SX-9のオペレーティングシステム SUPER-UXの概要

外丸 浩子・宮崎 恵美子・大谷 敦久
佐竹 康司・坂井 智・北川 信亨

要　旨

スーパーコンピュータSX-9のオペレーティングシステムSUPER-UXは、SX-6/7/8シリーズで信頼と実績を築きあげてきたSUPER-UXを更に強化したものです。
SUPER-UXでは、従来からの特長である、高速・大規模システム、高信頼性を引き継ぎながら、いっそうの大規模システムでの使い勝手の良さ、運用管理の充実を追求しています。
本稿では、SUPER-UXの特長とGFS、NQSIIの強化内容について述べます。

キーワード

●スーパーコンピュータ ●オペレーティングシステム ●マルチノード
●クラスタ ●gStorageFS ●JobManipulator ●SCACCT

1. はじめに

ハードウェア技術の進歩による、スーパーコンピュータの計算能力の向上・コストパフォーマンスの向上は、官公庁や大学の計算センターなどの大規模ユーザから、民間企業、更には研究室単位の導入まで、その利用分野を大きく拡大させています。

近年、HPC分野では、ノード内の共有並列では並列数の上限があるため、ノード数を増やすことによって並列数を上げるマルチノードシステムが主流になり、そのノード数も増えつつあります。

このため、単なる大規模化、高速化への対応だけではなく、システムの導入から運用、プログラムの開発をより容易にかつ柔軟に行えるとともに、標準化・オープン化に対応していくことが、スーパーコンピュータにとってますます重要となっています。

ここでは、SUPER-UXの特長である、大規模・高信頼性、高速性、使い勝手の良さ、充実した運用管理、新標準やオープンシステムへの対応と、最近の強化内容を紹介します（**図1**）。

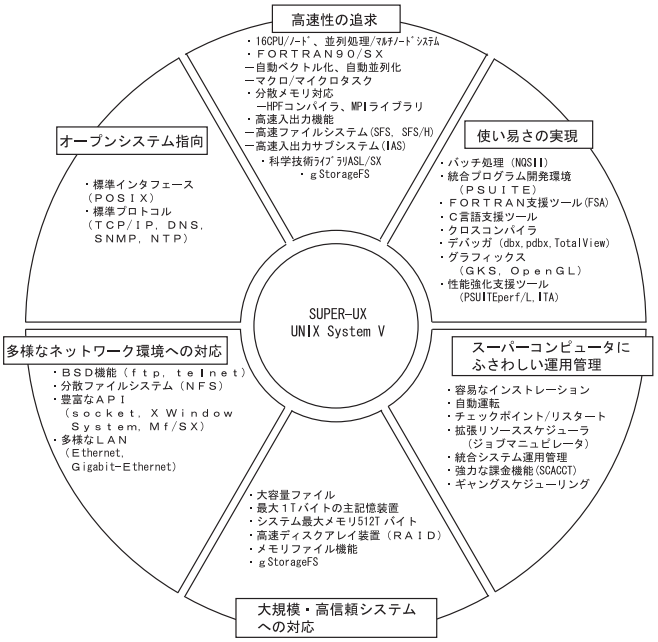


図1　SUPER-UXの特長

2. SUPER-UXの特長

SUPER-UXは、オペレーティングシステムとしてUNIX System Vをベースに、BSD及びSVR4.2MPからも機能を取り込み、更にスーパーコンピュータに必要な機能を強化したオペレーティングシステムです。

SX-9のSUPER-UXは、ハードウェアの相違による制限を除いて、SX-6/7/8シリーズでもそのまま動作することが可能です。これにより、最新の機能をより多くのSXシリーズで利用

できると同時に、SX-6/7/8シリーズからの互換性を維持しながら SX-9への移行、従来機種との併設運用の容易性を保証しています。

またSUPER-UXでは、カーネル自身の高並列化によりシングルノードシステムモデルで最大16CPUまでをサポートし、マルチノードシステムにおいては最大512ノードをクラスタ接続することにより、最大8,192CPUまでをサポートしています。

