ホワイトペーパー

FUJITSU Server PRIMERGY & PRIMEQUEST Xeon E7-8800／4800 v2（Ivy Bridge-EX）搭載システムのメモリパフォーマンス

PRIMEQUEST 2000 シリーズおよび PRIMERGY RX4770 M1 の Xeon E7-8800／4800 v2（Ivy Bridge-EX）搭載モデルは、2 世代のシステムで実証済みの QuickPath インターコネクト（QPI）のメモリアーキテクチャーの拡張により、先行世代と比較してパフォーマンスが 2 倍も向上します。このホワイトペーパーでは、アーキテクチャーの変更されたパラメーターについて説明し、それが商用アプリケーションのパフォーマンスに与える影響を数量化します。

バージョン
1.1
2014-05-16

目次

ドキュメントの履歴

バージョン 1.0（2014 年 3 月 7 日）

初版

バージョン 1.1（2014 年 5 月 16 日）

PRIMERGY RX4770 M1 を追加

はじめに

Intel Xeon E7-8800／4800 v2（Ivy Bridge-EX）プロセッサが搭載された PRIMEQUEST 2000 シリーズおよび PRIMERGY RX4770 M1 のモデルでは、大部分の負荷シナリオでパフォーマンスが実に 2 倍に向上したハイエンドサーバの製品セグメントを継続しています。Westmere-EX 搭載の旧機種（PRIMEQUEST 1800E2、PRIMERGY RX900 S2、PRIMERGY RX600 S6）と比較してパフォーマンスが向上した 2 つの理由を以下に示します。

- プロセッサチップの製造テクノロジーが 32 nm から 22 nm へ進化したことにより、従来のプロセッサあたり最大 10 コアが 15 コアまで可能になりました。同時に、Sandy Bridge-EP 世代のデュアルソケットサーバですでに行われていたマイクロアーキテクチャーの更新が、ここでも適用されました。これらの手段はパフォーマンス向上に 50 %程度貢献しています。
- 前述したパフォーマンス向上の残りの半分の理由は、メモリシステムの拡張によるものです。

2009 年以来実証済みの QPI（QuickPath Interconnect：QuickPath インターコネクト）システムアーキテクチャーの主要機能が維持されています。プロセッサには統合されたメモリコントローラーがあるため、ローカルメモリモジュールの高性能な制御が可能になります。同時に、QPI リンク経由でメモリの内容を隣接プロセッサに提供し、隣接プロセッサからの情報を要求することもできます。ローカルメモリとリモートメモリのアクセスを区別するこのアーキテクチャーは、NUMA（Non-Uniform Memory Access：非均等型メモリアクセス）タイプのアーキテクチャーです。

最大メモリ容量と最適な RAS（Reliability、Availability、Serviceability：信頼性、可用性、サービス性）を設計目標とするハイエンドサーバクラスでは、メモリバッファーがコントローラーとチャネル間に配置されるだけでなく、プロセッサあたり 2 つのメモリコントローラーのそれぞれに 4 つの DDR3（Double Data Rate：ダブルデータレート）メモリチャネルがあります。この構成により、プロセッサあたりの DIMM（Dual Inline Memory Module：デュアルインラインメモリモジュール）スロット数を、メモリバッファーなしで動作するデュアルソケットサーバの場合よりも増やすことができます。最新バージョンのバッファー（Jordan Creek 1）では、24 個の DIMM スロットと、プロセッサあたり 1.5 TB の最大構成が可能です。その半数は、メモリバッファーを持たない現行世代の Ivy Bridge-EP ベースデュアルソケット PRIMERGY サーバ [関連資料 3] に適用されます。

メモリバッファー搭載の QPI ベース NUMA アーキテクチャーのプロファイルには変更がありませんが、Xeon E7-8800／4800 v2 搭載サーバのメモリパフォーマンスの向上は、次の機能から生じたものです。

- Jordan Creek 1 は、先行世代（Mill Brook 2）では最大 1066 MHz だった DDR3 メモリ周波数を、1600 MHz までサポートします。
- 旧システムの DDR3 チャネルは常にロックステップモード（2 つのチャネルの動作が同期するモード）であり、これによって RAS 機能が向上しました。今回もこのモードを継承しています。ただし、このモードを使用するには、独立したメモリチャンネルの帯域幅をさらに広くする必要があります。つまり、RAS 機能とパフォーマンスとの間に新たなトレードオフが生じます。
- 最大 QPI 周波数が、6.4 GT/s（ギガトランスファー/秒）から 8.0 GT/s に増加しました。
- キャッシュコヒーレンシプロトコルが、スヌーピングベース（QPI 1.0）からディレクトリベース（QPI 1.1）に変わりました。

メモリパフォーマンスの最も基本的な指標であるメモリ帯域幅は、PRIMEQUEST 2800E の測定で 110 GB/s から 393 GB/s に向上し、PRIMERGY RX4770 M1 では 102 GB/s から 244 GB/s に向上しました。

このホワイトペーパーでは、強力なシステムを構成するために必要な、メモリアーキテクチャーに関する基本的な知識を提供しています。ここでは、次の点を取り上げます。

- NUMA アーキテクチャーであるため、すべてのプロセッサのメモリを可能な限り同等の構成にする必要があります。これは、各プロセッサが原則としてそのローカルメモリ上で動作するためです。
- メモリアクセスを並列化するために、物理アドレス空間の隣接する領域をメモリシステムの複数のコンポーネントに分散させます。これは技術用語で *インターリーブ*と呼ばれます。インターリーブは 2 つの次元で行われます。まず、各プロセッサにあるメモリコントローラーと DDR3 チャネルが含まれる横方向においてです。そして、ロックステップ動作モードの影響を受けるのは、メモリパフォーマンスのこの側面です。また、個々のメモリチャネルの中でもインターリーブを実現しています。このためのリソースがランク数です。ランク数は、DIMM の下位構造で、ここに DRAM

（Dynamic Random Access Memory：ダイナミックランダムアクセスメモリ）チップのグループが統合されています。個々のメモリアクセスでは、常にこのようなグループを参照します。

- メモリ周波数はパフォーマンスに影響を与えます。こうした周波数は、DIMM のタイプ、数、省エネモードによって、1600 MHz、1333 MHz、1066 MHz と変わります。DIMM のタイプと数は必要なメモリ容量に関わるため、パフォーマンス、容量、消費電力の側面を相互に比較検討する必要があります。

比較検討しやすいように、影響を与える要因を取り上げ、数量化しています。数量化には、STREAM と SPECint_rate_base2006 のベンチマークを使用します。STREAM でメモリ帯域幅を測定します。SPECint_rate_base2006 は、商用アプリケーションのパフォーマンスのモデルとして使用されます。

ミラーリングやスペアリングなど、冗長性を考慮する場合のメモリパフォーマンスについては、本書の最後にまとめています。

メモリアーキテクチャー

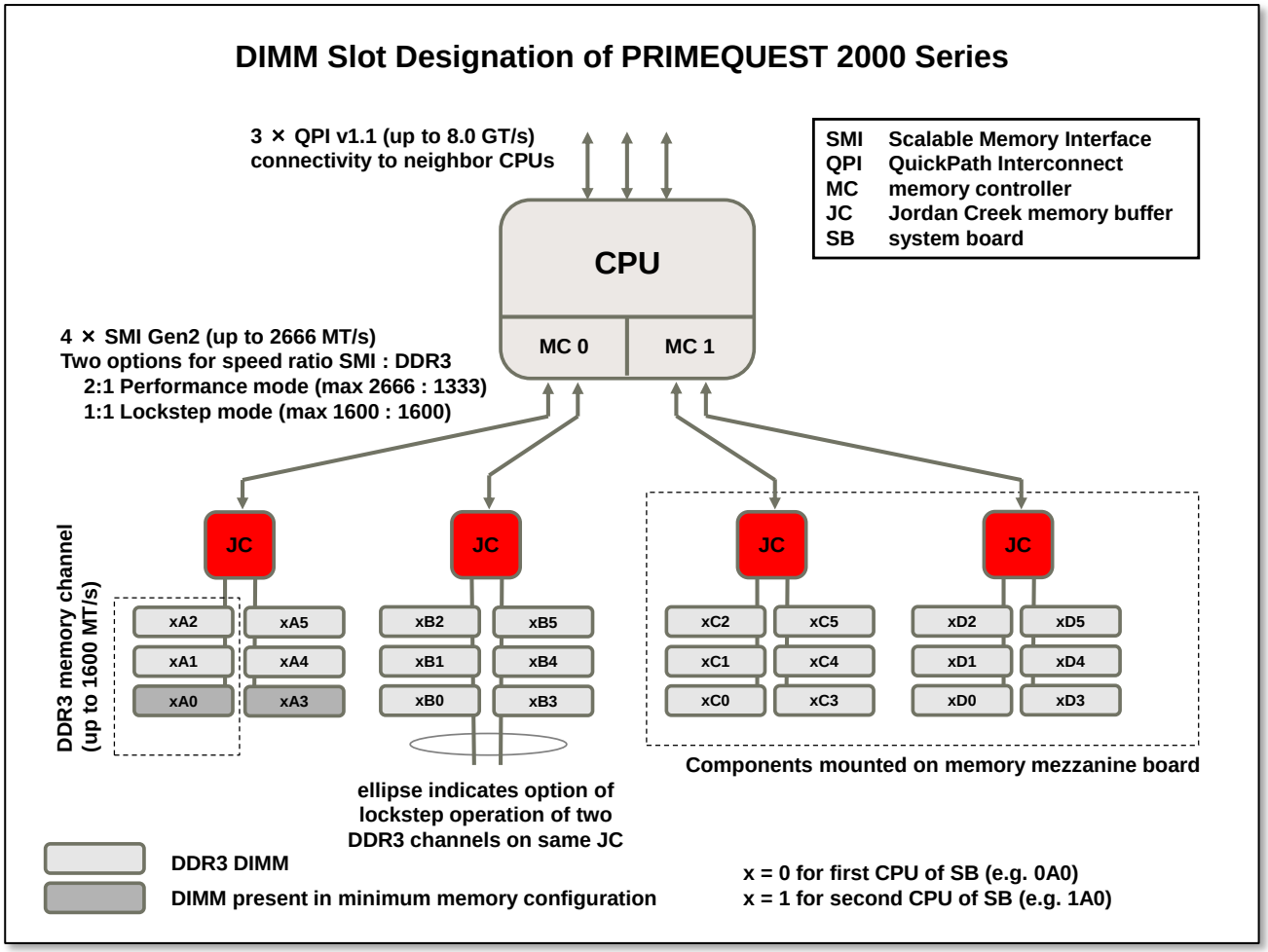
ここでは、5 部構成でメモリシステムの概要を説明します。まずブロック図で、利用可能な DIMM スロットの配置を説明します。2 つ目のセクションでは、使用可能な DIMM タイプを示します。続く 3 つ目のセクションでは、ファームウェアと、メモリシステムに影響を与える BIOS パラメーターについて説明します。4 つ目のセクションでは、有効なメモリ周波数への影響について説明します。最後のセクションでは、メモリパフォーマンスがある程度「理想的」になるメモリ構成表を示します。

