WHITE PAPER

SGI® UV™ 300H

シングルノード・アーキテクチャを大規模な
SAP HANAテーラード・データセンター統合に
適応して、リアルタイム・オペレーションを実現

INDEX

1.0 はじめに

SGI® UV™ 300Hは、SAP HANA上で実行する大規模あるいは将来に拡大が見込まれるアプリケーション向けに開発されたインメモリー・コンピューティング・アプライアンスです。ハイパフォーマンス・コンピューティング分野で豊富な実績と経験を有するSGIが開発したこのシステムは、インテル® Xeon® E7 プロセッサー とSGI NUMAlink® ASICとを技術基盤としています。SAP HANAテーラード・データセンター統合(TDI: Tailored Datacenter Integration)向けのソリューションとして、UV300Hは、4、8、12、16ソケットのシステムとしてSAP認定を受けており、シングルノードとして32ソケット、24TBの共有メモリまでスケーリングできるよう設計されています。本紙では、お客様が最大級のスケーラビリティと低いTCOでリアルタイム・オペレーションを実行できるように、将来のビジネス拡大まで見据えたSGI UV 300Hのモジュラー型アーキテクチャについて説明します。

2.0 SAP HANAプラットフォームのためのSGIインメモリー・コンピューティング

SGIが長年培ったスケーラブルなインメモリー・コンピューティング技術をベースとしたSGI UV 300Hにより、お客様はシングルノード・システムを必要とするミッションクリティカルなアプリケーションや負荷の高いマルチエンジン解析において、SAP HANAのパワーを最大限に活用できるようになります。また、クラスタが抱えるオーバーヘッドの問題を低減し、サービスレベルを向上させることができます。フルインテグレートされたHANAアプライアンスが必要となる大規模または成長が見込まれるSAP環境では、SGI UV for SAP HANAが優れた選択肢となります。既存のEMC VNXまたはVMXのようなストレージ資産の継続利用を検討する場合、アプライアンス・ビルディングブロックの中核であるSGI UV 300HがSAP HANA TDIとして理想的なソリューションです。

革新的なコヒーレント共有メモリを備えたスケールアップ型シングルノード・アーキテクチャにより、SGI UV 300Hは、大企業におけるSAP Business Suiteおよび他のSAPアプリケーションの利用効果を最大化させます。システムはOLTPとOLAPのワークロードを統合し、時間のかかる抽出、変換、ロード(ETL)処理を排除して、オンデマンドのリアルタイム・レポートを生成します。ユーザは大規模で非常に複雑な結合処理を実行し、またマルチ解析エンジンを利用してテキスト、地理空間情報およびライブデータのストリーミングを同時に取り込むことができます。アプリケーションとインフラストラクチャを統合し、管理コストを増大させる情報リソースのサイロ化を排除することで生産性の向上が可能になります。

図1に示すように、SGI UV 300Hのシングルノード・アーキテクチャにより、お客様はクラスタ構成におけるアプリケーションの複雑さやオーバーヘッドに頭を悩ますことなく、シンプルな環境でアプリケーションを実行できます。構成や管理が難しいクラスタノード、クラスタネットワーク、ストレージエリアネットワークは必要ありません。加えて、性能がほぼリニアかつ自動的に向上するため、SGIアプライアンスのサイズを大きくした場合でも、データベースのパーティション