更に、大規模なメモリへの対応として、 SX-9では、従来シングルノードシステムで最大256Gバイトであった実装メモリを、1Tバイトに拡大しました。これにより、512ノードのマルチノードシステムで最大メモリサイズ512Tバイトを実現しています。

このようにSUPER-UXは、シングルノード16CPUから大規模ノードまでの高スケーラビリティを保障するために柔軟な資源管理、カーネル・I/Oの高い並列処理性を持っています。

またノード数が多くなってもプログラムの実行や管理が煩雑にならないように各種運用ツールの強化を図っています。

2.1 大規模・高信頼性システムへの対応

(1) 大規模メモリへの対応

1) 大規模ページサポート

一般コマンド用の32Kバイト、コンパイラやシステムコマンドのための4Mバイトに加え、大規模なユーザプログラムのために64Mバイトの3種類のページサイズをサポートしてします。これにより、大配列を使用するプログラムの実行性能の向上を図るとともに、メモリ管理のオーバヘッドを削減しています。

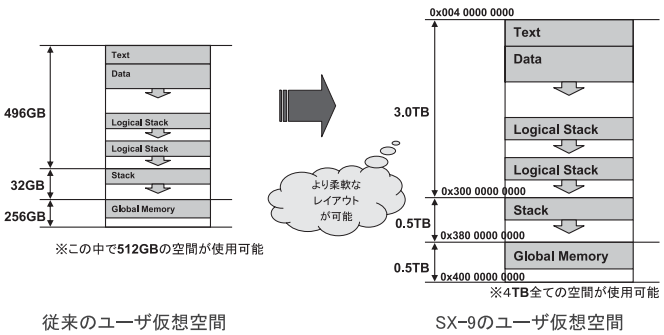


図2　4Tバイトのユーザ仮想空間

2) 仮想空間の拡大

SX-9 のSUPER-UXでは、従来、約800Gバイトであったユーザ空間を4Tバイトに拡大しました（**図2**）。

これにより、単体のプロセスで1Tバイトの主記憶を限界まで使用する並列プログラムや、巨大なグローバルメモリを使用するMPIプログラムを効率よく配置することが可能です。

(2) メモリファイル機能

大容量のメモリ上に、通常のファイルシステムを構築して高速の入出力を可能にすると同時に、ディスクキャッシュとして利用することもできるSX-MFF(SX Memory File Facility)を提供しています。

2.2 高速性の追求

(1) IXS制御機能

マルチノードシステムでは、IXS(Internode Crossbar Switch)という超高速のノード間接続装置をサポートしています。

通常、MPI(Message Passing Interface)やHPF(High Performance Fortran) でのノード間のメッセージ転送が発生すると、データ転送のためにはOS を呼び出す必要があります。しかし、SUPER-UX がサポートしているIXS制御機能を用いることにより、ユーザプログラムが他のノードに、OS を介することなく直接データを転送することができるため、MPI や HPF などの分散並列プログラムを効率良く動作させて、高性能を達成しています。

また、IXS上でTCP/IP(Transmission Control Protocol/Internet Protocol)を実現することにより、ftp(file transfer protocol)によるファイル転送やNFS によるファイル共有も、高速に実現しています。

(2) 大規模SANにおける高速ファイル共有

ファイバチャネルによる大規模SAN(Storage Area Network)環境において、SXシリーズやLinuxなど、マルチプラットフォームでの高速ファイル共有を可能にしました。

2.3 使いやすさの追求

- **IOXソフトウェア**

IOX(Integrated Operation Station for SX)ソフトウェアのインストールでは、シングルノードはもちろん、マルチノードに

おいてもWebによるインストールが可能であり、導入から立ち上げまでを短時間で行うことができます。
また、修正物件配布ツールにより、マルチノードにおいて修正物件を全ノードに一度に配布・適用および適用状況を参照することが可能であり、より高い保守性を実現しています。
IOXとしては、SUSE Linux Enterprise Serverを採用し、CLUSTERPRO Xによる二重化をサポートしました。この冗長構成により、IOXソフトウェア 及び IOX上で動作可能なジョブスケジューラ等の高い運用継続性も実現しました。

