ホワイトペーパー
FUJITSU Server PRIMERGY & PRIMEQUEST Xeon E7 v4（Broadwell-EX）搭載システムのメモリパフォーマンス

PRIMEQUEST 2000 タイプ 3 シリーズおよび PRIMERGY RX4770 M3 の Xeon E7 v4（Broadwell-EX）搭載モデルでは、QuickPath インターコネクト（QPI）のメモリアーキテクチャーの能力により、前世代に比べてパフォーマンスが大幅に向上します。これは 4 世代のシステムで証明されています。このホワイトペーパーでは、メモリ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
1.0a
2016-07-29

目次

ドキュメントの履歴

バージョン 1.0（2016 年 6 月 30 日）

初版

バージョン 1.0a（2016 年 7 月 29 日）

軽微な訂正

はじめに

従来世代の Haswell-EX では 22 nm 製造技術を使用していたのに対し、PRIMEQUEST 2000 タイプ 3 シリーズおよび PRIMERGY RX4770 M3 に採用されている Intel Xeon E7 v4（Broadwell-EX）プロセッサでは、新たな 14 nm 製造技術を使用しています。Brickland-EX プラットフォームおよび Intel C602 チップセットについては、変更はありません。

新世代のパフォーマンスは、ほとんどの負荷シナリオで旧世代に比べて約 20 - 30 %の向上を果たしています。この成果の大きな要因は、プロセッサあたりの最大コア数が 18 から 24 に増えたことによります。従来世代の実績のある機能をベースに、さらに優れたメモリシステムとなっています。革新的なのは PRIMERGY RX4770 M3 のキャッシュコヒーレンスプロトコルです。

Broadwell-EX ベースのサーバは、省電力の 1.2 V で稼働する DDR4 メモリテクノロジーも使用しています。Haswell-EX ベースのシステムと同様、最大 1866 MHz のメモリ周波数および最大 9.6 GT/s の QPI 転送速度をサポートします。メモリパフォーマンスの最も基本的な指標である帯域幅は、PRIMEQUEST 2800E3 でおよそ 450 GB/s、および PRIMERGY RX4770 M3 で 270 GB/s を若干下回る程度となっており、それぞれの先行モデルとほぼ同じ値となっています。

革新の 1 つとして挙げられるのは、PRIMERGY RX4770 M3 のキャッシュコヒーレンスプロトコルが、過去 2 世代のデュアルソケット PRIMERGY サーバで実績のある機能の *Cluster-on-die*（COD）に対応していることです。プロトコルのこのような拡張は、NUMA の特性に非常に合った負荷に対するオプションとして使用できます。

一方、旧世代で使われていた QPI ベースのメモリアーキテクチャーの基本機能は、以下のハイエンドサーバクラスの特定の特性を含め、変更はありません。

- プロセッサあたりの DIMM スロット数は 24 で、最新のデュアルソケット PRIMERGY サーバの 2 倍です。スロットは、プロセッサごとに 8 つの DDR4 メモリチャネルで分散されます。各プロセッサには、4 つのチャネルそれぞれに 2 つの統合メモリコントローラーが装着されています。コントローラーとチャネルの間に、デュアルソケットサーバには未搭載の Jordan Creek メモリバッファーがあります。
- プロセッサとそのメモリコントローラーは、QPI リンク経由でメモリの内容を隣接プロセッサに渡し、隣接プロセッサにメモリの内容を要求します。システム内のすべてのメモリモジュールが、整合のとれたアドレス域を形成します。ただし、この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。
- これらのシステムでは引き続きディレクトリベースの QPI 1.1 キャッシュコヒーレンスプロトコルを採用しており、PRIMERGY RX4770 M3 ではそれとともに上記の COD 機能に対応しています。COD は、主にローカルメモリアクセスのスペキュレーションに適した負荷の場合は一考に値します。有効化されると、2 つのメモリコントローラーに基づいて各プロセッサに 2 つの NUMA ノードが作成されます。PRIMEQUEST 2000 のような、プロプライエタリな接続チップがないプロセッサを直接結合するシステム設計（いわゆる *グルーレス設計*）の場合、NUMA ノードの総数は 8 つに制限されます。これがつまり、PRIMEQUEST 2000 タイプ 3 では COD が利用できない理由です。
- RAS（信頼性、可用性、保守性）とパフォーマンスとのトレードオフは依然として存在しています。メモリチャネルのモードは、ロックステップモードまたはパフォーマンスモードのいずれかです。ロックステップは、それぞれの場合で 2 つのメモリチャネルの同期演算モードで、これによって RAS 機能が向上します。一方、パフォーマンスモードまたは独立モードでは、メモリチャネルは互いに独立しています。