DIMM スロット

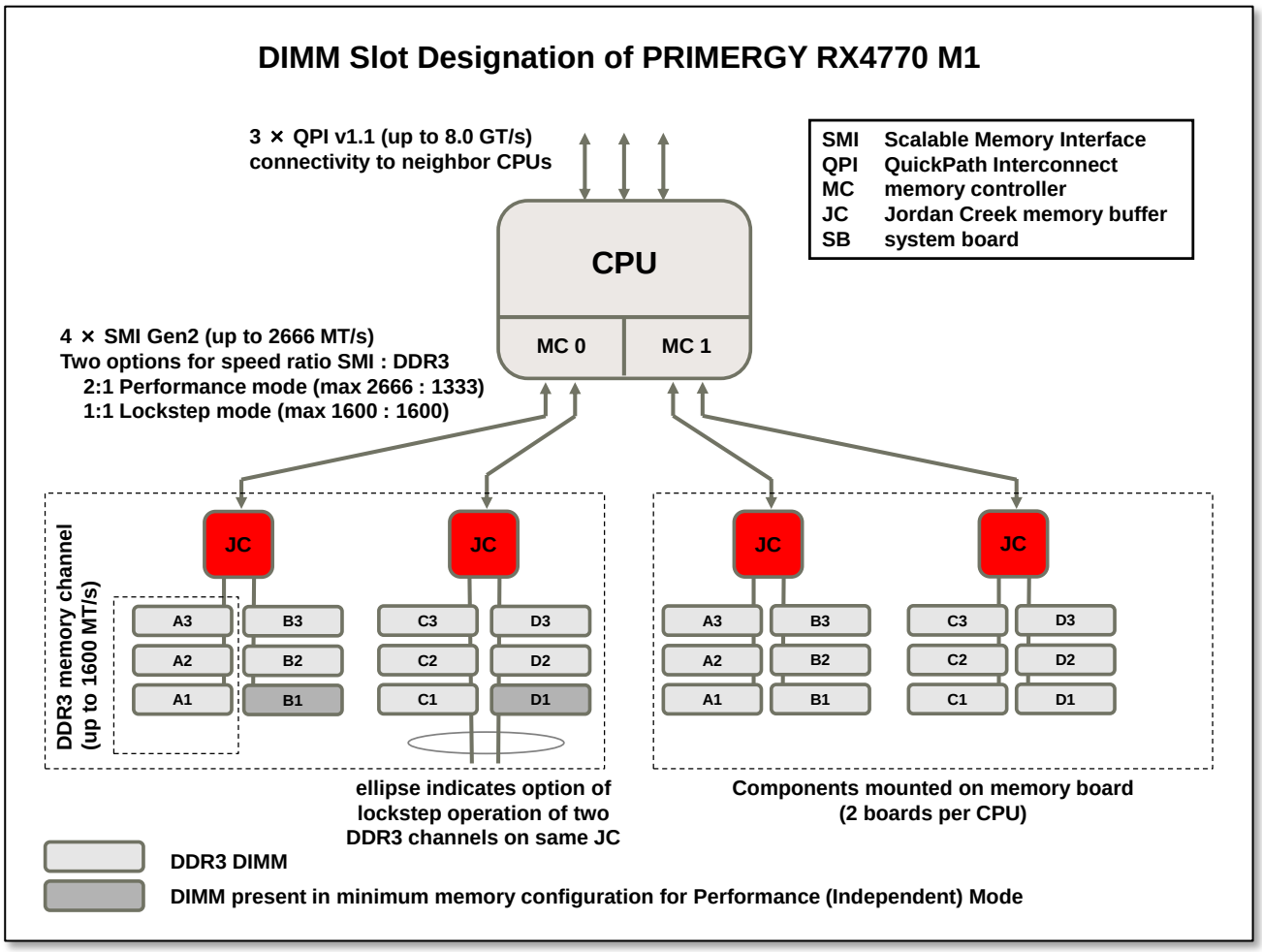
次の図には、個々の Ivy Bridge-EX プロセッサから見たメモリ接続を示しています。各プロセッサには、2 つの統合メモリコントローラーがあります。各コントローラーは、双方向のシリアル SMI Gen2（Scalable Memory Interface）リンク経由で 2 つの Jordan Creek 1 メモリバッファーに接続しています。各メモリバッファーには、それぞれ DIMM スロットを 3 つ備えた 2 つの DDR3 メモリチャネルがあります。したがって、プロセッサあたり合計 24 個の DIMM スロットが装備されています。

Jordan Creek 1 の新機能では、メモリチャネルあたり 3 個の DIMM がサポートされています。旧モデルの Mill Brook 1 と 2 では、チャネルあたり 2 個しかサポートされていませんでした。チャネルごとに構成された DIMM の数は、その構成の DPC（DIMMs per channel：チャネルあたりの DIMM 数）値と呼ばれます。この値は、パフォーマンスに一定の影響を与えます。チャネルが等しく構成されていない場合は、最大 DPC 値がシステム全体に影響を与えます。

PRIMEQUEST 2000 シリーズのシステムは、それぞれに 2 基のプロセッサとそのメモリリソースを備えたシステムボードから構成されます。図の下に示しているように、DIMM スロットの x に置き換わる数値は、1 つ目のプロセッサのスロットの場合は 0 となり 、2 つ目のプロセッサの場合は 1 となります。各プロセッサでは、24 個のスロットの半分がシステムボード自体にあります。残りの半分は、インストールされたメザニンボード上にあります。



4 つのプロセッサはすべて、PRIMERGY RX4770 M1 の 1 つのシステムボード上にあります。DIMM スロットは、それぞれに 12 個のスロットを持つメモリボード上にあります。つまり、各プロセッサに最大 2 つのメモリボードがあります。コンフィギュレータは、プロセッサあたり 1 つのメモリボードか 2 つのメモリボードかで構成を区別します。スロットの名前は、メモリボード内にのみ明記されています。完全な名前には、メモリボードの追加の仕様が必要です。



この図でメモリバッファーの 2 つの DDR3 チャネルの例に表示されている楕円は、ロックステップモードで 2 つのチャネルを動作させるためのオプションを示しています。この動作モードでは、すべてのメモリアクセスが両方のチャネルを介して同時に発生します。つまり、読み取りまたは書き込みが行われるブロックが 2 つのチャネルで分割されます。これは、メモリエラーの修復機能を向上させるために行われます。そのため、ロックステップモードでは、x4 SDDC（Single Device Data Correction）よりも強力な機能である x4 DDDC（Double Device Data Correction）が、独立したメモリチャネルでサポートされています。ロックステップ動作モードは、常にシステム全体（つまり、すべてのメモリチャネル）に適用されます。

ロックステップモードの強化された RAS 機能は、メモリ帯域幅を消費します。プロセッサの 8 個の物理メモリチャネルが 4 個の論理メモリチャネルに減るためです。これによって並列化される容量が制限されるため、メモリアクセスのパフォーマンスも制限されます。コンポーネントである Jordan Creek 1 と SMI Gen2 の技術革新により、このモードが選択できるようになりました。システムまたはパーティションは、ロックステップモードまたはパフォーマンス／独立モードのいずれかに設定できます。Nehalem-EX と Westmere-EX の 2 つの先行世代のシステムは、常にロックステップモードでした。

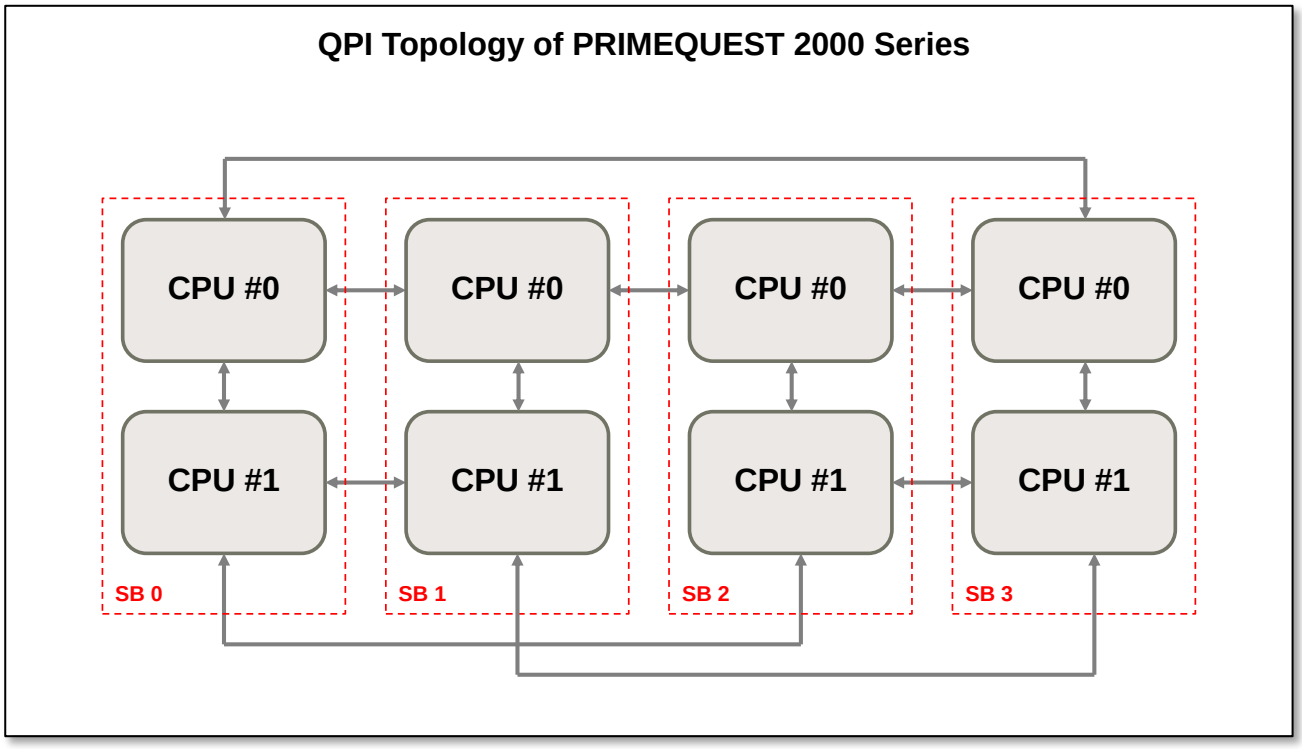
動作モードの適合性は、リソース SMI Gen2 リンクと DDR3 チャネルの周波数に影響を与えます。8 つのチャネルが対する SMI Gen2 リンクが 4 つだけのため、パフォーマンスモードでは、最大メモリ帯域幅を実装するリンクは、メモリチャネルの 2 倍の速度となります。それに対し、ロックステップモードの場合の周波数は同じです。図では、両方の場合で考えられる最大の周波数を示しています。パフォーマンスモードの場合は SMI Gen2 によって上限が 2666 MT/s に、ロックステップモードの場合は、Jordan Creek 1 によって DDR3 周波数の上限が 1600 MHz になります。したがって、パフォーマンスが低い方のモード（ロックステ

