



White Paper

SGI® Altix® UVアーキテクチャにおける 先進技術

日本SGI株式会社

第1.1版
2010年3月31日

改訂履歴

版番号	日付	内容
第1.0版	2010年2月22日	初版
第1.1版	2010年3月31日	プロセッサ名記述をコードネームから「インテル® Xeon® プロセッサ 7500 番台」に変更

本ホワイトペーパーには、リスクおよび不確定要素を持つSGI®サーバ・ファミリのロードマップ、その他のSGI®テクノロジーおよびサードパーティ・テクノロジーの将来の見通しに関する記述が含まれています。これらのリスクおよび不確定要素により、実際の結果が記載内容とは大きく異なる場合があります。将来見通しの記述は現在および将来のパフォーマンスを保証するものではありません。読者はその点に注意して、本ホワイトペーパーをお読みください。リスクおよび不確定要素に含まれるものとしては、長期計画のコミットメント、サードパーティ製品の性能、現在および将来の製品の持続的性能、金融リスク、競争市場の影響、複数のプロバイダおよびユーザに関わる複雑なテクノロジー・ソリューションの統合およびサポート、市場および顧客による適用可能テクノロジー、などがあります。これらの将来に対する見通しは、SGI最新のForm 8K、およびForm 10-QおよびForm 10-KでのSECレポートに記載されているリスクおよび不確定要素の影響を受けることがあります。SGIは、将来の見通し、新情報、将来のイベントなどに関して、更新や変更を公的に行う責務を負いません。

製品計画、説明内容および日付はあくまで予想であり、予告なく変更されることがあります。SGIは、本ホワイトペーパーに記載されるいかなる製品や計画に関しても、一般への情報提供等を行わないことがあります。そのため、本ホワイトペーパー記載の情報を過信して事業活動を変更することのないようにしてください。

ホワイトペーパー

目次

1. はじめに.....	4
2. Altix UVプラットフォーム.....	6
2.1. Altix UV Xeon®ベースの計算ブレード.....	6
2.1.1. UV_Hub ASIC機能.....	7
2.1.1.1. グローバル・レジスタ・ユニット.....	8
2.1.1.2. アクティブ・メモリ・ユニット (AMU).....	9
2.2. Altix UVインターコネクต์・トポロジ.....	10
3. 統合されたシステム機能.....	12
3.1 MPIオフロード・エンジン (MOE).....	12
3.1.1. 軽減されたCPUオーバーヘッドおよびMPI_Sendのレイテンシ.....	12
3.1.2. ハードウェア内の高速バリアおよびリダクション.....	12
3.1.3. MOEへのアクセス.....	13
3.2. ペタバイト・クラスのメモリおよびデータ・インテンシブな計算.....	13
3.3. マッシブI/O機能.....	13
3.4. ペタバイト対応の計算機能強化.....	14
3.5. 向上した信頼性.....	14
4. おわりに.....	15

1. はじめに

グローバル共有メモリ（Global Shared Memory, GSM）アーキテクチャでは、全プロセッサがすべてのメモリにダイレクトにアクセスできます。これにより、科学アプリケーションやエンジニアリング・アプリケーションに加え、リアルタイムでの複雑なイベント処理、大規模データベース（Very Large Database, VLDB）や先進のビジネス・アプリケーションといった幅広い分野で大きな力を発揮します。SGIは、GSMプラットフォームの開発における業界リーダーとして、ハイパフォーマンス・コンピューティング（HPC）およびデータ・インテンシブ（データ集約型）タスクに対応する圧倒的な処理能力、メモリおよびI/O機能を提供します。

SGI® Altix® UVは、SGIが生み出した第5世代のスケラブルなグローバル共有メモリ・アーキテクチャです。SGIのGSMシステムの歴史は1996年に世に送り出した64コアのSGI Originシステムに遡ります。SGIの現行GSMプラットフォームであるSGI® Altix® 4700では、標準Linuxディストリビューションの単一OS下で最大1,024コアまでのスケラビリティが実現されています。これらのシステムの構築にはSGIが特許を有するNUMalink™インターコネクが使用されており、GSMシステムに必須となる高バンド幅、低レイテンシを実現し、またコヒーレンスに最適化された機能が搭載されています。また、NUMalinkインターコネク・ファブリックを使用することで、OS間の通信も効率化されます。これにより、MPI、OpenMPやUnified Parallel Cなどの並列アプリケーションにおいて、数千ものCPUコアに対応したスケラビリティを実現します。

Altix UVに搭載された新しいアーキテクチャ機能により、アプリケーションのスケラビリティとパフォーマンスは、プログラミング・モデルに依存することなく強化されます。また、Altix UVシステムは、インテル® Xeon® プロセッサ・ファミリーおよび標準Linuxが実現する高いコストパフォーマンスを活かすとともに、既存のアプリケーションとの互換性を維持します。