本書では、メモリシステムにおける技術革新について説明します。また、これまでの号と同様に、強力なシステムを構成するうえで不可欠な QPI メモリアーキテクチャーの基本的な知識についても説明しています。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 DDR4 チャネルが

含まれる横方向においてです。次に、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。メモリチャネルの動作モード、DIMM のタイプと数、構成されたプロセッサモデルに応じて、1866、1600、または 1333 MHz です。

メモリのパフォーマンスに影響を与える要因を挙げ、数量化します。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

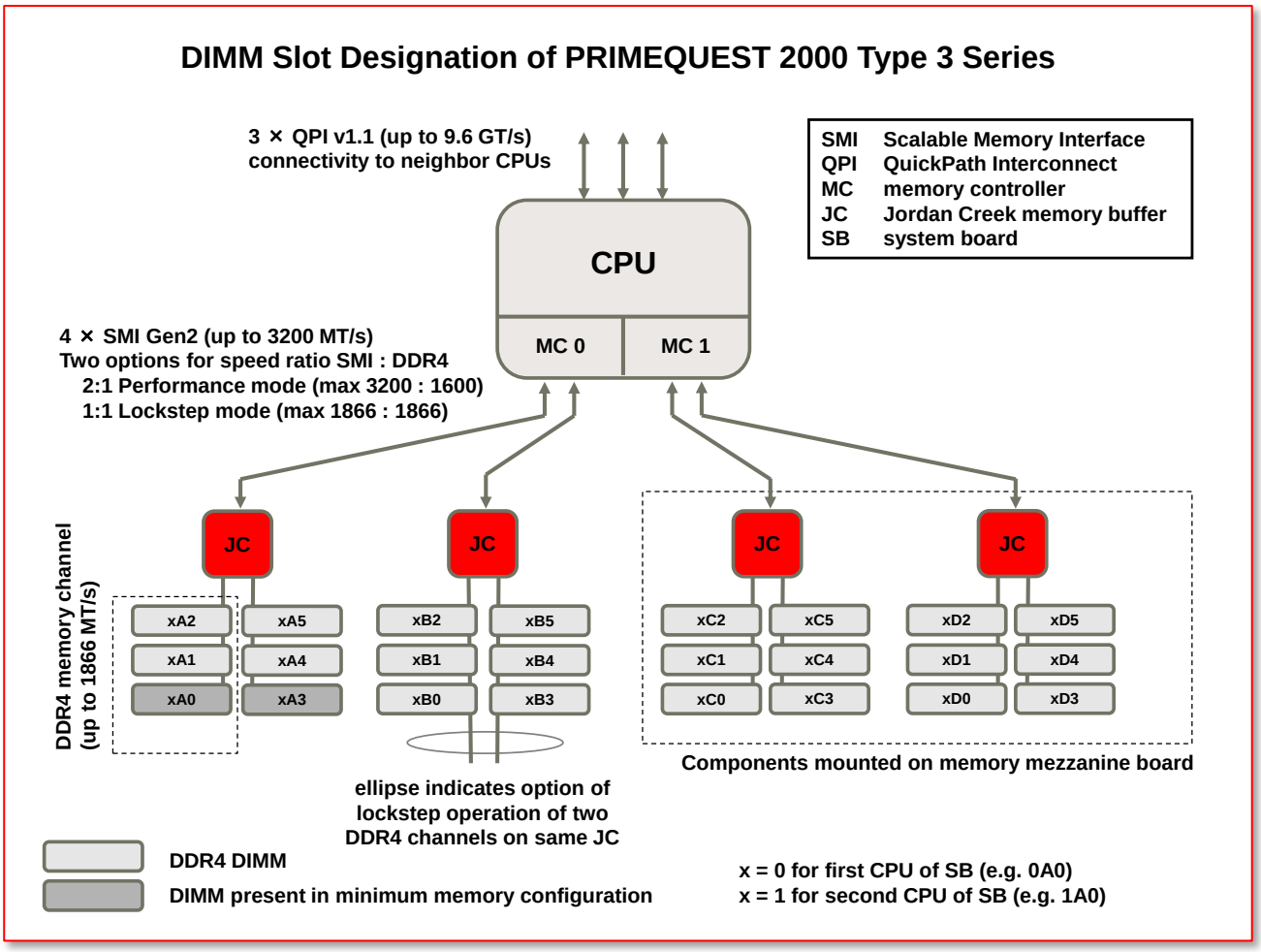
ここでは、5 部構成でメモリシステムの概要を説明します。まずブロック図で、利用可能な DIMM スロットの配置を説明します。2 つ目のセクションでは、使用可能な DIMM タイプを示します。続く 3 つ目のセクションでは、ファームウェアと、メモリシステムに影響を与える BIOS パラメーターについて説明します。4 つ目のセクションでは、有効なメモリ周波数への影響について説明します。最後のセクションには、メモリパフォーマンスに関してある程度まで「理想」を含めたメモリ構成の表を掲載しています。