3. 高速なI/Oシステム

HPCシステムでは、中心となるSXシリーズだけではなく、スカラマシンなど他の計算ノード、フロントエンドサーバなど、多種多様なマシン環境にSXシリーズも組み込まれて利用されることが一般化しています。このため、I/Oシステムにおいては、計算ノードとフロントエンドサーバなど異なるマシン間での大容量データの移動に要する時間や処理の煩雑さが問題とされる場合があります。

また、一方でCPUの処理能力やメモリ容量に見合ったより高速かつ高効率なI/Oが行えることも重要になります。

本章では、HPCのI/OシステムにおけるNECのアプローチについて紹介します。

3.1 GFSの目的

上記のような大容量データの取り扱いに関する問題を根本から解決するため、NECのgStorageFS（以下、GFSと略す）ではファイバチャネルによるSANを前提とした高速ファイル共有の機能[1)]を提供しています。このため、**図3**のようなヘテロジニアスな環境においても大容量データを移動させることなく、シームレスかつ高速なデータアクセスが可能です。

例えば、エンドユーザは、フロントエンドサーバにログインし、ジョブに必要な入力データをGFS上に置いてジョブ投入します。その後、SXシリーズなどの計算ノードでその入力データをGFSから読み込み、計算結果をGFSに書き込みます。そして、ジョブの実行が終わると、エンドユーザは、計算結果をフロントエンドサーバからGFSにより参照することができるので、データを移動させることなく、可視化など次の処理に移ることができます。これらGFSに対するアクセスは、ファイバチャネルをベースとしたSANを経由して実施するので、ネットワークを経由した場合より遥かに高速なアクセスが可能です。

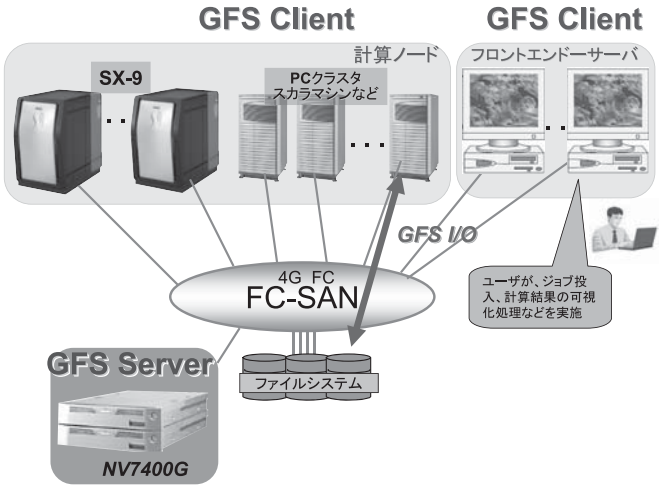


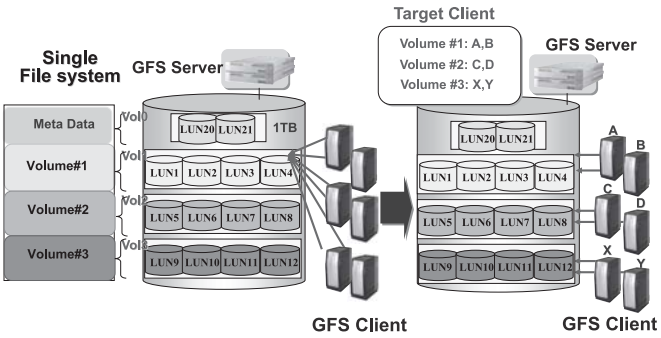
図3　ファイル共有の構成の概念図

3.2 大規模ファイルシステムと並列I/Oの効率化への試み

GFSでは、ストライピングとストライピングディスクのコンカチネート（連結）を組み合わせて、単一のファイルシステムとして128Tバイトまでの大規模ファイルシステムを構築して運用することが可能です。

