には，結合するPCの台数が増加すると多くのPCと通信を行う必要があるため，高性能なインタコネクトの採用が望まれる。

ジョブスケジュール機能

たくさんのPCを結合したPCクラスタではノードPCを一つのシステムと見せる実行環境のほか，効率的にジョブを割り当てるスケジュール機能が重要である。

代表的なスケジューリング手法としては，バッチスケジューリングとTSS（時分割スケジューリング）がある。商用スーパーコンピュータでは，バッチシステムを導入し複数のジョブ間の資源管理を実行している場合が多い。

耐故障性のための機能

PCクラスタは，一般にノードPC数が増えるにつれて信頼性が低下するため，数週間にわたる大規模計算を行う場合，途中で機器故障により計算が異常終了することも考慮しなければならない。

異常終了のリスクを減らす方式には，複数PCクラスタで多重実行する方式や，チェックポイント機構を用いる方式がある。

多重実行は容易に信頼性の向上を実現する方式である。ノードPC数が最低でも2倍必要になるが，PCクラスタは安価に実現できるので，機器故障がクリティカルな処理に適応するには有効な手法である。

チェックポイント機構は，実行中のプログラム実行イメージ（プログラム，データ，スタックなど）を一時的にディスクに書き出して，故障から復帰後にディスクに格納したプロセス実行イメージよりプロセス実行を再開する方式である。PCクラスタ用のチェックポイント機構は，上記のプロセス実行イメージに加え，送受信バッファや通信プロトコルの状態を含めたネットワークの状態をプログラム実行イメージとしてディスクに格納する必要がある。

SCoreクラスタシステムソフトウェア

PCクラスタとしてBeowulfが良く知られている。1990年代半ば，当時NASAの研究者であったドナルド・ベッカー（Donald Becker）とトーマス・スターリング（Thomas Sterling）がLinuxとSocket上のMPI（メッセージ通信インタフェース）ランタイムシステムを用いてPCクラスタを構築し，そのシステムをBeowulfと名付け米国内に広めて回った。とくにLinuxを含めフリーソフトウェアだけでPCクラスタが構築可能であったため，PCのハードウェアの性能向上とともにPCクラスタは急速に広まった。

Beowulfで提供されている機能は，MPIランタイムシステムであり，MPIランタイムシステムから進んで，ジョブスケジューリング機能などPCクラスタをより活用する機能を統合したものをクラスタシステムソフトウェアと呼ぶ。

SCoreはRWC（リアルワールドコンピューティング）プロジェクト（http://www.rwcp.or.jp）で開発されたLinux上で稼働するオープンソースのクラスタシステムソフトウェアである。RWCプロジェクトは2001年度にそのプロジェクト活動を終了したが，プロジェクト終了後のSCoreの研究開発と普及活動はPCクラスタコンソーシアム（http://www.pccluster.org）に引き継がれている。

SCoreは，複数種類のネットワークをサポートした高い通信性能を実現しているだけでなく，商用スーパーコンピュータに匹敵するジョブスケジュール機構と耐故障性機構を含む運用管理機構を備えている点，100台を超える大規模クラスタの稼働実績が多い点がBeowulfなどのMPIランタイムシステムとの大きな違いである。

ここまでの機能をトータルに備えたクラスタシステムソフトウェアは，現時点ではSCore以外には存在しない。

SCoreに採用されている技術

SCoreにはこれまで述べてきた高性能通信，スケジューラ，耐故障性を実現する機能が実現されている。SCoreの構成を**図-1**に示す。PMv2通信機構，

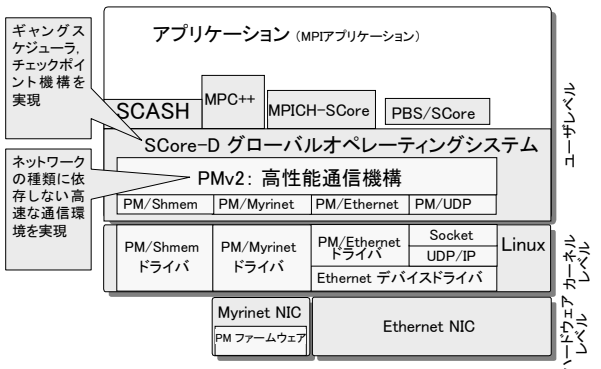


図-1　SCoreの構成
Fig.1-Architecture of SCore.