インテル® Xeon® プロセッサ 7500 番台を搭載したAltix UV アーキテクチャの主な利点を以下に示します。

- ・ **大規模インコア計算**： Altix UVは、大規模モデルや詳細モデル、大規模データセットなどを完全にメモリ常駐にすることができます。
- ・ **大規模メモリマップドI/O**： 大規模なデータセット上のランダムI/Oアクセスを必要とするアプリケーションに対して、インテル® Xeon® プロセッサ搭載のAltix UVはデータセット全体をメイン・メモリに置くことで、高い性能を実現します。
- ・ **高効率のアプリケーション・スケリングおよびメッセージ・パッシング**： Altix UVは、複数の先進ハードウェア機能およびソフトウェア機能を利用することで、スレッドの同期、データ共有およびメッセージパッシングによるオーバーヘッドの負荷をCPUから軽減し、大規模なタスクの処理を高速化します。これらの機能はすべてのプログラミング・スタイルに恩恵をもたらします。MPIアプリケーションでは、この機能を総称して「MPIオフロード・エンジン」またはMOEと呼びます。
- ・ **大幅に簡素化されたアプリケーション負荷分散**： クラスタ・コンピューティング環境では、各ノードは自ノードに割り当てられたすべてのスレッドを完了しても、他のすべてのノードが割り当てられたタスクを完了するまで待機します。Altix UVに搭載されたインテル® Xeon® プロセッサのグローバル共有メモリでは、ひとつのスレッドの処理を完了したプロセッサは、他のスレッドの処理を開始します。これが実現できるのは、各プロセッサが、グローバル共有メモリを介してすべてのデータおよび同期ポイントにアクセス可能であるためです。

・ **アプリケーション・サイズおよび複雑性に対するスムーズな拡張性**: 大部分のクラスタ環境では、アプリケーションは一定数のノードで稼働します。各ノードには一定数のCPUコアおよびメモリが備えられています。アプリケーションが「壁」にぶつかるのは、クラスタ内のコアまたはノードごとに一定のメモリを超えた場合です。しかし、Altix UV上で稼働するアプリケーションは、このような「壁」は存在しません。システム全体に分散されたメモリを利用することで、アプリケーションはスムーズにその規模を拡張することが可能です。

・ **システムおよびアプリケーションのペタスケールのスケーラビリティ**: グローバル共有メモリのサポートにより、Altix UVではすべてのリソースがオペレーティングシステムの単一OSにより共有できます。さらに、Altix UVでは、より大規模な、グローバルにアドレス指定可能なメモリ (Globally Addressable Memory, GAM) を提供し、任意のプロセッサまたはOSの共有メモリを超えてシステムを構築できます。また、GSM環境の効率を強化する先進のキャッシュコヒーレンス機能およびアトミック操作によって、単一のMPIアプリケーションやパーティショント・グローバルアドレススペース (Partitioned Global Address Space, PGAS) アプリケーションをアーキテクチャ上最大で262,144コア／8ペタバイト・メモリまでスケール可能です。ここではAltix UVの効率的なGAM機能が活用されています。

・ **効率的なアプリケーション開発**: マルチスレッド・アプリケーション、MPIアプリケーションまたはPGASアプリケーションをAltix UVシステムで開発することで、並列化プロセスの早期ステージでの迅速な開発および大規模な問題解決が可能となります。若干の並列プログラミングを行うだけで、数千ものCPUコア上で稼働し数十テラバイトのメモリにアクセスできるアプリケーションを実現できます。開発サイクル初期から並列アルゴリズムの検証が可能であり、結果として、最初の結果を得るまでの時間を短縮します。

・ **効率的なメモリ利用による低コスト化**: クラスタ・システムでは、各ノードには、それぞれ、オペレーティングシステム、I/Oバッファ・キャッシュに加えて、メッセージパッシングのオーバーヘッドおよび複製されたデータ構造のための追加領域が不可欠となります。また、各ノードは、通常、大量のメモリを必要とするアプリケーションがノードに割り当てられた場合を想定して、各コアが大量のメモリを持つように構成されます。しかし、この2つの要因によって、クラスタ・システムでは大量のメモリを余分に購入することが必要となり、システムのコスト増大を招きます。これに対し、Altix UVのグローバル共有メモリ・アーキテクチャで必要とされるのは、単一のOSと単一のバッファ・キャッシュのみとなり、クラスタ・システムでは不可欠であったメモリのオーバーヘッドが削減されます。また、全アプリケーションがすべてのメモリにアクセス可能なため、自ノード上で利用可能なメモリ容量を超えるメモリを必要とするスレッドは他ノードに搭載されたメモリをダイレクトに利用することが可能です。結果として、購入するメモリ総容量が大幅に削減されます。

・ **簡素化された管理**: Altix UVプラットフォームでは、計算、メモリおよびストレージのそれぞれのリソース群の一元的な管理が可能となります。これにより、システム管理の複雑性やコストを軽減します。

2. Altix UVアーキテクチャ

SGI Altix UVは次世代のスケラブルなグローバル共有メモリシステムであり、SGI NUMAflex®アーキテクチャをベースとしています。物理的には、SGI Altix UVシステムは既存のSGI Altix 4700およびSGI Altix 450と似たコンパクトなブレード設計となっており、グローバル共有メモリ構成によって標準Linuxの単一OS下で多数のプロセッササポートを可能にするNUMAflexアーキテクチャが採用されています。またこのアーキテクチャは、SGI NUMALinkで接続された複数のOS間でのメモリ共有もサポートします。

Altix UVは、業界をリードしているSGI Altixのスケラビリティを、すべての次元（プロセッサ、メモリ、インターコネクトおよびI/O）でさらに拡大するように設計されています。最初のリリースでは、SSIごとに最大で2,048コアおよび16TBのメモリをサポートしています。アーキテクチャ上は最大16,384ノード／32,768プロセッサ・ソケットまで拡張可能であり、コア数では最大262,144コア（ソケットあたり8つのコア）までのスケラビリティを実現します。これにより、業界をリードするSGI Altixのスケラビリティがさらに向上します。

構成オプションにより、さまざまな用途に最適化されたシステムを構築することが可能です。たとえば、計算機能（最大のコア数）用に最適化されたシステムや、最大の総メモリ容量、最大のシステム・バイセクション・バンド幅または最大のI/Oに対応したシステムを実現できます。

