



White Paper

SGI のグリッドへの取り組み

グリッドコンピューティングのためのそのユニークな特長と機能

日本SGI株式会社

目次

1	はじめに.....	1
1.1	グリッドコンピューティングとは.....	1
1.2	SGI®のグリッドへの取り組み.....	3
1.3	スーパーコンピュータとサーバ.....	4
2	ORIGIN プラットフォーム: 共有メモリのスーパーコンピューティング.....	5
2.1	CAPABILITY COMPUTING のための SGI ORIGIN 3000 シリーズ.....	5
2.2	CAPACITY COMPUTING のための SGI ORIGIN 300.....	6
3	ALTIX 3000: スケーラブルなハイパフォーマンス LINUX システム.....	7
3.1	LINUX オペレーティング環境での卓越したパフォーマンス.....	7
3.2	複数かつ大規模の 64 ビットクラスタノード間で共有可能なグローバル共有メモリ.....	7
3.3	超高速ビルトイン・クラスタ・インターコネクティブファブリック.....	7
3.4	標準 LINUX でのハイプロダクティビティ・コンピューティング.....	7
3.5	かつてないモジュラー性を備えたシステムデザイン.....	8
4	ヘテロジニアスな SAN をサポート.....	8
4.1	CXFS—ハイパフォーマンス共有データアクセス.....	8
4.2	DATA MIGRATION FACILITY (DMF).....	10
5	グリッドのためのセキュリティ.....	12
6	ビジュアル・エリア・ネットワーキング (VAN).....	12
7	まとめ.....	14

1 はじめに

1.1 グリッドコンピューティングとは

グリッドコンピューティングは、ミッションクリティカルなコンピューティング・リソースや高度の周辺機器や測定機器などに対する迅速でトランスペアレントなアクセスを科学者や技術者に提供することを目的として進化してきました。グリッドユーザは、リソースの物理的な位置をほとんど意識することなく、コンピュータ、ソフトウェア、データなどのクリティカルなリソースに直接アクセスすることが可能です。グリッドを利用することで、難問の解決能力を高め、組織を超えたコラボレーションを可能にし、高価なリソースを有効に活用することができるようになります。

複数のコンピュータ・システムを使用して一つの大規模な問題を解決する分散コンピューティングは、グリッドコンピューティングにおける重要な側面ですが、現在構築されているグリッドの多くは、固有な能力や機能に対する分散アクセスが大きな原動力となっています。特にヨーロッパでは、リソース共有のためのグリッド構築に多大な努力が注がれています。グリッドの接続によりシステムの利用効率が上がり、高度なコンピューティング・リソースの投資効果を最大限に引き出すことができるのです。

「グリッド」を1つの独立したエンティティとして捉えるのではなく、ネットワークコンピューティングを構築するための1つのモデルであると認識することが重要です。例えば、英国のケンブリッジ大学 (Cambridge University) における「COSMOSプロジェクト」では専用のコスモロジー (宇宙論) グリッドを使い、ビッグバン後の銀河系形成に関する研究を行っています。ケンブリッジ大学には、Capability Computingに取り組むために、英国最大のコスモロジーのためのスーパーコンピュータが設置されています。異なるシステム・アーキテクチャの小規模なサーバを所有している他の大学や研究機関は、各自のアーキテクチャに最適な方法でこのグリッドに接続し、Capacity Computingに取り組んでいます。同様の小規模なグリッドが世界中で次々と構築されています。

現在のところ、グリッドの主な用途は科学的あるいは政府が資金提供するプロジェクトに限られていますが、民間企業の間でもグリッドコンピューティングという概念は受け入れられ始めています。多くの企業が、優れたコストパフォーマンスでコンピュータに対するニーズを満たすために、イントラ・グリッド、エクストラ・グリッド、インター・グリッドの導入を検討しています。イントラ・グリッド (別名 エンタープライズ・グリッド) は、企業内のコンピュータに対するニーズに応えるために設計されました。エクストラ・グリッド (別名 パートナー・グリッド) は、信頼できるパートナーやサプライヤーと手を組むことにより、コラボレーションを強化することができます。一方インター・グリッドは、いくつかの企業による広範囲のリソース共有を促進します。インター・グリッドでは転送メディアとしてインターネットを使用することがあるため、インターネットグリッドとも呼ばれます。

上記のどのグリッドコンピューティングについても、その発達には高速ネットワークを欠かすことはできません。主要な研究施設をつなぐ高速バックボーン (英国のSuperJANET、米国のNSFnet、カナダのCANARIEなど) の可用性が、広域にわたるグリッドコンピューティングの実現に必要なバンド幅を提供するのです。よりローカルなグリッドや、グリッド接続のコンピュータセンターについては、標準的なLANテクノロジーが提供するバンド幅で十分です。

グリッドコンピューティングが取り組む問題は、次の2つのクラスに分類されます：

- **Capacity Computing** では、小規模なジョブが多数同時に実行されます。各ジョブの実行には、少数のプロセッサが必要です。負荷がピークとなった場合には多数のシステムが必要になることもあります。また、小さいジョブへのパーティション化が可能な大規模な問題も、このカテゴリーに分類されます。自己完結型の (self-contained) 性質を持つ各ジョブは一般的に高いコンピューティング性能を必要としますが、システムの I/O リソースやシステム内部のバンド幅に対しての要求は厳しくありません。
- **Capability Computing** では、極めて難解な計算問題に取り組むために莫大な数のプロセッサ、大規模な共有メモリ、複数の I/O チャンネルが必要となります。計算能力が重視される問題は多くの場合、計算性能だけでなくプロセッサ相互接続の能力の限界まで使用することになります。

1.2 SGI®のグリッドへの取り組み

HPC(ハイパフォーマンスコンピューティング)と先進のビジュアルライゼーションにおける先導者として評価されているSGIは、初期の段階からグリッドコンピューティングに関わってきました。グリッドコンピューティングの歴史において、SGIは次のような成功を収めています:

- SuperComputing '97にてグリッドコンピューティングの最初の公開デモンストレーションが行われました。アルゴンヌとUSCが率いたこのデモには、SGIのシステムだけが使用されました。
- Globus Toolkitの開発は、SGIのIRIX® オペレーティングシステムと開用関連ツールを使い、SGIシステム上だけで行われました。
- NASAの分散型HPCグリッドであるInformation Power Gridは、これまで作られた中で最大のSSI(単一システムイメージ)の共有メモリシステムを含むSGI製品で構成されています。
- ヨーロッパ、北米、日本、オーストラリアの各国におけるほとんどすべての主要なグリッド導入事例で、SGIシステムが採用されています。