ップ）がより高い DDR3 周波数をサポートするという変則性が生じます。ただし、一段階高い DDR3 周波数よりも、より広いメモリ帯域幅の方が価値があります。

前に示した図では、各ケースで濃い灰色のものが 2 つの DIMM で構成される最小構成を表しています。これは、PRIMEQUEST 2000 シリーズと PRIMERGY RX4770 M1 の違いです。

PRIMEQUEST 2000 シリーズでは、ミッションクリティカルなサーバとして、ロックステップ動作が可能なメモリ構成のみとする原則があります。このために、2 つのメモリチャネルに関して Jordan Creek 1 メモリバッファーでは常に対称になっています。マークされた最小構成では、このモードが考慮されます。2 番目に構成されるスロットペアは xC0／xC3 であり、それに応じて xB0／xB3、xD0／xD3 と続いていきます。既存のメモリチャネル全体を順を追って構成することにより、使用可能なすべてのメモリリソースを均等に活用することができます。また、こうした構成はパフォーマンスにも関係します。

各メモリ構成のロックステップ機能は、PRIMERGY RX4770 M1 には存在しません。2 つの DIMM で構成される最小構成は、2 つ目のメモリバッファーを組み込むことでこのケースで考え得る最高のパフォーマンスを前提としています。この構成では、パフォーマンスモードのみを使用できます。PRIMERGY RX4770 M1（マークなし）のロックステップ対応の最小構成は、1 つ目のメモリボードの A1、B1、C1、D1 の位置にある 4 つの DIMM で構成されます。

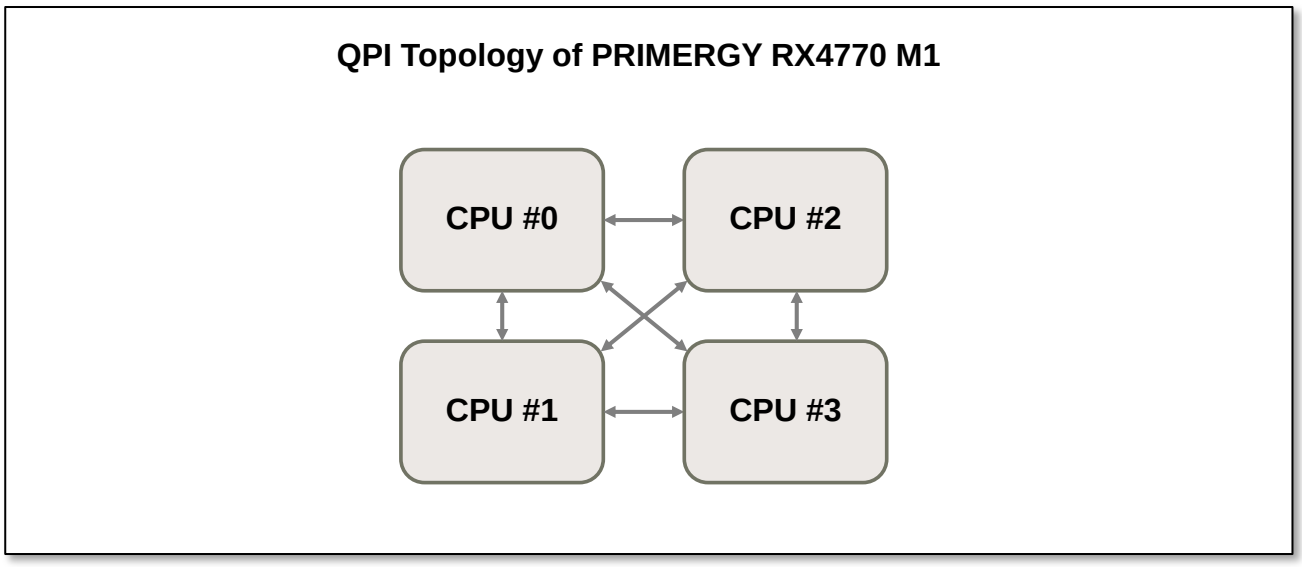


次の図に、PRIMEQUEST 2000 シリーズの QPI トポロジー（プロセッサとそれに関連したメモリコンポーネントのネットワーキング）を示します。ネットワーキングは、プロセッサごとに 3 つの QPI リンクのみを経由しているため、SMI Gen2 リンク、メモリバッファー、DIMM スロットなどの前述したコンポーネントの説明は省略します。また、すべての図で、プロセッサあたりの 32 オンチップ PCIe Gen3 レーンはメモリアーキテクチャーには直接関係しないため、省略されています。

8 基のプロセッサを搭載した完全構成の PRIMEQUEST 2000 シリーズのすべてのプロセッサは、隣接する 7 基のプロセッサのうち、3 基のプロセッサのみと直接接続します。この 3 つのプロセッサは、直接接続されていないプロセッサと通信する場合、中継器としての機能を果たします。必要な中継器は 1 つだけです。このようなアクセスで生じる遅延は、直接結合の場合と比べて大きくなりますが、ソフトウェア対応の NUMA アーキテクチャーではローカルアクセスが主流であるため、このような追加機能が正当と認められます。

プロセッサを最大 4 基備えた PRIMEQUEST 2400E モデルでは、システムボードは 0 と 1 しかありません。この場合と、システムボードが 4 枚未満の PRIMEQUEST 2800E パーティションの場合は、未使用の QPI インターフェースが生じます。

PRIMERGY RX4770 M1 は、最初から 4 つのプロセッサに制限されています。これにより、プロセッサあたり 3 つの QPI リンクによって各プロセッサが互いに接続されるシステム設計が可能になります。したがって、次の図に示す QPI トポロジーは、PRIMEQUEST 2000 シリーズのトポロジー、特に PRIMEQUEST 2400E モデルのトポロジーとは異なります。



この QPI トポロジーの図は、システム全体のネットワーキングに対するプロセッサチップの重要な役割を示しています。最大構成でない場合、存在しないプロセッサに割り当てられた DIMM スロットは使用できません。

DIMM タイプ

メモリ構成にあたっては、次の表に示す DIMM 数が考慮されます。DIMM には、*レジスタード*（RDIMM）、*ロードリデュースド*（LRDIMM）があります。この 2 つの DIMM タイプを組み合わせた構成はできません。これは、DDR3 メモリを備えたすべてのシステムに常に適用されます。

メモリモジュール（システムリリース以降）										
メモリモジュール	タイプ	容量 [GB]	ランク数	メモリチップのビット幅	周波数 [MHz]	低電圧	Load reduced	Registered	ECC	GB あたりの相対価格
16GB (2x8GB) 1Rx4 L DDR3-1600 R ECC (2 × 8 GB 1Rx4 PC3L-12800R)	RDIMM	16	1	4	1600	✓		✓	✓	**1.0**
32GB (2x16GB) 2Rx4 L DDR3-1600 R ECC (2 × 16 GB 2Rx4 PC3L-12800R)	RDIMM	32	2	4	1600	✓		✓	✓	0.9
64GB (2x32GB) 4Rx4 L DDR3-1600 LR ECC (2 × 32 GB 4Rx4 PC3L-12800L)	LRDIMM	64	4	4	1600	✓	✓	✓	✓	1.5
128GB (2x64GB) 8Rx4 L DDR3-1333 LR ECC (2 × 64 GB 8Rx4 PC3L-10600L)	LRDIMM	128	8	4	1333	✓	✓	✓	✓	3.3

この表は、DIMM がそれぞれ 2 枚単位で順番に提供される事実と、PRIMEQUEST 2000 シリーズおよび PRIMERGY RX4770 M1 の構成プロセスを考慮に入れています。その理由は、ペアでの構成仕様です。

どの DIMM タイプでも、データは 64 ビット単位でメモリコントローラーと DIMM 間で転送されます。これは、DDR3-SDRAM メモリテクノロジー（Synchronous Dynamic Random Access Memory：同期型 DRAM）の機能です。64 ビットの帯域幅のメモリ領域は、DRAM チップのグループから DIMM 上に設定されます。この個々のチップが 4 ビットまたは 8 ビットを受け持ちます（タイプ名のコード x4 を参照してください。Ivy Bridge-EX 搭載サーバでの x8 モジュールのサポートは現在のところ予定されていません）。このようなチップグループを*ランク*と呼びます。表に示すように、1 ランク、2 ランク、4 ランク、または 8 ランクの DIMM タイプがあります。メモリチャネルあたりの利用可能なランク数は、パフォーマンスに一定の影響を及ぼします。これについては後述します。4 ランクまたは 8 ランクの DIMM のメリットは最大容量にありますが、その一方で、DDR3 の仕様は、メモリチャネルあたり最大 8 ランクのみをサポートしています。

そのことを踏まえると、2 つの DIMM タイプの重要な特徴は、次のようになります。

- RDIMM：メモリコントローラーの制御コマンドは、DIMM 上の独自のコンポーネントにあるレジスター内でバッファーされます（これが名前の由来です）。メモリチャネルの負荷が軽減されることで、最大 3 DPC（チャネルあたりの DIMM）での構成が可能になります。規模の小さいサーバクラスで見られるアンバッファード（UDIMM）DIMM の場合は、2DPC 構成のみが可能です。
- LRDIMM：制御コマンドとは別に、データ自体も DIMM 上のコンポーネントにバッファーされます。さらに、この DIMM タイプのランク増加機能により、いくつかの物理ランクを論理ランクにマップできます。したがって、メモリコントローラーは論理ランクを監視するだけです。ランク増加機能は、メモリチャネル内の物理ランクの数が 8 を超える場合に有効になります。

RDIMM または LRDIMM のうち、どのタイプグループが望ましいかは、通常、必要なメモリ容量によって決まります。周波数とランク数のパフォーマンスへの影響は、2 タイプとも同じです。こうした影響は、タイプとは関係がありません。タイプごとにパフォーマンスに影響が出ることもありますが、非常に小さいため、たいていの場合は無視できます。タイプごとの影響として、ここで例を 2 つ挙げておきます。ただし、大きな影響ではないため、システムによる定量的評価には表れません。

- Ivy Bridge-EX ベースサーバのローカルメモリアクセスでは、未ロードシステムの場合に RDIMM で約 110 ナノ秒のレイテンシがあります。この値は、メモリ周波数 1333 MHz に該当します。LRDIMM の場合はさらに 5 ～ 10 ns 高くなり、これも 1333 MHz に該当します。理由は、DIMM ではバッファーコンポーネントがより複雑になるためです。この差異を推定するには、負荷のあるシ

ステムでは遅延が大きくなることを考慮する必要があります。その結果、前述の差異のパーセンテージを低くすることができます。

- LRDIMM を使った構成でメモリチャネルあたりの物理ランク数が 8 を超える場合、ランク増加を行うと、最大メモリ帯域幅とアプリケーションパフォーマンスが、RDIMM を使った構成と比べて約 5 %減少します。

すべての DIMM タイプは、1.5 V または低消費電力の 1.35 V で動作します。ただし、1.35 V または LV（Low Voltage：低電圧）の動作ではメモリ周波数が低くなるため、メモリパフォーマンスも低下します。ファームウェアと BIOS パラメーターに関する次のセクションでは、このトレードオフのための管理パラメーターについて説明します。

