特集

Linux搭載PCクラスタによるHPCへの取組み

Fujitsu's Work on Linux PC Clusters for High-Performance Computing

あらまし

PCクラスタは，パソコン（PC）のCPU性能向上と搭載メモリの大容量化，およびLinuxなどのオープンソースソフトウェアの充実を背景にHPC分野に登場した。PCクラスタは当初，安価だが手間のかかるシステムとして小規模な専用システムとして利用されていた。しかし現在では，“TOP500”（www.top500.org）のランキングの上位を含めた多くのシステムに採用されており，従来のメーカ製スーパーコンピュータと同様に，大規模な計算機センタの中核システムとして利用されるようになってきている。

本稿では，PCクラスタの要素技術を紹介し，PCクラスタを計算機センタの中核システムへ適用する場合の要件と課題を示す。その上で，富士通のPCクラスタへの取組みについて，今後の課題を含め述べる。

Abstract

Recently, there have been significant improvements in the CPU performance and installed memory size of PCs. Also, open-source software such as Linux has recently been enhanced. Because of these developments, PC clusters are now being applied in high-performance computing (HPC). PC clusters were initially used as inexpensive, but troublesome, small-scale private systems and are currently used in many systems that are ranked in the TOP500 (www.top500.org). Similarly to off-the-shelf supercomputers, they have been used as main systems in large-scale computing centers. This paper outlines the trends of PC clusters and discusses some of the issues regarding their use as the main system of a computing center. It also describes Fujitsu's efforts to provide PC clusters to customers in the HPC area.

井口裕次（いぐち　ゆうじ）

Linuxソフトウェア開発統括部 所属
現在，HPC分野向けソフトウェアの開発に従事。

まえがき

パソコン（PC）の高性能化・大容量化とともに，Linuxなどのオープンソースソフトウェアの充実に伴い，複数台のPCをネットワークで接続した並列計算システムがHPC（High Performance Computing）分野で広まっている。この並列計算システムをPCクラスタ[1]と呼ぶ。PCクラスタは，PC本体およびネットワークにいわゆるコンシューマ製品とLinuxなどのオープンソースソフトウェアを利用することにより，安価で比較的高性能なシステムを構築することができる。ただし，システム構築や日々のシステム運用に複数台のPCを意識した独特のノウハウが必要であったことから，登場当初は安価ではあるものの手間のかかるシステムとして，一部の研究者が利用する専用システムであった。

しかし，Webフロントシステム向けなどのラックマウントタイプやブレードタイプのIAサーバの登場によるハードウェア面での運用性の向上，MyrinetやInfiniBandなどの高速ネットワーク（インタコネクト）の普及，および各種ディストリビューションの登場によるオープンソースソフトウェアの使い勝手の向上などにより，コンピュータが得意な一部の研究者の手から離れ，一般的なユーザにも受け入れられるようになった。この結果，数年前からTOP500[2]の上位に常連としてPCクラスタがランクインしており，従来のメーカ製スーパーコンピュータのように，大規模な計算機センタの中核システムとしても利用可能な性能を得るまでになってきている。

本稿では，PCクラスタを構成する要素技術を振り返るとともに，計算機センタの中核システムとして利用される上で重要と思われる技術について検討する。さらに，これら技術に対する富士通の取組みを紹介し，今後の展望・課題について述べる。

PCクラスタの要素技術

PCクラスタは，複数台のPCを並列で使用し計算を行わせることにより高い演算性能を得る，いわゆる並列計算システムである。このため，複数台のPC上で並列して計算を行わせるための基本的な技術が必要となる。並列計算を行うための要素技術としてPCクラスタでは，一般にBeowulf[3]とSCore[4]のいずれかが用いられる。

BeowulfはNASAの研究者による，コモディティ製品およびオープンソースソフトウェアにより，安価で高性能なシステムの構築を試みたプロジェクトである。まさに，PCクラスタの生みの親的な存在である。Beowulfプロジェクトでは，Linuxと汎用的なプロトコル（TCP/IP）上に実装した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組み合わせることによりPCクラスタを構築した（以下，Beowulf型クラスタ）。Beowulf型クラスタの登場により，安価で比較的高性能なシステムの構築が実証されたことから，米国を中心に急速に広まっていった。

Beowulf型クラスタは，MPIプログラムを複数台のPC上で実行する環境だけを提供するものであり，大きく二つの課題があった。一つは，TCP/IPを利用しているために高い通信性能が得られないということ。もう一つは，複数台のPCを効率的に運用するための機能が備わっていないことである。

