



White Paper

SGI® Altix® UV システムの 信頼性、可用性、保守性について

世界最速のスーパーコンピュータ向けに RAS を最適化

目次

1.0 はじめに	2
2.0 信頼性を備え持つ Altix UV	4
2.1 信頼性	5
2.2 可用性	7
2.3 保守性	8
3.0 信頼性の高いシステム・コンポーネント	8
3.1 業界標準のインテル® Xeon® プロセッサ	8
3.2 革新的な NUMALink® 5 Hub ASIC	10
3.3 環境モニタリング、電源、冷却	11
4.0 可用性のための機能	11
4.1 システム・パーティショニング	12
4.2 メモリ機能強化	13
4.3 信頼性の高い I/O	14
4.4 可用性の高いクラスタ構成	14
5.0 最大の保守性	15
5.1 サービスの継続時間を最大化する SGI Altix UV の機能	15
5.2 シャーシマネジメントコントローラ (CMC) ネットワーク	16
6.0 アップタイムを保障する SGI オプションサービス	17
7.0 まとめ	17

SGI® Altix® UV は最大 2,048 コア (256 ソケット) という、シングルシステムイメージ (SSI) の拡張性を提供します。システムの規模が大きくなる中、そこで稼働するアプリケーションの重要性を考慮すると、確実に継続運用を行うには信頼性、可用性、保守性 (RAS) の機能が必要不可欠です。SGI はサポートするプロセッサ・コアおよびグローバル共有メモリの数を新たな高みにまで引き上げると同時に、革新的な RAS 機能をハイエンドの Linux システムにもたらしめています。SGI の RAS に対する取り組みでは、インテル® Xeon® プロセッサ 7500 番台にみられるような先進的な RAS 機能をベースに、世界最大かつもっとも堅牢なサーバシステムを構築した同社ならではの経験を活用しています。SGI と Intel Corporation は、大規模な共有メモリシステムの信頼性を高めるための集中的な投資を行っています。

1.0 はじめに

SGI Altix UV は、要求レベルが非常に高い科学技術および商用アプリケーションが直面する計算やデータ・アクセスの課題に対応することと、シン・ノード小型サーバによる従来の計算クラスタにある限界の克服を意図して設計されています。科学技術およびエンタープライズ分野でのアプリケーションで、きわめて大量のデータセットの操作と分析を行う能力を求めるものが増加を続けており、多数のプロセッサと非常に大量のメモリを必要としています。これらのアプリケーションの多くは、大量のデータセットやデータベースをシステムメモリに適合させることができれば、多大な恩恵を受けます。それと同時に、従来型の計算クラスタの規模 (計算ノードの数、およびプロセッサの数) が拡大するなかで、必要な計算時間に占める通信処理の割合は相当なものになっています。SGI は Altix UV サーバ製品ファミリーにより、この両方の問題に対応します。

SGI Altix UV では、Altix UV 計算ブレード上の NUMalink® 5 Hub ASIC がハードウェア内の通信処理を実行しますが、これにより非常に多数のプロセッサを単一の Linux システムとして使用することが可能となっています。大規模グローバル共有メモリアーキテクチャは、科学技術から、エンジニアリング、小規模な計算ノードのクラスタでは不可能なビジネスまでの幅広い問題の解決にまったく新しい方法を提供し、以下のような分野に飛躍的な進歩をもたらします。

- 不正検知/サイバーセキュリティ
- バイオインフォマティクス (生物情報科学)
- データ解析
- 大量データセットを用いるリアルタイムインタラクション

- イン・メモリデータベース（例：Oracle TimesTen データベース）
- データフロー・アプリケーション



図1. SGI Altix UV 1000 システム (左) では4 ラック構成のSSIにおいて最大256個のIntel® Xeon® プロセッサ 7500 番台と最大16 テラバイトのメモリをサポートします。またSGI Altix UV 100 システム (右) は2 ラック構成のSSIにおいて最大96個のプロセッサをサポートします (42U ラック使用時。写真は20U ラック)。

Altix UV システムのアーキテクチャと基本設計は、大規模かつデータ集約型のテクニカル・コンピューティングおよびエンタープライズ・コンピューティングに直接的な利益をもたらします。Altix UV は、グローバル共有メモリ、演算パワー、I/O 帯域幅を小規模構成から実質的に無制限の規模にまで拡張する能力を提供します。Altix UV 1000 システム (図1) は、Intel® Xeon® プロセッサ 7500 番台を最大 256 ソケット (2,048 個のプロセッサ・コア) と、最大 16 テラバイトの共有メモリを 4 ラック構成にて実現し、この単一システム・イメージ内で最大 18.5 テラフロップスの演算パワーを提供します。より小規模な構成のためには、業界標準の 3 ラック・ユニット (3U) ラックマウント・フォームファクタにより構成される Altix UV 100 システムが、ミッドレンジ・マーケットに対応します。Altix UV 100 システムは 2 ラック構成において、Intel® Xeon® プロセッサ 7500 番台を最大 96