その後に続くセクションでは、特定の構成の有効なメモリ周波数について説明します。エネルギー消費量とのトレードオフを別にすれば、有効なメモリ周波数は、影響を与えるその他の多くの要因によって決まります。DIMM タイプの表内の最大周波数は、こうした有効な周波数の上限を表しているにすぎません。

DIMM 表の最終列は、相対的な価格差を示しています。2014 年 5 月現在の PRIMERGY RX4770 M1 の料金表をベースにしています。ここでは 8 GB の RDIMM を基準とし（1.0 として強調表示）、GB あたりの価格比を示します。メモリ容量が非常に大きい LRDIMM では、特に同等の新しい 64 GB LRDIMM の場合、コストがより高くなります。さらに、相対価格の状況は絶えず変化しています。そのため、この表は一時的なものであると解釈してください。

PRIMEQUEST または PRIMERGY モデルによっては、一部の DIMM タイプを利用できない場合があります。常に最新のコンフィギュレータを参照してください。また、販売地域によっても、利用できない DIMM タイプがあります。

ファームウェアと BIOS パラメーター

このセクションで説明するパラメーターは、Ivy Bridge-EX プロセッサの機能の結果であり、基本的に PRIMEQUEST 2000 シリーズと PRIMERGY RX4770 M1 は同じです。ただし、ファームウェアメニューと BIOS メニューでは、命名、デフォルトの割り当てと配置に違いがあります。これは、サーバクラスのさまざまな機能の要望によるものです。

構文の詳細に進む前に、ここで取り上げる影響を与える要因の概要を説明します。

- 高パフォーマンスの独立メモリチャネル（パフォーマンスモードまたは独立モード）またはフェイルセーフのロックステップモード（ロックステップモードまたは通常モード）の選択。
- RAS 機能のミラーリングまたはスペアリングの有効化。ここで、PRIMEQUEST 2000 シリーズと PRIMERGY RX4770 M1 の違いは、ミラーリングとスペアリングが PRIMEQUEST 2000 ではロックステップモードだけで可能であるのに対して、PRIMERGY RX4770 M1 ではパフォーマンスモードでもサポートされていることです。
- DIMM の低消費電力 1.35 V 動作と 1.5 V 動作のトレードオフ。これで、高いメモリ周波数とそれに伴う高いメモリパフォーマンスを実現できます。ただし、メモリ構成の一般的な条件を説明すると、このトレードオフはすべてのメモリ構成に前提のものではなく、該当のパラメーターが無効になることがあります。したがって、DIMM タイプや使用する 64 GB LRDIMM に関係なく、3DPC 構成での 1.35 V の動作はありません。一方、1.35 V のときよりも高い周波数を実現できた場合、DIMM は 1.5 V のみで動作します。このようなより高い周波数が適切となる状況（ごく稀に発生）については、次のセクションで扱います。
- パトロールスクラブの場合、メインメモリ全体では修正可能なメモリエラーが 24 時間サイクルで検索され、必要に応じて修正が開始されます。これにより、エラーが修正不能になる可能性を低くすることができます。動作はメモリコントローラーによって制御されます。感度の高いパフォーマンス指標がある場合は、この機能を一時的に無効にすることもできます。ただし、パフォーマンスに及ぶ効果を実証するのは難しい場合があります。このパラメーターのサポートは、Ivy Bridge-EX ベースサーバでは遅れる場合があります。
- リフレッシュレートは、DRAM の基本機能に関するものです。1 つのビットの情報値 1 または 0 を決定する電荷が放電するため、すべてのメモリセルはマイクロ秒単位のサイクルで継続的にリフレッシュする必要があります。これもメモリコントローラーによって制御されます。サイクルのタイミング（リフレッシュレート）設定は BIOS が行います。管理インターフェースでのパラメーターの可用性に関する背景を説明します。稀なアクセスパターンを持つ一部のメモリタイプでは、修正可能なメモリエラーが累積して表示されることがあります。これは、パスゲート効果と呼ばれるものです。これを排除するために、そのような DIMM に対して BIOS は 2 倍のリフレッシュレートを設定します。レートを倍にすることで、パフォーマンスが 2 %程度低下します。メモリコントローラーと DIMM 間の信号線のリフレッシュ制御などによって生じる特定のオーバーヘッドがその原因です。蓄積される修正可能なメモリエラーの可能性を受け入れても、精度の高いパフォーマンス測定においてパフォーマンスの低下を受け入れられない場合、2 倍に設定したものを戻すことができます。

この導入の説明の後は、PRIMEQUEST 2000 シリーズと PRIMERGY RX4770 M1 の具体的な構文設計が続きます。PRIMEQUEST 2000 シリーズの場合、パラメーターは 2 つの異なる管理インターフェースにあります。

PRIMEQUEST 2000 シリーズの Web-GUI インターフェース

次のオプションを持つパラメーターであるメモリ動作モードは、管理ボード（MMB）の Web-GUI の［パーティション］/［パーティション#］/［モード］下（分離可能な PRIMEQUEST 2000 モデル）およびシステム（PRIMEQUEST 2800B）内にあります。

- パフォーマンスモード
- 通常モード
- 部分ミラーモード
- 完全ミラーモード
- スペアモード

製品起動時のデフォルトには下線が引いてあります。オペレーティングシステムでは、構成された物理メインメモリ全体を通常モードとパフォーマンスモードで利用できます。通常モードとは、より高度な RAS 機能を持つメモリチャネルのロックステップ動作モードのことを意味します。パフォーマンスモードとは、独立したメモリチャネルの高性能動作モードのことを意味します。

オペレーティングシステムでは、構成したメモリ容量の一部のみ（完全ミラーの 50 %など）を、部分ミラー、完全ミラー、スペアの 3 つの冗長モードとともに利用できます。スペアリングの場合、実質容量のパーセンテージは DIMM タイプによって決まります。8 GB の RDIMM が使用されている場合、実質容量は構成された容量の 3 分の 2 となり、その他の DIMM タイプの場合は 6 分の 5 になります。常に 3DPC で構成するスペアモードの仕様もこの計算に含まれます。

3 つの冗長モードは、メモリチャネルのロックステップ動作モードに基づいています。これらはロックステップモードに追加されたモードです。パフォーマンスモードの独立したメモリチャネルに関連するミラーリングおよびスペアリングはありません。

PRIMEQUEST 2000 シリーズのデバイスマネージャーのインターフェース

BIOS には（具体的には［デバイスマネージャー］/［メモリ構成］下に）、その他のパラメーターがあります。このインターフェースには、パーティションまたはシステムのコンソールを介してアクセスできます。次のオプションを持つ 3 つのパラメーターがあります。繰り返しますが、製品起動時のデフォルト値には下線が引いてあります。

- DIMM 速度：パフォーマンスモード／通常モード
- パトロールスクラブ：無効／有効
- リフレッシュレート：1x／自動

1 つ目のパラメーターは、DIMM のエネルギー消費量とのトレードオフに関連するものです。同じ名前のメモリ動作モードとは関連性がありません。ここでの通常モード設定は、DIMM の 1.35 V の省電力動作（可能な場合）を意味します。パフォーマンスモード設定は、より高いメモリ速度が有効になる場合、1.5 V での動作のことです。詳細は次のセクションで説明します。

2 つ目と 3 つ目のパラメータについては説明済みです。

PRIMERGY RX4770 M1 の BIOS のインターフェース

PRIMERGY RX4770 M1 には、BIOS の Advanced の下 に、次のパラメーターに関するメモリ設定サブメニューがあります。

- DDR パフォーマンス：パフォーマンスに最適化／低電圧に最適化／電力に最適化
- メモリモード：通常／ミラー／スペアリング
- DRAM メンテナンス：自動／手動
- リフレッシュレートマルチプライヤー：無効／有効
- メモリ RAS ポリシーをグローバルに適用：無効／有効
- VMSE ロックステップモード：独立／ロックステップ

ここでも、製品起動時のデフォルト値には下線が引いてあります。

最初のパラメーター DDR パフォーマンスは、発生した場合に 1.35 V での動作と 1.5 V での動作のトレードオフに関係します。繰り返しますが、詳細については次のセクションを参照してください。3 つ目のオプション 電力に最適化では、メモリ周波数が Ivy Bridge-EX で最小の 1066 MHz に低下します。ただし、1.35 V での動作は、メモリ周波数自体ほどではないにしても、DIMM のエネルギー消費量に決定的な影響を与えることを指摘しておく必要があります。低電圧に最適化の設定を超えて省エネが発生するかどうかは不明です。

2 つ目のパラメーターメモリモードは、RAS 機能のミラーリングとスペアリングの有効化に関係するものです。ミラーリング設定には追加のサブ項目があり、この項目では個々のメモリコントローラーレベルで有効化ができます。

リフレッシュレートマルチプライヤーパラメーターは、上で説明したように無効に設定するとリフレッシュレートが 2 倍になることを取り消しますが、「DRAM メンテナンス = 手動」の場合にのみ表示されます。パトロールスクラブのサブパラメーターも、「DRAM メンテナンス = 手動」の場合に表示されます。

DRAM メンテナンス は、リフレッシュレートとパトロールサブクラブのパラメーターが適切な判断でのみ変更される指標として理解しておくべきです。

「メモリ RAS ポリシーをグローバルに適用」は、RAS の機能をシステム全体で有効にするのか、またはまったく無効にするのかを指定するときに使用できます。

最後のパラメーター VMSE ロックステップモードは、独立したメモリチャネル（Independent）とロックステップモードの選択に関係します。PRIMERGY RX4770 M1 の場合のロックステップモードは、RAS モードのミラーリングとスペアリングの任意の有効化から独立しています。

メモリ周波数の定義

構成の有効なメモリ周波数（メモリパフォーマンスに関する主要なパラメーター）は、一般的な条件の範囲で決まります。Ivy Bridge-EX 搭載サーバには、1600、1333、1066 MHz の 3 つの値が適しています。システムまたはパーティションに電源が入ると、周波数が BIOS によって定義され、プロセッサごとではなく、システムまたはパーティションごとに適用されます。

前述の全般的な条件には、前のセクションで取り上げた BIOS 設定の一部が含まれます。PRIMEQUEST 2000 シリーズと PRIMERGY RX4770 M1 の構文の違いのためと、わかりやすくするために、意味的なレベルでのみ説明をします。このセクションでは、メモリチャネルの動作モードの範囲で、ロックステップと独立、および低電圧（LV）とパフォーマンス間のエネルギー効率のトレードオフについて、区別されています。この場合、PRIMERGY RX4770 M1 には 3 つ目の設定である「Energy Efficient」は意味的な違いだけであり、さらなる依存関係なしに、最小周波数 1066 MHz への低下をもたらし、したがって、以下のようなケースバイケースの分析では考慮する必要がなくなります。