SGIは「グローバル・グリッド・フォーラム」(Global Grid Forum)のメンバーであり、同フォーラムに属する複数の「ワーキンググループ」や「リサーチグループ」の活動に積極的に参加しています。SGIは2002年6月、分散コンピューティングやグリッドコンピューティングのためのソフトウェアソリューションのリーダーであるPlatform Computing社との提携を発表しました。SGIとPlatformの両社はこの提携契約のもとに、SGIのグリッド・ソリューションに合わせてPlatform Grid SuiteとPlatform Globus™を導入するために協業していきます。SGIのグリッド・ソリューションには、SGI® Origin® サーバシリーズ、SGI® Onyx® 3000 シリーズ、SGI Onyx 300 ビジュアルライゼーション・システム、SGI Altix 3000 シリーズ、OpenGL Vizserver™、SGI CXFS™などがあります。また両社は、データ、コンピューティング・リソース、ビジュアルライゼーション・リソースを統合する新たなグリッド・ソリューション開発のために協業していく予定です。

SGIは、スケジューリング、バンド幅、データの供給、整合性など、グリッドコンピューティングにおける多くの重要な課題に常に取り組み、そして解決しています。SGIのコンピュータ・システム、ストレージ・システム、ネットワーク・テクノロジーおよびソフトウェアは、グリッドコンピューティングのための比類ない機能を提供します。

SGIは、今日最大のSSIをUNIX® とLinux® の双方のオペレーティングシステム環境において実現します。実績あるスケラビリティを備えた共有メモリ・スーパーコンピュータによって、最も困難なCapability Computingに取り組むことができます。また、より規模の小さいコンフィギュレーションは、Capacity Computingに適しています。

64ビットLinuxシステムであるSGI Altix 3000シリーズは、SGI NUMAflex™アーキテクチャおよびIntel® Itanium® 2 プロセッサをベースにしています。また、すべての64ビットLinuxアプリケーションについて完全なバイナリ互換となっています。このソリューションは、最高のスーパーコンピューティングを実現するために64ビットNUMAテクノロジーとクラスターリング・テクノロジーの特性が融合したものであり、Linuxユーザならびにアプリケーション開発者により多くの可能性を提供します。

SGIのCXFSが、SAN(ストレージ・エリア・ネットワーク)上のデータに対するハイパフォーマンスな共有アクセスを提供します。IRIX、Solaris™、Windows NT®、いずれのOS環境のシステムからでもSAN上のデータに直接アクセスす



NUMAflex

モジュラー式 HPC アーキテクチャのコンセプトで、次のエレメントから様々な校正が可能です。

- CX-brick 各ブリックの最大プロセッサ数は16、最大メモリは32GB
- D-brick 各ブリックに最大2TBのハイパフォーマンス・ディスクストレージ
- PX-brick PCI 拡張
- X-brick XIO 拡張
- G-brick 最新のグラフィックスとビジュアルライゼーション
- R-brick 複数のCXブリック間のプロセッサとメモリをつなぐルーティングと桁外れのバンド幅
- IX-brick システムとディスク、CD-ROM、ネットワークのためのベースI/OとI/Oポート

ることが可能です。また、これ以外のプラットフォームのサポートも予定されています。
Trusted IRIX™はB1レベルのセキュリティに対して、UNIXオペレーティングシステムのプロテクションの範囲を超えた内部・外部の脅威に対応するセーフガード(safeguards)を提供しています。Trusted IRIXは、セキュリティが非常に重要となる組織間でのグリッドコンピューティングに理想的です。

SGIはビジュアル・エリア・ネットワーキング(VAN)のテクノロジーにより、最先端のリモート・ビジュアライゼーションを持つパワーをグリッドの領域へもたらします。グリッドユーザは、グリッド上の物理的位置に関係なく、ビジュアライゼーション・システムの出力結果を見ることが可能です。また、離れた場所に居る複数のグリッドユーザが、同時に同じ画像を見ながらインタラクティブなコラボレーション作業をすることもできます。

要求が極めて厳しい科学技術計算分野のアプリケーションであっても、SGIのソフトウェア・パッケージとツールによって最大限のパフォーマンスを発揮します。プロダクティビティ(生産性)を向上させる為には、単に高速コンピュータを導入するだけでは不十分です。真のプロダクティビティ向上には、高速で効率性が高く、安全性にも優れたオペレーティング環境が求められます。SGIは、ハイプロダクティビティ・コンピューティング環境に100%フォーカスしたユニークなIRGO™ソリューションとSGI ProPack™によるハイパフォーマンスLinuxのための最適化ソフトウェア・ライブラリを提供します。

これらのテクノロジーとそのグリッドコンピューティングへの適用について以下にご説明します。

1.3 スーパーコンピュータとサーバ

HPCシステムの基幹は、スーパーコンピュータとサーバであり、SGIは、2002年11月に、SGI Origin 3000シリーズの新モデルと、Intel ItaniumプロセッサとLinux オペレーティングシステムを搭載したスケーラブルなLinuxスーパークラスタ SGI Altix 3000シリーズを発表しました。これらの製品は、「ハイ・プロダクティビティ・コンピューティングの実現」に対し、よりフレキシブルでバランスのとれたハードウェア設計、テラバイトのデータにアクセスする高密度実装を可能とする計算機システムです。

SGIは、「1つのスケーラブルシステムアーキテクチャ」による、2つのシステム - SGI Origin 3000 スーパーコンピュータ・システムとLinuxスーパーコンピュータ・システム - を提供します。どちらの製品も、優れた拡張性、高い生産性をもたらすグローバルな共有メモリアーキテクチャであるSGI独自のNUMAflexを採用しています。NUMAflexは、最先端の技術者や研究者向けに卓越したパフォーマンスとスケーラビリティを提供すると同時に、容易な導入・プログラミング・管理を実現するように設計されています。

NUMAflexは、お客様が目指している技術革新へと迅速に到達することを可能にします。これは、NUMAflexのスケーラブルな低レイテンシのメモリ・アクセスと効率的なリソース管理に加え、プログラミングモデルやオペレーティング環境を自由に選択できることによって達成されるものです。このような機能と導入の容易性のコンビネーションにより、NUMAflexはハイプロダクティビティ・コンピューティングの理想的なアーキテクチャとして高く評価されています。