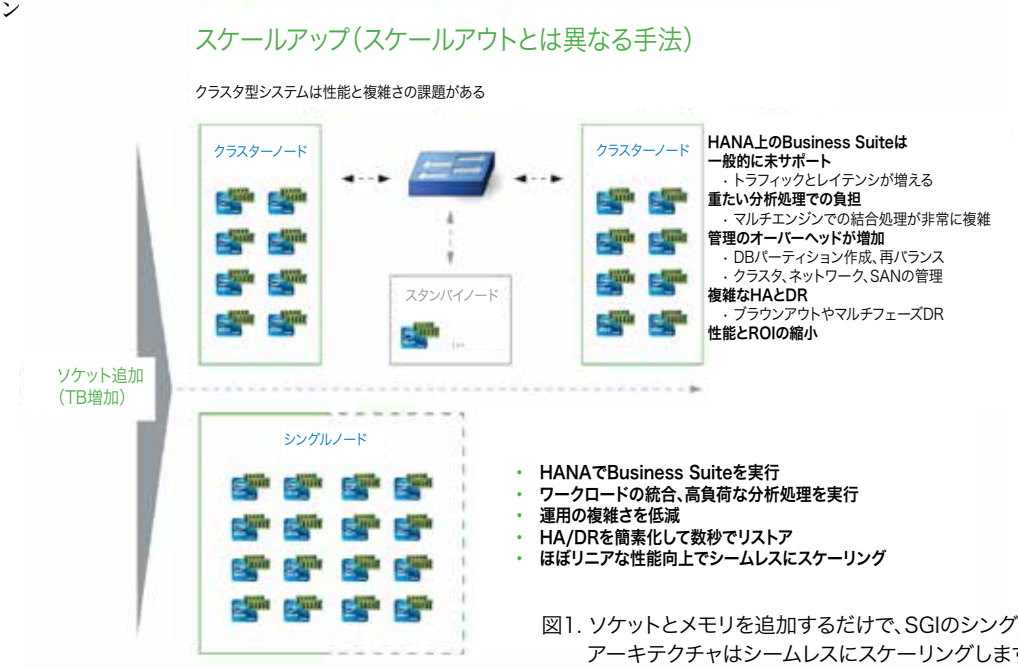

図1. ソケットとメモリを追加するだけで、SGIのシングルノードアーキテクチャはシームレスにスケーリングします

3.0 アーキテクチャの概要

SGI UV 300Hは、シングルノードとして4～32ソケット、24TBまでのキャッシュコヒーレントな共有メモリにスケールするように設計された、先進の対称型マルチプロセッシング(SMP)システムです。SGI UV 300Hのモジュラーシャーシ・アーキテクチャにより、ユーザは複雑さを増すことなく、シングルノード・システムを4ソケット単位で拡張できます。シャーシは第7世代SGI NUMAlink 7 ASIC技術を用いて、All-to-Allトポロジでインターコネクト接続し、超低レイテンシのネットワーク帯域幅をもたらします。システムには次の特徴もあります。

- インテル® Xeon® E7プロセッサーと高密度メモリ
- SUSE® Linux® Enterprise Server for SAP Applications上で稼働するSAP HANA®
- エンタープライズクラスの信頼性、可用性、保守性(RAS)

システムは、標準の19インチ42Uラックに設置可能で、あらかじめプリインストールとテストが行われて出荷されます。システムごとに1つのラック管理コントローラ(RMC)も搭載されます。

3.1 SGI UV 300H

インメモリー・コンピューティング向けSGI UVサーバラインの一翼を担うSGI UV 300HはSGI UV 300スーパーコンピュータをSAP HANA専用にデザインしたモデルです。SGI UV 300Hは5Uモジュラーシャーシを特徴とし、120までのスレッドを処理する4つのプロセッサと、モジュール、プロセッサ、共有メモリをインターコネクト接続するためのNUMAlink ASICを備えています。SGI UV 300Hは、追加シャーシ(標準19インチラックごとに8つまで)を組み合わせることにより、32ソケット、960スレッド(ハイパースレッディング有効時)までスケールアップするように設計されています。インターコネクト接続したシャーシはすべて、単一のオペレーティングシステム・インスタンスで実行するシングルノード・システムとして動作します。それぞれのSGI UV 300Hモジュラーシャーシには次の特徴があります。