SCore-Dグローバルオペレーティングシステム，MPIライブラリ（MPICH-SCore），バッチシステム（PBS/SCore）などから構成される。

● **PMv2高性能通信機構**

PMv2は複数種類のネットワークハードウェアをサポートし，ネットワークの種類に依存しない共通のAPIを上位のSCore-DやMPICH-SCoreに対して提供している。このため，上位のアプリケーションはネットワークの違いを意識することなく実行することができる。現在，PMv2がサポートするネットワークは，Myrinet，Ethernet，UDP/IP，共有メモリである。この中でEthernetについては複数のEthernetのカードを用いてバンド幅の向上を実現するNetwork Trunking機構が実現されている。Gigabit EthernetとMyrinet上でのPMv2レベルの通信性能を**表**-1に示す。

● **SCore-Dグローバルオペレーティングシステム**

SCore-DはPCクラスタ上でのランタイムシステムやジョブスケジューリングの機能のほか，ギャングスケジューラによる複数ジョブでのTSS実行機能を実現している。

また，信頼性を高める機構としてチェックポイント機構をも実現している。SCore-Dの提供するチェックポイント機構ではユーザプログラムへの変更は一切不要で，プログラム起動時にコマンドオプションでチェックポイントを採取する時間間隔を指定する。さらに，1台のノードPCのディスクが故障するとチェックポイントからのリスタートができなくなるため，複数のノードPCにデータを冗長に分割して格納している。

表-1　PMv2レベルの通信性能

	ラウンドトリップ時間	バンド幅
PRO1000XT×1枚	23.4 μs	119.1 Mバイト/秒
PRO1000XT×2枚	23.5 μs	226.0 Mバイト/秒
Myrinet 2000	11.1 μs	245.8 Mバイト/秒

（注）Pentium III，66 MHz 64ビットPCI，Linux 2.4.18上でのGigabit EthernetとMyrinet 2000の結果

- 2001/5　PCクラスタコンソーシアム設立準備委員会発足に協力
- 2001/5　ハイエンドLinuxクラスタ構築サービス（プレスリリース）
- 2001/10 世界初InfiniBandインタコネクトによるクラスタ実現（SS研デモ）
- ➢ （2001/10 PCクラスタコンソーシアム発足）
- 2002/2　世界初のPCクラスタ上での高性能ファイルシステムDAFS（Intel Developer Forum）
- 2002/2　GbE ×2接続によるブレードサーバPCクラスタ（Intel Developer Forum）
- 2002/4　バイオインフォマティクス分野でクラスタ向け高性能相同性検索プログラム開発（Hi-per Blast）
- ➢ （2002/6 グリッド協議会発足）

現在PCクラスタコンソーシアム技術開発部会でSCore技術開発，PCクラスタ市場拡大に向け，技術支援，普及活動を推進中

図-2　富士通のPCクラスタに関する取組み
Fig.2-Fujitsu activities about PC cluster.

このギャングスケジューラによる時分割実行機能とチェックポイント機構はすべてユーザレベルで実現されており，SCoreが世界で初めて実現した機能である。

富士通のPCクラスタへの取組み

富士通のPCクラスタへの取組みを**図**-2に示す。いち早くLinux上でのクラスタ構築サポートを開始した。また，PCクラスタコンソーシアムの設立準備から係わり，その活動においても，SCoreの開発，技術支援や普及活動と広範囲に貢献している。

富士通研究所は上記に述べた富士通全体としての取組みのほか，PCクラスタ上でのインタコネクトに関する研究開発，PCクラスタ向けのアプリケーションの開発，最適化を進めている。

ここで，富士通研究所で実施したPCクラスタ向けアプリケーションの開発と最適化に関する成果としてPCクラスタ向けに最適化したHi-per BLASTとMASPHYCの実行性能を紹介する。

図-3はバイオ分野で利用されるBLASTをPCクラスタ向けに最適化したHi-per BLASTの性能のグラフである。Hi-per BLASTは，PCクラスタの持つ台数分のメモリ容量を使いディスクアクセスを減ら