これらの課題を克服するために，SCoreとよばれるクラスタ向けソフトウェアが開発された。SCoreは，経済産業省のRWC（リアルワールドコンピューティング）プロジェクトで開発されたLinux上で稼働するオープンソースソフトウェアである。RWCプロジェクトは2001年度に活動を終了したが，SCoreの研究・開発・普及活動は，PCクラスタコンソーシアム[4]に引き継がれている。

SCoreは，汎用プロトコルの利用も可能であるが，ネットワーク（インタコネクト）の特性に応じたプロトコルの利用により，高い通信性能を実現している。また，MPIプログラムを複数台のPC上で実行する環境を提供するだけでなく，複数台のPCを一括して操作するための各種運用機能を提供している。さらに，実行完了までに非常に長い時間を要するプログラムのために，途中で何らかの異常によりプログラムが停止させられても，中断点などからプログラムの実行を再開可能とするチェックポイントリスタート機能などを有している。

計算機センタ要件とPCクラスタの現状

先に述べたようにPCクラスタは，性能面・機能面の充実に伴い，従来の専用システム的な利用形態から，計算機センタの中核システムとして，不特定

多数の利用者による共用システムに位置付けられつつある。また，SCoreの登場により，Beowulf型クラスタの課題を克服してきている。ここでは，共同利用センタで求められる基本的な以下の要件について検討し，SCoreにより構築されたPCクラスタの課題を考察する。

（1） ISVアプリケーションによるサービスの提供
（2） バッチシステムによる共用環境の実現
（3） 課金・統計情報の収集
（4） クラスタ内ファイル共用
（5） 容易な運用管理の実現
（6） 計算サービスの安定供給

これらの要件は，計算機センタを運用していく上で最低限の要件であり，富士通のスーパーコンピュータPRIMEPOWER HPC2500をはじめ，各社のスーパーコンピュータ・HPCサーバで既に実現されている機能である。

これらの機能要件に対し，現在のPCクラスタがどの程度対応できているのかを，以下に示す。

（1） ISVアプリケーションによるサービス提供

PCクラスタの普及に伴い，HPC分野で利用される著名なISVもPCクラスタをサポートしてきている。しかし，現状では各ISVがサポートしているのは，一般にBeowulf型クラスタである。これは，SCoreは日本国内のプロジェクトから誕生した経緯を持ち国外での知名度が低いこと。また，高い通信性能を持つが，そのかわりにSCoreが提供するMPIライブラリを再リンクしなければならないことから，ISVやアプリケーション開発者から敬遠されやすいためと考えられる。

（2） バッチシステムによる共用環境の実現

計算機センタが，提供するサービスは特定の利用者に偏らず，公平に提供する枠組みが必要である。このためには，なんらかの単位によりサービスの提供量を数値化する必要がある。一般に計算機センタではバッチジョブを一つの尺度としてサービスの数値化を行っている。すなわち，バッチジョブの実行環境が必要となる。

現状のPCクラスタは，Beowulf，SCoreとも，オープンソースソフトウェア（OpenPBSなど），または商用のバッチ処理システム（NQS，PBSpro，LSFなど）を利用することができ，バッチジョブによる計算機利用サービスを比較的容易に提供することができる。

（3） 課金・統計情報の収集

計算機センタではその運営費を回収するために，提供したサービスへの対価を利用者へ要求する場合がある。対価請求は（2）で述べたようなサービスの提供単位であるバッチジョブを単位として行うこととなる。すなわち，個々のジョブへ提供したサービス量であるジョブが使用したシステム資源の量を課金データとして収集・蓄積・閲覧する機能が必要となる。また，この課金データは，将来のシステム拡充あるいは縮小に向けた重要な統計データともなる。

PCクラスタでは，一つのジョブが複数PCを利用して並行に実行される。このため，ジョブの課金データもこのような並列ジョブに対応して複数PC上の資源使用量を一括集計できる仕組みも必要である。現状，SCoreなどのクラスタソフトウェアにおいて並列ジョブに対応した課金・統計情報の採取機能をサポートしている。しかし，一般的な実装としてLinuxのプロセスグループを各ジョブの識別に利用しており，悪意のある利用者が不正を行うことが可能である。公平・公正な課金が求められる計算機センタでは，大きな問題となる。

（4） クラスタ内ファイル共用

多数のPCから構成されるPCクラスタでは，実行するプログラムが格納されているファイルや入力データを各PCで共有する必要がある。一般にはNFS（Network File System）によってPC間でファイル共用を行うか，プログラム実行前に必要なファイルを転送する手法をとっている。いずれの手法も，PCクラスタのシステム規模が大規模化するほど，現実的な解とは言えなくなってくるという問題がある。