- 4つのインテルXeon E7 8890 v2 15コアプロセッサーが、インテルQuickPathインターコネクト(QPI)により内部でリング状に接続
- 8つのメモリライザのそれぞれに2つのSGI Jordan Creek ASICを実装し、12のDDR3メモリDIMMをサポート
- シャーシごとに32GB DIMMを使って3TBまでのメモリ拡張が可能
- 2つのSGI HARP ASICを用いてプロセッサをSGI NUMAlink 7ネットワークファブリックに接続
- BaseIOカード1枚(システムごとに1枚)
- 4台の2.5インチSSDドライブ(4ソケットシステムのみ)
- 高さフル、長さ6/7(最長10.5インチ)Gen3 ×8 PCIeスロットを最大で8つサポート
- 高さフル、長さ6/7(最長10.5インチ)Gen3 ×16 PCIeスロットを最大で4つサポート
- 4つの1600ワット・パワーサプライ
- 8つの80mm × 38mm冷却ファン
- それぞれのパワーサプライに2つの36mm × 28mm冷却ファン(シャーシごとに4つ)
- ラックマウント可能な19インチフォームファクタ

3.2 インテル Xeon E7 プロセッサー

SGI UV 300Hは、シャーシあたり4つのインテル Xeon E7 8890 v2 プロセッサーを搭載しています。各プロセッサは、インテル®QuickPathインターコネクト1.1テクノロジ(QPI)により、高速にPoint-to-Point接続されています。インテル Xeon E7 プロセッサーの主要な特徴には次のものがあります。

- ソケットごとに15個のプロセッサコア
- プロセッサごとに3つのフル幅インテルQPIリンク(QPIリンクごとに、最大転送速度8.0GT/s、総帯域幅25.6GB/s)
- ハイパースレッディング可能なコア、コアごとに2つのスレッド
- 64ビット演算で、48ビット仮想アドレス指定と46ビット物理アドレス指定をサポート
- 単一ビットのエラー訂正を備えた32kBレベル1命令キャッシュ、タグにエラー訂正データとエラー検出を備えた32kBレベル1データキャッシュ
- 256KBのレベル2命令/データキャッシュ、ECC保護(SECDED)
- 37.5MB命令/データ最終レベルキャッシュ(LLC)、ECC保護(タグに2ビットエラー訂正、3ビットエラー検出(DECTED))およびSECDEC)
- コアごとに2.5MBまでの命令/データLLC(これはすべてのコア間で共有)
- 32レーンのPCIe 3.0
- DDR3メモリ

3.3 SAP HANAに最適化された SUSE Linux Enterprise Server for SAP Applications

価値を生むまでの時間を短縮するため、システムはSUSE® Linux® Enterprise Server for SAP ApplicationsとSAP HANAが、単一インスタンスとしてあらかじめ構成された状態で出荷されます。SAPはSAP HANAのライセンス許諾と一次システム・サポートを提供します。データベース、データ処理、アプリケーション・プラットフォームの機能を、インメモリーでSAP HANAプラットフォーム上に統合し、リアルタイム・ビジネスを加速させるノウハウについては次を参照してください:http://hana.sap.com/abouthana.html

SUSE Linux Enterprise Server for SAP Applicationsの詳細を知るには次を参照して下さい:
http://www.suse.com/products/sles-for-sap/

3.4 ラック管理コントローラ

SGI UV 300Hシステムは、ラック管理コントローラを使ってシステム全体の制御ができます。このコントローラはスタンドアロンの1Uラックマウント・シャーシです。24ポートEthernetスイッチを経由して、1台のラック管理コントローラから1つまたは2つのラックに構成された32ソケット(8シャーシ)までのSGI UV 300Hシステムを制御します。

3.5 カスタム設計ラック

カスタム設計の42Uラックは、最大で8つのSGI UV 300Hシャーシと1つのラック管理コントローラを収容できます。ラックは、空冷または冷却水を補助的に用いる水冷のどちらもサポートするように設計されています。ラックの残りのスペースは他の19インチラックマウント装置に使用できます。SGI UV 300Hシステムは、SGI工場であらかじめラックに構成されて出荷されます。既存の標準19インチラックを利用する場合は、SGIサポートエンジニアがオンサイトでSGI UV 300Hをインストールします。