ソケット(768個のプロセッサ・コア)、共有メモリを最大12テラバイトまで拡張でき、単一システム・イメージ内で最大6.9テラフロップスの演算パワーを提供します。

大規模グローバル共有メモリシステムは、さまざまな問題に対応できるうえ使い勝手がよいという二つの価値をもっています。こうした価値への期待を果たすためには、システムの信頼性がきわめて高くなければいけません。Altix UVは、何年、または何十年にもわたって安定稼働する多数の高性能サーバを提供し続けてきたSGIの25年の実績をさらに前進させます。こうした実績は、設計、製造、テストから導入、運用、サービス・サイクルまで、製品のあらゆる側面に及びます。このホワイトペーパーでは、SGI Altix UVプラットフォームのRAS機能について述べます。

2.0 信頼性を備え持つ Altix UV

大規模グローバル共有メモリシステムの構築に要求されるインテグレーションのレベルを考慮して、SGIはIntel社と強固な協力関係を築いています。SGI Altix UVプラットフォームは、25万6千個以上のコアと8ペタバイトのメモリに拡張可能なように設計され、それに必要とされる信頼性を実現するために、ハードウェアとソフトウェアの様々な機能強化をインテル® Xeon® プロセッサ向けに実装しています。ペタスケールシステム向けの信頼性を強化するために、このアーキテクチャは広範囲に渡る障害切り分け、データパス保護、モニタリング、デバッグ機能を備え、データ整合性の保証とシステム停止の防止を図っています。また問題のあるノードやメモリを識別し、スケジュール済みリソースの実行中プールからそれらを除去できるよう、システム・ソフトウェアも強化されています。

インテル® Xeon® プロセッサ 7500 番台は、ハイエンドへの拡張性と先進的なRASサポートにより、ミッションクリティカルコンピューティングのソリューションに向けた能力と価値を大幅に高めています。インテル® Xeon® プロセッサ 7500 番台は、最大8個の高性能コア、16個の実行スレッド、24メガバイトのレベル3キャッシュに加え、データ保護、可用性の向上、計画的ダウンタイムの最小化を実現するテクノロジーを提供します。これらの先進的プロセッサの利用により、Altix UVは独自仕様のメインフレームやRISCアーキテクチャの数分の1のコストで最高水準の拡張性、可用性、データ整合性を実現します。Altix UVは表1にまとめたとおり、インテル® Xeon® プロセッサ 7500 番台およびIntel 7500 チップセットが提供する基本機能を超える大きな利点をもたらします。

システム要素	RAS 機能
システム	<ul style="list-style-type: none"> ・ 全面的なデータパスの整合性 ・ ファームウェアのプロビジョニング ・ FRU障害解析 ・ オンライン障害診断 ・ アップタイム管理
ブレード相互接続	<ul style="list-style-type: none"> ・ 全面的なデータパスの整合性 ・ 障害検知時の自動再試行 ・ 耐アルファ線ラッチ
プロセッサ	<ul style="list-style-type: none"> ・ 起動時の切り離し
メモリ	<ul style="list-style-type: none"> ・ DRAM障害解析 ・ ページ・マイグレーション ・ 起動時の無効化 ・ 多段の障害封じ込め
電源および冷却	<ul style="list-style-type: none"> ・ 冗長かつホットスワップ可能な電源と冷却ファン ・ オンラインの障害検知と ACPI サポート

表 1. 豊富な RAS 機能を提供する SGI Altix UV システム

2.1 信頼性

システムおよびメモリの規模が大きくなるにつれ、信頼性はますます重要になります。もっともよくあるエラーの種類はメモリ・エラーであることが、私たちの経験からわかっています。

- アルファ粒子または宇宙線がコンポーネントを直撃すると、メモリのソフトエラーが起きる可能性があり、トランジスタの状態が予想外に変化します。
- メモリのハードエラーは、通常メモリビットやリンク、機器の障害に伴って発生します。

Altix UV の設計にあたり、ソフトおよびハードに起因するメモリエラーを回避するために、SGI のエンジニアは、包括的な検証とテストはもとより、回路の設計ルールに細心の注意を払いました。各コンポーネントは、SGI の厳しい性能基準と信頼性基準に合致するよう入念に選択が行われています。製造の過程では、お客様固有の構成に応じた特別な品質保証テストを用いて極めて厳格な品質管理手続きを行うことにより、出荷時の品質をより高めています。これにより、導入後にシステム全体が確実に稼働することを保証します。

物理的な複雑さを減少させることは、システムの信頼性に直接貢献します。最新の CMOS VLSI 設計、製造、ブレード・パッケージングを利用することで、SGI は Altix UV システム内のコンポーネント数を大幅に減らしました。コンポーネント数が少なければ複雑性も低減し、機械的な接合部や相互接続部も減り、システムの可用性がはるかに高まります。Altix UV には、必要があれば冗長性のあるコンポーネントを利用できます。たとえば、すべての冷却ファンと電源装置はホットプラグ可能で n+1 の冗長性を備えています。また、すべてのディスクと I/O カードにはホットスワップ機能があります。

