

高度 HPC システムの動向と PFU の取り組み High-Level HPC System Trend and PFU's Approach

高橋 真*
Makoto Takahashi

青柳真澄**
Masumi Aoyagi

今任嘉幸**
Yosiyuki Imato

廣川 新**
Arata Hirokawa

* システム基盤グループ IT ソリューション事業部

** システム基盤グループ IT ソリューション事業部 プロフェッショナルサービス部

科学技術計算分野での HPC システムの利用ユーザーは学術機関、生命科学関連研究所だけでなく、製造系企業の設計部門においても利用が進んでいる。HPC は成長著しい IT 分野の一つである。ハードウェア各社が大型で使いやすい高性能クラスタ製品の開発に意欲的に取り組んでいる。今回は、更なる高性能を実現する近年の HPC システムの動向、PFU が取り組む HPC インフラ基盤技術についての技術解説、PFU の取り組みを示す。

HPC systems spread widely in the field of scientific and technological computing; the use of HPC systems has been extended from academic institutions and life science related laboratories to the design departments of manufacturing corporations. HPC is one of the fastest-growing IT fields. Hardware makers are very actively developing large-scale, easy-to-use, high-performance cluster products. This paper describes the recent trend for HPC systems that realize higher performance, the HPC infrastructure technology PFU will work on, and PFU's approach to the technology.

1 まえがき

HPC は High Performance Computing の略で膨大な計算を高速に処理することを指している。古くはスーパーコンピュータを使用していた分野である。現在では、PC クラスタシステムという、ネットワークで複数の計算ノードを相互接続し、高速な並列処理を行うシステムが広く利用されるようになった。PC クラスタシステムは、Intel^{注1)}アーキテクチャの CPU を搭載するサーバを大量に利用したシステム (MassiveParallel) の上に、高速ネットワーク技術、並列化用ミドルウェア、並列化アプリケーションを搭載している (図-1 参照)。HPC 関連製品を開発している各企業は得意な技術力を鍛えながら競争力を向上させている。