真に高い生産性をもたらすシステムには、以下の3点が必要不可欠です。

・低レイテンシのメモリ・アクセスによるシステムのスケーラビリティ

NUMAflexのアーキテクチャでは、低レイテンシで高バンド幅のインターコネクが採用されており、システムを拡張してもパフォーマンスが維持できるように設計されています。これによって、スイッチやバックプレーンなどの商用のインターコネク技術にはつきもののボトルネックが解消され、ローカルおよびリモートメモリへの効率的なアクセスを可能にしています。

・効率的なリソース管理

NUMAflexのアーキテクチャは、複雑で大規模なモデルの実行を容易に実現できるように設計されています。NUMAflexではメモリ領域全体が共有されているため、大規模なモデルでもプログラミングモデルの制限なしにメモリに格納することができます。共有メモリシステムでは、動的なリソースの割り当てが可能であり、より複雑で解析リソースを必要とする部分に、より多くの計算機リソースを割り当てることで遊休するリソースを減らし、結果的に問題解決までの時間を短縮することになるのです。

・導入展開の容易性

容易なシステムの拡張を実現するNUMAflexは、費用対効果が高いアーキテクチャで、他に例のない多種多様なオペレーティング環境、プロセッサ、およびコンポーネントが使用可能です。計算処理能力、システムバンド幅、ストレージ、I/Oバンド幅、そしてグラフィックスの独立した拡張性を持ち、コスト削減へと導きます。MIPS[®] またはIntelのプロセッサ、PCIまたはXIO™のインターコネクト・スキーム、IRIXまたはLinuxのオペレーティングシステムがサポートされ、独自仕様のソリューションとスタンダードなソリューションの間のギャップを解消することができます。これによって、ユーザ各々のハイ・プロダクティビティ・コンピューティングの要件に合わせてシステムをカスタマイズでき、必要なパフォーマンスに応じてシステムを簡単に拡張することが可能となります。

SGIのNUMAflexアーキテクチャでは、コンピュータが必要とする最高のリソースでシステムを構築し、導入することを可能とします。SGIのモジュール化コンポーネントは、要求される多種多様なコンピューティング環境に最適なコンポーネントの選択、および組み合わせを可能にします。同時に、必要な計算能力に応じてシステム環境を増強したり、逆に分割することも可能です。

2 Origin プラットフォーム: 共有メモリのスーパーコンピューティング

SGIのOriginプラットフォームは、特許取得のNUMAflexアーキテクチャをベースにしており、2プロセッサから512プロセッサ¹構成のシステムに対しSSI(単一システムイメージ)と共有メモリを実現しています。Originシステムでは、すべてのプロセッサがすべてのシステムメモリに対して直接アクセスすることができます。これは、クラスター・ソリューションとは非常に対照的な特長です。クラスター・ソリューションでは、数個のプロセッサ毎に個別にオペレーティングシステムが必要となる上に、各プロセッサが直接アクセスできるのはメモリ全体の一部に限定されます。様々なタイプのケイバリティ問題を解決するためのプログラミングでは、共有メモリ・プログラミングモデルを使用することでより簡単にパフォーマンスを実質的に向上させることができます。OriginはSSIによって、同等サイズのクラスター・システムと比べて管理が容易なシステムとなります。

特許取得のSGI NUMAflexアーキテクチャは、Originシステムのプロセッサ数をこれまでの他の共有メモリ設計をはるかに超えるレベルまでスケールアップする可能にしました。各プロセス・ノードには、最大4個のプロセッサと最大8GBのメモリを持ったローカルプールがあります。従来のバックプレーン設計に代わってNUMAflexではクロスバースイッチと高速ケーブルを使用しているため、各ノードは他のノードのメモリに対して直接アクセスすることが可能です。その際の待ち時間(レイテンシ)の増加は、ローカルメモリへのアクセスと比較してもわずかです。アーキテクチャは、"NUMA" (Non-Uniform Memory Access: 非一様メモリ・アクセス)と呼ばれています。

NUMAflexは、ブリックと呼ばれる標準のモジュラー式ブロックを採用しています。これにより、システムを個々のエレメントに対して独立して段階的に拡張することが可能となり、これまでにはなかった高レベルの柔軟性、耐障害性、投資保護を実現しています。各種ブリックを必要に応じて追加し、アプリケーションが要求する通りの性能を持ったシステムへと再構成していくことができます。くわえてSGI Originには、他の同等のシステムに比べ、極めて少ない専有面積しか必要としないという利点もあります。これはNUMAflexによってもたらされた効率的なモジュラー方式によるものです。

2.1 Capability Computing のための SGI Origin 3000 シリーズ

Origin 3000は、最大512プロセッサ、1TBのメモリ、毎秒716GBのシステムバンド幅をサポートする、SGIの最上位のスーパーコンピューティング・プラットフォームです。すべてのOriginシステムが共有メモリ・プログラミングモデルをサポートしています。また、クラスターに使用されるプログラミングモデルである、MPI: Message Passing Interfaceもサポートしています。

SGIシステムで利用可能な共有メモリ・プログラミングモデルは、様々なクラスのコンピューティング問題にとって大きなメリットを提供します。例えば、近年NASAではプロダクションで利用するスーパーコンピュータの供給元として、SGIの比重

¹ 1,024 プロセッサのシステムが NASA Advanced Supercomputing の施設内で使用されています。特別注文でのみ購入が可能です。

がますます大きくなっています。最新鋭の航空機・宇宙船の計算流体力学(CFD)シミュレーションや気象シミュレーションなど、NASAのワークロードにおいて、共有メモリそしてSSIであるOriginファミリは大規模なクラスタ・スーパーコンピュータにも優る性能を発揮しています。SGIシステムはNASAのプロジェクトにおいて、プログラミングの負担を小さくし、全体のコストも大幅に削減しながらも、遥かに優れたパフォーマンスを提供します。かつてNASAが使用していたCrayのベクトル・スーパーコンピュータC-90では数ヶ月を要したジョブも、現在では数時間のうちに完了させることができます。これにより、NASAの科学者たちはコンピュータプログラミングに掛ける時間を削減することができ、その分本来の専門の研究分野にフォーカスすることができます。

SGI Origin 3000シリーズは、設置スペース当たりのプロセッサ・メモリの拡張性、およびアプリケーション処理性能を従来では実現困難なレベルまで向上させた、ラックあたり世界で最もパワフルな計算機能力を提供します。SGI Origin 3000シリーズはさらに、画期的な高いリアルタイム処理機能、マルチレベルセキュリティおよびHPC開発環境などにおいてリーダーシップを発揮します。