データ整合性の維持は SGI による RAS の取り組みの最優先事項です。エンド・ツー・エンドのエラー修正コード (ECC) データチェックに加え、メモリ、キャッシュ、レジスタ、インターコネクタデータパス・チェックなどの伝統的なエラー検知メカニズムが広範囲に採用されています。また、Altix UV システムは以下のようなサイレントデータ破損を予防する機能を備えています。

- 最大 8 ビットを訂正するよう強化されたメモリ・エラー訂正
- メモリ・モジュール上の複数の DRAM 障害を検知する機能
- エラーの累積的な増加を最小に食い止めるエラーチェック
- さまざまなデータ整合性保証ツール
- 適切なエラー検知とリカバリを検証するハードウェアエラー・インジェクション機能

また、優れた信頼性は、システム内のコンポーネントの利用や、使用されるオペレーティング・システムおよびストレージ・サブシステムからも図られています。

- Linux の信頼性: Altix UV システムは Linux オペレーティング・システムで動作しており、Linux の仕様がシステム信頼性の鍵となります。実際に SGI は Linux 上での共有メモリコンピューティングのあらゆる側面 (RAS を含む) で、Linux コミュニティの主要な貢献者となっています。サイレントデータ破損を予防する具体的機能には、内部状態および整合性のチェック、アプリケーション強制終了/システム停止機能、幅広い回帰テストなどがあります。
- NUMalink 5 テクノロジー: NUMalink 5 Hub ASIC は、システムにおいてきわめて重要な中心的な役割をもっているため、システム信頼性の鍵となります。NUMalink および XIO™ のすべての結合部には CRC (Cyclical Redundancy Checking) エラー検知が使用され、すべてのメッセージを再実行します。NUMalink 5 は、ECC または全てのメモリ・マップドレジスタと内部メモリに対するパリティ、アドレス・パス・パリティ保護、データ送信中のソフトエラー訂正、耐アルファ線ラッチによって、データ保護を推進

します。

- 信頼性の高いストレージ・テクノロジー: 信頼性の高いシステム運用を行うには、信頼性の高いストレージが不可欠です。すべての SGI Altix UV サーバは、信頼性向上の機能を提供する様々な高信頼性ストレージ製品をサポートしています。XVM ボリューム・マネージャーは、重要データの自動的なミラーリングを可能とし、ディスク障害の際にもデータ損失が起きないようにします。全面的な耐障害性を備えた RAID ユニットも利用できます。すべての SCSI またはファイバ・チャネル・ストレージ・ユニットは冗長性のある電力供給、冷却、コントローラを備え、障害時にもデータの供給を継続できます。RAID およびファイバ・チャネル・ストレージではディスクドライブのホットプラグがサポートされており、ストレージをオフラインにすることなく、故障したドライブの交換ができます。当システムはディスクのウォーム・プラグインをサポートしており、ディスクの抜き取りまたは挿入の前に管理コマンドを使ってバスをシャットダウンします。

2.2 可用性

可用性とは、障害を起こしたコンポーネントや予期せぬ動作の問題に、システム全体がどのように巧みに対応できるかの尺度です。SGI は要求度の高い環境での高性能システムの導入に豊富な経験をもっており、この分野における機能を強化しています。Altix UV の場合、問題やコンポーネント障害を検知してそれをシステムから安全に切り離すと同時に、電源や環境的状况にまつわる問題を効果的に処理する能力によって、可用性が強化されています。

問題発生時には、高性能の Altix UV ソフトウェアがシステム利用への影響を最小化します。システム開始時には毎回、パワーオン診断が実行され問題のチェックを行います。検知された問題の箇所（CPU やメモリを含む）をシステム側から切り離すことができるので、システムは起動し、継続して機能します。コンポーネントが再構成されると、Altix UV のシャーマネジメントコントローラ（CMC）が、フィールド交換可能ユニット（FRU）分析に必要な特定データの取得を支援します。

正常な運用が行われている間は、システムの環境センサが温度と電圧をモニターし、問題を特定してクラッシュの発生前にシステムを正常にシャットダウンします。Altix UV 上の NUMalink 5 はケーブル切断/再接続および信号の自動リタイアのほか、データおよびアドレス・バスの全面的な保護機能を備えています。無停電電源装置（UPS）ソリューションも利用可能で、単一のソリューションが UPS システム、電源モニタリング・ソフトウェア、サポートをすべてこなします。また SGI は、マシンの熱を効果的に取り除く水冷ドアをオプ

ションで提供しています。

2.3 保守性

保守性とは、システムが稼働を続けながら保守できる能力に関連しています。Altix UV システムには、平均復旧時間 (MTTR) を短縮するために数多くの強化が行われています。システム内の仕切りやケーブルを取り外す必要なく、コンポーネントに容易にアクセスでき、またシステムからの取り外しが行えるなど、SGI Altix UV ブレードのフォームファクタは保守を非常に容易なものにしています。モジュール内、またはラック内の独立した電源により、他の部分の動作を続けつつ保守のためにシステムの一部をシャットダウンできます。高性能のオンライン診断も継続中のシステム運用を監視しています。さらに Altix UV システムには、技術者が現場に到着する前に素早い診断と修理を行えるサービス・ノードが統合されています。さまざまなオプションサービスも用意されており、それらについては当資料の後半で説明します。