DIMM スロット

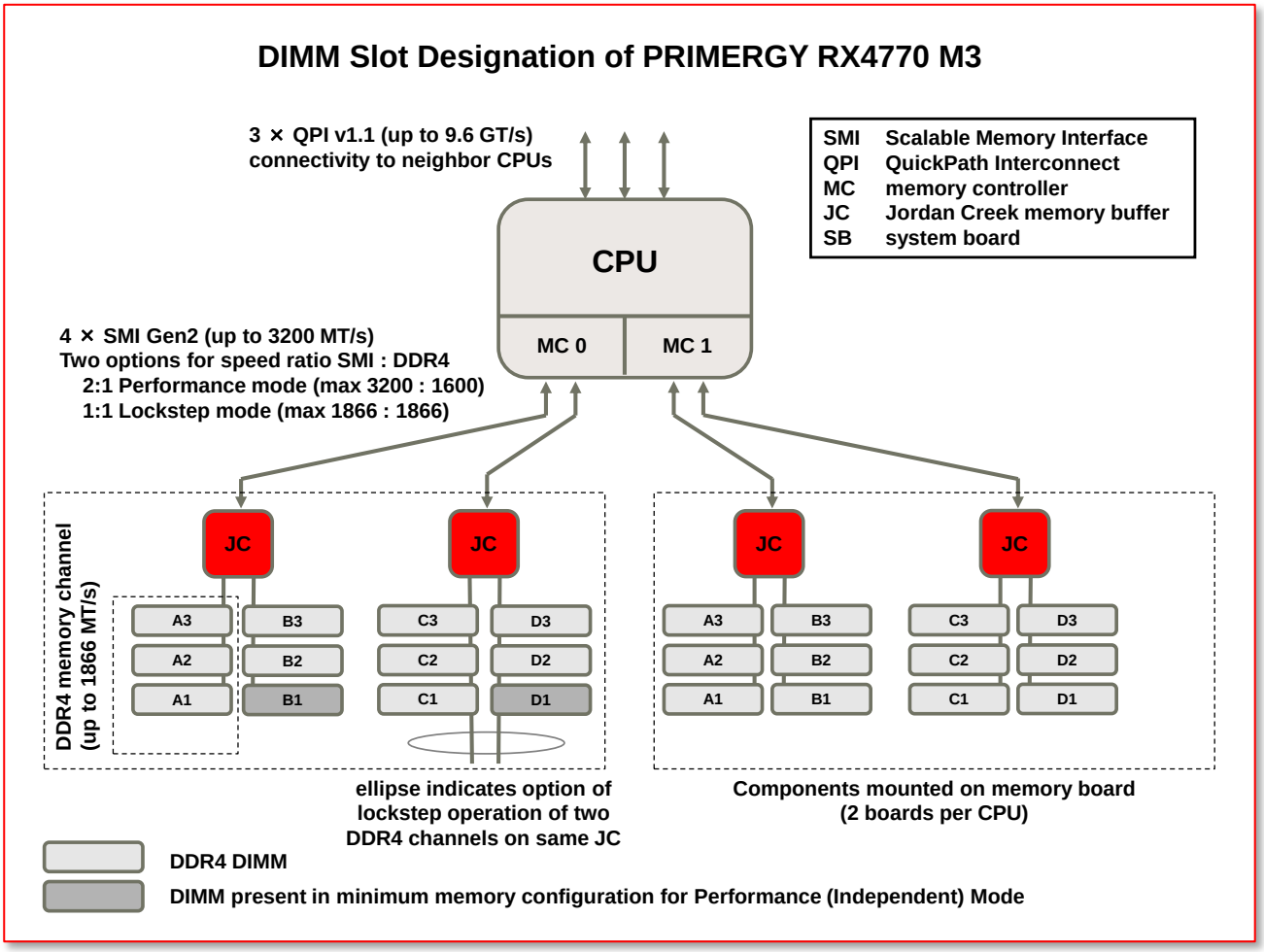
次の 2 つの図は、個々の Broadwell-EX プロセッサでのメモリ接続を示しています。各プロセッサには、2 つの統合メモリコントローラーがあります。各コントローラーは、双方向のシリアル SMI Gen2（Scalable Memory Interface）リンク経由で、2 つの Jordan Creek 2 メモリバッファーに接続されています。各メモリバッファーには、DIMM スロットが 3 つずつ付いた DDR4 メモリチャネルが 2 つ接続されています。したがって、プロセッサあたり合計 24 本の DIMM スロットが装備されています。