SGI Origin 3000シリーズは、HPCワークフロー最適化のために設計されているオペレーティング環境以外にも、他のどんなシステムよりラック当たりの高い計算処理能力を持ち、高い生産性を提供します。

1つの4U(7インチ)モジュールに、SGI NUMalink™ルータを加え、MIPSマイクロプロセッサの低消費電力を利用し、最大16CPU、32GBメモリまで拡張可能な新しい「SuperBrick」を開発しました。SGI Origin 3000スーパーコンピュータの新しいモデル - SGI Origin 3900にこの「Super Brick」は搭載されており、標準の19インチラックのサイズで、最大128プロセッサ、256GBメモリまで搭載可能です。SGI Origin 3000シリーズの高いアプリケーション性能によって、ラックあたりの実効性能は、非常に高いシステムとなります。Super-Brickは既存のSGI Origin シリーズのC-brickと互換性を持っているので、容易にアップグレードが可能です。

同時に、SGIは新しいIRIXオペレーティング環境:SGI IRGOを発表しました。プロダクティビティ(生産性)を向上させるためには、単に高速コンピュータを導入するだけでは不十分です。真のプロダクティビティ向上には、高速で効率性が高く、安全性にも優れたオペレーティング環境が求められます。SGIは、ハイ・プロダクティビティ・コンピューティング環境に100%フォーカスしたユニークなIRGOソリューションを提供します。

IRGOは、特にランタイムおよび開発環境の最適化を目的として設計され、IRIXオペレーティング環境に対応した一連のHPCワークフロー効率化機能を提供します。同時に、お客様の知的財産も安全に保護します。IRGOにより、SGIは従来のeビジネスオペレーティングシステムよりも遥かに優れた方法でHPCに要求される困難な課題を解決し、プロダクティビティの向上に大きな付加価値を提供します。

2.2 Capacity Computing のための SGI Origin 300

SGIのOrigin 300は、Origin 3000と変わらない機能をより小さくパッケージングした製品です。2プロセッサから32プロセッサ構成まで、そしてシステムメモリは32GBまで拡張可能なOrigin 300は、Capacity Computingや小規模のCapability Computingに取り組むのに最適なシステムです。

カーディフ大学(Cardiff University)では、エンジニアリング、物理学、地球科学、生命科学、化学など、多種多様な研究分野に身を置いている研究者とビジネスユーザの両者が、HPCとビジュアルライゼーションのためにグリッドリソースにアクセスしています。グラフィックスパイプを3パイプ搭載した32プロセッサのOnyx 300²がCapability Computingの用途に使用されています。そして、Capacity Computingの用途では、8プロセッサのOrigin 300サーバが利用されています。また、SGIのOpenGL Vizserver(詳細はセクション5の「ビジュアル・エリア・ネットワーキング」参照)により、グリッドにおけるリモート・ビジュアルライゼーションが実現しています。

² Origin システムに InfiniteReality3 または InfinitePerformance グラフィックスパイプを搭載することにより、Origin を Onyx ビジュアルライゼーション・システムに再構成することが可能です。

3 Altix 3000: スケーラブルなハイパフォーマンス Linux システム

3.1 Linux オペレーティング環境での卓越したパフォーマンス

オープンソース・コンピューティングでのハイパフォーマンスを望むテクニカル分野のユーザを対象に、SGI Altix 3000シリーズは、Linux OSベースのクラスタ構成で、グローバル共有メモリによる卓越したパフォーマンスを発揮します。SGI Altix 3000スーパークラスタは、テクニカルアプリケーション向けに最適化された64ビット環境において、Intel Itanium 2 プロセッサを数百の規模まで拡張し、従来のLinuxクラスタより遥かに優れたパフォーマンスを提供します。



Model 3700 Superclusters

3.2 複数かつ大規模の64ビットクラスタノード間で共有可能なグローバル共有メモリ

SGI Altix 3000スーパークラスタは、大規模データを単一の共有メモリスペースで管理することで、テクニカルアプリケーションの処理に必要な時間とリソースを大幅に削減しました。また、グローバル共有メモリによってアプリケーションは同一ノード内のメモリに対しても、他のノード内のメモリに対してもダイレクトにアクセスすることができます。SGI Altix 3000シリーズは、より複雑なジオメトリの処理や一連のワークフローを共有メモリ内で完結できるので、従来のLinuxクラスタでは成し得なかったテクニカルアプリケーションの処理のブレイクスルーをもたらします。



Model 3300 Servers

第3世代のSGI NUMAflexアーキテクチャは、大規模なノードをスケーラブルかつ柔軟に構成することで、ノード間でテラバイト級のグローバル共有メモリを実現します。各ノードは、シングルイメージのLinuxオペレーティングシステムでItanium 2プロセッサを4CPUから最大64CPUまで拡張可能です。これにより、ソフトウェアならびにシステム管理・運用コストを削減することができます。グローバル共有メモリを実現したこの強力なクラスタノードにより、SGI Altix 3000シリーズは標準Linux環境での卓越したパフォーマンスを提供します。

3.3 超高速ビルトイン・クラスタ・インターコネクトファブリック

SGI Altix 3000シリーズのSGI NUMAlinkインターコネクトファブリックは、クラスタノード間を高いバンド幅で接続し、標準クラスタリング・スイッチよりも最大で200倍もの高速通信を可能にしました。データがSGI NUMAlinkスイッチを介してノード間を折り返すラウンドトリップ時間は、わずか50ナノ秒で、これはスーパーコンピュータがローカルメモリにアクセスする時間よりも短い時間です。これにより、テクニカルアプリケーションであっても、安定した実行性能を提供することができます。

3.4 標準Linuxでのハイプロダクティビティ・コンピューティング

SGI Altix 3000スーパークラスタは、業界標準の64ビット版Linux環境を最適化し、優れたデータ処理、パフォーマンス監視、リソース管理を提供します。SGI Altix 3000シリーズは、ハイパフォーマンスなCXFS共有ファイルシステムにより、ヘテロジニアスなネットワーク環境であってもローカルのファイルシステムと変わらない速度でのデータアクセスを可能にします。先進の階層型ストレージ管理ソリューションと拡張ボリューム・マネージャ(XVM)は、エクサバイト規模のデータ利用をサポートしており、I/O ボトルネックを解消します。

Performance Co-Pilot™、CPU Sets、Message Passing Toolkit (MPT) といったユニークなツールが、大規模システムにおけるパフォーマンスを効率化します。