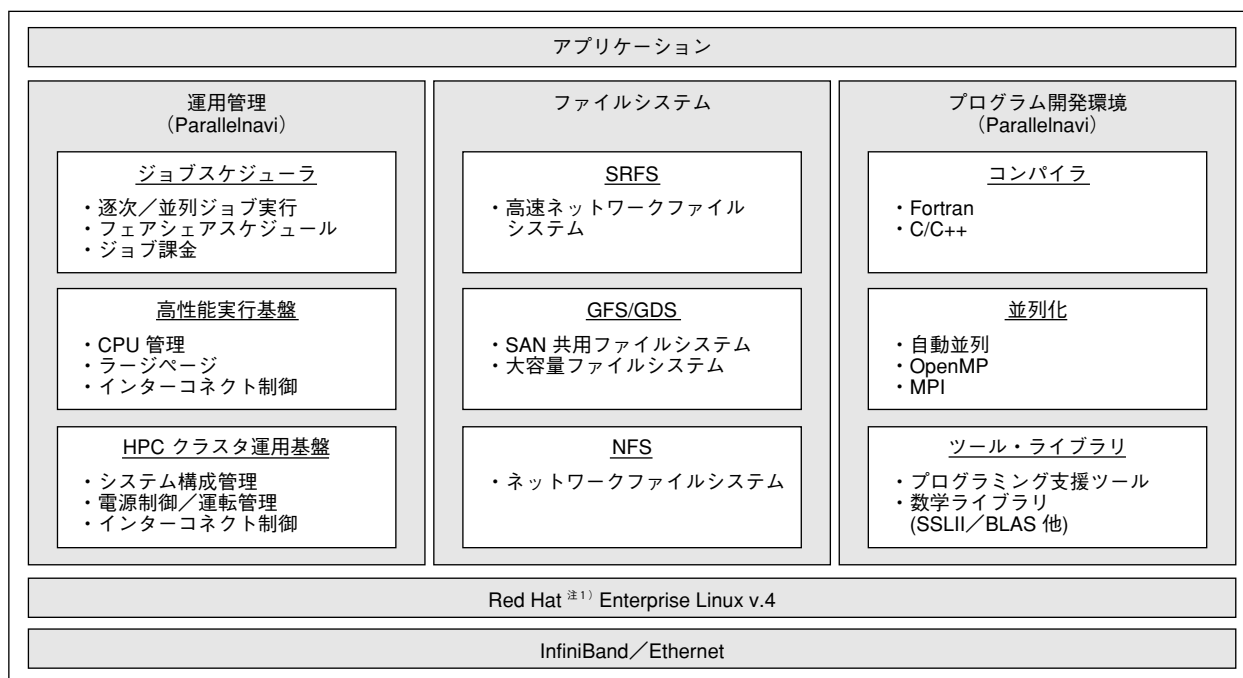
注1) Intel は、Intel Corporation の登録商標である。

本稿では、HPC 技術の市場動向、顧客が必要とする技術の解説を行い、PFU が注力する HPC システムへの取り組みについて解説する。

2 市場動向

2.1 市場動向

HPC 技術は大学、科学系研究所で最先端研究分野 (気象、環境、バイオ、航空など) での利用が中心であったが、産業界においても厳しい企業間競争を勝ち抜くための武器として、CAE、バイオ、化学、金融等の市場で高精度な設計を推進するために利用が進んでいる。CPU 性能の向上がもたらす高精度なシミュレーション環境の差が製品の差別化を生み出すことから新たな HPC システムへの投資も行われている。



注1) Red Hat は米国その他の国で Red Hat, Inc. の登録商標若しくは商標である。

●図-1 HPC システム構成●
(Fig.1-HPC system configuration)

2.2 顧客ニーズ

HPC システムは、CPU を単に高速とするだけで最適なシステムとできる訳ではない。個々の部品が高速で、しかもその高速性が発揮できる環境が必要である。メモリへのアクセス、計算ノードを接続するためのインターコネクト、並列化ミドルウェア、そして並列化アプリケーションが全体としてバランスがとれていなければならない。

さらに各種シミュレーションを実行するためには、膨大な計算時間が必要であり、計算時間中の故障は、顧客の作業効率の悪化につながる。そこで、個々の部品の故障率が低いことと、故障時の計算への影響を最小限にする対策も必要である。

このような問題を解決することにより、最適な HPC インフラシステムを構築し、安定した運用が可能となる。

2.3 技術解説

ここでは、HPC システムを構成する代表的な要素技術を説明する。

(1) 並列コンピューティング/分散コンピューティング

並列コンピューティングは、複数のプロセッサで一つの計算をさせる。多くの問題を解く過程でより小さな計算に分割することができることを利用した計算処理方

法である。並列コンピューティングでは複数搭載されている物理的な CPU を利用した並列計算だけでなく、マルチスレッドなどの仮想的な資源多重での計算も可能である。

一方、分散コンピューティングはネットワークを介した複数のコンピュータを利用して全体のスループット性能を向上させる。一台のコンピュータで処理する場合に比べ、解析するデータの配付、計算結果の収集や集計のためのネットワークが必要になる。プロセッサの増加への対応が容易であるが、ネットワークの負荷が高くなることや、データを格納するファイルサーバの性能がボトルネックになることなどが課題となる。また、分散コンピューティングでは、故障が発生しても故障部分以外の稼働が可能なが多く、高信頼性を有している。

(2) プロセッサ

当社の HPC 商談は富士通の IA サーバである Primergy による PC クラスタで対応している。Primergy は Intel の Xeon プロセッサを採用し、優れたパフォーマンスを実現している。Primergy で利用可能なプロセッサは、以下のとおり。

Dual-Core Xeon 理論演算性能：24 GFLOPS
Quad-Core Xeon 理論演算性能：42.6 GFLOPS

Xeon プロセッサは 1 CPU 内に複数の CPU を組

み込むことが可能で組み込まれた CPU をコアという。近年 CPU 性能は飛躍的に上昇し、ムーアの法則を充たしてきたが、CPU はクロックの向上が限界に近づいたことで複数コアで性能を向上させる方式になっている。

コア数は現在 2 と 4 が存在する。これらのコアを有効に利用するにはソフトウェアがマルチスレッドに対応していることが必要である。1 コアのみで動作するソフトウェアでも、OS (Linux などの Operating System) レベルで、並列動作が可能である。ただし、科学技術計算ではメモリキャッシュが有効とならないアクセスも多く、実メモリへのアクセス性能が、アプリ性能に影響を及ぼす。メモリアクセス性能はコア数には比例しないので、コア数の選択はベンチマークテスト等でのアプリレベル動作の評価で決定する必要がある。