GFSクライアントより直接DISKアクセスを行うGFSにおいて、このような大規模ファイルシステムで効率よく性能を出すには、GFSクライアント全ノードから、ファイルシステムを構成するDISK全体にアクセスを満遍なく分散させることが必要になります。ファイル分散配置機能は、“GFSクライアント”と“GFSクライアントが優先的にファイルを作成するボリューム”を定義することにより、特定のボリュームへのI/O集中を防ぎ、システム全体のI/Oスループット向上を実現しています。

図4は、ファイル分散配置機能を使用する場合と使用しない場合において、複数ノードから同一ファイルシステムに同時にI/Oを行った場合を示しています。(a)では、同一のボリュームにアクセスが集中してしまいますが、(b)では各ボリュームにアクセスが分散するので、性能低下がありません。



(a) 分散配置機能なし　(b) 分散配置機能あり

図4　GFSのファイル分散配置機能

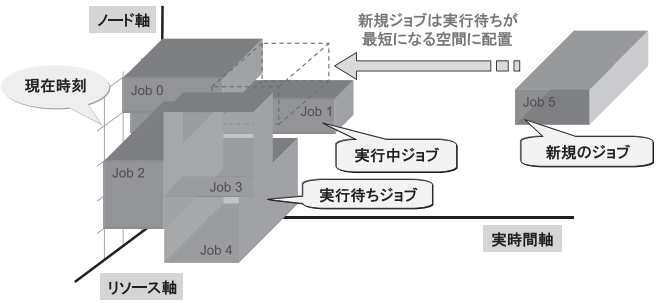


図5　バックフィルの概念

4. 高度なジョブ実行環境

SXシリーズでは、小規模から大規模システム構成において高性能なシステム資源の高稼働率の達成やアプリケーションの高速、高信頼な実行を容易に実現できる高度なジョブ実行環境を提供しています。ジョブ実行環境では、SXシリーズの超高速な演算性能を極限までアプリケーション実行に割り当てる高度なリソーススケジューリング機能とマルチノードに対応した課金システムの強化を図っています。

4.1 バッチスケジューラ（JobManipulator）

JobManipulator はNQSII(Network Queuing System II)のスケジューリング機能を担当します。NQSIIと連携してシステムの高稼働率とリソース最適化を両立するジョブ管理システムを実現しています。

(1) バックフィル方式のスケジューリング

JobManipulatorは利用者がジョブ投入時にジョブに必要な資源量（CPU数やメモリ量など）と必要な計算時間（経過時間）を宣言することによりバックフィル方式のスケジューリングを実現しています。このバックフィル方式のスケジューリングによってシステムの高稼働率の達成とジョブの高速な実行を保証する資源占有を実現します。バックフィル方式のスケジューリングとは有限の未来時間に対しシステム資源のジョブへの割り当てを最適化するスケジューリングです。システム資源の管理を実時間方向への軸と計算ノードを表す軸、及びリソース量を表す軸（JobManipulatorではCPU数とメモリ量の2軸）からなる多次元空間（リソース空間）に対して、ジョブを効率良く配置することで計算ノードの稼働率を向上させると同時に、ジョブに必要な計算リソースの占有利用を保証し、かつ、システム資源全体のリソース利用を最適化するスケジューリングを実現しています（**図5**）。

(2) 多様なスケジューリング機能

JobManipulatorは多種多様なユーザニーズに対して柔軟に対応できるジョブ管理ソリューションを提供します。

・ フェアシェアスケジューリングは、ユーザあるいはグループ単位に設定されるリソース使用比率と過去のリソース使用実績から、リソースの公平な配分を実現します。

・ スケジューリングプライオリティは、10種類以上の項目とそれに対する重み付けの総和として動的に算出されるプライオリティによって、ジョブの実行順序をダイナミックに制御することができます。