4.0 SGI NUMSlink 7を用いたスケールアップ

SGI UV 300Hが備えるシングルノードでの高いスケーラビリティとリニアな性能向上は、システムに統合されたSGI NUMAlink 7インターコネクト技術により実現されます。

4.1 4ソケットから32ソケットまでのスケーラビリティ

それぞれのSGI UV 300Hシャーシに実装される革新的なSGI HARP ASICと、第7世代のSGI NUMAlink 7ネットワークインターコネクトにより、SGI UV 300Hはシャーシを追加するだけでシングルノード・サーバとしてスケールアップできるよう設計されています。システムを拡張する場合にも、必要なラック管理コントローラは1つだけです。図2に示すように、将来に備えのあるSGIのアーキテクチャは、4ソケット単位で32ソケット、24TBの共有メモリまでスケールアップできるよう設計されています。本紙執筆時点で、最大12TBの共有メモリをサポートする、4、8、12、16ソケットのシステムとしてSAP認定されています。(12、16ソケットは限定出荷)

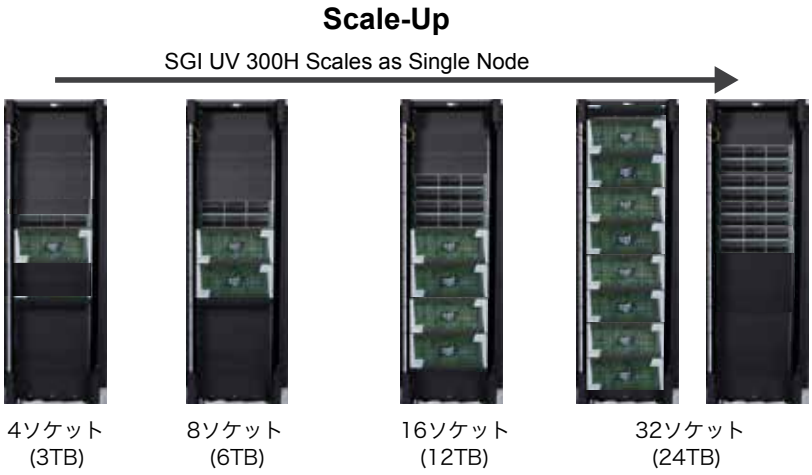

図2. SGI UV 300Hは4ソケット構成から32ソケットまで、4ソケット単位で容易にスケーリングできます

4.2 SGI HARPベースのマザーボード

革新的なSGI HARP ASICは、SGI UV 300Hシャーシのコア・コンポーネントです(図3)。これらのリンクは、複数のシャーシにまたがってプロセッサを接続し、高帯域幅で低レイテンシのSGI NUMAlink 7ネットワークファブリックで接続してシングルノード・システムを構成します。各SGI UV 300Hにはシャーシあたり2つのSGI HARP ASICが実装されており、そのシャーシ内の4つのプロセッサの内の2つに対してQPIチャネルを経由して接続します。SGI HARP ASICには4レーンのNUMAlink 7チャネルが16本あります。HARPインターフェースボードは2つのHARP ASIC接続用に2本のリンクを割り当て、残りの14本のリンクは別のシャーシのHARP ASICへと配線されます。それぞれのリンクは双方向での転送速度のピーク値が14GT/s(7.47GB/s)と高速なため、大容量のHANAデータベースと高負荷なアプリケーションに対し高いスループットをもたらします。