まず、構成されたプロセッサーモデルはメモリ周波数の定義で重要になります。本書では、Ivy Bridge-EX モデルを次の表に従って分類することをお勧めします。

CPU タイプ	QPI (GT/s)	チャネル動作モードでの最大メモリ周波数（MHz）		Xeon E7-8800 v2 モデル	Xeon E7-4800 v2 モデル
		独立	ロックステップ		
Advanced	8.0	1333	1600	E7-8890 v2 E7-8880 v2 E7-8870 v2 E7-8893 v2 E7-8857 v2	E7-4890 v2 E7-4880 v2 E7-4870 v2
Standard	7.2	1066	1333	E7-8850 v2	E7-4850 v2 E7-4830 v2 E7-4820 v2

このセクションのその他の表では、これら 2 つの CPU クラスが区別されています。メモリ周波数に与える別の影響は、メモリチャネルの動作モードです。独立モードとロックステップモードの場合で区別する必要があります。この区別の理由は前述したとおりです。つまり、8 個の独立したメモリチャネルの最大周波数を最大限に活用するために、SMI Gen2 リンクの周波数速度は、DDR3 チャネルの 2 倍になります。SMI Gen2 周波数には上限があり、DDR3 周波数に対して遡及効力を持ちます。

チャネル動作モードに関する区別では、BIOS パラメーターによる区別も行い、エネルギー消費量とのトレードオフを考慮する必要があります。割り当てパフォーマンスモードの表のフィールドに数値が入力されてさえいれば（実際には低電圧モードの表の値と差異がありますが）、パラメーターの有効性はより明確になります。空の構成がある場合、そのパラメーターは無効となります。つまり、メモリ周波数と DIMM 電圧の両方の値が、低電圧設定の場合と同じ値になります。

プロセッサモデルと、チャネル動作モードおよびエネルギーのトレードオフに関連するパラメーターを扱った後に、メモリ周波数に影響を与えるものとして最後に考慮すべき内容は、DIMM タイプと DPC 値です。そこで、特定の構成における有効なメモリ周波数と DIMM 電圧を次に示します。

ロックステップチャネル動作モード

	エネルギーのトレードオフ：低電圧 グレーの網掛け：1.5V – 網掛けなし：1.35V								
	RDIMM			32 GB LRDIMM			64 GB LRDIMM		
CPU タイプ	1DPC	2DPC	3DPC	1DPC	2DPC	3DPC	1DPC	2DPC	3DPC
Advanced	1333	1066	1066	1333	1333	1333	1066	1066	1066
Standard	1333	1066	1066	1333	1333	1333	1066	1066	1066

	エネルギーのトレードオフ：パフォーマンス [1] グレーの網掛け：1.5V – 網掛けなし：1.35V								
	RDIMM			32 GB LRDIMM			64 GB LRDIMM		
CPU タイプ	1DPC	2DPC	3DPC	1DPC	2DPC	3DPC	1DPC	2DPC	3DPC
Advanced		1333		1600	1600				
Standard		1333							

[1] フィールドが空欄の場合は無効

独立チャネル動作モード

	エネルギーのトレードオフ：低電圧 グレーの網掛け：1.5V – 網掛けなし：1.35V								
	RDIMM			32 GB LRDIMM			64 GB LRDIMM		
CPU タイプ	1DPC	2DPC	3DPC	1DPC	2DPC	3DPC	1DPC	2DPC	3DPC
Advanced	1333	1066	1066	1333	1333	1333	1066	1066	1066
Standard	1066	1066	1066	1066	1066	1066	1066	1066	1066

	エネルギーのトレードオフ：パフォーマンス [1] グレーの網掛け：1.5V – 網掛けなし：1.35V								
	RDIMM			32 GB LRDIMM			64 GB LRDIMM		
CPU タイプ	1DPC	2DPC	3DPC	1DPC	2DPC	3DPC	1DPC	2DPC	3DPC
Advanced		1333							
Standard									

[1] フィールドが空欄の場合は無効

メモリ周波数がメモリパフォーマンスの主要なパラメーターであることは最初に言及しました。考え得る最高の周波数 1600 MHz は、ロックステップチャネル動作モードで達成されます。それにも関わらず、もう一方の動作モードはパフォーマンスモードまたは独立モードと呼ばれています。これはどのように正当化できるでしょうか？

周波数レート 1066、1333、1600 MHz のパーセンテージの違いはレベルあたり 20 - 25 %であり、GB/s で表現されるメモリ帯域幅に置き換えると、4 分の 3 以上の違いになります。このような規模であれば、主要なパラメーターと言えるでしょう。

しかし、同じメモリ周波数でも、8 つの独立したメモリチャネルと、ロックステップモードでの 4 つの論理チャネルの帯域幅には約 80 %の違いがあります。独立メモリチャネルでは、ロックステップ動作が多くの場合高いメモリ周波数を低減する可能性があるものの、等しくはならない帯域幅の利点が有効になります。さらに、チャネル動作モードとしての適合性は常に与えられますが、このセクションで説明したように、より高いメモリ周波数のオプションは全般的な条件によって決まります。

パーセンテージの割合をメモリ帯域幅からアプリケーションパフォーマンスへ 1 対 1 で置き換えられないことは言うまでもありません。後者への影響は著しく低くなります。しかし、前者の場合は、影響を与える要因は同じです。

理想的なメモリ容量

ここまでに、Ivy Bridge-EX 搭載サーバのメモリパフォーマンスに与える主な 2 つの影響について説明してきました。1 つ目は、RAS（ロックステップ）と、メモリ動作モードのパラメーターが制御するパフォーマンスとのトレードオフです。2 つ目は、メモリ周波数に影響を与える依存関係の範囲です。そこでは、ファームウェアの影響と細かい調整、そして、それに影響を与える BIOS について取り上げました。パフォーマンスにおけるそれぞれのパーセンテージの違いは、本書の 2 部で扱います。

3 つ目の主な影響は、構成される DIMM の数です。これは、必要なメモリ容量に直接関係します。最小構成（プロセッサごとに DIMM 2 枚）と最大構成（プロセッサごとに DIMM 24 枚）について、はすでに説明しました。最小構成と最大構成は、メモリアーキテクチャーを最適に使用するための理想的なメモリ構成範囲を示しています。理想的なメモリ構成を行うには、プロセッサごとに 8 枚、16 枚、または 24 枚の DIMM が必要です。この構成を次の表に示します。PRIMERGY RX4770 M1 では、プロセッサごとに 2 つのメモリボードが構成されていることが必要です。

Ivy Bridge-EX 搭載システムのさまざまな CPU 構成での理想的なメモリ容量									ベンチマーク
2 CPU の GB	4 CPU の GB	8 CPU の GB	DPC	DIMM タイプ（CPU と DPC ごとに DIMM 8 枚）	メモリ動作 = 通常		メモリ動作 = パフォーマンス		
					MHz 1.35V	MHz 1.5V	MHz 1.35V	MHz 1.5V	
128	256	512	1	8GB RDIMM	1333		1333		
256	512	1024	2	8GB RDIMM	1066	1333	1066	1333	+
			1	16GB RDIMM	1333		1333		
384	768	1536	3	8GB RDIMM		1066		1066	
512	1024	2048	2	16GB RDIMM	1066	1333	1066	1333	+
			1	32GB LRDIMM	1333	1600	1333		
768	1536	3072	3	16GB RDIMM		1066		1066	
1024	2048	4096	2	32GB LRDIMM	1333	1600	1333		+
			1	64GB LRDIMM		1066		1066	
1536	3072	6144	3	32GB LRDIMM		1333		1333	
2048	4096	8192	2	64GB LRDIMM		1066		1066	
3072	6144	12288	3	64GB LRDIMM		1066		1066	

これらの構成では、各プロセッサにある 8 つのメモリチャネルが等しく扱われます。これは、メモリシステムに生じる負荷を理想的に分配または並列化できる決定的な機能です。表に示した構成では、メモリコントローラー、SMI Gen2 リンク、Jordan Creek メモリバッファー、DDR3 チャネルなどの既存のメモリリソースが未使用のままになることはありません。同時に、すべてのメモリチャネルに統一性があるため、すべてのアルゴリズムが都合よく「均等に動作」し、メモリコントローラーのマイクロコードのメモリアクセスが並列化されます。これは技術用語でインターリーブと呼ばれます。ここでその詳細を説明します。

この表は、システムまたはパーティションの GB 総容量でソートされています。構成レベルの値は、すべてのプロセッサが等しく構成されていることを前提に、各行で 2 基、4 基、または 8 基のプロセッサに指定されています。この前提については、本書の「はじめに」で、強力なシステムのメモリ構成の基本的なルールとして言及しました。技術的背景は、NUMA システムアーキテクチャーのローカルメモリアクセスとリモートメモリアクセス間の違いです。実際の経験では、残念ながら、このルールは当然のこととみなされていません。

プロセッサのすべてのメモリチャネルを均等に扱うと、8 枚の DIMM でグループで構成が完了します。チャネルごとに 3 つの DIMM スロットがあるため、プロセッサごとに、1 つ、2 つ、または 3 つのそうしたグループに接続できます。これは、構成の DPC（DIMMs per channel：チャネルあたりの DIMM 数）値と呼ばれます。そのため、表に示した総容量は、次の式で計算されています。

容量（GB）= 8 メモリチャネル × DPC × DIMM サイズ（GB）× CPU の数

表には各構成の可能なメモリ周波数が示されていますが、ここでは、前述した区別を考慮に入れる必要があります。これらのメモリ構成では、RAS（ロックステップ）とパフォーマンス間のトレードオフや、エネルギー消費量とパフォーマンス間のトレードオフがどのように決定されたかに関わらず、最適なチャネルインターリーブ機能を使用できます。このようなトレードオフの決定がパフォーマンスに悪影響を及ぼすものであっても、この構成では、可能な限り最適なインターリーブを実現する機能を維持できます。さらに、実稼動環境では、基本方針として、是が非でも最高のパフォーマンスを実現するよりも、バランスの取れたメモリパフォーマンスを実現する方が明らかに価値があります。本書の 2 部に属する以下の定量的影響についての説明は、これらの影響を相互に調整する際に役立ちます。

このような検討事項は、エネルギー効率指標に関わるベンチマークにも存在します。データベースのベンチマークの場合は、非常に大きなメモリ容量で I/O 処理を減らしても、得られるメモリ周波数が低いという問題があります。

PRIMEQUEST 2000 シリーズおよび PRIMERGY RX4770 M1 の標準のベンチマークで使用されるメモリ構成も、言うまでもなく、この表の最適な構成の中にあります。最後の列で + 記号でマークされているものがそれに該当します。実際にはコスト上の理由から、メモリ構成はサポートされている容量スケールの最下位にあることが多いため、表にある最小構成が精度の高いパフォーマンス測定で避けられる理由を強調する必要があるようです。この構成では、メモリチャネルで 8 GB RDIMM のみがシングルランクの設計のため、パフォーマンスが数パーセント低下します。それには以下に示す理由があります。これは通常、実稼働環境で機能するものではありません。しかし、このようなパフォーマンスの低下は、ベンチマークでも、特別なパフォーマンスが期待される場合でも、望まれるものではません。