3.0 信頼性の高いシステム・コンポーネント

すべてのハードウェア障害が回避可能というわけではないものの、入念な設計とコンポーネントの選定を行うことで、自動エラー検知および訂正によって発生するエラーの数および影響を低減できます。たとえば、Altix UV システムはシステム・メモリ、ディレクトリ、データバスで発生するすべてのエラーを、確実に検知するよう設計されています。SGI はすべてのシステムバスおよびメモリでエラー修正コード (ECC) を利用しており、シングル・ビットエラーの検知と修正、およびダブル・ビットエラーの検知を行います。効果的な環境コントロールはシステムの信頼性にも貢献しています。

3.1 業界標準のインテル® Xeon® プロセッサ

すべての SGI Altix UV システムは、インテル® Xeon® プロセッサ 7500 番台を搭載しています。ハイエンドシステムとしての拡張性と先進的な RAS 機能をサポートする最新のインテル® Xeon® プロセッサ 7500 番台は、Altix UV のようなシステムの能力と価値を大幅に向上させます。インテル® Xeon® プロセッサ 7500 番台は、新しく先進的な RAS 機能を半導体レベルでサポートします (表 2)。これらのプロセッサの RAS に関連する主な利点は以下のとおりです。

- 確実なデータの保全性：データエラーの防止、検出、訂正、隔離を包括的かつ効果的に行い、データの保全性を維持します。訂正不能なエラーが発生した場合は、タグを付けて隔離し、他のシステムやアプリケーションに影響が及ぶのを防ぎます。

- ・ システム可用性の向上： マシン・チェック・アーキテクチャー・リカバリー（MCA リカバリー）により、OS のアシストで行われるシステムリカバリーが可能になり、前世代のサーバでは稼働停止を引き起こしかねなかった訂正不能なエラーにも対応します。
- ・ 保守性の向上： エラーロギングとレポート機能の強化により、事前障害予知機能（PFA）が可能になり、稼働停止や訂正不能なエラーを引き起こす前に、問題のあるコンポーネントを特定できます

利点	半導体の特長
データ保護 <ul style="list-style-type: none"> ・ 回路レベルのエラーを低減 ・ システム全体のデータ・エラーを検知 ・ エラーの影響を制限 	<ul style="list-style-type: none"> ・ パリティ・チェックおよびエラー修正コード（ECC） ・ メモリ・サーマル・スロットリング ・ メモリ・デマンド&パトロール・スクラビング ・ 破損データ隔離モード ・ 巡回冗長検査（CRC）によるインテル® QuickPath Interconnect（インテル® QPI）プロトコル保護：8ビットまたは16ビット・ローリング
可用性の向上 <ul style="list-style-type: none"> ・ データ接続失敗の修復 ・ 主要システム・コンポーネントの冗長性およびフェイルオーバーのサポート ・ 訂正不能なデータエラーからのリカバリー 	<ul style="list-style-type: none"> ・ マシン・チェック・アーキテクチャー・リカバリー（MCAリカバリ） ・ インテル® スケーラブル・メモリー・インターコネクト（インテル® SMI）レーン・フェイルオーバー ・ インテル® SMI クロック・フェイルオーバー ・ インテル® SMIおよびインテル® QPIパケットリトライ ・ インテル® QPI クロック・フェイルオーバー ・ インテル® QPI 自己障害回復 ・ Single Device DRAM Correction（SDDC）およびランダム・ビット・エラー・リカバリー ・ ダイナミック・メモリー・マイグレーション
計画的ダウンタイムの短縮 <ul style="list-style-type: none"> ・ 障害発生前の予測 ・ システムに代わりパーティションをメンテナンス ・ 障害が発生したコンポーネントの能動的交換 	<ul style="list-style-type: none"> ・ 電子的に絶縁された（静的）パーティショニング ・ MCAエラー・ロギング（CMCI） ・ CPU オンライニング

表 2. 豊富な RAS 機能を提供するインテル® Xeon® プロセッサ 7500 番台

3.2 革新的な NUMAlink® 5 Hub ASIC

Altix UV 計算ブレードは、インテル® Xeon® プロセッサ 7500 番台に対応するソケットを 2 個備えており、それぞれが 64GB の RAM 用のソケットに接続されています。2 個のプロセッサ・ソケットは NUMAlink 5 Hub ASIC を介して残りのシステムに相互接続されています (図 2)。これらのコンポーネントは、Altix UV システム・アーキテクチャの基本的構成要素となっているとともに、エラー検知および修正の機能を備えており、システム全体の信頼性を向上させています。また NUMAlink 5 Hub ASIC は、ハードウェアアクセラレートされた通信処理および接続性を提供するほか、SGI Altix UV システムの RAS 機能向上に直接貢献するイノベーションを実現しています。