(3) ネットワーク

分散コンピューティングでは、ネットワークで計算ノードを接続しているため、ネットワーク性能が演算性能にも大きな影響を与える。高速なネットワークとして、Giga-bit-Ether, Myrinet, 10 G-Ethernet などがある。近年はこれに Infiniband が加わり、20 Gbps の高速ネットワークを利用することができる^{※1)}。

(4) ファイルサーバ

PC クラスタ用のファイルシステムは、NAS, SAN などが利用可能である。計算処理の中の作業領域として使用されることが多いが、膨大な計算の結果として大量データが発生すると、ファイルサーバ能力がボトルネックになることもある。計算処理の途中での書き込み処理は、全計算ノードから一斉に要求されることが多いためである。

(5) 並列化用ミドルウェア

PC クラスタでの並列コンピューティングを支えるミドルウェアとして、MPI (Message Passing Interface) が重要である。MPI は並列機能を実現するためのライブラリ規約である。実装はベンダーが提供する他にもフリー版も多く存在する。アプリケーションは使用できる MPI 実装があらかじめ決まっている場合があるので注意が必要である。

また、並列化アプリケーションを作成するためのコンパイラが必要となる。現在は、富士通コンパイラ、Intel^{※1)}、PGI、Absoft などが有名である。言語は Fortran と C++ が多い。

(6) HPC アプリケーション

弊社で提供する PC クラスタでは、富士通が販売・サポートしているアプリを主に扱っている。衝撃解析、落下解析、プレス解析などは、米国 LSTC 社開発の非線形動的構造解析ソフトウェア LS-DYNA^{※2)}での実績が多い。自動車業界や電機業界で主に利用されている。また、富士通社開発の電磁波解析ソフトウェアの Poynting^{※3)}も他社製品を差別化する独自機能をもつ優れたソフトウェアである。

3 PFU の注力する HPC インフラ基盤技術

PFU は、インターコネクトとして業界標準である Infiniband を今後とも重点的に扱い、高速性、接続ノード数、拡張性を確保するとともに、移行性を維持する。また、Infiniband の高速性を生かした大容量高性能ファイルシステムとして SRFS を採用する。

3.1 Infiniband

Infiniband はコンピュータの種類やデバイスなどに依存しない汎用アーキテクチャであり、コンピュータ同士、ストレージとの接続、ネットワーク機器との接続を可能にする。接続は銅線ケーブルまたは光ファイバケーブルが用いられる。

規格では、信号帯域幅が 2.5 Gbps と規定されている。更に高速化を図るため、4 本束ねる「4X」では 10 Gbps、12 本束ねる「12X」では 30 Gbps の帯域を確保できる。さらに全二重通信をサポートしているので、双方向の合計帯域幅は、それぞれ、5 Gbps、20 Gbps、60 Gbps となる。

また、サーバ本体が PCI-X に対応していることで、DDR 転送 (2 倍速、最大 2.6 GB / 秒) の性能が得られるため、Infiniband 4X の性能を引き出すことができる。

そして、Infiniband では、サーバ上のユーザープロセス間で直接通信が可能となる RDMA (Remote Direct Memory Access) を実現しているため、メモリ間のデータ転送時には、ホスト CPU への割込みが発生しないこと、カーネル空間からプロセス空間へのオーバヘッドが発生しないために、HPC の世界では CPU を計算処理に専念させることが可能となった。並列化用ミドルウェアである MPI の富士通製 FJ-MPI は、この RDMA 機能を内部で利用しているため、通

常の TCP/IP の socket を利用した他 MPI 実装よりも高速な並列演算環境を提供している。

HPC 分野では、Infiniband の高速性（転送速度、低レイテンシー）、安定性、拡張性がフィットするため、多くのベンダーが採用し始めている。

3.2 SRFS

SRFS (Shared Rapid File System) は高速インターコネクで接続した Linux システム上で動作する分散ファイルシステムのソフトウェアである (図-2 参照)。以下の特長を持っている。