Altix UVは、インテル® QuickPath インターコネクト（QPI）と密接に統合されています。この密な統合によって、膨大なブレード数で構成されるシステム全体を横断してのキャッシュライン・レベルでの動作が可能となり、グローバル共有メモリのサポートとグローバル共有メモリへの効率的アクセスを実現します。このキャッシュライン指向のアプローチは、クラスタベースのアプローチとは大きく異なります。クラスタベースのアプローチは、大量のデータを、I/Oチャンネルに接続しているInfiniBandインターコネクト上に転送するように最適化されています。

また、Altix UVアーキテクチャは、業界標準であるインテル® Xeon® プロセッサをベースとした単一プラットフォーム環境に、複数のコンピューティング・パラダイムを統合します。この設計は、ベクトル・メモリ操作、豊富なアトミック・メモリ操作に加えて、GPU、デジタル・シグナル・プロセッサおよびプログラマブルなハードウェア・アクセラレータなどのアプリケーション固有ハードウェアの密統合を実現することで、スカラー型のインテル® Xeon® プロセッサを効率化します。

2.1. Altix UV Xeon計算ブレード

Altix UVの標準計算ブレードには、1つまたは2つのプロセッサ・ソケットが含まれており、それぞれのソケットは4コア、6コアまたは8コアのインテル® Xeon® プロセッサをサポートします。図1に示されるように、各ソケットはメモリへのダイレクトなアクセスを実現するほか、総バンド幅が100GB/秒を超える4つのインテル® QuickPath インターコネクト接続を提供しています。これにより、SGIが開発したUV_Hub ASICおよびI/Oハブ（IOH）を介したオプションの外部I/O接続によって、プロセッサは相互に通信することができます。また、インストールするプロセッサ、メモリおよびI/Oの数および種類を変更することで、各ブレードは、4コアから16コアまでコア数を自在に構成できることに加え、大規模から小規模までの最適なメモリ容量およびI/Oを備え

ることができ、完璧な柔軟性を持つ構成を実現できます。

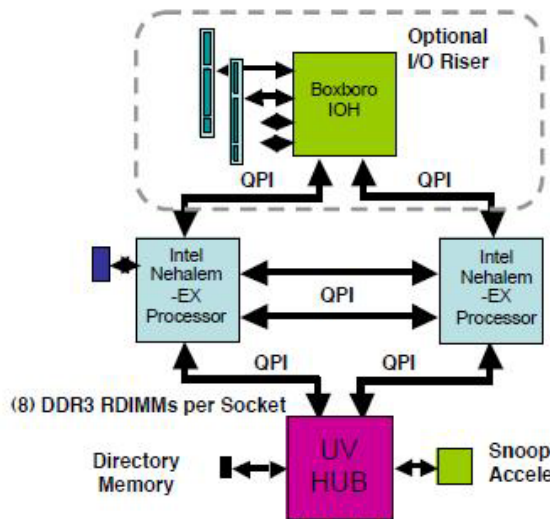


図1: Altix UV計算ブレードのブロックレベルの図。

UV_HubはSGIが開発したカスタムASICであり、スケーラブルなグローバル共有メモリ・システムを実現するAltix UVプラットフォームのベースです。UV_HubはNUMalink 5プロトコル、メモリ操作および関連するアトミック操作を実装し、次のシステム機能を提供しています。

- ・ キャッシュコヒーレントなグローバル共有メモリ — 業界標準のLinux OSおよび業界標準アプリケーションを修正することなく実行できます。
- ・ 時間のかかるデータ集約型処理のプロセッサからのオフロード化 — 処理効率およびスケーラビリティを向上します。
- ・ スケーラブルで信頼性の高いインターコネクション — NUMalink 5を介して他のブレードと相互接続します。
- ・ グローバルにアドレス指定可能なベタスケール・メモリ — 低レイテンシの同期により実現します。
- ・ 各種の先進機能を統合するMPIオフロード・エンジン（MOE） — プロセッサ上のメッセージパッシングオーバーヘッドを軽減するほか、アプリケーションおよびシステムのスケーラビリティを向上します。

2.1.1. UV_Hub ASIC機能

SGIが開発したASICであるUV_Hubは、インテル® Xeon® プロセッサ 7500 番台上のキャッシュコヒーレントなインテル® QuickPath インターコネクトをAltix UVシステム全体にわたる大規模でキャッシュコヒーレントなNUMalink環境に接続します。UV_Hubは、キャッシュコヒーレンシをより多くのプロセッサコアに拡張するだけでなく、ベタスケールシステムの操作を効率化するための機能も提供します。

UV_Hub ASICは4つの主要ユニットと追加機能を備え、ディレクトリの管理とスヌープの高速化を行います。4つの主要ユニットは次のとおりです。

- ・ グローバル・レジスタ・ユニット（GRU） — キャッシュコヒーレンスをブレード・レベルからNUMAflex環境全体に拡張し、他のメモリ関連の機能を提供します。
- ・ アクティブ・メモリ・ユニット — アトミックなメモリ操作をサポートし、主要な同期処理を高速化します。
- ・ プロセッサ・インターコネクト — 2つのインテル® QPI インタフェースを実装します。総バンド幅約50GB/秒でインテル® Xeon® プロセッサ間を接続します。接続されたプロセッサにとって、プロセッサ・インターコネクトはシステム上の他の物理メモリおよび仮想メモリの残りすべてを管理するメモリ・コントローラのように見えます。
- ・ NUMALinkインターコネクト — NUMALink 5のポートを4つ実装します。NUMALink 5の双方向バンド幅は、リンク当たり15GB/秒であり、総バンド幅約60GB/秒でブレード間を接続します。