チャネルごとに構成された DIMM の数は、構成の DPC（DIMMs per channel：チャネルあたりの DIMM 数）値と呼ばれます。この値は、パフォーマンスに一定の影響を与えます。チャネルが同等に構成されていない場合、最大 DPC 値がシステム全体を決定します。

PRIMEQUEST 2800E3 などの PRIMEQUEST 2000 タイプ 3 シリーズのシステムは、それぞれのケースで 2 つのプロセッサとメモリリソースが搭載されたシステムボードをベースにしています。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。各プロセッサの 24 スロットの半分がシステムボード上にあります。残りの半分は、組み込ま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 M3 の 1 つのシステムボード上にあります。DIMM スロットそれぞれに 12 個のスロットがついて、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 DDR4 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その都度 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、各メモリアクセスが両方のチャネル経由で同時に行われます。つまり、読み取られるまたは書き込まれるブロックは、2 つのチャネルに分割されます。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。これにより、並列化される容量が制限され、そのためにメモリアクセスのパフォーマンスも制限されます。Broadwell-EX は、この動作モードがオプションとなっている第 3 世代です。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。これに対し、旧世代の Nehalem-EX と Westmere-EX のシステムは、常にロックステップモードでした。

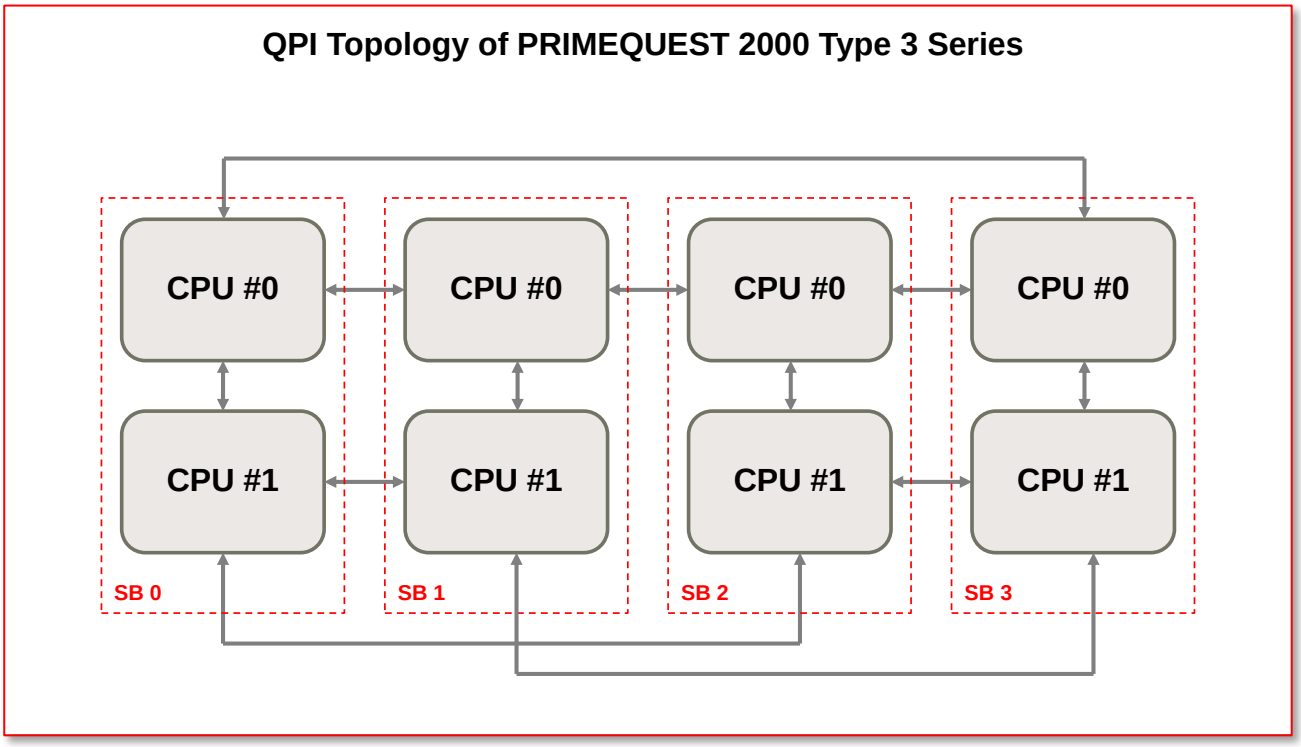
動作モードの適合性は、リソース SMI Gen2 リンクと DDR4 チャネルの周波数に影響を与えます。8 つのチャネルが対する SMI Gen2 リンクが 4 つだけのため、パフォーマンスモードでは、最大メモリ帯域幅を実装するリンクは、メモリチャネルの 2 倍の速度となります。一方、ロックステップモードでの周波数は同じで