メモリパフォーマンスに対する定量的影響

メモリシステムの機能とその定性的情報を説明した後は、メモリ構成の違いがパフォーマンスに与える影響を、パーセンテージベースで説明します。その準備として、最初のセクションでは、メモリパフォーマンスの特徴を表すために使用する 2 つのベンチマーク（STREAM および SPECint_rate_base2006）について説明します。後者のベンチマークは、商用アプリケーションパフォーマンスのモデルとして機能します。

続くセクションでは、メモリコントローラーとメモリチャネルへのインターリーブについて説明します。ここでは、独立チャネル動作モードとロックステップチャネル動作モードの違いもトピックとして扱います。その後に続くセクションでは、ランクとメモリ周波数でのインターリーブについて説明します。ミラーリングやスペアリングなど、冗長性を考慮する場合のメモリパフォーマンスについてのセクションは、本書の最後にあります。個々の機能をテストする際には、影響を混同しないように、その他の機能を非表示にしています。

測定の構成を次の表に示します。PRIMEQUEST 2000 シリーズでは、それぞれ 2 個のプロセッサが搭載された 1 つおよび 4 つのシステムボードで構成されるパーティションでテストが実施されます。結果はパーティションサイズに大幅に依存するものではなかったため、以降のセクションでは、この点の差異を省略しました。

SUT（System Under Test：テスト対象システム）		
ハードウェア		
モデル	PRIMEQUEST 2800E	PRIMERGY RX4770 M1
CPU 種類	Xeon E7-8890 v2	Xeon E7-4890 v2
メモリタイプ	16GB (2x8GB) 1Rx4 L DDR3-1600 R ECC 32GB (2x16GB) 2Rx4 L DDR3-1600 R ECC 64GB (2x32GB) 4Rx4 L DDR3-1600 LR ECC	32GB (2x16GB) 2Rx4 L DDR3-1600 R ECC 64GB (2x32GB) 4Rx4 L DDR3-1600 LR ECC
ディスクサブシステム	1 × RAID Ctrl SAS 6G 1GB 1 × HD SAS 6G 300 GB 15K HOT PL 2.5” EP	1 × RAID Ctrl SAS 6G 1GB 1 × HD SAS 6G 300 GB 15K HOT PL 2.5” EP
ソフトウェア		
ファームウェア	統合ファームウェア 14012 （BIOS、BMC、MMB）	BIOS R1.3.0、BMC 7.24F
オペレーティングシステム	Red Hat Enterprise Linux Server release 6.5	Red Hat Enterprise Linux Server release 6.5

以降の表では、常に相対的なパフォーマンスが示されます。理想的なメモリ条件下での STREAM および SPECint_rate_base2006 のベンチマークの絶対測定値は、通常、表では 100 %の値に相当します。この値については、さまざまなプロセッサモデルの観点からさらに差別化した内容が、PRIMEQUEST 2800E のパフォーマンスレポート［関連資料 6］および PRIMERGY RX4770 M1 のパフォーマンスレポート［関連資料 7］に記載されています。

メモリパフォーマンスのテストには、最も強力なプロセッサモデルである Xeon E7-8890 v2 と Xeon E7-4890 v2 を使用します。これにより、パフォーマンスの違いを最も明確に把握することができます。パワーの低いプロセッサでは、パフォーマンスの違いが少しわかりづらくなるため、こうした構成にパーセンテージベースでその内容を転記する際には、そのことを考慮に入れる必要があります。

通常、ベンチマークの測定は、システム使用率を 100 %に近い状態で行うことが特徴的です（STREAM および SPECint_rate_base2006 がこれに該当します）。これは実稼働環境において一般的なことではありません。パーセンテージベースでシステムを評価する際には、この緩和要因も考慮に入れる必要があります。ただし、使用率を考慮する際には、簡単な式はありません。

測定ツール

測定は、STREAM および SPECint_rate_base2006 ベンチマークを使用して行いました。

STREAM ベンチマーク

STREAM ベンチマーク（開発者：John McCalpin 氏）［関連資料 4］は、メモリのスループットを測定するツールです。このベンチマークは、double 型データの大規模な配列でコピーおよび算術演算を実行して、Copy、Scale、Add、Triad の 4 種類のアクセスの結果を提供します。Copy 以外のアクセスタイプには、算術演算が含まれています。結果は、常に GB/s 単位のスループットで示されます。一般に、Triad の値が最もよく引用されます。以下で使用されるメモリパフォーマンスを定量化する STREAM のすべての測定値は、この手法に基づいていて、アクセスタイプ Triad での値です。

STREAM は、サーバのメモリ帯域幅を測定するための業界標準で、シンプルな方法を使用してメモリシステムに大規模な負荷を与えることができます。特にこのベンチマークは、複雑な構成でのメモリパフォーマンスに対する影響を調査する場合に適しています。STREAM は、構成によるメモリへの影響とそれによって生じるパフォーマンスへの影響（低下または向上）を示します。後述する STREAM ベンチマークに関する値は、パフォーマンスへの影響度を示しています。

アプリケーションのパフォーマンスに対するメモリの影響は、各アクセスの遅延時間とアプリケーションが必要とする帯域幅に区別されます。帯域幅が増加すると遅延時間は増加するため、両者は関連しています。並列メモリアクセスによって遅延時間が相殺される度合いは、アプリケーションや、コンパイラーによって作成されたマシンコードの質にも依存します。このため、すべてのアプリケーションシナリオでの全般的な予測を立てることは非常に困難です。

SPECint_rate_base2006 ベンチマーク

SPECint_rate_base2006 ベンチマークは、商用アプリケーションパフォーマンスのモデルとして追加されました。これは、Standard Performance Evaluation Corporation（SPEC）の SPECcpu2006 ［関連資料 5］の一部です。SPECcpu2006 は、システムのプロセッサ、メモリおよびコンパイラーを評価するための業界標準です。大量の測定結果が公開され、販売プロジェクトおよび技術調査に使用されているため、サーバ分野で最も重要なベンチマークとなっています。

SPECcpu2006 は、大量の*整数*演算および*浮動小数点*演算を使用する独立した 2 つのテストセットで構成されています。整数演算部分は商用アプリケーションに相当し、12 種類のベンチマークから構成されます。浮動小数点演算部分は科学アプリケーションに相当し、17 種類のベンチマークで構成されます。いずれの場合も、ベンチマークの実行結果は、個々の結果の幾何平均です。

さらに、それぞれのテストセットには、単体実行時の処理性能を評価する*速度*測定と、並行処理の性能を評価する*スループット*測定があります。多数のプロセッサコアとハードウェアスレッドを持つサーバにとっては、後者が重要です。

また、測定の種類により、コンパイラーに許可される最適化が異なります。*ピーク*値の測定では、各ベンチマークを個別に最適化できますが、*ベース*値の測定では、コンパイラーフラグがすべてのベンチマークで同一である必要があり、特定の最適化は許可されません。

以上が SPECint_rate_base2006 の概要です。PRIMERGY サーバでは商用アプリケーションの使用が主流であるため、整数演算を使用するテストセットである SPECint_rate_base2006 でスループットを測定しました。

本来のルールに準拠した測定では 3 回の実行が必要であり、各ベンチマークに対して平均の結果が評価されます。しかし、ここで説明している技術調査では、このルールに準拠していません。効率化のために、測定は 1 回にしています。

インターリーブ

インターリーブとは、同じタイプの複数のメモリリソース間で切り替えを行う、物理アドレス空間の設定を意味します。まず、Ivy Bridge-EX 搭載サーバの場合は、プロセッサの 2 つのメモリコントローラーが適しています。ローカルアドレス空間セグメントの最初のブロックは最初のコントローラーで使用し、2 番目のブロックは 2 番目のコントローラーで使用し、3 番目のブロックは最初のコントローラーに戻って使用するという具合に続いていきます。この原則は、コントローラーあたり 4 つのメモリチャネルのレベルにも引き継がれ、最終的に個々のメモリチャネル内のランクのレベルにも引き継がれます。

それぞれのリソースのメモリ容量が同一であることが、このパターンの決定的な前提条件です。切り替え作業はその条件が満たされている場合のみ実行されます。この条件が満たされていない場合の手順については、次のセクションで説明します。このパターンでは、切り替えを行うために、ブロックサイズに一定の柔軟性が必要になります。ブロックサイズは、コントローラーとチャネルのレベルでも、ランクのレベルでも同一ではありません。

メモリアクセスは、局所性原理より主に隣接するメモリ領域に行われ、インターリーブの結果、メモリシステムのすべてのリソースに分散されます。このようなパフォーマンスの向上は、並列化によるものです。多くのメモリコントローラーやメモリチャネルで現在見られるインターリーブは、メモリ周波数よりも、メモリパフォーマンスに最も重要な影響を与える可能性があります。

メモリコントローラーとメモリチャネルへのインターリーブ

前述したように、理想的なメモリ容量は、プロセッサごとに 8 枚、16 枚、または 24 枚の同じタイプの DIMM で構成されます。この場合、コントローラーとチャネルへのインターリーブは、最適な効果を得て展開していきます。次の表にある別の数の DIMM を使用した構成、特に、プロセッサあたりの DIMM 数が 8 枚未満の構成から最小構成までは、パフォーマンスが低下します。インターリーブ、メモリ帯域幅、商用アプリケーションパフォーマンスの 3 つの各カテゴリの最良条件は、太字で示されています。

PRIMEQUEST 2000 シリーズのチャネルインターリーブ				
	動作モード	**CPU ごとに DIMM 8 枚（およびその倍数） 理想的な容量**	**CPU ごとに DIMM 4 枚**	**CPU ごとに DIMM 2 枚 最小構成**
インターリーブ（コントローラー／チャネル）	独立	**2-WAY／4-WAY**	2-WAY／2-WAY	1-WAY／2-WAY
	ロックステップ	2-WAY／2-WAY	2-WAY／1-WAY	1-WAY／1-WAY
メモリ帯域幅（STREAM）	独立	**100 %**	56 %	28 %
	ロックステップ	57 %	30 %	16 %
商用アプリケーションパフォーマンス（SPECint_rate_base2006）	独立	**100 %**	93 %	73 %
	ロックステップ	94 %	76 %	54 %

表の一番上の横ブロック（インターリーブ）は、さまざまな構成のインターリーブを示しています。ここでの n-WAY は、n コントローラーとチャネル間で切り替えができる構成を意味しています。この切り替えのブロックサイズは、64 バイトのプロセッサのキャッシュラインサイズに基づいています。

この時点で、メモリ動作モードである通常（ロックステップ）モードのメモリパフォーマンスに関する「問題」がどこにあるのかが分かります。この場合の切り替えは、2 つの物理チャネルがそれぞれのケースで組

み合わされる、論理メモリチャネルのレベルで行われる必要があります。64 バイトのブロックは、切り替えが不可欠な要素となる、アドレスレベルの下位レベルで 2 つの物理チャネルに分割されます。ロックステップモードを有効にすると、メモリチャネルのインターリーブは半分になります。そのため、この動作モードはパフォーマンスに影響を与えません。