2.1.1.1. グローバル・レジスタ・ユニット

UV_Hub GRUは重要な機能を多数サポートしており、これらの機能によってインテル® Xeon® プロセッサ搭載のAltix UVをアーキテクチャ上262,144 CPUコア／8ペタバイト・メモリまで拡張することが可能となります。サポートされる主要機能は次のとおりです。

- ・ **拡張アドレッシング:** Altix UVアーキテクチャ内のメモリ管理では、2階層のアプローチが使用されます。各計算ブレード内では、メモリ管理はXeonチップ上の統合メモリ・コントローラにより処理されます。これに対して、UV_HubのGRU操作はキャッシュコヒーレントなメモリ・アクセスをブレード間で実現し、利用可能なアドレス範囲を拡張します。Altix UVシステム内で使用されるインテル® Xeon® プロセッサは44ビットの物理アドレス空間および48ビットの仮想空間を備えています。UV_Hubにより、物理アドレス空間は53ビット、仮想空間は60ビットまで拡張されます。UV_Hubメモリ管理機能は、計算ブレード内の高速なメモリ・アクセスを阻害することはありません。
- ・ **ラージページ・サポート機能付き外部TLB:** TLB（Translation Lookaside Buffer）により、仮想アドレスから物理アドレスへの高速な変換が可能となります。インテル® Xeon® プロセッサにはTLBが搭載されており、TLBによってすべての仮想参照をそのプロセッサに接続されたメモリの物理アドレスに変換します。UV_Hub内の外部TLBは、NUMALinkインターコネクトを介して物理的に接続されているメモリに対して、仮想アドレスから物理アドレスへの変換を行います。Altix UVのブレード上のインテル® Xeon® プロセッサにとっては、UV_Hubは大量のメモリをマップできるもう一つのメモリ・コントローラのように見えます。また、UV_Hub上のTLBは、TLB Shoot Down機能を提供しており、これにより、大規模なメモリを使用するジョブ（大規模なデータベースなど）の起動や終了処理が高速化します。
- ・ **ページの初期化:** ペタスケールのシステムでは、メモリの初期化はパフォーマンスの大きなボトルネックになることがあります。UV_Hubおよび関連するNUMALink 5の機能強化により、メモリ・ページを分散方式で初期化できます。システムのCPUの関与はほとんどまたはまったくありません。メモリ初期化の負荷をCPUからUV_Hub ASICの分散ネットワーク上にオフロードすることで、初期化時間が大幅に短縮されます。
- ・ **BCOPY操作:** ブロックコピー（BCOPY）操作は、データを1つのメモリ位置から別の位置にコピーする非同期メカニズムです。この操作が行われるのは、競合を回避するためにマルチスレッド・アプリケーション内の読み取り専用データ構造を複製する場合、または、メッセージパッシング環境で実際のメッセージを1つのブレードから別のブレードに送信する場合です。SGI Altix システムは、一部の通信に関しては既に最小のメッセージパッシングレイテンシを実現していましたが、これはBCOPYが

転送されるデータの大部分を占めるアライメント境界に整列されたデータの移動に対応していたためです。UV_HubのBCOPY機能は、整列されていないデータのコピー負荷も軽減し、同時にBCOPYをユーザ空間からアクセス可能とすることで、メッセージパッシングのオーバーヘッドを軽減します。

Step	Life of an MPI Send Message in SGI Altix 4700 Systems	Step	Life of a Message with Altix UV Xeon Systems
1	Send fetchop fetch-and-increment request to remote node	1	Place payload in GRU register and issue message send instruction
2	Obtain response, ID queue slot location		
3	Issue queue slot store to the remote node. This sends read-before-write request to remote node		
4	Read/hold remote cache line in local node, insert data in cached line		
5	Upon queue cache-line polling, write data back to remote node		

図2: Altix UVプラットフォーム上のMPIのメッセージ送信の高速化をSGI Altixシステムと比較。BCOPYを使用するとCPUの関与なく、実際のデータを1つのブレードから別のブレードに送信できます。

・ **スキャタ／ギャザ操作:** 固定ストライドおよびリストドリブンのスキャタ／ギャザ・メモリ転送操作などのベクトル・メモリ操作は、UV_Hub内のGRUにより直接的に処理されます。これらの操作により、ネットワーク全体のすべてのキャッシュラインを保持することなしに、ランダム・アドレスがアクセス可能となります。これにより、有効バンド幅を改善し、レイテンシを低減します。ベクトル・メモリ操作は、非連続の関連データ要素ベクトルをグローバル共有メモリのアドレス空間内の連続ロケーションにまとめるか、または、連続ロケーションから共有アドレス空間の任意の特定メモリ・ロケーションへ値を分散します。ベクトル・スキャタ／ギャザによってメモリ転送パターンを最適化し性能向上を図る手法には長い歴史がありますが、UV_Hubに実装されているスキャタ／ギャザ操作では、これらの最適化がインテル® Xeon® プロセッサで利用可能となり、CPUストールがほとんど発生しない処理を可能とするキャッシュフレンドリなデータ構造が作成されます。

・ **AMO (Atomic Memory Operations: アトミック・メモリ操作)のためのアップデート・キャッシュ:** アップデートキャッシュによって、SGIマネージドAMO変数のコピーをローカルには置くことで、バリアや競合ロックのような変数をグローバルにアクセス可能にします。これにより、ホーム・メモリ内のホットスポットが削減されます。競合ロックおよびリダクション変数は、アクセス時間の高速化を実現します。これは、数千から数万ものプロセッサ・コアを利用する場合に強力な武器となります。