・ 最優先で実行したいジョブのために、緊急実行の機能を有しています。緊急実行機能は低優先ジョブの実行を中断して緊急実行するモード、あるいは実行中ジョブが終了次第実行を開始するモードが選択できます。

・ アドバンスリザベーションによって、あらかじめリソース空間を予約しておき、特定ジョブのリソース確保と開始時刻を保証します。

・ ランリミット/アサインリミット機能は、同時に実行できるジョブ数の制限（ランリミット）、及びリソース空間へのジョブの配置数制限（アサインリミット）を、システム全体、またはキュー単位に設定することが可能です。また、複数のキューを束ねてランリミット/アサインリミットを設定するコンプレックスキュー機能も搭載しています。

4.2 マルチノード対応課金システム：SCACCT

従来の課金集計では日次や月次の課金レポートを集計するには大量の課金レコードを扱う必要があり、その集計に各SXノードのCPUを消費していました。SCACCTでは計算ノードの高性能なCPUを極限までアプリケーションに割り当てるため、集計機能を計算ノードからSCACCTサーバにオフロードすることにより、日次や月次の課金レポート作成などを各SXノードのリソースを使わずに提供しています。また、大規模なマルチノード構成においても容易に課金運用可能となるように強化を図っています。

(1) マルチノード対応課金処理

SCACCTサーバで課金情報を集中管理することによりマルチノードシステム全体の課金機能を実現しています。例えば複数の計算ノードにまたがって実行されたバッチリクエストの課金情報をまとめて表示させることも容易に行えます。

(2) マルチノード対応予算管理

ユーザ単位/グループ単位/プロジェクト単位にあらかじめ設定した使用量（=予算）を超えないように管理する予算管理機能のマルチノード対応を行っています。例えばNQSIIと連動することで、任意の計算ノードでジョブを実行させたときに予算を超えると、マルチノードシステム全体へのジョブ投入を禁止することができます。

5. むすび

以上、スーパーコンピュータSX-9のオペレーティングシステムSUPER-UXについてご紹介いたしました。ハードウェアの持つ性能を最大限に引き出すためには、ハードウェアの技術の進歩に合わせて、オペレーティングシステムなどソフトウェアの技術の進歩が必要です。今後も、スーパーコンピュータの適用分野がますます広がり、オペレーティングシステムへの要求も更に高くなるものと思われます。ユーザ要求を先取りし、市場・技術動向を見極めつつ、一歩先んじたSUPER-UXの開発に努力する所存です。

*UNIXは、The Open Groupの登録商標です。

*Ethernetは、米国XEROX社の登録商標です。

*Linuxは、Linus Torvalds氏の米国およびその他の国における登録商標あるいは商標です。

*NFSは、米国サンマイクロシステムズ社の商標です。

*SUSEは日本におけるNovell, Inc.の商標です。

参考文献

1) Ohtani, A. et al, “A File Sharing Method for Storage Area Network and Its Performance Verification”, NEC Res. & Develop., Vol.44, No. 1, pp.85-90, Jan. 2003.

執筆者プロフィール

外丸 浩子
コンピュータソフトウェア事業本部
第一コンピュータソフトウェア事業部
エキスパート

宮崎 恵美子
コンピュータソフトウェア事業本部
第一コンピュータソフトウェア事業部
主任

大谷 敦久
コンピュータソフトウェア事業本部
第一コンピュータソフトウェア事業部
主任
電子情報通信学会
日本リモートセンシング学会各会員

佐竹 康司
NECソフトウェア東北
ソフト開発事業部
主任

坂井 智
NECシステムテクノロジー
プラットフォーム事業本部
サーバソフトウェア事業部
技術マネージャー

北川 信亨
NECシステムテクノロジー
プラットフォーム事業本部
サーバソフトウェア事業部
技術エキスパート