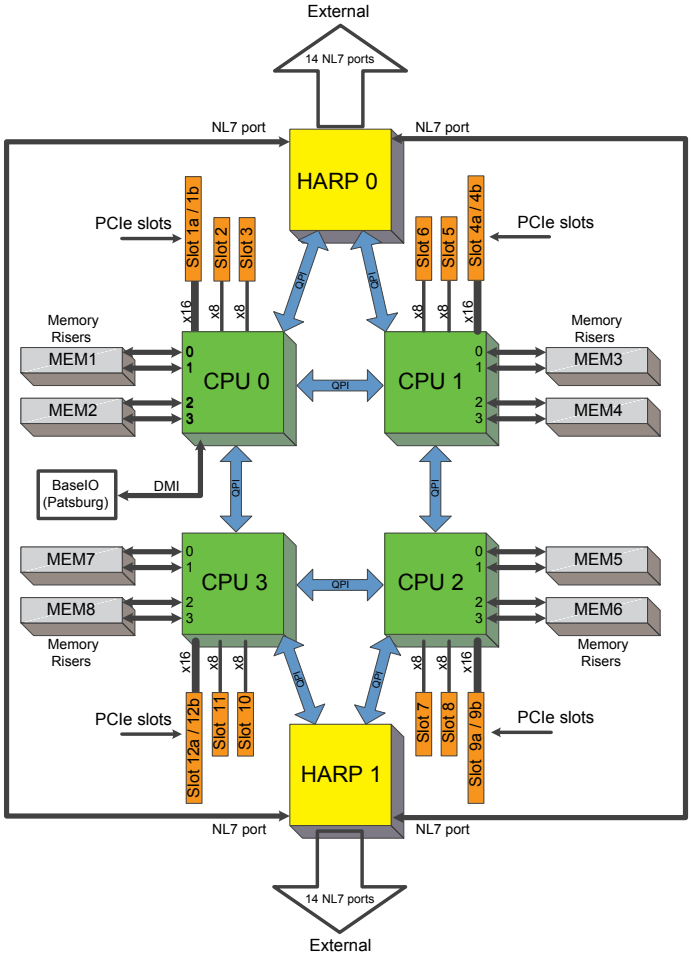

図3. SGI HARP ASICは複数のシャーシを結合してシングルシステムを構成し、
インテルQuickPathインターコネクトはシャーシ内でプロセッサを接続します

図3のとおり、インテルQPIリンクは各SGI UV 300Hのシャーシ内で4つのインテルプロセッサをリング状に接続します。プロセッサソケットがシャーシ内の他の3つのプロセッサソケットと通信するには最大でQPI 2ホップとなります。インテルQPIは次のような特徴を備えています。

- キャッシュコヒーレンシ
- 高速な片方向リンクと同時双方向トラフィック
- CRCによるエラー検出と、リンクレベルリトライによるエラー訂正
- パケットベースのプロトコル

また、インテルQPIには次のようなRAS機能があります。

- リンク幅低減による自動修復
- リンクレベルのリトライ機構
- 8ビットCRCまたは16ビットローリングCRC
- データポイズニングの指摘やウイルスビットなどのエラー報告機構
- QPIリンクでのレーンリバースや極性リバースをサポート
- 高帯域幅ECC保護されたクロスバールータとルートスルー機能

4.3 ccNUMAアーキテクチャ

メモリはSGI UV 300Hシャーシ内でもシャーシ間でも物理的に分散されており、NUMAlinkファブリックに接続しているすべてのプロセッサからアクセスでき、それらのプロセッサすべてで共有されます。SGI NUMAlink 7はメモリのキャッシュコヒーレンシをサポートするccNUMAです。

非均一メモリアクセス(NUMA)

分散された共有メモリシステム内で、メモリはプロセッサから様々な物理的距離に位置しています。結果として、メモリアクセス時間(レイテンシ)が異なり、すなわち不均一となります。例えば、プロセッサはリモートのメモリを参照するよりも、ローカルに配置したメモリを参照するほうが短時間となります。NUMAlinkファブリック内の総メモリはグローバルメモリと呼ばれますが、数多くの異なるサブタイプのメモリがSGI UV 300Hのシステム内に存在します。

- **ローカルメモリ:** プロセッサソケットに直接接続しているメモリにプロセッサがアクセスする場合、そのメモリをローカルメモリと呼びます。
- **オフプロセッサソケットメモリ:** 別のソケットにより管理されているシャーシ内のローカルメモリに対しては、最大でQPI 2ホップとなります。
- **リモートメモリ:** プロセッサが別のシャーシに位置するメモリにアクセスする場合、そのメモリをリモートメモリと呼びます。このパスは、最大でQPI 2ホップ、NUMAlink 7 1ホップとなります。