2.1.1.2. アクティブ・メモリ・ユニット (AMU)

2種類のアトミック・メモリ操作 (AMO) がAMU内でサポートされています。最初のもは、Intel Xeon CPUが実装するAMOの代わりとして使用されます。ユーザは、Intel AMOとUV_Hub上に実装されているAMOを自在に使い分けることが可能です。これは、競合しない変数が最適にアクセスされるのはプロセッサ・ソケット上のAMOを介した場合であり、また最適に再利用されるのはCPUのキャッシュ内にキャッシュされた場合であるのに対し、競合する変数およびアプリケーション全体のバリアおよびリダクションに関する変数は、UVプラットフォームの一部として実装されるAMOを介してアクセスし、UV_Hubアップデート・キャッシュ内で再利用されるべきである、という考え方に基づくものです。これにより、キャッシュコヒーレンス・オーバーヘッドが軽減されます。

- ・ **コヒーレントメモリ内のAMOキャッシュ:** すべてのAMOはキャッシュコヒーレントな標準ユーザ・メモリ内に実装されており、特別なリソースは不要です。標準メモリを使用することで、アプリケーションはより多くのAMOを使用できるようになり、UV_Hub上のAMOキャッシュによって、NUMAlinkを介するアクティブな変数へのアクセスを高速に実行できるようになります。
- ・ **アップデート・マルチキャスト:** アップデート・マルチキャストによって、アップデート・キャッシュ・ベースのAMOのレイテンシおよび反復率を大幅に改善します。これはバリア、コレクティブおよび競合ロックに対して特に役立ちます。
- ・ **コヒーレントメモリ内のメッセージ・キュー:** AMOと同様に、メッセージキューは標準システム・メモリ内に実装されているため、ユーザはキューの数を自由に設定できます。

2.2. Altix UVプラットフォームのインターコネクト・トポロジ

Altix UVプラットフォームは、システム・サイズおよび目的に応じて多様なトポロジを使って相互接続できます。大規模システム用のAltix UVブレード・シャーシは最大16ブレードをサポートし、任意のノード間での最大NUMAlinkホップ数が3となるスイッチド・デュアルプレーン・トポロジで相互接続されます（図3aを参照してください）。図3bに表示されるように、8ラック／16ブレード・シャーシ（256ブレード、512ソケット、4,096コア）までの規模のシステムは、ファットツリー・トポロジで相互接続され、各ブレード・シャーシで利用可能な4つの16ポートNUMAlink 5ルータと、追加の外部NUMAlink 5ルータが使用されます。より大規模な構成を実現するには、2Dトーラス内で接続される256ソケットのファットツリー・グループを使用します。図3cを参照してください。