最新版のSGI NUMALinkアーキテクチャでは、インターコネクットのバンド幅ならびにレイテンシが一層向上しています。ご提案するシステムは、このSGI NUMALinkアーキテクチャに基づいて作成されたものです。プロセッサには最先端のIntel Itanium 2 プロセッサが、そしてオペレーティングシステムには同じく最先端のHPC Linuxが採用されています。

3.5 かつてないモジュラー性を備えたシステムデザイン

対称型マルチプロセッシング(SMP)の機能を持つこの共有メモリ・コンピュータシステムは、まず900MHzあるいは1 GHz Intel Itanium 2マイクロプロセッサを搭載して提供されます。また、その後発表される次世代マイクロプロセッサへのフィールド・アップグレードを可能にしています。世界初の、そして唯一の64ビットLinuxスーパーコンピュータであるSGI Altix 3000シリーズは、非常に高いスケーラビリティと共有メモリを備えており、システムデザインにおける卓越したモジュラー性を提供します。4プロセッサから最大512プロセッサまでシームレスに拡張することができるので、不安定でコストのかかるボックス・スワップによるアップグレードを行なう必要がありません。また、SGI Altix 3000シリーズのモジュラー性は、I/Oサブシステムにおける事実上無限のスケーラビリティも提供します。SGIはお客様の投資を最大限に活かし、柔軟性に富み、最高のパフォーマンスを発揮するアーキテクチャを最小のリスクで実現します。

SGI Altix 3000シリーズは、現在そして将来お客様が取り組む大規模な計算やデータを処理するのに最適な設計となっています。SGI Altix 3000シリーズには、計算、メモリ、そしてI/Oの各要求をバランス良く処理する柔軟性が備わっています。モジュール、プロセッサ、あるいはメモリ容量を追加してシステム拡張を行なう場合でも、I/Oバンド幅およびシステム全体のバンド幅を簡単に増加できます。従来のようなバス・ベースのコンピュータ・アーキテクチャとは違い、SGI Altix 3000シリーズはデータや計算要求がどれだけ大きくてもボトルネックが生じることはありません。

4 ヘテロジニアスな SAN をサポート

大規模なケイパビリティ問題の解決にとっては、コンピューティング性能と同様にデータI/Oも重要です。SGIは、最も厳しいストレージ要求にも応えることができるSAN(ストレージ・エリア・ネットワーク)ソリューションと業界トップのソフトウェアを提供しています。SGIはハイパフォーマンスSANテクノロジーの導入において、他のシステムベンダーと比較して、より豊富な経験を持っています。SGIは、ファイバーチャネル・ストレージを最初に出荷したシステムベンダーです。また、完全なSANファブリックを最初に出荷したのもSGIでした。さらにSGIは初めてのSAN対応の共有ファイルシステム(CXFS)を開発し、最近では2Gb/secのファイバーチャネル・テクノロジーを出荷した最初のシステムベンダーにもなりました。SGIはSANテクノロジーを使い、シングルシステムで、7Gb/sec I/Oスループットのデモも実際に行ないました。

コンピューティング・リソースへの共有アクセスに加え、グリッドは多くの場合重要なデータに対する共有アクセスを提供することを目的としています。SGIが提供するストレージ・ソリューションは、これの実現を促進します。

4.1 CXFS—ハイパフォーマンス共有データアクセス

データに対する迅速かつ信頼性が高い共有アクセスは、グリッドコンピューティングの成功にとって必要不可欠です。ハイパフォーマンスな共有アクセスが実現されなければ、データが必要となるたびにデータをグリッドからローカルストレージまでコピーしてこなければならず、データのセキュリティや整合性に問題が発生する可能性が出てきます。あるいはNFSなどの低速ネットワークファイル共有技術を使用しなくてはならなくなります。どちらの場合でもデータを待つだけのために貴重な時間が費やされる結果となるのです。

共有データアクセスがクリティカルであるのにLAN(ローカルエリアネットワーク)では十分なバンド幅を提供することができないというアプリケーションをサポートするために、SGIはCXFSを設計しました。CXFSは、あらゆるSAN接続システムに対し、同一のファイルシステムや同一のファイルへの高速同時アクセスを可能にしています。1つのシステムが複数の接続を持つことができ、毎秒数ギガバイトにも及ぶデータ転送速度を実現できるようになります。

CXFSは、SAN上にあるすべてのシステムに対し、同一ファイルやファイルシステムへのハイパフォーマンスな共有デー

タアクセスを提供します。SANの範囲は数十キロメートルにも及ぶことがあるため、キャンパス環境のようなローカルグリッドでCXFSを使用するで、グリッドコンピューティングはより完全なものとなります。

さらに距離が広がると、多くの場合CXFSを使用しても待ち時間が長くなり過ぎるため実用的とは言えません。この場合データのコピーが必要となりますが、ほとんどのグリッドは各ローカルのアクティビティに対応するセンターが集まって構成されているため、これはそれ程大きな問題ではありません。大規模なグリッド上にある各ノードが、あるデータセットのローカルコピーを持っていれば、CXFSを使ってそのデータをローカルに共有することが可能となります。その結果、存在するコピー数を最小限に抑えることができ、各ノードのI/Oパフォーマンスも最適化されます。

CXFSは、ケイパビリティ問題の解決に必要なデータを、コピーすることなく、アクセス可能にする機能を提供します。1つのデータセットに対する共有アクセスを必要とするケイパビリティ問題も、CXFSによるハイパフォーマンスなデータ共有によって、即座にその恩恵を受けられます。カーディフ大学では、グリッドコンピューティングのために、2TBのRAIDストレージとCXFSを組み込んだSANを構築しました。このストレージ・システムは、32プロセッサ構成のケイパビリティ・システムと4台の8プロセッサ構成のキャパシティ・システムに対して、ダイレクトな高速データアクセスとデータ共有機能を提供しています。