キャッシュコヒーレンシ

SGI UV 300Hはキャッシュを使用してメモリレイテンシを低減します。データはローカルまたはリモートメモリに存在しますが、データのコピーはシステム全体のさまざまなプロセッサキャッシュに存在します。キャッシュコヒーレンシは、キャッシュしたコピーの一貫性を維持するものです。この機能を実現するため、ccNUMA技術ではディレクトリベースのコヒーレンスプロトコルを使用し、そのプロトコルでは64バイトブロックのメモリのそれぞれにテーブル(ディレクトリ)のエントリがあります。エントリが表しているメモリブロックと同じように、ディレクトリはシャーシ間で分散されています。1ブロックのメモリはキャッシュラインとも呼ばれます。

各ディレクトリエントリは、それが表すメモリブロックの状態のインジケータとなります。例えば、ブロックがキャッシュされていない場合、それは「unowned」(所有者不在)ステートです。1つのプロセッサだけにメモリブロックのコピーがある場合は「exclusive」(排他)ステートであり、2つ以上のプロセッサにそのブロックのコピーがある場合は共有ステートとなります。ビットベクトルは、どのキャッシュにコピーが含まれうるかを示します。あるブロックのデータをプロセッサが変更する場合、同じブロックのデータをキャッシュに持つプロセッサに変更が通知されます。一般に、SGI UVシステムはキャッシュコヒーレンスを維持するための無効化方法を用います。無効化方法は、そのブロックのデータのすべてのキャッシュコピーをフラッシュし、ブロックを変更したいプロセッサがそのブロックの排他的所有権を受け取ります。

4.4 All-to-All NUMAlink 7トポロジ

SGI UV 300HはAll-to-Allネットワークトポロジを特徴とし、すべてのSGI HARP ASICが他のすべてのHARP ASICとダイレクトに接続されます。このトポロジはNUMAlink 7高速インターコネクトチャネルと業界標準ケーブルを使って実装しています。All-to-Allトポロジは、最大のレイテンシが500ns未満で、1つから8つのシャーシまで1シャーシ単位でスケーリングします。図4に、32ソケットのフルシステムのAll-to-Allトポロジを図解します。図では各SGI UV 300H のすべてのNUMAlink 7ポートが使われており、赤線は内部の接続を表します。システムは図に描かれているとおり、8つのSGI UV 300Hシャーシと112本のNUMAlink 7ケーブルで構成されます。