- (1) Linux OS の標準ファイルシステムインターフェース
- (2) 高速インターコネク制御ソフトによる、高いデータ転送性能
- (3) 高可用：インターコネク冗長化、I/O ノード冗長化によるサービス継続性
- (4) 整合性：複数ノードからのデータ更新に対するファイルデータの一貫性・整合性の保証

HPC 環境では多くの計算ノードからファイルアクセスが同時集中的に発生するため、ピーク性能の良好な分散ファイルシステムが要求される。SRFS は Infiniband のような、RDMA 機能を含む高速データ転送が可能なインターコネクを利用して、キャッシュ制御や通信バッファの処理を効率化している。

表-1 に示すように、小規模クラスタでは簡便なファイルシステムとして NAS で十分であるが、規模が大きくなり接続台数が増加することで、高速 I/O が可

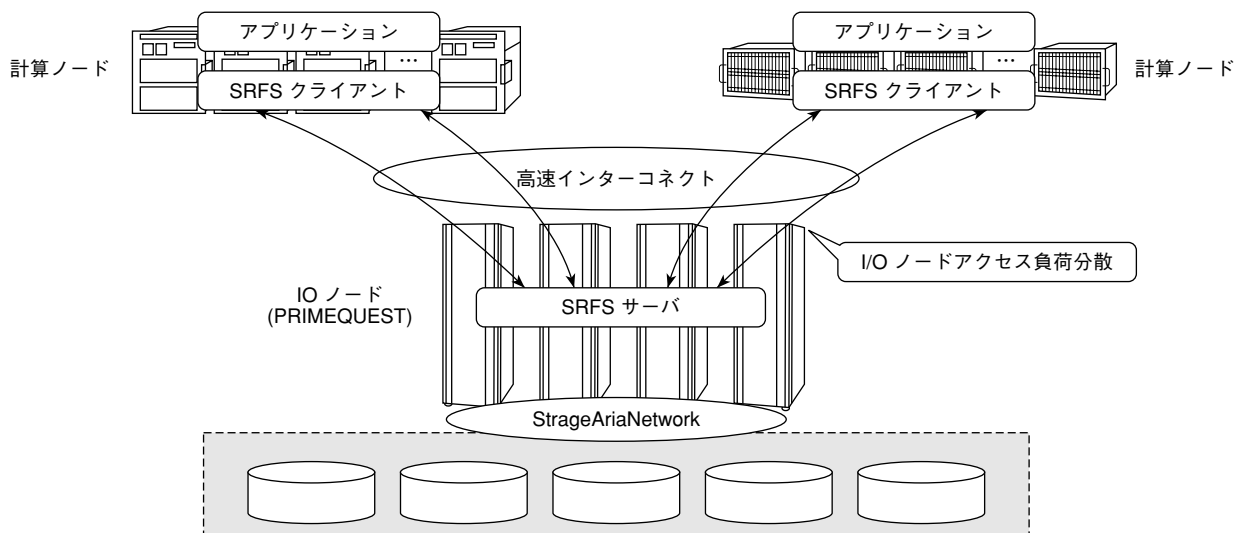
能なファイルシステムが要求される。大規模でかつ高速なファイルシステムには SRFS が最適である。

4 PFU の取り組み

近年の HPC 市場のニーズは、高い演算能力の CPU を持つサーバを大量に並べ、それを並列性の高い論理で動作させて性能向上を狙っている。高信頼性も含め、より高精度の計算能力が必要とされている。

●表-1 ファイルシステムの比較●

共有手段	メリット・特徴	評価と推奨
NAS (NFS)	<ul style="list-style-type: none"> ・多様なクライアントからファイル共有可能 ・既設 LAN 環境へ導入しやすい 	性能面：△ ノード接続数：○ 小規模 PC クラスタ
GFS / GDS	<ul style="list-style-type: none"> ・SAN による高速ファイル共有 ・FC ストライピングによる高速 IO アクセス 	性能面：○ ノード接続数：× 小規模 SMP クラスタ
SRFS + GFS / GDS	<ul style="list-style-type: none"> ・インターコネクやメモリ上のデータ・キャッシュを利用した高速 I/O (GB / 秒オーダー) ・高速インターコネクが必要だが、GigaEther でも大 I/O 長で NFS より高速 ・多数ノードでの SAN 共有は FC コストが増大 (SRFS + GFS が優位) 	性能面：◎ ノード接続数：◎ 中～大規模 SMP クラスタ (～数十台規模) 大規模 PC クラスタ (～数千台規模)