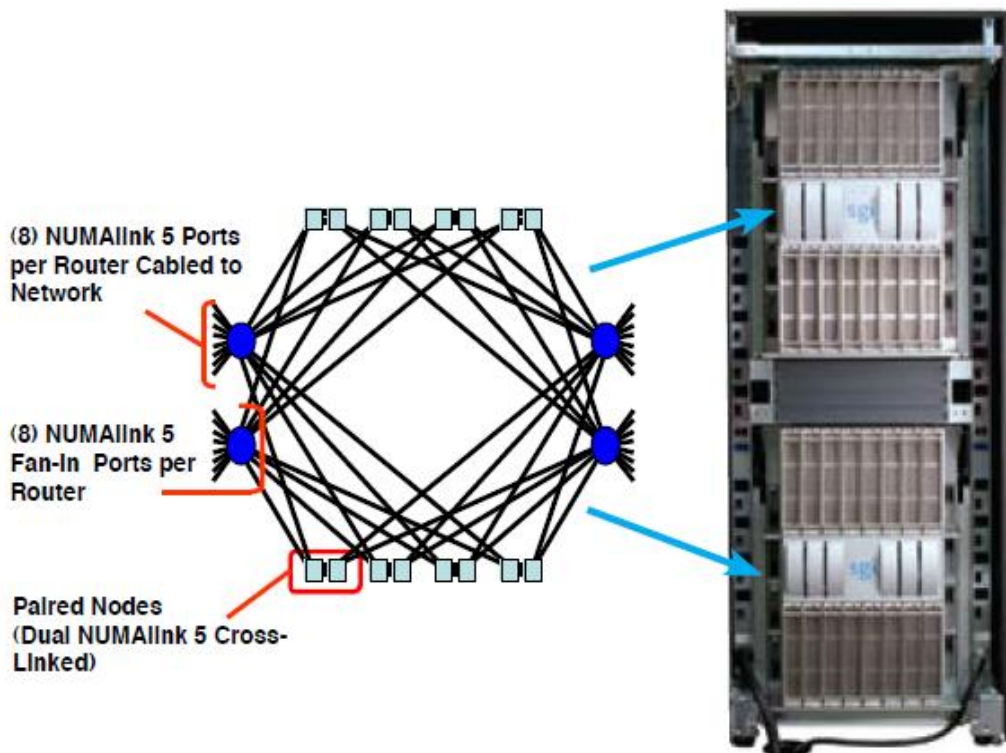


図3a: Altix UVプラットフォーム・ブレード・エンクロージャの32ソケット・インターコネクト・トポロジ。ラックあたり2つのエンクロージャ

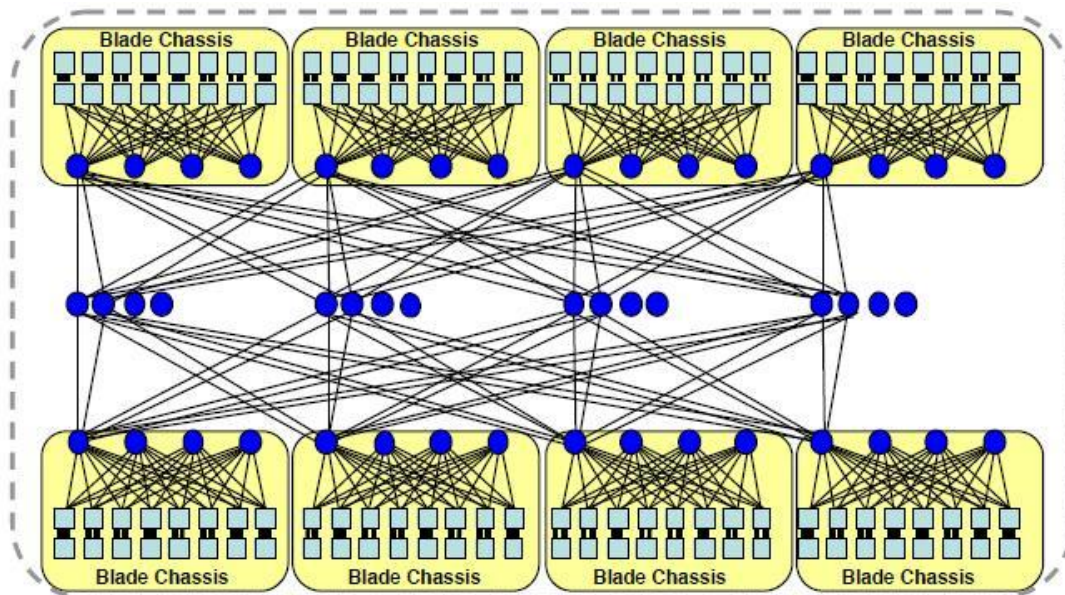


図3b: Altix UVプラットフォームの256ソケット(最大2,048コア)ファットツリー構成。16の外部16ポートNUMalink 5ルータを使用(実際のケーブルのうち、4分の1が表示されています)。1,024ソケット構成システムを構築する場合は、同一の3レベルルータ・トポロジで4倍のブレード・シャーシ数と2倍の外部ルータ数が必要になります。

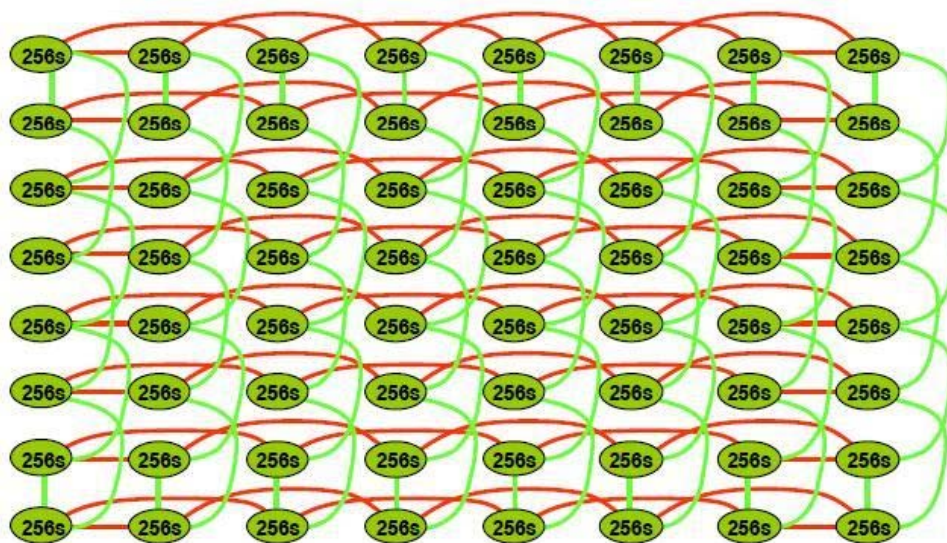


図3c: 16,384ソケット(最大131,072コア)Altix UVプラットフォーム・システム構成。64個の256ソケット・ファットツリー基本ブロックを8 x 8 トラスで相互接続。それぞれの赤および緑のラインは2つのNUMalink5双方向接続を示しています。

システム・サイズに応じて異なるインターコネクト・トポロジを使用することで、パイセクション・バンド幅を最大化し、最遠ノードのレイテンシを最小化できます。また、同時に、導入コストを抑えることができます。多くの場合、メモリはローカル・ブレードまたはトポロジ上に最も近いブレードから読み取られます。しかし、最も遠い場合でも、32,768ソケット(262,144コア)システム上の最大MPIレイテンシは2マイクロ秒以内であると予測されます。

3. 統合されたシステム機能

本ホワイトペーパーのセクション1では、Altix UVプラットフォームを利用することでエンドユーザが得られる利点について説明し、セクション2では、アーキテクチャの機能強化について説明しました。本セクションでは、各アーキテクチャ機能強化が5つの高レベル機能にどのように組み合わせられ、セクション1で説明した利点が生まれるのかを説明します。

- ・ MPIオフロード・エンジン (MOE)
- ・ ペタスケールのメモリおよびデータ・インテンシブな計算
- ・ 大規模 I/O
- ・ ペタスケール対応の計算機能強化
- ・ 向上した信頼性