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 3200 MT/s に、ロックステップモードの場合は、Jordan Creek 2 によって DDR4 周波数の上限が 1866 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 DDR4 周波数をサポートするという変則性が生じます。ただし、より高いメモリ帯域幅は、1 ステップ高い DDR4 周波数よりも貴重です。

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 タイプ 3 シリーズと PRIMERGY RX4770 M3 の違いです。

PRIMEQUEST 2000 タイプ 3 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 2 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 つ目の構成スロットペアは xC0／xC3 で、同様に xB0／xB3、xD0／xD3 と続きます。既存のメモリチャネル全体での構成シーケンスにより、利用可能なすべてのメモリリソースを均等に使用でき、良好なパフォーマンスを得られます。

各メモリ構成のロックステップ機能は、PRIMERGY RX4770 M3 にはありません。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 M3（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 の各位置にある 4 つの DIMM で構成されます。



この図は、PRIMEQUEST 2000 タイプ 3 シリーズの QPI トポロジ、つまりプロセッサのネットワーキングとその適切なメモリコンポーネントを示しています。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、省略しています。

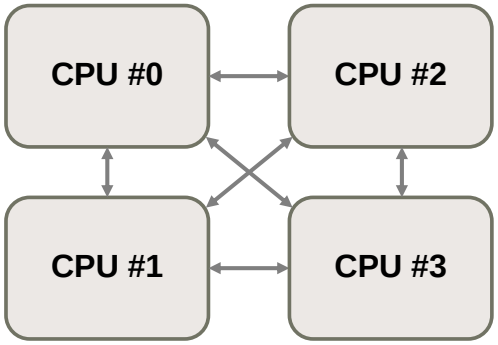
8 つのプロセッサを搭載した PRIMEQUEST 2000 タイプ 3 シリーズの完全構成での各プロセッサは、7 つの隣接プロセッサのうち 3 つだけに直接接続されています。この 3 つのプロセッサは、直接接続されていないプロセッサと通信する場合、ブローカーとしての機能を果たします。必要なブローカーは 1 つだけです。このようなアクセスで生じる遅延は、直接結合の場合と比べて大きくなりますが、ソフトウェア対応の

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。

最大 4 つのプロセッサを搭載した PRIMEQUEST 2400E3 モデルには、システムボードは 0 と 1 だけです。この場合と、システムボードが 4 枚未満の PRIMEQUEST 2800E3 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 M3 は、最初からプロセッサ 4 つ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 トポロジは、PRIMEQUEST 2000 タイプ 3 シリーズのトポロジ、特に PRIMEQUEST 2400E3 モデルのトポロジとは異なります。

QPI Topology of PRIMERGY RX4770 M3



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。

DDR4 トピックと使用可能な DIMM タイプ

Broadwell-EX 搭載システムでは、DDR4 SDRAM メモリモジュールを使用します。従来世代の Haswell-EX で、DDR3 から DDR4 SDRAM メモリモジュールへの移行が行われました。DDR3 および DDR4 の名称の JEDEC (Joint Electron Device Engineering Council：電子機器技術評議会）規格では、メモリメーカーとシステムメーカーの橋渡しとなるインターフェースを定義しています。