●図-2 SRFS の特長●
(Fig.2-Characteristics of SRFS)

構築期間も短期間を要求されている。我々は独自の SE ツールを利用することで、数百ノードにもなる計算サーバの OS 配付を短期間で実施している。このツールはノード数には依存せず、20 分程度でシステムの配付ができ効果を発揮している。

また計算サーバや CPU 負荷率が 100 %状態で長期間に渡って動作するため、メモリへの負荷、ネットワークへの負荷、ファイルサーバへの負荷を最大限に与える試験を実施することで、初期障害の早期発見に努め、顧客先でのトラブルを未然に防ぐようにしている。

HPC のシステムで重要となる高速インターコネクと SRFS を使った事例を以下に紹介する。事例の 2 テーマとも富士通殿と共同構築した事例である。

4.1 Infiniband 適用事例

Infiniband は既に多くの適用事例がある。PFU が導入作業を実施したシステムでの計算ノード数は 8 台～450 台と大規模から小規模のシステムまで幅広い。最近、システム構築を実施した 450 台規模の科学技術計算システムについて以下に述べる。

(1) 設計のポイント

長時間 JOB に対応するため、JOB を管理する管理ノードを冗長化し、計算ノードも予備ノードを持たせることで冗長化し、ノンストップシステムを目指した。

(2) システム構成

450 ノードは二つのシステムに分割され、200 ノードと 250 ノードに分かれている。ノードは富士通の Primergy RX200S3 を採用し、最新の Intel CPU

である WoodCrest (Xeon5160, 2 Core) を搭載した計算ノードをそれぞれ有している。ユーザーの計算プログラムはバッチジョブ投入システム「Parallelnavi NQS」から起動される。Infiniband 用のスイッチは接続ポート数の多さと動作実績や安定性から Silverstorm 社の SilverStorm9240 を採用した。

(3) 効果

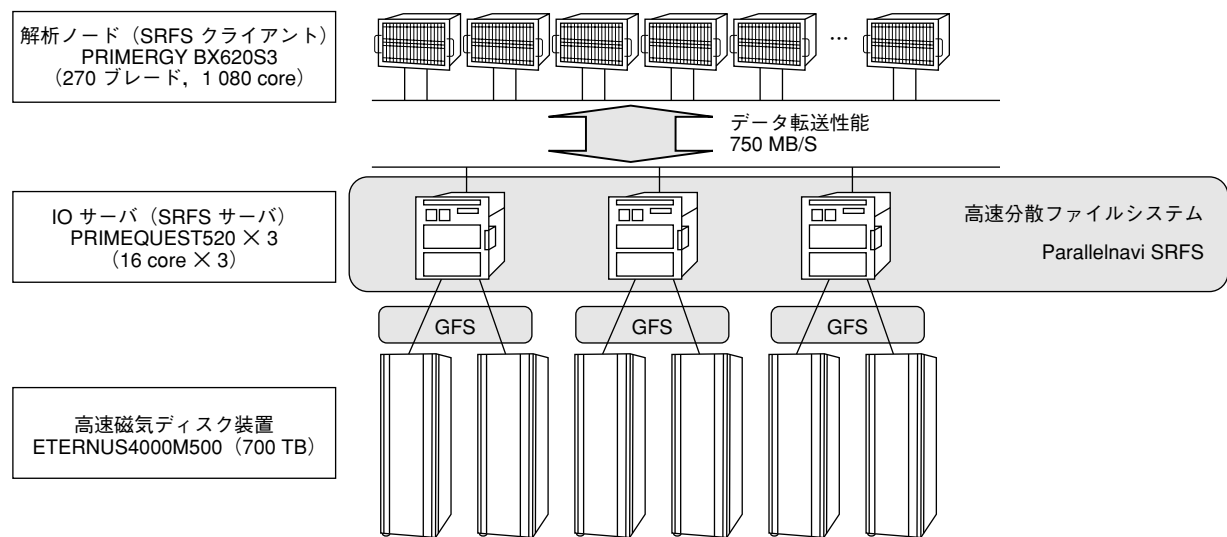
ハードウェアの初期不良を早期に検出し、450 ノードでの運用をスムーズに開始することができた。管理ノードの冗長化、計算ノードの冗長化は非常に有効であり、ハードウェアトラブル発生時の影響範囲の局所化と、運用のノンストップ化に多に貢献した。