3.1 MPIオフロード・エンジン (MOE)

MPIオフロード・エンジン(MOE)は、CPUからAltix UV Hub ASIC (UV_Hub) へのMPI通信の負荷をオフロードする一連の機能です。GSMやGAMアドレス空間上で実行されるバリアやリダクションなどの一般的なMPIタスクが高速化されます。MOEは、概念的には、システムCPUからTCP/IPプロトコル処理の負荷をオフロードするTCP/IPオフロード・エンジン (TOE) に似ています。MOEを搭載することによって、CPUのオーバヘッドおよびメモリ・アクセス・レイテンシは小さくなり、MPIアプリケーションのパフォーマンスが向上するほか、非常に多くのプロセッサへの拡張が可能となります。

3.1.1. 軽減されたCPUオーバヘッドおよびMPI_Sendのレイテンシ

Altix UVプラットフォームにおいてMPIメッセージの送信に必要なのは、CPUがペイロード情報をGRU上のレジスタ内へ格納することとGRU message_send命令を発行することのみです。すると、次にGRUが処理を引き継ぎ、適切なリモート計算ブレード上のメッセージキューズロットにデータを送信します。これにより、CPUの負荷が大幅に軽減し、メッセージのレイテンシは他のハードウェア・プラットフォームに比べて大幅に下がり、単一ストリームの転送速度が向上します。ネットワーク・パスの長さが極端に長い場合でも、同様の効果が生まれます。

3.1.2. ハードウェアによる高速バリアおよびリダクション

MPI-1とMPI-2の両仕様には、コミュニケータ・グループ内の全プロセッサ間での同時通信を行うような集団通信が含まれています。Altix UVはAMUの機能とGRUのアップデート・キャッシュを組み合わせることで、バリアやリダクションなど多数の重要な機能を高速化します。

バリアはすべての通信プロセッサを同期するために使用され、これによってすべてのMPIタスクがマスタに対してバリアに到達したことを通知し、その後、マスタはすべてのタスクに対して、バリアが完了したタスクがデータ処理に戻ることができることを通知します。AMUおよびアップデート・キャッシュによりCPUからバリア更新の負荷を軽減し、更新速度が最大で100倍も高速化します。同時に、GRUにより、同期変数の更新がコミュニケータ・グループ内のプロセッサすべてに対してマルチキャストされ、これらの機能によって、バリアの完了が大幅に高速化されます。

リダクション機能は、グローバル操作（SUMやMAXなど）をコミュニケータ・グループのすべてのメンバー間で実行するために使用されます。UV_Hubは、多数の機能をハードウェアとして提供しており、これらの機能を活用すると、リダクションの速度が一般的なクラスタに対して2～3倍に向上します。

3.1.3. MOEへのアクセス

アプリケーションは、MPIを実装するSGI® Message Passing Toolkit（MPT）ライブラリを使用するだけで、MOEの機能を利用できます。また、ユーザアプリケーションにおいてGRUやAMUの主要な機能を使用可能にする低レベルAPIをインターフェースにすることで、他のMPIライブラリでもMOEを使用することができます。

また、低レベルのAPIを使用すると、共有メモリ・アプリケーションやPGAS機能を利用する言語を最適化できます。SGIではこの機能を利用し、コンパイラ・レベルでPGASプログラミング環境のひとつであるUnified Parallel C（UPC）をサポートすることが可能となっています。

3.2. ペタスケールのメモリおよびデータ・インテンシブな計算

Altix UVアーキテクチャは、Linuxの単一OS下で稼働する最大で16TB（インテル® Xeon® プロセッサの物理アドレス限界）のコヒーレントなグローバル共有メモリ、および、ユーザによるPUT/GET操作、PGASプログラミング環境、またはMPIでダイレクトにアクセスすることが可能な最大8PB（UV_Hubの物理アドレス空間）のグローバルにアドレス指定可能なメモリに対応します。

GRUはPUT/GET操作およびPUT操作をダイレクトにサポートし、データのコヒーレントなスナップショットをGSMドメイン間で移動することを可能にします。また、GRUはメモリ参照の数を増加できます。その数は、個々のプロセッサがアクセスできる範囲を超えて、個々のブレードから大規模なNUMALink 5インターコネクでサポートされます。これは、スキャタ／ギャザのパフォーマンスを最大限に引き出す場合に特に重要となります。

さらに、Altix UVは、ペタスケールのメモリを使っている場合の重大な問題のひとつである「変数初期化」に対応しています。

UV_Hubは、メモリページの初期化に、CPUではなくUV_Hub ASICを使用します。これにより、大規模なメモリ・ジョブの起動時間が大幅に短縮されます。

3.3. マッシュアップ機能

インテル® QuickPath インターコネク（QPI）をプロセッサ・ソケット、I/O、およびUV_Hub間で使用すると、ボトルネックが解消され、UVシステム内で1TB/秒を超える総I/O速度が実現されます。

Altix UVプラットフォームのペタスケール対応計算機能／メモリ機能は、外部ストレージや外部ネットワークとの間でデータを移動するために、それに見合う外部I/O機能を必要とします。システム内の各計算ブレードはI/Oライザを搭載することができ、これによって多様なI/Oオプションが利用できます。これらには2つのPCIe Gen2カード、または4つのフルハイトPCIeカードをサ

ポート可能な外部I/O拡張シャーシへの2つの接続が含まれます。Altix UVのGSM設計により、すべてのI/Oデバイスはすべてのプロセッサにより共有が可能となります。1TB/秒を超える高い総I/Oレートをサポートするために、I/Oまたは共有計算機能、I/Oおよびメモリ機能専用の複数のブレード間で物理接続を分散させます。