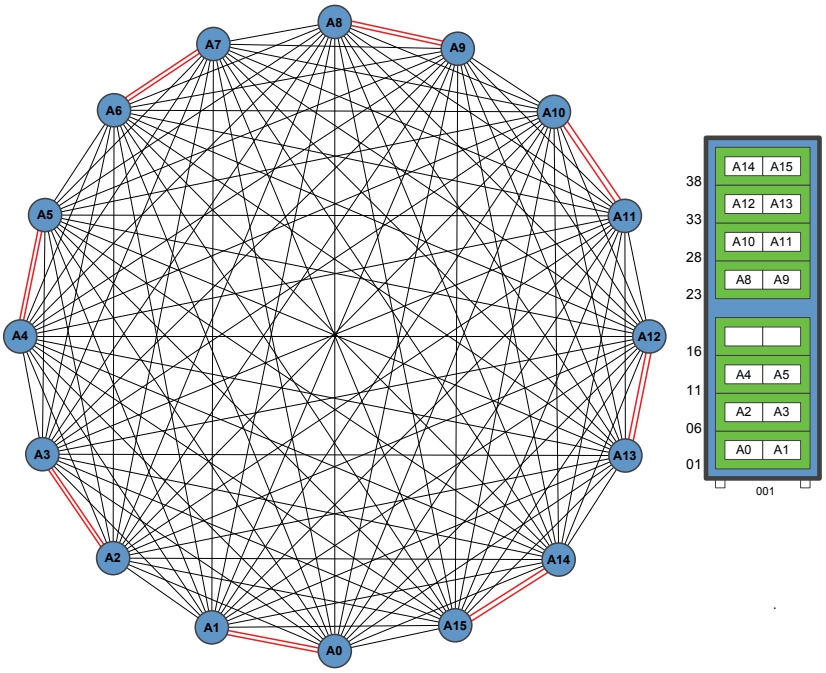

図4. 8シャーシで構成されたシングルノード・システムのAll-to-Allトポロジ

4.5 アダプティブ・ルーティング

NUMAlink 7は、輻輳状態のネットワークと故障したリンクを迂回するアダプティブ・ルーティングに対応し、システムの至るところで高帯域幅と低レイテンシを可能にします。ネットワーク輻輳を判断する手段として、HARP ASICはNUMAlink 7リンクのトラフィックを監視し、どのリンクの使用量が最も多いかを把握します。また、パケットが送信されるまでの待ち時間も監視します。2つのシャーシ間のパケットのルーティングには、1次パスと3つまでの2次パスがあります。

1次パスは最も短いパスであり、最も少ないホップ数を表します。2次パスはホップ数が多くなり、そのためにレイテンシが増える可能性があります。1次パスと2次パスの両方を使用することにより、2つのノード間で利用できる総帯域幅が増えます。パケットを送る前に、HARP ASICは次の基準に基づき、パケットが取るべき最良のパスを選択します。

- 最も短いパス
- 輻輳が最も少ないパス
- パケットが送信されるまでの待ち時間

5.0 エンタープライズクラスの信頼性、可用性、保守性

システムで使用するインテル Xeon E7 プロセッサ同様に、SGI UV 300Hもまた、memlogや、ホットプラグ可能な冗長コンポーネント、全体のシステム設計において様々なRASをサポートしています。

SGIのmemlogユーティリティは、アプライアンスの性能問題や予期せぬダウンタイムにつながりうるメモリDIMMのエラーを解決するのに役立ちます。訂正されたメモリエラーが記録され、解析されます。DIMMページに欠陥があると考えられる場合は、透過的にデータを新しいページにリロケートして古いページから退避させる試みが行われ、中断なしにアプライアンスが稼働し続けられるようにします。また、故障しているDIMMに関してシステム管理者に警告を出すことで、次の計画メンテナンスを待たずにDIMMを交換できます。

コンポーネントの冗長性としては、ホットプラグ可能なファンN+1個、およびホットプラグ可能なパワーサプライN+N個またはN+1個をサポートし、オンラインでの障害検知にも対応しています。すべてのコンポーネントは、アクセスし易いシャーシ前面から保守可能です。

SGI UV 300Hシステムには、20年間にわたるSGIのインメモリー・コンピューティング技術が随所に取り入れられています。信頼性や安定性が高く、スケーラブルなシステムを提供するため、SGIは細部にわたってエンジニアリング開発を進めてきました。それらには、インターコネクト・コントローラ・ハードウェアの設計、高速インターコネクト、高速プリント回路基板(PCB)設計、プラットフォームソフトウェア開発が含まれています。

6.0 まとめ

SAP HANAプラットフォームにより、データベース、データ処理、アプリケーション・プラットフォームをインメモリー・コンピューティング環境に統合することは、現在の流れを真に変えることを意味します。お客様の業務オペレーション、財務、調査、あるいはマーケティングといった各部門が、業務ワークフローを大幅に効率化し、リアルタイムに現状分析ができ、必要なアクションを迅速に行うことが可能になるのです。SGIは、SGI UV 300Hにより、お客様のリアルタイム・ビジネスの実現を支援いたします。

日本SGI株式会社
http://www.sgi.co.jp

〒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 31F

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
つくば営業所	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：011-806-3570（代表）	FAX：011-806-3501