DDR4 テクノロジーは、現在も比較的新しく、DDR3 と比べて重要な違いがあります。DDR3 から DDR4 への移行は進化という点では自然なことであり、1 回限りのパフォーマンス向上にはとどまりませんでした。

- DDR4 では DIMM あたりのピン数が増えたため、DDR3 と DDR4 の DIMM ソケットに互換性はありません。古い DDR3 メモリモジュールを DDR4 ベースのシステムで使用することはできません。
- DDR4 では、3200 MHz までのメモリ周波数をサポートします。この周波数範囲は、今後、数世代のサーバで使用されることになります。現在はこのテクノロジーが使用されて、Haswell-EX および Broadwell-EX 搭載システムで最大周波数 1866 MHz がサポートされています。DDR3 ベースのサーバ世代と同様に、周波数は 266 MHz ずつ継続的に上昇していきます。DDR4 への移行は進歩的です。1 回限りのパフォーマンス向上ではありません。
- DDR4 の重要なメリットは、DIMM がわずか 1.2 V で動作する点です。DDR3 では、1.5 V または 1.35 V（低電圧版）でした。これは、データ転送速度が同じ場合、約 30 % の消費電力の節約に相当します。
- DDR3 テクノロジーの最初のフェーズのときと同様に、現在のところ DDR4 に低電圧版はありません。したがって、BIOS におけるパフォーマンスと消電電力の構成トレードオフもほとんど関係ありません。

一般リリース時点の Xeon E7 v4 搭載システムのメモリ構成に適している DIMM は、次の表のとおりです。後でこの表に追加が発生する可能性があります。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不可です。

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]	Low voltage	Load reduced	Registered	ECC	GB あたりの相対価格
16GB (2x8GB) 1Rx4 DDR4-2400 R ECC	RDIMM	16	1	4	2400			✓	✓	1.2
32GB (2x16GB) 1Rx4 DDR4-2400 R ECC	RDIMM	32	1	4	2400			✓	✓	1.0
32GB (2x16GB) 2Rx4 DDR4-2400 R ECC	RDIMM	32	2	4	2400			✓	✓	**1.0**
64GB (2x32GB) 2Rx4 DDR4-2400 R ECC	RDIMM	64	2	4	2400			✓	✓	1.1
128GB (2x64GB) 4Rx4 DDR4-2133 LR ECC	LRDIMM	128	4	4	2133		✓		✓	1.3

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 タイプ 3 シリーズおよび PRIMERGY RX4770 M3 の構成プロセスを考慮に入れています。その理由は、ペアでの構成ルールです。

データは、すべての DIMM タイプで 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。これは全 DDR 世代の機能です。この幅のメモリ領域は、DRAM チップのグループから DIMM に設定されます。このとき、個々のチップは 4 または 8 ビットを担当します（タイプ名のコード x4 を参照。x8 モジュールは現在のところ、Broadwell-EX 搭載サーバには計画されていません）。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、または 4 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 ランクの DIMM のメリットは、最大容量である点にありますが、同時に DDR4 の仕様ではメモリチャネルあたり最大 8 ランク以外はサポートされません。

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。より小規模なサーバクラスで見られるアンバッファード DIMM（UDIMM）では、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「ランク乗算」 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。ランク乗算は、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。

特定のサーバ構成における効率的なメモリ周波数は、2 つ先のセクションで説明する一連の影響によって異なります。DIMM タイプの表で説明されている最大周波数は、他のサーバクラスでも使用されている部品の機能で、効率的なメモリ周波数の上限を示しているにすぎません。この表の 2133 MHz および 2400 MHz は、Broadwell-EX 搭載サーバで理論上変わりません。これらのサーバの有効周波数は最大で 1866 MHz です。

DIMM タイプの表の最終列は、価格の相対的な差異を示しています。この価格は、2016 年 6 月現在の PRIMERGY RX4770 M3 の料金表を使用しています。ここでは 16 GB の 2Rx4 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。8 GB RDIMM および 64 GB LRDIMM でコストの上昇が見られますが、メモリの大容量化が原因です。さらに、相対価格の状況は常に変化しています。この表は 1 つのスナップショットとして理解してください。