しかし、最も高速な究極のI/OはI/Oを行わないことです。Altix UVプラットフォームの大規模なメモリ機能はいくつかの方法で利用でき、物理的なI/Oを解消または軽減します。1つ目は、非常に大規模なI/Oバッファ・キャッシュをシステム・メモリ内で定義可能です。これにより、アウトオブコアソルバなどのアプリケーションや大規模なスクラッチ・ファイルを必要とするソルバのI/O効率を向上させます。2つ目は、GRUおよびGAMのマルチキャスト機能を利用すると、Linuxの異なるOSイメージの下で実行されるミラーリング機能およびライトスルー機能によってマルチテラバイトのRAMディスクを生成できます。ミラーリング機能およびライトスルー機能は信頼性向上を実現し、アプリケーションおよびオペレーティングシステムをクラッシュから保護します。

3.4. ペタスケール対応の計算機能強化

ペタスケールのアプリケーションには、数千のノード間に分散した計算構造やデータ構造が含まれ、メモリへの効率的アクセスおよびスレッドまたはプロセッサ間での迅速な同期が必要です。UV_Hubにより、分散ギャザ／スキッタ、コヒーレントなAMOアップデート・キャッシュ、および非同期のユーザレベルのメモリからメモリへのコピーなどの新しいメモリ・アクセス機能が追加され、分散データ構造へのアクセスが効率化するほか、キャッシュライン指向のCPUの動作効率が最も高くなります。また、先進のフェアネス機能もNUMALinkプロトコルに追加され、大規模メッセージおよび小規模メッセージのパフォーマンスがペタスケール環境の高負荷下でも維持されます

ハイパフォーマンス・アプリケーションをサポートするため、Altix UVプラットフォームは、複数階層の16ポートNUMALink 5ルータによる高バンド幅のNUMALink 5接続を利用します。これにより、総バイセクション・バンド幅は15 TB/秒を超え、MPIレイテンシは2マイクロ秒未満に抑えられます。その結果、極めて低いレイテンシかつ高いバンド幅でのアクセスが同一ブレード上でもブレード間のメモリ・リソースに対しても実現できるようにアーキテクチャが最適化されています。

3.5. 向上した信頼性

ハードウェアおよびソフトウェアの多様な機能向上により、Intel XeonプロセッサのAltix UVプラットフォームでは、256,000コアおよび8PBメモリを超えるシステム拡張に不可欠な信頼性を実現しています。ペタスケールのシステムにおける信頼性を向上するため、Altix UVプラットフォームのアーキテクチャには広範な障害隔離、データパス保護、モニタリング機能およびデバッグ機能が備えられており、データ整合性の維持や業務中断の防止に役立ちます。

1つ目に、NUMALink 5プロトコルおよびUV_Hub ASICは機能強化により、追加のエラー・チェック機能およびリトライ機能が備わっており、一時的な通信エラーを2桁も軽減します。2つ目に、UV_Hubへのリモート・メモリの読み取り機能をオフロードにすることにより、プロセッサをハングさせる障害は、リトライや正常に処理をすることが可能になります。3つ目に、UV_Hub ASICは、ノード、メモリまたはインターコネクに障害がある場合でもノード間で通信を可能とする安全なメカニズムを提供します。最後に、システム・ソフトウェアの機能強化により、問題のあるノードおよびメモリを特定し、指定されたリソースのアクティブ・プールから削除することが可能です。

4. おわりに

Altix UVプラットフォームはSGIが生んだ第5世代のグローバル共有メモリ・アーキテクチャであり、一から設計することにより、極めてスケーラビリティの高いシステムでのアプリケーション効率向上を実現しました。また、ハイパフォーマンスな多数のテクニカル・アプリケーションおよびビジネス・アプリケーション、多数のチップ、プロトコルおよびシステム・レベルの機能強化を詳しく調査し、その結果必要であると判断された特定の機能強化が実現されました。これにより、CPUパフォーマンス、システム・スケーラビリティ、信頼性が向上し、管理が簡素化されます。

本ホワイトペーパーでは、Altix UVプラットフォーム用に開発された機能の概要を説明しました。これらの機能を組み合わせ、効率的かつ高いスケーラビリティのシステムを構築するにより、大幅の性能向上が実現し、現在最も大規模かつ最も求められている、計算、メモリおよびI/Oインテンシブな問題を解決します。

(C)2010 SGI Japan, LTD. All rights reserved.

Silicon Graphics International, SGIおよびSGI のロゴマークは米Silicon Graphics International /日本SGI 株式会社の登録商標です。AltixはSilicon Graphics Internationalの登録商標です。またNUMAflex、NUMAlink は米Silicon Graphics International / 日本SGI 株式会社の商標です。Intel および Xeon は Intel コーポレーション、またはその子会社の商標、または登録商標です。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。

日本SGI株式会社

〒150-6031 東京都渋谷区恵比寿4-20-3 恵比寿ガーデンプレイスタワー31階



TEL:0120-161-086 FAX:0120-161-087 <http://www.sgi.co.jp>

本 社	TEL : 03-5488-1811 (大代表) FAX : 03-5420-7201
西 日 本 支 社	TEL : 06-6479-3918 (代表) FAX : 06-6479-3919
中 部 支 社	TEL : 0565-35-2561 (代表) FAX : 0565-35-2189
つくば・東北事業所	TEL : 029-858-1551 (代表) FAX : 029-858-1071
東 北 営 業 所	TEL : 022-221-2301 (代表) FAX : 022-221-2304
北 海 道 営 業 所	TEL : 011-708-1511 (代表) FAX : 011-758-2789