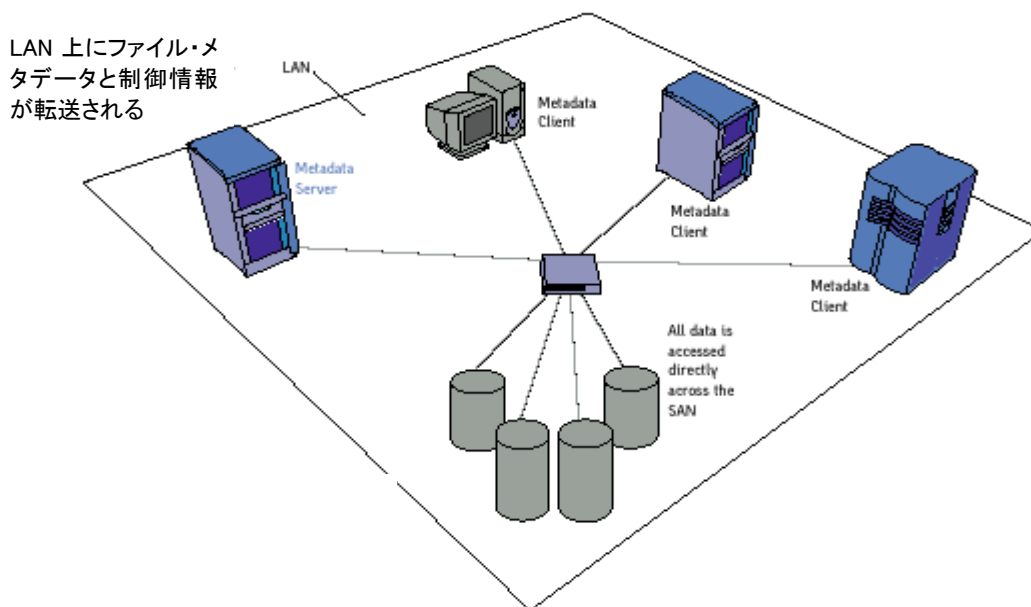


図2 CXFS は SAN のスピードでのデータ共有を可能にします。そのため、イントラ・グリッドや HPC センター間でのグリッド接続に最適のストレージ・ソリューションです。

SAN 上の1つのシステムは、メタデータサーバとしての役割を果たします。メタデータサーバはファイル許可を制御し、共有アクセスの調整を行いません。すべてのデータがファイルサーバを経由するのが原因となってボトルネックがしばしば発生するネットワークファイル共有とは異なり、メタデータサーバが一旦アクセスを許可した後は、CXFS を搭載したシステムは SAN 上においてディスクに対する読み取りと書き込みをダイレクトに実行することができます。

万一メタデータサーバに障害が発生した場合には、指定されたバックアップ・メタデータサーバがCXFSファイルシステムの管理を自動的に引き継ぎます。この機能は、完全な冗長性を備えたSANやRAIDストレージに組み合わせることで、極めて高い可用性と並外れたパフォーマンスをもたらします。システム障害発生した場合にも、CXFSは常にデータへのアクセスパスを確保します。

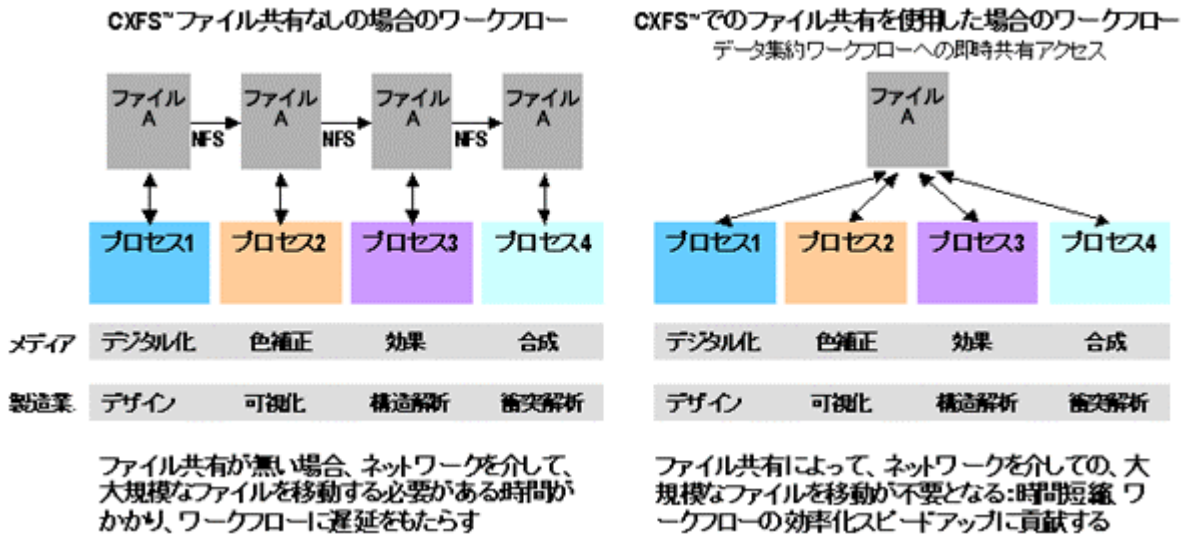


図 3 CXFS:SAN 上での共有ファイルシステムによるワークフローの改善

SGI は、CXFS共有ストレージ・エリア・ネットワーク(SAN)システムを利用することで、IRIX、Solaris、Windows®を含むマルチ・プラットフォームからスーパーコンピュータ・クラスのデータへのアクセスを実現します。SGIのCXFSは、SANの高い拡張性とパフォーマンスをネットワーク・アタッチド・ストレージ(NAS)の接続性・共有性の双方を実現し、プラットフォームの違いを意識することなく大規模な情報へのスムーズなアクセスを可能にしています。

CXFSは、SAN上でのデータ共有を可能にし、ファイルの複製や大きなファイルの移動といったデータ処理を解消することでシステム・ワークフローの改善と運用コストの削減を実現します。SANでは複数のホストとディスクストレージ間の直接かつ高速の物理的な接続が行われています。CXFSはこのストレージへの同時アクセスと共有を可能にするソフトウェアインフラを提供します。これによって全てのシステムから全てのデータへ直接アクセスすることができるようになります。システムはSANの持つ広いバンド幅を有効活用し、ディスクの設置してある場所で直接データの読み書きを行うことが可能になります。もはやネットワークの混雑やファイルサーバのオーバーロードによるボトルネックに悩まされる必要はなくなったのです。

4.2 Data Migration Facility (DMF)

現存するグリッドは、既に膨大な量のデータを管理しています。グリッドでは、コンピューティング・リソースを最大限に利用しているため、より高い比率で新しいデータが生成され、ストレージのリソースを消費する可能性が非常に高くなります。その結果、ストレージ容量とストレージ管理がクリティカルな問題となると考えられます。SGIのData Migration Facility (DMF)は、階層型ストレージ管理(HSM)システムです。多数のHPC施設でオンライン・ストレージの付属システムとして使用されています。業界をリードするHSM製品であるSGIのDMFを使用することによって、ほぼ無限の容量を持つストレージプールを作成することが可能となります。これはストレージを切実に必要としているグリッド接続のHPCセンターにとって、理想的なソリューションです。このようなセンターにおいては、様々なシステムがDMFを使って、使用されていないデータをテープへとトランスペアレントに移動させることができます。移動されたデータは、アクセスが要求された時点または他のグリッド接続ロケーションへ転送が要求された時点で、ローカルシステムによって再び呼び出され、アクセスできるようになります。