- NUMAlink 5 プロトコルおよび NUMAlink 5 Hub ASIC は、エラー・チェックおよび再試行機能が強化されており、一時的な通信エラーが 2 桁低減されます。
- NUMAlink 5 Hub ASIC がリモートメモリ読み込みをオフロードする事により、前世代のシステムではプロセッサのハングを引き起こしかねなかった障害も、正常に再試行または対処されます。
- NUMAlink 5 Hub ASIC は、ノードやメモリ、インターコネク트에障害があった場合でも、ノード間通信を行う安全機構を備えています。
- Altix UV ブレード間の全ての NUMAlink 5 チャンネルで巡回冗長検査 (CRC) が提供されます。

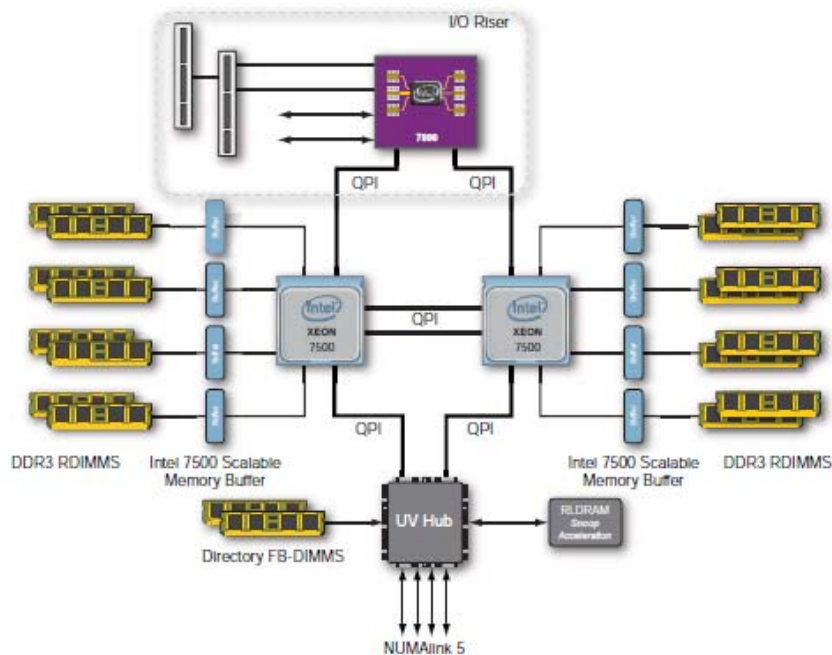


図2 NUMAlink 5 Hub ASICは、Altix UV 計算ブレード上のインテル® Xeon® プロセッサ 7500 番台の2個のソケットに接続しています。

3.3 環境モニタリング、電源、冷却

環境コントロールはシステムの信頼性において重要な要素です。クリーンな電源と適切な冷却が、大規模システムの実際の信頼性を大幅に向上させ、急速な不安定化と、複雑かつ予測不能な障害につながりかねないコンポーネントの加熱を回避します。Altix UV は、ハードウェアの運用を保護する、幅広い環境モニタリング、およびコントロール・システムを備えています。

- 冗長性のある電源装置および冷却ファンにより、これらのコンポーネントの障害からシステムを守ります。
- 速度可変な冷却ファンにより、システムは常に最適な温度で稼働します。
- 過熱した場合（起こる見込みはまずありませんが）には、損害を予防するためシステムの自動シャットダウンが提供されます。
- 電力効率の高さが、低温での運用と障害率の低下に貢献します。
- 電力効率のよいコンポーネントにより、性能を最大に引き出すとともに、物理的なサーバ設置面積、消費電力、冷却の必要性を最小化します。

Altix UV システムには、AC 電源を 90%以上（他社の多くは 60%~70%程度）の効率性で DC（直流）電圧に変換する電源装置および変換アーキテクチャが搭載されています。そのうえ SGI Altix UV 計算ブレードは電力損失を最小化するように設計されており、12 ボルト DC ブレードの入力電圧として使用可能な論理回路レベルの電圧へ、わずか 1 回で変換します。そしてインテル® Xeon® プロセッサ 7500 番台は市場でもっとも効率性のよい製品で、場合によっては他社の RISC CPU の消費電力の数分の 1 で稼働します。

Altix UV の優れたエネルギー効率は、導入したお客様の大部分が水冷による冷却を必要としないことを意味します。とはいえ過密状態のデータセンター施設へ大規模システムを長年導入してきた SGI は、従来型の空冷式サーバに対して早くから水冷テクノロジーの採用を進めてきました。Altix UV 向けの水冷オプションは、各ラックにたまった熱気を遮断するラジエータ状の冷却コイルを持っています。これにより、空気の効率的な冷却とマシンルームの温熱スポット発生の予防、およびホットアイルからコールドアイルへの再循環というよくある問題を予防することが可能です。SGI の水冷システムは、周囲の吸気温度を安定化し、信頼性の向上を図ります。