PRIMEQUEST モデルおよび 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、Broadwell-EX プロセッサの機能の結果であり、基本的に PRIMEQUEST 2000 タイプ 3 シリーズと PRIMERGY RX4770 M3 で同一です。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、配置に違いがあります。これは、サーバクラスのそれぞれの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示します。

- 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 機能のメモリミラーリングまたはスペアリングの有効化。ここで、PRIMEQUEST 2000 タイプ 3 シリーズと PRIMERGY RX4770 M3 の違いは、ミラーリングとスペアリングが PRIMEQUEST 2000 タイプ 3 シリーズ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 M3 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。Haswell-EX 世代から、スペアリングが刷新され、マルチランクスペアリングを利用できます。前世代では、メモリチャネルあたり予備ランクが 1 つしかありませんでした。
- メモリシステムの消費電力節約。前世代では使われていた、メモリモジュールの低電圧動作は、現在のところ DDR4 に低電圧版がないため、Broadwell-EX 搭載のサーバでは除外されていますが、消費電力節約に関係する最低限の側面が 2 つ残されています。PRIMERGY RX4770 M3 では、メモリ周波数を最小値の 1333 MHz に設定できるのが通常です。この周波数では消費電力節約に一定の効果があります。また、PRIMEQUEST 2000 タイプ 3 シリーズの場合は、起動時間の短縮を優先してメモリの電力状態を軽減できます。つまり、パフォーマンスを優先して電力節約機能 Memory Power States を抑えることができます。メモリモジュールの省電力状態（プロセッサの C 状態と同様）は、メモリアクセスなしで、フェーズで有効になります。
- 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、修正できなくなったエラーの発生率を低減します。この動作はメモリコントローラーが制御します。感度の高いパフォーマンス指標がある場合は、この機能を無効にすることもできます。
- Broadwell-EX プロセッサの世代では、信号線でパリティエラーが発生した場合に備え、メモリコントローラーに再試行オプションが搭載されています。この機能によるパーツのパフォーマンスへの微細な影響は負荷に依存すると考えられているため、PRIMEQUEST 2000 タイプ 3 シリーズでこの機能はオプションとなっています。

この導入の説明の後は、PRIMEQUEST 2000 タイプ 3 シリーズと PRIMERGY RX4770 M3 の具体的な構文設計について説明します。PRIMEQUEST 2000 タイプ 3 シリーズでは、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 タイプ 3 シリーズの MMB Web-GUI のインターフェース

次のオプションを持つパラメーターである *Memory Operation Mode* は、管理ボード（MMB）の Web-GUI の Partition/Partition#/Mode 下（分離可能な PRIMEQUEST 2000 タイプ 3 モデル）と System/Mode（PRIMEQUEST 2800B3）の下にあります。

- Performance Mode（パフォーマンスモード）
- Normal Mode（通常モード）
- Partial Mirror Mode（部分ミラーモード）
- Full Mirror Mode（完全ミラーモード）
- Spare Mode（スペアモード）
- Address Range Mirror Mode（アドレスレンジミラーモード）

デフォルトには下線が引いてあります。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のオペレーティングシステムが利用できます。通常モードは、高度な RAS 機能のあるメモリチャネルのロックステップ動作モードです。パフォーマンスモードは、独立メモリチャネルの高パフォーマンス動作モードです。

構成されたメモリ容量の一部、たとえば完全ミラーの 50 % は、部分ミラー、完全ミラー、スペア、アドレスレンジミラーの 4 つの冗長モードで、オペレーティングシステムが利用できます。

スペアリングはランクスペアリングを意味します。このため、構成した DIMM タイプとランクの数によって正味容量の割合が異なります。Broadwell-EX 搭載のシステムは、スペアとして複数のランクを保持するオプションをサポートしています。これをユーザーが制御できるように、オプション付きの *Memory Sparing Mode* というパラメーターが同じメニューにあります。

- 1Rank
- 2Rank
- Auto

最初の 2 つのパラメーターは読んで字のとおりです。Auto を指定すると、メモリチャネルに存在するランクのうち最大で半分が保持されます。スペア 1 つを保持するデフォルトの 1Rank は、Ivy Bridge-EX までの前世代のシステムのスペアモードに対応しています。

スペアモードの正味容量を計算する際は、異なった DIMM 構成ルールも考慮する必要があります。必ず 3DPC で構成します。8 つのメモリチャネルのうちの 2 つで、プロセッサあたり DIMM 6 つの最小構成で開始します。冗長性を考慮する場合のメモリパフォーマンスに関する次のセクションでは、再度この点について説明します。