DMFは、ユーザが定義した条件に従って、オンライン・ストレージからテープベース・ストレージへとデータを自動的に移動します。ファイルに対してのアクセスが要求された際に、自動的にオンライン・ストレージへと呼び出されます。データの移動には、ユーザあるいはシステム管理者が一切介入する必要はありません。これにより、DMFでは、ほぼ無限のデータプールに対してアクセスすることが可能です。また、データがどのメディアに保存されているかを配慮する必要もありません。

せん。これに比べてマニュアルのデータアーカイブ・ソリューションでは、アーカイブ化するデータの決定やその後のテープへのコピーのために貴重な時間が浪費されてしまいます。また、使用する際にテープからデータをリストアする時にも同様に時間がかかります。DMFによってユーザはデータ管理にまつわるマニュアル作業から解放され、今取り組んでいる科学的問題や技術的問題に専念することが可能になります。

ケンブリッジ大学のCOSMOSプロジェクトでは、共有データアクセス用としてCIFSを搭載した1.6TBのSANベースRAIDストレージが使われています。バックエンドにある4 TBのテープライブラリはDMFによって管理されており、データ管理のタスクを大幅に削減する大規模な仮想ストレージプールが作成されています。また、宇宙学者たちは以前のようにデータストレージについて頭を悩ませることなく、より多くの時間を宇宙に思いをはせるために費やすことが可能となりました。

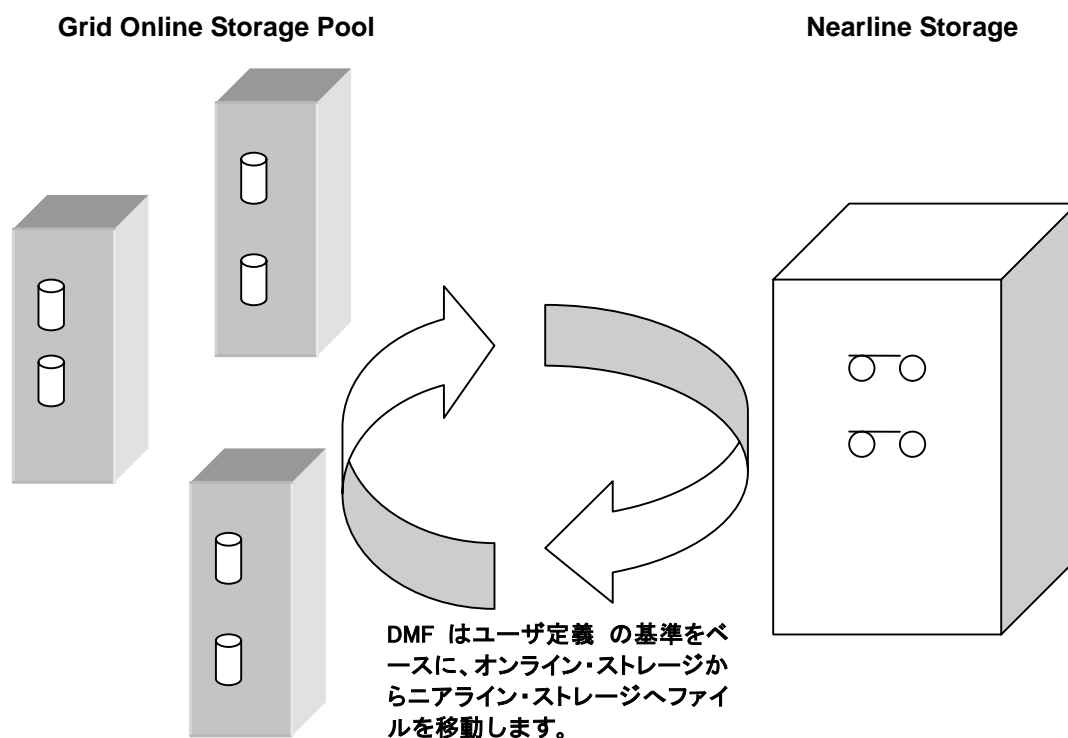


図 4 DMF はグリッドベースのストレージに対応した階層型ストレージ管理を提供します。オンライン・ストレージを遥かに超える容量を持った仮想ストレージプールを作成します。

5 グリッドのためのセキュリティ

グリッドコンピューティングでは異なる組織のシステム間でクリティカルなデータを共有することもあるため、そのセキュリティは極めて重大です。民間企業の間でもグリッドテクノロジーが広く受け入れられるにつれて、セキュリティの必要性はますます高まります。グリッドのどこかに安全でない個所があれば、グリッド全体のセキュリティレベルも落ちてしまいます。つまりグリッド上のシステム自身が安全に保護されていなければ、どんなに信頼できるミドルウェアを導入してもほとんど意味がなくなってしまいます。従って、オペレーティングシステムのセキュリティがグリッド全体のセキュリティにとって非常に重要なのです。

SGIIには長年にわたって連邦政府向けに信頼性の高いシステムを提供してきた専門家としての豊富な経験があります。Trusted IRIXは、SGIのスタンダードなIRIXオペレーティングシステムが持つ機能を何一つ損なうことなくセキュリティ機能を追加することにより、信頼性を高めたIRIX OSです。

Trusted IRIXのオペレーティング環境は、標準のTCSEC (Trusted Computer Security Evaluation Criteria) のB3機能セットに準拠するよう開発されたオプションのアドオン製品です。B1レベルのセキュリティまで保証しています。IRIXの標準インストールは、C2レベル³です。標準のIRIXのセキュリティ機能は、次の通りです。

- 識別と認証
- Capabilityベースの特権メカニズム
- スーパーユーザ・ベースの特権メカニズム
- 任意のアクセス制御
- オブジェクト・アクセス制御リスト
- ハードウェア・オブジェクトスクラビング
- アクティビティ・オーディットトレール

Trusted IRIXには、強制アクセス制御の機能 (mandatory object sensitivity/integrity) が追加されています。B1レベルには、標準的なUNIXシステムには無い幾つかの機能が必要であるほか、既存のコードのチェックと修正が必要となります。追加機能には、改良されたユーザ識別・認証プロシージャ、全システムアクティビティのオーディット記録、ファイルとデバイスに対するより厳しいアクセス制御などが含まれています。強化されたこれらのセキュリティ機能を備えたTrusted IRIXは、高水準のセキュリティを必要とするグリッドにとって最適なソリューションとなりました。