4.0 可用性のための機能

SGI は、基本となるハードウェア・コンポーネントが提供する信頼性を超えて、全体的な

可用性を向上させる数々の機能を Altix UV システムに盛り込んでいます。たとえばハイエンド Linux ソリューションの開発におけるリーダーとして、SGI は Linux の性能および信頼性の確保に大きな貢献をしてきました。SGI は、一般的な適用が可能な場合はいつでも、可用性を向上する Linux の強化機能を幅広い Linux コミュニティにリリースしてきました。RAS 機能に関する SGI の貢献には次のようなものがあります。

- 修正不可能なエラー (UCE) からの回復機能の強化
- ハードウェア・エラー・レポーティングの向上
- ダブルビット・エラーに起因するパニックの低減
- パーティションを越えたジョブに対する障害隔離の向上

4.1 システム・パーティショニング

Altix UV ハードウェアのパーティショニングにより、1台の物理システムは再ケーブルリングすることなく複数の論理システムに分割できます。Altix UV ハードウェアにはパーティショニング機能が盛り込まれており、信頼性の非常に高い運用ができるとともに、利便性が向上しています。またパーティショニングにより OS 構成が変わる一方で、アプリケーションが使用するメモリは、Altix UV のグローバル共有メモリ機能を使用してパーティション間でのデータ共有を選択的に行えます。この技法により、Altix UV システム上の領域が確保され、大規模メモリアーキテクチャの利点を失うことなく、かつクラスタとしての稼働が可用性を高めます。Altix UV ハードウェアのパーティションは、他のパーティションの運用に影響を与えることなく独立して再起動できるため、可用性の上で多くの利点があります。

- ひとつのパーティションに必要なハードウェアの修理は、他のパーティションの運用を混乱させることなく実行できます。
- アップグレードが必要な場合は、コンピュータ・インフラ全体を停止する代わりに、カーネルのローリング・アップデートによって各パーティションを順番にアップデートするだけで済みます。
- ソフトウェアの開発やテストを行っている企業は、パーティショニングを使用することで、本番環境に非常に近い開発環境およびテスト環境を構築できます。これらの開発環境およびテスト環境は、必要に応じて他の本番用パーティションに影響を与えることなく再起動可能であり、またソフトウェアの不良によって停止されたとしても他のパーティションに影響しません。

他のシステムと比較すると、Altix UV 上のハードウェア・パーティショニングの堅牢性は、

さまざまな独自のハードウェア機能によって向上しています。

- **メモリ保護:** Altix UV は、その計算ブレードに搭載された SGI 設計のチップセットにメモリ保護機能が組み込まれています。この保護機能により、各パーティションは他のパーティションによる予想外の書き込みから保護され、障害を回避します。このハードウェア機能がない他のシステムの場合、間違った構成のカーネルや不正なアプリケーションが不適切なメモリ・アクセスを行うと、メモリ破損に陥る可能性があります。たとえば、SGI MPI ライブラリ内の XPMEM サポートにより、ハードウェアは他のパーティションと共有しているメモリの保護を変更できます。その結果、1つのパーティション内のグローバル・リファレンス・ユニット (GRU) は、他のパーティションからのアクセスのためにメモリをオープンすることなく、共有メモリに直接ロードおよびストアできます。
- **リセットフェンス:** この機能も NUMALink 5 Hub ASIC に組み込まれており、他のパーティションで発生したハードウェアのリセットからパーティションを保護します。リセットフェンスにより、他のパーティションの再起動やハードウェアおよびソフトウェアの障害発生時にも、各パーティションが独立して確実に動作することを保証し、また同時に複数のシステム・モジュールを交換することをサポートします。
- **グローバル・リファレンス・ユニット (GRU) :** Altix UV ハブ・チップに組み込まれた GRU により、パーティション間のデータ転送を高い信頼性で実行できます。この機能により、パーティションは必要に応じてリプリケーションを行い、高速なデータ共有を行うことができます。障害を確実に分離するため、リモートパーティションの障害によってリモート参照を実行中のパーティションがクラッシュまたはハングしないよう、GRU を設計しています。

4.2 メモリ機能強化

メモリ・エラーは、もっともよくあるサーバのエラーです。標準的な Linux 環境のメモリ構成は、Altix UV システムのそれに比べて比較的控え目なものになっています。そのため SGI は、大規模メモリ・システムに向けた Linux の堅牢性向上に取り組んでいます。

メモリ・ロケーションがシングル・ビット・エラーのしきい値を超え、劣悪と判断されたときは、Altix UV メモリ無効化機能により OS がデータを別のページに移動させ、そのメモリを含むページを欠陥としてマークします。それを受けて OS が欠陥のあるページの使用を避けます。SGI は Linux の強化も行い、ハードウェア障害に起因する MCA イベントがシステムを停止させた場合でも、全体的なハードウェア状態を把握できるようにしています。この機能により、障害の根本原因究明を可能とし、適切なコンポーネントを迅速に交換して

システムを完全な運用状態に戻すことができます。

4.3 信頼性の高い I/O

SGI はファイバ・チャンネルおよび高性能ストレージ分野を長らくリードしてきました。SGI は冗長性の高いファイバ・チャンネル・ストレージ・インフラと、その効果的な利用に必要なソフトウェアの導入における先駆者です。Altix UV は以下のような機能を備え、堅牢なストレージ・システムによって実現した信頼性から直接メリットを受けています。