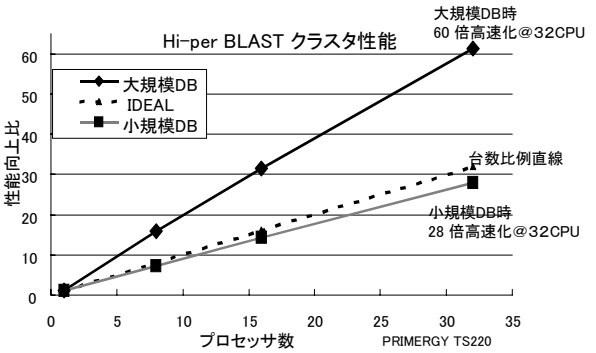
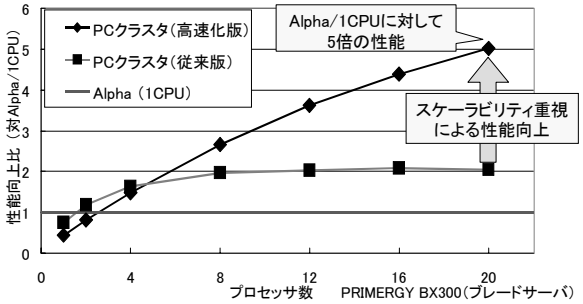


図-3　PCクラスタ向け高性能BLAST：Hi-per BLAST
Fig.3-High performance BLAST for PC cluster: Hi-per BLAST.



（注）次世代微細CMOS用高誘電率ゲート絶縁膜シミュレーション
（富士通研究所シリコンテクノロジ研究所データ）

図-4　PCクラスタ向けMASPHYC高速化
Fig.4-MASPHYC performance tuning for PC cluster.

し，さらに動的なスケジューリングを行い高い実行性能を実現している。

図-4は分子動力学のプログラムである富士通のMASPHYCをPCクラスタ向けに最適化した性能のグラフで，富士通研究所シリコンテクノロジ研究所の実データを用いた結果である。PCクラスタでは台数が増えるに従ってノードPC間の通信オーバヘッドが問題になる。そのため，この最適化では計算の一部を各ノードPCで冗長に行い，通信負荷を減らすことによって高い実行性能を実現している。

この結果，従来Alphaプロセッサを1台用いて1週間掛かっていた計算がPRIMERGY BX300の20プロセッサを用いて1日半で実行可能になった。

PCクラスタの課題

PCクラスタをより広く使ってもらうための課題としては以下の事柄が挙げられる。

・利用できるISVアプリケーションの充実
・ビジネス利用に必要なトータルなソリューションサービスの提供
・PCクラスタの認知度の向上

とくにISVアプリケーションの充実とトータルなソリューションの提供が重要である。従来のスーパーコンピュータではハードウェア価格が高価であったためハードウェアが重要視され，アプリケーションやソリューションは重要とみなされていなかった。しかし，PCクラスタは安価に高い性能が実現可能であるため，相対的にアプリケーションの重要性が高まっている。そのため，例えば，コンパクトなブレードサーバとアプリケーションを組み合わせたサービスの提供が必要とされている。

以上の課題を解決するために，富士通はPCクラスタコンソーシアムでの活動だけでなく，自社内でも研究開発とソリューション開発を推進している。

む　す　び

本稿では，複数のPCを用いて高い処理性能を実現するPCクラスタについて，その概要と技術，富士通の取組みについて述べた。

PCクラスタが従来の商用スーパーコンピュータと大きく異なるのは，Linuxを含めオープンソースソフトウェアをベースとしているため，ハイパフォーマンスコンピューティング領域での共通のオープンプラットフォームになり得ることである。標準化によりユーザが得られるメリットは計り知れなく，また，アプリケーションやツール，サービスなども相互に流通し，ユーザも飛躍的に増加していくと考えている。

富士通と富士通研究所はPCクラスタコンソーシアムの活動を通じてオープンプラットフォームとしてのPCクラスタの技術開発と普及を推進していくとともに，トータルなPCクラスタソリューションの提供を推進していく。

参 考 文 献

(1)　石川裕，住元真司ほか：Linuxで並列処理をしよう－SCoreを用いたクラスタ構築－．共立出版，東京，2002.