6 ビジュアル・エリア・ネットワーキング (VAN)

サイエンス/エンジニアリングの分野における難解な問題を理解するためには、シミュレーションとビジュアライゼーションは非常に重要なツールです。最先端のビジュアライゼーションのリーダーとして認められているSGIIは、ビジュアライゼーションの恩恵をできる限り多くのユーザに届けるための新たな方法を常に模索しています。ビジュアル・エリア・ネットワーキング (VAN) のためのコアとなるSGIの製品であるOpenGL Vizserverによって、最新のSGI Onyxシステムが持つビジュアライゼーション性能を、グリッド上に存在するすべてのクライアントシステムにもたらしように拡張することができるようになりました。これにより、先進のビジュアライゼーション・システムが存在する場所まで物理的に移動をしなくても、その先進のシステムを使用することができるのです。また、ローカルシステムでビジュアライゼーションを実現するためにデータをコピーしてくる必要もありません。

³ B1 と C2 は、TCSEC 標準規格 (別名「オレンジブック」) として規定されたセキュリティレベルである C1 (最も低い信頼度) から A1 (最も高い信頼度) の範囲に含まれます。定義されたセキュリティレベルには、C1、C2、B1、B2、B3、そして A1 があります。



図 5 グリッド上のビジュアル・エリア・ネットワーキング。ユーザはグリッド上のどこにいても、グリッド・ビジュアライゼーションサーバの機能を提供する SGI Onyx システムによって生成された先進のビジュアライゼーション・アウトプットを見ることができます。ユーザが今いる場所にデータはコピーすることは一切必要ありません。

OpenGL Vizserverは、SGI Onyxシステムをビジュアライゼーションサーバとして使用する、クライアント／サーバ・アプリケーションです。グラフィックスは、Onyxシステムの最新のケイパビリティによってレンダリングされ、さらに圧縮された後、グリッド上であれば事実上いかなるクライアントディスプレイに対してもそのグラフィックスを送ることができます。クライアントはレンダリング前のグラフィックスを表示するので、グラフィックスのアクセラレーションは不要です。OpenGL Vizserverは、単独のセッションにも、あるいは複数のグリッドユーザが同じビジュアライゼーション・ストリームを受け取ってインタラクティブな作業を共同で実施するコラボレーション型のセッションにも使用することが可能です。

OpenGL Vizserver の各機能	
サポートされているビジュアル・サーバ	IR グラフィックス搭載の SGI Onyx2、IR3 グラフィックス搭載の SGI Onyx 300、IR3 または InfinitePerformance グラフィックス搭載の Onyx 3000
サポートされているクライアント	Silicon Graphics ワークステーション (IRIX 6.5 6.5.5 以上)、Sun ワークステーション (Solaris 2.6 以上)、Intel Pentium III 以上および Red Hat Linux 6.2、Microsoft Windows NT 4.0、Windows 2000、または Windows XP のワークステーション
圧縮	4:1、8:1、16:1、32:1 または圧縮用 API
フレーム・スポイリング	
認証用 API	各種認証環境への統合が可能。
予約システム/予約用 API	ユーザはビジュアライゼーションサーバの時間を予約できます。API で既存のカレンダーシステムと統合が可能。
動的パイプ割当て	パイプの使用が終了すると直ちに待機しているセッションが開始されます。
課金	ユーザごとの使用履歴を完全に記録し、計画立案や請求書発行などに適用できます。

OpenGL Vizserverには、「圧縮」、「認証」、「時間指定(reservation)」、「課金」など、グリッドを可能にするための特長的機能がいくつも含まれています。次の表は、OpenGL Vizserverの機能を要約したものです。最新のビジュアライゼーション・システムは通常、固定された制限付きのリソースであることがほとんどです。OpenGL Vizserverとグリッドとを組み合わせることで、ビジュアライゼーションの有用性は大きく高まり、その結果大きなメリットも得られます。

グリッド環境におけるビジュアル・エリア・ネットワーキングについての詳細は、当ホワイトペーパーと対になっているもう一つのホワイトペーパー、「SGIのグリッドへの取り組み：グリッド環境におけるビジュアル・ネットワーキング」をご覧ください。

7 まとめ

グリッドコンピューティングは、ネットワークコンピューティングの進化における次のステップとなります。SGIは、初期の段階からネットワークコンピューティングとグリッドコンピューティングに携わってきました。そして、今日存在する大規模なグリッドのほとんどにSGIのシステムが採用されています。SGIは、グリッドコンピューティングの将来的展望を現実のものとするためのHPC、先進のビジュアライゼーション、データ管理、セキュリティに関する先進のテクノロジーを提供していきます。

© 2003 SGI Japan, Ltd. All rights reserved. Silicon Graphics, SGI, IRIX, Origin, Onyx, およびSGIロゴはSilicon Graphics, Inc.の登録商標です。OpenGL Vizserver, CXFS, Performance Co-Pilot, ProPack, XIO, IRGO, NUMalink, およびNUMaflexは、Silicon Graphics, Inc.の商標です。MIPSは、Technologies, Inc.の登録商標であり、Silicon Graphics, Inc.の米国および他の国におけるライセンスの元に使用されています。WindowsおよびWindows NTは米国および他の国のMicrosoft Corporationの登録商標および商標です。UNIXは米国内および他の各国におけるThe Open Groupの登録商標です。LinuxはLinus Torvaldsの登録商標です。IntelとItaniumはIntel Corporationの登録商標です。Solarisは、Sun Microsystems, Inc.の米国内および他の各国における登録商標です。Platform GlobusはPlatform Computing, Inc.の商標です。その他の商標については、商標の所有者に所有権が属しています。[03/2003]

日本SGI株式会社

〒150-8001 東京都渋谷区恵比寿4丁目20番3号 恵比寿ガーデンプレイスタワー

TEL : 03-5488-1811 (大代表)

東京本社 TEL : 03-5488-1800 (代表) FAX : 03-5420-7030



TEL : 0120-161-085 FAX : 0120-161-087

西日本支社 TEL : 06-6343-6700 (代表) FAX : 06-6343-6713

中部支社 TEL : 0565-35-2561 (代表) FAX : 0565-35-2189

つくば東北営業所 TEL : 029-858-1551 (代表) FAX : 029-858-1071

テクノロジーサポートセンター TEL : 045-682-3700 (代表) FAX : 045-682-0850