表の一番下の横ブロックには、メモリ帯域幅とベンチマーク SPECint_rate_base2006 の相対的なパフォーマンス効果が示されています。このベンチマークは、商用アプリケーションパフォーマンスのモデルとしての機能を果たします。STREAM と PECint_rate_base2006 の両方のカテゴリにおける最良条件は、パフォーマンスが 100 %の場合です。その他の構成の場合は、表に示されているように、それより低い数値になります。STREAM で示されているように、メモリ帯域幅の関係は、特に HPC（High-Performance Computing：高性能コンピューティング）環境では、特定のアプリケーション領域において除外できない極端なケースとして理解する必要があります。ただしこうした動作は、ほとんどの商用のワークロードでは見られません。STREAM および SPECint_rate_base2006 に関する解釈の質は、このセクションで取り上げているパフォーマンス面だけでなく、以降のすべてのセクションにも当てはまります。

測定値を見ると、ロックステップモードを、（純粋なパフォーマンスの観点から、そして、追加された RAS 値を除外して）チャネルへのインターリーブが半分になる独立モードとして論理的に理解できます。ロックステップのテスト結果は、基本的に、1 列右にある独立の テスト結果と一致しています。

前の表は PRIMEQUEST 2000 シリーズに関するもので、それぞれ許可されるメモリ構成はロックステップ対応です。ロックステップ機能は、各 Jordan Creek 1 メモリバッファーの 2 つのメモリチャネルの対称処理から生じたものです。全般的なロックステップ機能は、PRIMERGY RX4770 M1 の許可されるメモリ構成には適用されません。さらに、プロセッサごとに注文されるメモリボード数に関して、このシステムでの差別化があります。これらのもっと複雑な構成ルールの再現は、本書では取り扱いません。したがって、PRIMERGY RX4770 M1 のコンフィギュレーターに関する知識は、次の表を理解するうえでの前提条件です。

PRIMERGY RX4770 M1 のチャネルインターリーブ					
	動作モード	**CPU あたり：2 つのメモリボード全体で 8 枚の DIMM 理想的な容量**	**CPU あたり：2 つのメモリボード全体で 4 枚の DIMM**	**CPU あたり：1 つのメモリボード全体で 4 枚の DIMM**	**CPU あたり：1 つのメモリボード全体で 2 枚の DIMM 最小構成**
インターリーブ（コントローラー／チャネル）	独立	**2-WAY／4-WAY**	2-WAY／2-WAY	1-WAY／4-WAY	1-WAY／2-WAY
	ロックステップ	2-WAY／2-WAY		1-WAY／2-WAY	
メモリ帯域幅（STREAM）	独立	**100 %**	56 %	51 %	28 %
	ロックステップ	57 %		29 %	
商用アプリケーションパフォーマンス（SPECint_rate_base2006）	独立	**100 %**	94 %	92 %	78 %
	ロックステップ	94 %		79 %	

この表は、プロセッサあたり 1 つまたは 2 つのメモリボードでのメモリ構成におけるパフォーマンスの違いを評価するとき、特に役立ちます。例えば、最適なメモリパフォーマンスは、8 枚の DIMM とプロセッサあたり 2 つのメモリボードで実現されます（左から 3 列目）。一方、8 枚の DIMM とプロセッサあたり 1 つのボードを注文した場合、達成できるチャネルインターリーブは右から 2 列目です。プロセッサあたり（4 枚ではなく）8 枚の DIMM でも、1 つのメモリボードの 4 つのメモリチャネル容量を満たしますが、その場合チャネルインターリーブの向上は見られません。

アプリケーションパフォーマンスへの効果（両方の表の SPECint_rate_base2006 の横ブロックを参照）に関する簡単な評価を以下に示します。ベンチマークは常に 品質 100 %の構成を目標としています。90 %を超えるケースは実稼働環境では重大な状態ではありません。通常は、システム使用率に関するセキュリティの相違によるものです。70 %周辺のケースは、仮想化環境で高い使用レベルを目標としている場合などに重大な状態となります。50 %を少し超えたケースの場合は、プロセッサの演算処理パフォーマンスとメモリパフォーマンスとの間に不一致があることが想定できます。

表には、プロセッサごとに 6 枚の DIMM を使用する場合と、DIMM の数が 8 の倍数ではないときの 8 枚を超える DIMM を使用する場合について、許可される構成が示されていません。これらのすべてのケースでは、当該リソースの一部の容量が同一ではないため、切り替えが機能しません。プロセッサごとに 6 枚の DIMM を使用する場合は、1 つ目のコントローラーに 4 枚、2 つ目のコントローラーに 2 枚という配分になります。この場合、切り替えパターンが同じである同種のローカルアドレス空間セグメント（まさにパフォーマンス品質を確認できる場所）は、コントローラーレベルの容量に相違があるため形成されません。その一方で、プロセッサごとに 12 枚の DIMM を使用する場合は、コントローラーに 6 枚づつ均等に配分されますが、コントローラーあたり 4 つのチャネルでは不均衡になります。

この問題は常に、物理アドレス空間を異なるインターリーブのいくつかのセグメントに分割することで解決されます。アプリケーションのパフォーマンスは、アプリケーションにメモリが提供されるセグメントによって異なる可能性があります。6 枚と 12 枚のどちらの DIMM ケースも、この表の 4 枚の DIMM ケースに相当するメモリパフォーマンスとなります。2 枚の DIMM を使用するケースも、（プロセッサあたり 10 枚の DIMM の場合のように）多くの状況で除外できないケースとなります。性能を重視するアプリケーションの場合、この動作は、こうした構成を避ける理由の 1 つになり得ます。

ランクでのインターリーブ

物理アドレス空間のセットアップ時にメモリリソースを切り替える方法は、コントローラーとチャネルでのインターリーブからチャネルのランクでのインターリーブまで継続できます。

ランクのインターリーブは、アドレスビットにより制御されます。この理由から、2 のべき乗でのインターリーブのみが問題となります。つまり、2-WAY、4-WAY または 8-WAY のランクインターリーブのみが存在します。メモリチャネルでの奇数のランク数は、1-WAY インターリーブとなりますが、これは分類上そのように呼ばれているだけです。1-WAY の場合、ランクは次のランクに変更される前にすべて利用されます。

ランクインターリーブの粒度は、前述したコントローラーとチャネルでのインターリーブよりも大きくなります。チャネルでのインターリーブは 64 バイトキャッシュラインサイズに使用されています。ランクインターリーブは、オペレーティングシステムの 4 KB ページサイズに向かい、DRAM メモリの物理特性に関係します。メモリセルは、大まかに言って 2 つの次元で行われます。行（ページとも呼ばれる）が開かれ、列項目が読み取られます。ページが開いている間、より大幅に低いレイテンシで他の列の値を読み取ることもできます。さらに大まかなランクインターリーブは、この機能に最適化されます。

メモリチャネルあたりのランク数は、構成の DIMM タイプおよび DPC 値に従います。

表には、理想的なメモリ容量グループの DIMM 構成（プロセッサごとに 8 の倍数の DIMM を使用）を例とした、ランクインターリーブの影響が示されています。表に示されたすべての構成でメモリ周波数 1333 MHz が許可されています。つまり、異なる周波数によるパフォーマンスへの影響は表示されていません。表示されているすべての測定は 1333 MHz で実施されました。メモリチャネル動作モードの影響も表示されていません。測定値は独立モードで判定されています。

	32GB 4R LRDIMM 2DPC	16GB 2R RDIMM 2DPC	32GB 4R LRDIMM 1DPC	16GB 2R RDIMM 1DPC	8GB 1R RDIMM 2DPC	32GB 4R LRDIMM 3DPC[1]	8GB 1R RDIMM 1DPC
ランクインターリーブ	**8-way**	**4-way**		**2-way**			**1-way**
メモリ帯域幅（STREAM）	95 %	**100 %**	99 %	99 %	96 %	85 %	80 %
商用アプリケーションパフォーマンス（SPECint_rate_base2006）	99 %	**100 %**	99 %	99 %	99 %	94 %	98 %

[1] 影響を与える他の要因のランク乗数

メモリパフォーマンスへのランクインターリーブの影響は、コントローラーとチャネルでのインターリーブの場合よりもはるかに少ないものとなります。これは、実稼働環境ではほとんど無視できるベンチマークトピックです。これは微妙な差異の問題であるため、測定精度の割合がほぼ同じであれば、チャネル動作モードの区別と同様に、理想的ではないメモリ容量の区別も省くことができます。

この表には、LRDIMM を使用した 8-way インターリーブで最高のパフォーマンスを達成できない理由を示す必要があります。DPC 値が高くランクの数が多いチャネル構成で、ランクインターリーブが生み出す利点と、高いリフレッシュレートによるオーバーヘッドが生み出す欠点とのトレードオフがその理由です。

メモリ周波数

メモリコントローラーとチャネルでのインターリーブによるパフォーマンスへの影響はかなり大きく、ランクでのインターリーブによるパフォーマンスへの影響はそれよりもかなり低くなります。そして、メモリ周波数の影響は、これらの中間にあります。

次の表には、100 %で表現される最大パフォーマンスに対して相対的なパフォーマンスが、STREAM と SPECint_rate_base2006 の 2 つのベンチマークに分けて示されています。周波数の影響を判断することを目的としたこの一連の測定は、プロセッサとごに 8 枚、16 枚、または 24 枚の RDIMM と LRDIMM を使用して実施されました。当該の周波数を得るには、異なる DPC 値と DIMM タイプが必要です。透明性を高めるために、平均値を算出し、ランクインターリーブの効果などによる小さな変動を非表示にしています。PRIMEQUEST 2000 シリーズと PRIMERGY RX4770 M1 の区別は必要ありません。

	動作モード	1600 MHz	1333 MHz	1066 MHz
メモリ帯域幅（STREAM）	独立		**100 %**	84 %
	ロックステップ	66 %	57 %	47 %
商用アプリケーションパフォーマンス（SPECint_rate_base2006）	独立		**100 %**	97 %
	ロックステップ	97 %	94 %	89 %

表の説明をより良く分類するためには、さらにケースを追加して、そこで異なる周波数に関する疑問を投げかける必要があります。

RDIMM の準最適周波数が 1066 MHz であり、1600 MHz でない理由は 2 つあります。どちらもチャネル動作モードの独立およびロックステップに関係します。

- 2DPC 構成には、パフォーマンスとエネルギー消費の間にトレードオフがあります。低電圧での動作では 1066 MHz ですが、DIMM は 1333 MHz で動作します。
- RDIMMS を使用した 3DPC 構成では、値は常に 1066 MHz になります。1333 MHz が可能な 1DPC または 2DPC 構成と比較すると、ストレージ容量の違いを計算に入れなければパフォーマンスが低下します。ただし、大容量の場合（3DPC）は、通常、I/O レートを低くすることでパフォーマンスが向上するため、公正な比較を行うためには、この事実を考慮に入れる必要があります。

LRDIMMs にも、エネルギー効率とストレージ容量に関する同様のトレードオフが適用されます。ただし、ランク数が多いこれらのバッファード DIMM がメモリチャネルに課す負荷が異なるため、詳細は異なります。