- マルチパス I/O: 複数のブレードにわたる、直接またはファブリックを介した SGI InfiniteStorage RAID アレイに接続されるファイバ・チャンネル・ホスト・バス・アダプタ (FC HBA) を複数備えるシステムでは、シングルポイント・フェイラのない I/O インフラを実現します。マルチパス I/O は、チャンネル全体の I/O 負荷のバランスをとり、障害の生じたポートや HBA から正常なポートに負荷を移行させます。
- InfiniteStorage ファイルシステム (XFS) : SGI が開発し、今では標準の Linux ディストリビューションで提供されている XFS は、ハイパフォーマンス・コンピューティング環境における I/O 要件に対応する一方、エラー回復および急速再起動のためのジャーナリングに信頼性を提供します。
- SGI InfiniteStorage 共有ファイルシステム (CXFS) : クラスタ内の共有データ・アクセスのため、CXFS は XFS を足掛かりに非常に信頼性の高い高性能ストレージ・インフラを形成しています。これによりクラスタ・メンバーは、データをディスクに直接アクセスし、SAN のフル速度で読み書きすることが可能です。パーティショニングされた Altix UV システムは CXFS を利用して、各パーティションがパフォーマンスを損なうことなく、同一のデータセットへのアクセスを共用できるようにします。

4.4 可用性の高いクラスタ構成

Altix UV システムの可用性をさらに強化するため、SGI InfiniteStorage Cluster Manager for Linux またはサードパーティ製の Linux 向け高可用性ソフトウェア・ソリューションを使って、クラスタ構成に設計できます。Cluster Manager によって、複数の Altix UV システムや、単一の Altix UV システム内の複数のパーティションにわたる、可用性の高いアプリケーション・サービスを作成できます。ひとつのクラスタ・メンバーから別のクラスタ・メンバーへのアプリケーションのフェイルオーバーは、稼働中サービスに何ら影響を与えません。

Real Application Cluster (Oracle 11g RAC) は、Altix UV において利用可能であり、高

可用性ソリューションを提供します。Oracle 11g RAC の機能には、ディスク障害に対処する自動ストレージ管理 (ASM) や、特定のアプリケーションや機能に対するフェイルオーバー機能を提供する透過アプリケーション・フェイルオーバー (TAF) があります。

5.0 最大の保守性

SGI の目標は、Altix UV プラットフォームの大半のコンポーネントを、最小のシステム中断で、あるいは中断なしにシステム管理者が保守できるようにすることです。Altix UV のシステム設計には、先進的なシステム・コントロール機能、システム健全性モニタリング、システムのオンライン管理および保守・障害解析を提供する、重要な保守機能が含まれています。本質的に Altix UV は、先進のモジュール型ブレード設計によって保守性を強化しており、保守やメンテナンス、アップグレードの際に個々のシステム・コンポーネントへ容易にアクセスできます。

5.1 サービスの継続時間を最大化する SGI Altix UV の機能

Altix UV の計算ブレードは、Individual Rack Unit (IRU) と呼ばれるシャーシに収められています。Altix UV 1000 IRU エンクロージャには、最大 16 枚の計算ブレードを収容できます。1 台の Altix UV 1000 ラックには、図 3 に示すとおり最大 2 個の IRU (32 ブレード、64 ソケット、512 コア) を搭載できます。Altix UV 100 システムは、最大 2 枚の Altix UV 計算ブレードを収容する小規模な IRU を備えています。Altix UV 100 システムは最大 96 基の Intel® Xeon® プロセッサ 7500 番台をサポートし、コア数は最大で 768 コアをサポートします。



Altix UV 1000 IRU



Altix UV 100 IRU

図3. Altix UV 1000 IRU (ラックあたり2台) は最大16枚のAltix UV計算ブレードをサポートします。Altix UV 100 IRUは最大2枚の計算ブレードをサポートします。

各 IRU 内のコンポーネントは電氣的に孤立しているため、IRU の電源を切らずに交換できます。たいていの場合、電源や個々の PCI カードは、システムやコンポーネントを含むパーティションの運用を中断せずに、ホットスワップすることが可能です。システムがパーティションで分割されている場合、あるパーティション内で障害を起こした計算ブレードは、他のパーティションの運用に影響を与えることなく交換できます。とはいえ、障害を起こした計算ブレードを含むパーティションは、システムの運用中にシャットダウンしなくてはなりません。その他にコンポーネント障害によるダウンタイムを最小化する Altix UV システムの関連する機能には、以下のものがあります。

- インテル® Xeon® プロセッサ 7500 番台の先進的 RAS 機能が、CPU 障害の可能性を最小化します。
- 定期保守が可能になるまで計算ブレードを無効化し、障害を起こした計算ブレードを除外したシステム運用が可能です。
- 個々のメモリ・ページは、欠陥のマーク付けが可能で、運用を継続しつつページを回避させることができます。
- プロセッサおよびメモリは、ブート時に必ず自己テストされ、障害が起きた場合は自動的に再割当が行われます。その後、システムは影響を受けたリソースなしにブートできるため、そのまま運用が継続できます。