（5） 容易な運用管理の実現

多数のPCから構成されるPCクラスタは，システムの演算能力に比例して，PCの台数が増えることとなる。これは，システムの管理コストの増大も意味する。数台規模であれば管理者が手作業によりシステム管理作業を実施できるが，数百台規模のPCクラスタにおいては，故障したPCの物理的な位置を探し出すだけでも大変な労力を必要とする。このようなPCクラスタの特性に対応して，運用管理コストを低く抑える機能が今後ますます重要となってくる。

(6) 計算サービスの安定供給

不特定多数の利用者が共同利用する計算機センタにおいては，安定したサービスを常に提供し続けることが重要である。このため，これまでのメーカ製スーパーコンピュータに求められたのと同様に，24時間365日の継続したサービスの提供がPCクラスタにも求められることが考えられる。SCoreのチェックポイントリスタート機能のような可用性を向上させる機能も重要であるが，発生した問題を速やかに解決するサポート体制の充実も重要である。

富士通のPCクラスタへの取組み

富士通は，PCクラスタコンソーシアム発足への協力や，国内でもいち早くPCクラスタの構築およびサポートのサービスを開始するなど，積極的な取組みを実施してきた。今後も，PRIMERGYで実績のある卓越したLinuxサポート技術を活用し，計算機センタの高度な要件への対応を進めていく予定である。本章では，これまでに述べた計算機センタの中核システムとしてPCクラスタを共同利用していく上での課題に対し，富士通がどのような取組みを実施していくのかについて述べる。

(1) ISVアプリケーションの充実

現状の課題は，Beowulf対応版のISVアプリケーションではSCoreほどの性能を得られない点にある。基本的には，SCoreへの対応をISVへ要請していくことをPCクラスタコンソーシアムとともに働き掛けていくこととなるが，短期的な対応も必要である。このため，Beowulf対応版ISVアプリケーションをSCoreと同等以上の性能で実行するための高速通信機能を今後提供していく（**図**-1）。本高速通信機能は，インタコネクトとしてInfiniBandを対象とし，TCP/IPを利用せず独自プロトコルを利用することにより，高い通信性能を実現するものである。

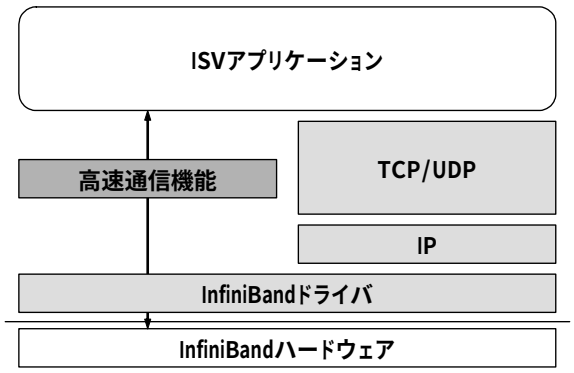

図-1 高速通信機能
Fig.1-High performance network interface.

(2) バッチシステムによる共用環境の実現

富士通は，これまでVPPシリーズやPRIMEPOWERシリーズで豊富な実績を持つ，NQS（Network Queuing System）をLinuxへ移植し，PCクラスタ向けバッチシステムとして提供している。商品名をParallelnavi NQS for Linuxと命名し，VPP/PRIMEPOWERでの豊富な経験と技術を継承していることを示し，これまでにいくつかのPCクラスタに採用いただいている。Parallelnavi NQS for Linuxの初版では，SCoreによる並列ジョブ実行環境を前提としていたが，現在ではBeowulf型PCクラスタでのジョブ実行も可能となっている。

富士通はこれまで，計算機センタの各種運用要件に応えるべく，NQSのエンハンスを実施してきた。しかし，現在のParallelnavi NQS for Linuxは必ずしもすべての要件を満たしているとは考えていない。とくに以下の課題については，今後早急に対応していく必要がある。

- システムの大規模化に伴う同時実行ジョブ数などを拡張する必要がある。数万・数十万規模でのジョブ投入・実行・制御に耐えられるバッチシステムでなければならない。
- フェアシェアスケジュールやデッドラインスケジュールなど，様々なジョブスケジューリング要件へ追従する必要がある。

(3) 課金・統計情報の収集

前章で述べたように複数PC上の資源使用量を一括集計するとともに，悪意のある利用者の不正を防止する，正確かつ堅ろうな課金・統計情報の収集処理が必要である。これには，現Linuxの課金システムに不足しているジョブの識別情報の付与が必要である。また，この識別情報は悪意のある利用者から変更されないような考慮も必要である。

富士通はLinuxの課金システムを改善すべく，Linuxコミュニティ，およびディストリビュータとの検討を行っている。この活動により，近い将来のディストリビューションでエンハンスされる予定である。