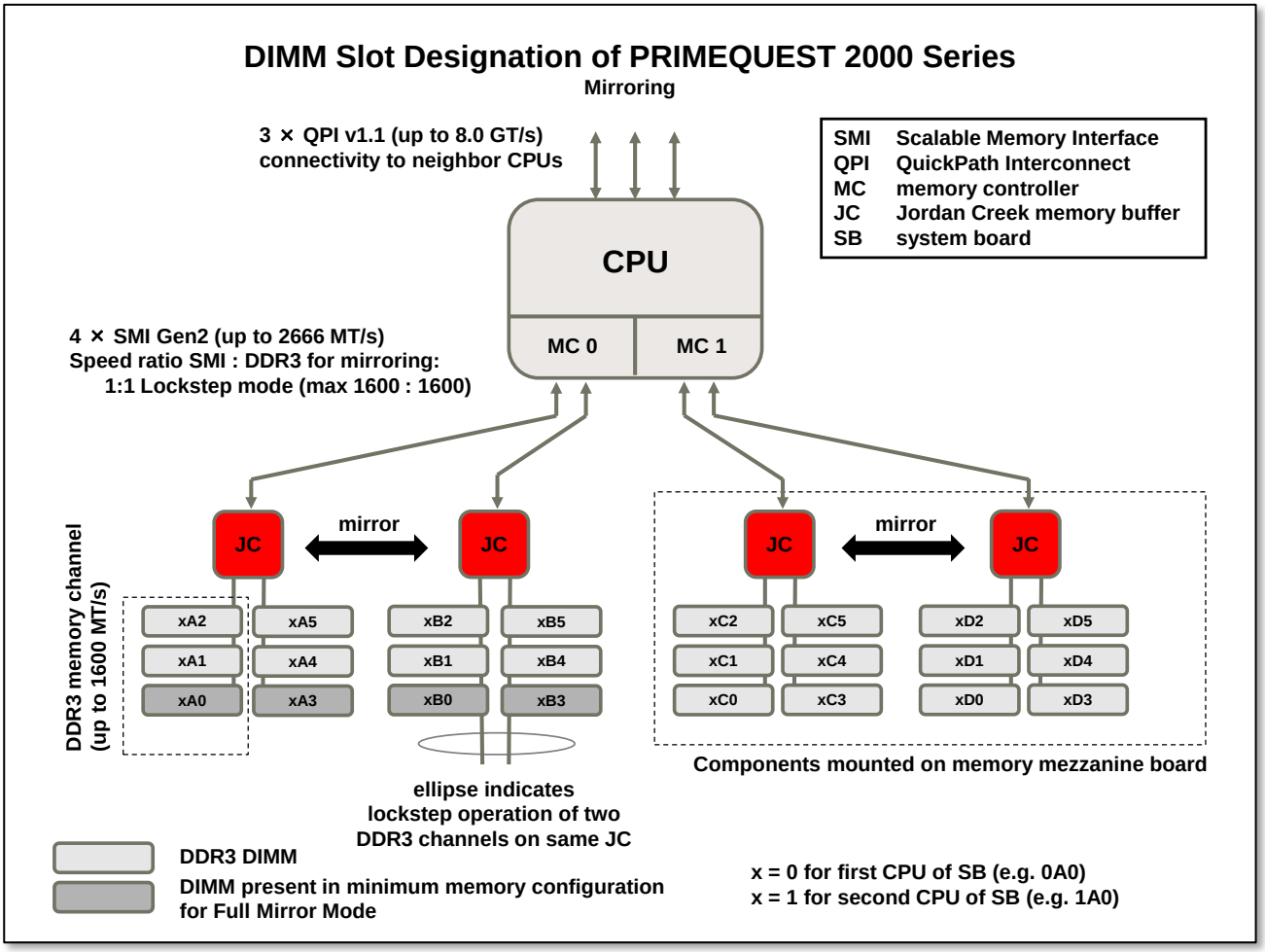
- ロックステップ動作モードの場合に限り、1DPC と 2DPC には、低電圧（1.35 V、1333 MHz）と 1.5 V（1600 MHz）の動作の間にトレードオフがあります。この表で、ロックステップのパフォーマンスが独立メモリチャネルの場合よりも低下することは、周波数を 1600 MHz にすることで修正できますが、この問題を完全には解決できないことがわかります。
- 可能な最大メモリ容量として 64 GB の LRDIMM を使用した場合、値は常に 1066 MHz になります。それに対して 32 GB の LRDIMM を使用した場合は、値は 1333 MHz または 1600 MHz までにもなります。そのため、1066 MHz によって生じるパフォーマンスの低下と、容量を大きくすることの利点を、再び比較検討する必要があります。

冗長性を考慮した際のメモリパフォーマンス

最後に、冗長性の下でのメモリパフォーマンス、つまり、メモリのミラーリングとスペアリングについて、少し説明します。

PRIMEQUEST 2000 シリーズのフルミラーモード

ミラーリングは、2 つの Jordan Creek 1 バッファーと、バッファーあたり 2 つの DDR3 チャネルを持つメモリコントローラー内で行われます。適切なメモリを備えた 2 つ目の Jordan Creek 1 が、1 つ目の Jordan Creek 1 をミラーリングします。この目的のためには、両方の Jordan Creek 1 を均等に構成する必要があります。プロセッサの 2 つのメモリコントローラー間でのミラーリング、さらにはプロセッサの境界を超えたミラーリングは行われません。すでに紹介したブロック図に補足と変更を加えたものを以下に示します。



変更は最小構成に関するものです。メモリ動作モードが通常（ロックステップ）モードとパフォーマンスモードの場合、最小構成は、xA0 と xA3 に配置した 2 枚の DIMM で構成されます。図に示すように、完全ミラーモードでは、4 枚の DIMM で構成されます。また、この変更された最小構成は、通常（ロックステップ）モードとパフォーマンスモードの 4 枚の DIMM 構成に相当するものではありません。この場合は、xA0 と xA3 の最小構成が、パフォーマンス上の理由で xC0 と xC3 まで拡張されます。これは、2 つ目のメモリコントローラーが同様の構成であるためです。この構成は、完全ミラーモードで、最小構成後に最初の追加を行う場合にのみ可能となり、それによって、8 枚の DIMM を xA0、xA3、xB0、xB3、xC0、xC3、xD0、xD3 に配置した構成となります。

次の表には、すでに説明した通常（ロックステップ）モードおよびパフォーマンスモードと比較した場合の完全ミラーモードのパフォーマンスが示されています。測定は、メモリ周波数を一貫して 1333 MHz にした

状態で実施されました。ここに示された値は「理想的な」パフォーマンスに関連するものです。これは、メモリ動作モードがパフォーマンスモードのときに、8 枚（またはその倍数）の DIMM を構成し、メモリコントローラーとチャネルでインターリーブを最大化することによって達成されます。

	メモリ動作モード	CPU ごとに DIMM 8 枚（およびその倍数）	CPU ごとに DIMM 4 枚 [1]
メモリ帯域幅（STREAM）	パフォーマンスモード	**100 %**	56 %
	通常モード（ロックステップ）	57 %	30 %
	完全ミラーモード	37 %	21 %
商用アプリケーションパフォーマンス（SPECint_rate_base2006）	パフォーマンスモード	**100 %**	93 %
	通常モード（ロックステップ）	94 %	76 %
	完全ミラーモード	85 %	68 %

[1] DIMM は、通常（ロックステップ）モードの場合は xA0、xA3、xC0、xC3 の配置になり、完全ミラーモードの場合は、xA0、xA3、xB0、xB3 の配置になります。

この表を理解するためには、完全ミラーモードにロックステップモードを含めることが不可欠となります。RAS 機能のミラーリングは、RAS 機能のロックステップに追加されています。そのため、ミラーリングによるパフォーマンスへの影響は、メモリパフォーマンスのその他すべての側面を無視すると、完全ミラーモードと通常モードを比較した場合のみ確認される可能性があります。

PRIMERGY RX4770 M1 のミラーモード

PRIMEQUEST 2000 シリーズとは異なる DIMM 構成ルールが PRIMERGY RX4770 M1 には適用されます。したがって、繰り返しになりますが、コンフィギュレーターを参照してください。相違点の 1 つについては、すでにメモリチャネル全体でのインターリーブに関するセクションで述べています。ロックステップ対応ではない構成がある点です。その他の違いは、PRIMEQUEST 2000 シリーズと同様にミラーリングをロックステップ動作モードに追加できるだけでなく、独立動作モードにも追加できる点です。この違いは、次の表に示されています。

	動作モード	CPU あたり：2 つのメモリボード全体で 8 枚の DIMM 理想的な容量	CPU あたり：2 つのメモリボード全体で 4 枚の DIMM	CPU あたり：1 つのメモリボード全体で 4 枚の DIMM	CPU あたり：1 つのメモリボード全体で 2 枚の DIMM 最小構成
メモリ帯域幅（STREAM）	独立	**100 %**	56 %	51 %	28 %
	独立 + ミラー	69 %	35 %	35 %	17 %
	ロックステップ	57 %		29 %	
	ロックステップ + ミラー	37 %		19 %	
商用アプリケーションパフォーマンス（SPECint_rate_base2006）	独立	**100 %**	94 %	92 %	78 %
	独立 + ミラー	97 %	87 %	85 %	67 %
	ロックステップ	94 %		79 %	
	ロックステップ + ミラー	85 %		68 %	

スペアモード

スペアモードは、Ivy Bridge-EX 搭載サーバの最初の販売リリースでは見送られました。パフォーマンスに関する説明は、本書の更新版で扱います。

関連資料

PRIMERGY & PRIMEQUEST サーバ

［関連資料 1］http://jp.fujitsu.com/platform/server/

メモリパフォーマンス

［関連資料 2］このホワイトペーパー：

http://docs.ts.fujitsu.com/dl.aspx?id=8ff6579c-966c-4bce-8be0-fc7a541b4a02
http://docs.ts.fujitsu.com/dl.aspx?id=9a7ec9d5-8140-4230-972b-2a04d76e43d6
http://docs.ts.fujitsu.com/dl.aspx?id=a9489f25-465a-48d6-80c0-e726809616ea

［関連資料 3］Xeon E5-2600 v2（Ivy Bridge-EP）搭載システムのメモリパフォーマンス
http://docs.ts.fujitsu.com/dl.aspx?id=43d136df-46f6-443f-9f79-56466daddd1d

ベンチマーク

［関連資料 4］STREAM
http://www.cs.virginia.edu/stream/

［関連資料 5］SPECcpu2006
http://docs.ts.fujitsu.com/dl.aspx?id=00b0bf10-8f75-435f-bb9b-3eceb5ce0157

パフォーマンスレポート

［関連資料 6］パフォーマンスレポート PRIMEQUEST 2800E
http://docs.ts.fujitsu.com/dl.aspx?id=94da09b9-97a6-4723-adca-d59cd164bf28

［関連資料 7］パフォーマンスレポート PRIMERGY RX4770 M1
http://docs.ts.fujitsu.com/dl.aspx?id=62f6f8c4-d55e-41c5-88cd-3cc08b17256e

お問い合わせ先

富士通

Web サイト：http://jp.fujitsu.com/

PRIMERGY のパフォーマンスとベンチマーク

mailto:primergy.benchmark@ts.fujitsu.com

2014-05-16 WW JA