5.2 シャーシマネジメントコントローラ (CMC) ネットワーク

すべての Altix UV 1000、および Altix UV 100 の Individual Rack Unit (IRU) エンクロージャはスタンバイ電源で稼働し、有効な電源に接続されていればいつでもオペレーション可能な、組込式のシャーシマネジメントコントローラ (CMC) を備えています。CMC ネットワークはシステム内のハードウェア・パーティションを管理し、ピンポイントの電源コントロール、システム・ブート、構成管理のサポートを提供します。CMC は、システム稼働中にすべての内部レジスタ状態と、接続されている Altix UV 計算ブレードの状態を抽出して豊富な入力データを提供します。これにより、障害アナライザが、フィールド交換可能ユニット (FRU) のレベルにまで落としこんだ障害データ・レポートを作成できるようにしています。

CMC は、完全なハードウェア構成を、個々の FRU シリアルナンバーのレベルまでリアルタイムで読み取ることができます。この機能が、システム・サービスに必要な情報の迅速かつ正確な通知と伝送を支援します。CMC は、エンクロージャ内の各 IRU に対するコントロールとモニター機能、および他の CMC への通信機能を提供します。CMC は、システムが起動して

いない場合や電源が入っていない場合でも動作可能です。全体として、システム・コントローラ・ネットワークは以下のような機能を提供します。

- システム全体に対する電力コントロール
- 個々の計算ブレードに対する電力コントロール
- 環境モニタリング
- ステータスとエラー・メッセージのモニタリング
- システム機能のモニターと変更を行う専用コマンド
- システム・ブート・コントロール

6.0 アップタイムを保障する SGI オプションサービス

長年のグローバル・リーダーとして、SGI カスタマ・サービス組織はミッションクリティカルな 24 時間 365 日のシステム・サポートに至るまで、幅広い SGI カスタマー・サポート・サービスを提供しています。SGI カスタマ・サービス部門は、第三者の評価指標で常に業界の上位に位置付けられています。その他のサービスには以下のようなものがあります。

- SGI MAS (マネージド・サービス) コンソール: コンソール・サーバ管理のための SGI ソリューションは、システム管理者がシステムダウン時に SGI サーバのモニターと管理を行えるようにする貴重なツールで、ネットワーク・アクセスができない場合でもシステムへのインターフェースを提供します。当ソリューションは、1 台以上の異種サーバを管理するためのハードウェア、ソフトウェア、サポート、そしてオンサイト導入パッケージを組み合わせたものです。
- SGI UpSafe UPS (無停電電源装置): UPS システムは、停電、電圧低下、天候または重機設備 (エレベータ、工場機械等) の電源停止等に由来する、電気サージ/サグといった電源の問題から電子機器を守るために非常に重要です。UPS は、電源需要が競合し瞬時の停電も珍しくないような複数テナントが入るビルなどの環境においてとりわけ重要です。SGI は UpSafe を通じて、データセンター環境や Altix UV サーバ構成の特定ニーズに合わせた総合的 UPS ソリューションを提供します。ひとつのソリューションで無停電電源システム、電源モニタリング・ソフトウェア、サポートを網羅しています。

7.0 まとめ

Altix UV 製品ラインは、SGI のパワフルで可用性の高い Altix システムの力強い歴史に立脚した、新たなレベルのパフォーマンス、保守性、総合的な機能を備えています。最大 2,048

個のプロセッサ・コアと最大 16 テラバイトのメモリのサポートに加え、Altix UV の RAS 機能は、継続的な改善の歩みを示しています。Intel Corporation との密接かつ実り多い協力関係をベースに、Altix UV 1000 および Altix UV 100 システムはインテル® Xeon® プロセッサ 7500 番台の RAS 機能と、革新的な NUMalink 5 インターコネクートを融合し、優れた信頼性、可用性、保守性を誇る SGI の Altix UV システム・アーキテクチャを生み出しています。

SGI は主力システム商品の RAS 機能を、信頼性の高いハイパフォーマンス・コンピューティング・ソリューションの不可欠かつ重要な構成要素であると捉えています。システムのモニターおよび管理の手法がお客様の環境で改善され最適化されるなかで、Altix UV システムの RAS インフラストラクチャ（システム・コントローラ、OS、診断法、内部ファームウェアを含む）には、たゆみない RAS 機能の向上が図られています。また SGI の製造部門も、信頼性の継続的強化のため、強力なフィードバックおよびプロセス・コントロール手法を採用しています。SGI は今後も、お客様のフィードバックや実環境での経験に基づいて RAS 機能を継続的に進展させ、最大規模のスーパーコンピュータ・システムとして結実させていきます。

©2010 SGI Japan, LTD. All rights reserved.

Silicon Graphics, SGI, Altix, および SGI のロゴマークは米 Silicon Graphics, Inc. / 日本 SGI 株式会社の登録商標です。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

<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