(4) 今後の取り組み

HPC システムは計算ノードの数が多くなることで、構築作業が複雑化する。初期ハード不良に伴う設定作業の後戻りとシステムの整合性をとるための管理作業は大変なものがある。短時間でシステムを組み上げるためには、ハードウェア出荷からのすべての工程において作業時間短縮と品質強化のためのツールの充実を進める。

4.2 SRFS の適用事例

富士通 (株) は東京大学宇宙線研究所附属神岡宇宙素粒子研究施設 (スーパーカミオカンデ) の実験解析システムの導入作業を実施した⁴⁾、⁵⁾。このシステムはスーパーカミオカンデが捉えたニュートリノに関する観測データをもとにニュートリノの蓄積および解析を行うシステムである (図-3 参照)。



●図-3 SRFS のシステム構成●
(Fig.3-SRFS system configuration)

(1) 設計のポイント

ニュートリノの捕捉は難しく、24 時間 365 日連続して観測が行われるため、データの収集、解析も連続運転が必要であり、その蓄積データ量は 1 日当り 50 GB にもなる。また、導入前のデータが 550 TB 存在しそのデータ移行も必要である。そのため、高速なネットワーク、高速で大容量なストレージが必要である。

(2) システム構成

この要求を満たすために、富士通製ブレードサーバ「Primergy BX620 S3」270 台 (540 CPU, 1 080 Core) からなる PC クラスタと、富士通製基幹サーバ「PrimeQuest 520」3 台、ディスクアレイ「Eternus 4000」、高速分散ファイルシステム「Parallelnavi SRFS for Linux」で分散ファイルシステムを構成した。

(3) 効果

PC クラスタを構成する 270 台のサーバからは Giga ビットのイーサネットに接続されている。トータルスループット 750 MB / 秒のアクセス性能を実現することができた。

(4) 課題

SRFS を利用した分散ファイルシステムは、大きなファイルのアクセスは非常に高速な転送が可能であるが、逆に小さなファイルへのアクセスが若干苦手である。今後この点の改善が課題である。

た需要が高まると予想される。CPU 技術ではマルチコア化がさらに進み、1 サーバのコア数が増加することによる MP 対応が行われ、サーバを複数台並べるマルチサーバと組み合わせたシステムへと発展し、更なる高速インターコネクトを用いた、ペタコンピューティングシステムへとその利用技術が進むと予想される。既に、文部科学省「次世代 IT 基盤構築のための研究開発」^{※6)}で採択されたペタフロップス級の計算能力の次世代システムの研究が進展しつつある。PFU は今後も積極的に HPC インフラ技術への取り組みを推進していく。

参考文献

- 参 1) 伊勢雅英の InfiniBand 探検隊 - 【中編】HPCC で高いパフォーマンスを発揮する InfiniBand
<http://enterprise.watch.impress.co.jp/cda/special/2004/07/08/2575.html>
- 参 2) 非線形動的構造解析ソフトウェア LS - DYNA 紹介ホームページ
<http://jp.fujitsu.com/solutions/plm/analysis/lodyna/>
- 参 3) 汎用 3 次元電磁波解析ソフトウェア Poynting 紹介ホームページ
<http://jp.fujitsu.com/solutions/plm/analysis/poynting/>
- 参 4) 事例概要 東京大学宇宙線研究所 [スピード] (麻木久仁子の富士通導入事例レポート)
http://jad.fujitsu.com/adver/produce/report/case_18/
- 参 5) 対談 東京大学宇宙線研究所 [スピード] (麻木久仁子の富士通導入事例レポート)
http://jad.fujitsu.com/adver/produce/report/case_18/details/?banner
- 参 6) 「次世代 IT 基盤構築のための研究開発」に関する研究開発課題の選定について (平成 17 年 5 月 24 日 文部科学省研究振興局情報課)
http://www.mext.go.jp/b_menu/houdou/17/05/05052401.htm

5 むすび

今後、科学技術計算分野は PC クラスタを中心とし