(4) クラスタ内ファイル共用

PCクラスタは複数のPCから構成されるが，各

PCから物理的に同一のファイルを一致した名前で参照できると，利便性が格段に向上する。富士通は，VPPシリーズ，PRIMEPOWERシリーズで蓄積した高速共用ファイルの技術を利用し，大規模なPCクラスタ環境でも高速なファイル共用を実現するSRFS（Shared Rapid File System）を提供する予定である。SRFSの概要を**図**-2に示す。SRFSは，InfiniBandまたはギガビットイーサネットを通信媒体として，高速でかつ複数のPCから同時に同一ファイルをアクセスしてもファイル内データの同一性を保証するネットワークファイル共有システムである。ギガビットイーサネットにより接続されたPRIMEPOWERをファイルサーバとすることも可能である。

(5) 容易な運用管理の実現

数百～数千台のPCから構成される大規模なPCクラスタを効率的に管理・運用していくには，多数のPCに対する一括操作などを容易に実現する各種ツールが求められる。富士通では，これまでにPC，IAサーバが有しているIPMI（Intelligent Platform Management Interface）を利用した一括電源投入・システム起動・停止・電源切断などを行う運用管理ツールを提供している。本運用管理ツールでは，各PCの操作とともに，各PCの動作状況監視も可能である。

なお，運用管理ツールのグラフィックユーザインタフェース（GUI）は，画面レイアウトなどが導入する計算機センタによって細かく要件が異なるため，要件に応じてカストマイズするサービスも提供している。

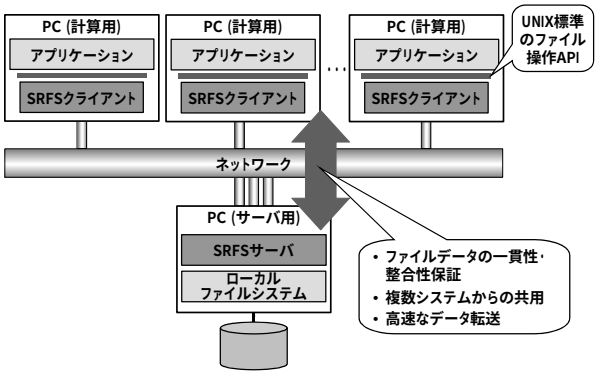

図-2　SRFSの概要
Fig.2-Shared rapid file system (SRFS).

(6) 計算サービスの安定供給

計算機センタの利用者に対し安定した計算サービスを提供するには，ハードウェア・ソフトウェアの信頼性だけでなく，各種問題を迅速に解決できるサポート体制の充実も重要である。とくにオープンソースソフトウェアは，安定して利用できる範囲とアグレッシブに新機能を取り込み不安定な範囲が共存しているため，各センタの運用内容に即した利用範囲を見極める必要がある。

富士通は国内でもいち早くPCクラスタの構築サービス，およびサポートサービスを提供している。構築サービスは，各種ソフトウェアのインストールの代行だけでなく，お客様の要件を十分に伺った上で各オープンソースソフトウェアの最新の状況を踏まえた最適な利用範囲の提案を行っている。これにより，運用開始後のトラブル発生を事前に回避することができる。

さらに，Linuxのサポートに対しても他社に先駆けて次のような内容を提供している。計算機センタの中核システムとしてPCクラスタを適用しても安心してご利用いただける内容となっている。

・ダンプ解析に基づく早期問題解決の支援
・応急修正提供による早期解決の支援（正式修正は別途ディストリビュータから提供）

一方，従来のPCクラスタ利用者にとっては，前述のような手厚いサポートを必要としない場合もある。このような場合には，セキュリティパッチなど重大な問題に関する修正のみの提供や，過去の事例に基づく回避策の提供など，安価かつ有用なメニューも用意している。

むすび

本稿では，PCクラスタを大型計算機センタの中核システムとして適用する際の要件・課題を示すとともに，それぞれの要件・課題への富士通の取組みについて紹介した。一部課題への継続的な対応は必要であるが，現時点でも十分に計算機センタの中核システムとしてPCクラスタが採用可能であることを示した。本稿では言及していないが，PCクラスタは汎用機やスーパーコンピュータのような万能システム（様々なアプリケーションがそれなりの性能で動作できる）ではない。PCクラスタに向くアプリケーション，向かないアプリケーションが存在す

るのは事実である。富士通はこれまでと同様に，PCクラスタとともにスカラSMP（Symmetric Multiple Processor）型のサーバをそろえて，HPC分野のお客様の要求に最適なサーバを提供していく。

参 考 文 献

(1) 住元真司ほか：PCクラスタ．*FUJITSU*，Vol.54，No.2，p.137-141（2003）．
(2) Top500 Supercomputer Sitesホームページ．http://www.top500.org/
(3) Beowulf Projectホームページ．http://www.beowulf.org/
(4) PCクラスターコンソーシアムホームページ．http://www.pccluster.org/