4 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。これらのモードは、ロックステップモードに追加されたものです。PRIMEQUEST 2000 タイプ 3 シリーズには、パフォーマンスモードの独立メモリチャネルに関連したミラーリングとスペアリングはありません。

PRIMEQUEST 2000 タイプ 3 シリーズのデバイスマネージャーのインターフェース

これ以外のパラメータは、BIOS の Device Manager／Memory Configuration の下にあります。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経由でアクセスできます。ここでは次のオプションを持つ 4 つのパラメーターがあります。一般リリース時点で有効だったデフォルトには下線が引いてあります。

- Patrol Scrub：Disabled／Enabled
- Refresh Rate：Auto／1x
- Memory Power States：Default／Performance Mode
- DDR4 Command / Address Parity Check and Retry：Disabled／Enabled

パフォーマンス上の理由から、PRIMEQUEST 2000 タイプ 3 シリーズの *Patrol Scrub* パラメーターのデフォルトは、Disabled になっています。ただし、パフォーマンスへの影響は通常非常に小さいです。

また、Memory Power States のパフォーマンスへの影響もささいなものです。レイテンシが低い用途シナリオの場合、Performance Mode を設定することで、測定可能な改善がみられる場合があります。STREAM および SPECint_rate_base2006 のベンチマークでは改善を確認できませんでした。本書では、メモリパフォーマンスの特徴を表すためにこれらのベンチマークを使用しています。

2 つ目の Refresh Rate パラメーターは DDR3 テクノロジーのなごりで、PRIMEQUEST 2000 タイプ 3 シリーズなどの DDR4 ベースのシステムでは廃止されています。今後の BIOS バージョンでは除外される可能性があります。

PRIMERGY RX4770 M3 の BIOS のインターフェース

PRIMERGY RX4770 M3 には、Advanced の下の BIOS に、次のパラメーターが付いたメモリ構成サブメニューがあります。

- Memory Mode：Normal／Mirroring／Sparing
- VMSE Lockstep Mode：Lockstep／Independent
- DDR Performance：Performance optimized／Energy optimized
- Patrol Scrub：Disabled／Enabled

一般リリース時点で有効であったデフォルトには下線が引いてあります。

1 つ目のパラメーター *Memory Mode* は、RAS 機能のミラーリングとスペアリングの有効化に関するものです。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。一般リリース時点で、予備ランクが 1 つの旧システムのランクスペアリングをサポートしているのは PRIMERGY RX4770 M3 のみです。

2 番目のパラメーター *VMSE Lockstep Mode* は、PRIMEQUEST 2000 タイプ 2 シリーズと対照的に、PRIMERGY RX4770 M2 が RAS モードのミラーリングとスペアリングの随意的な有効化と無関係であるため、独立したメモリチャネル（Independent）とロックステップ動作モード間での選択に関連しています。

3 番目のパラメーター *DDR Performance* では、Energy optimized を設定すると、結果的にメモリ周波数が全般的に 1333 MHz まで低下します。ただし、消費電力の節約になる可能性は高くはありません。メモリの消費電力は主に DIMM の電圧によって決まります。Broadwell-EX 搭載サーバでは常に 1.2 V です。

4 番目のパラメーター *Patrol Scrub* は上記のとおり処理されていました。

キャッシュコヒーレンスプロトコルに拡張機能として導入された COD（Cluster-on-die）オプションのパラメーターは、CPU 設定のサブメニューで選択できます。

- COD Enable：Disabled／Enabled／Auto
- Home Dir Snoop with IVT- Style OSB Enable：Disabled／Enabled／Auto

デフォルトの Auto 設定では、COD は有効化されません。有効化するには、*COD Enable* のパラメーターを *Enabled* に設定し、*Home Dir Snoop with IVT- Style OSB Enable* のパラメーターを Auto のままにします。

COD は、主にローカルメモリアクセスのスペキュレーションに適した負荷の場合は一考に値します。有効化されると、2 つのメモリコントローラーに基づいて各プロセッサに 2 つの NUMA ノードが作成されます。プロセッサコアの半数および L3 キャッシュの半数は、それぞれのケースにおいて、ノードに割り当てられます。適した負荷におけるパフォーマンスの利点はおよそ 1-2 %です。これは標準のベンチマークである SPECint_rate_base2006 によって検証されました。

PRIMEQUEST 2000 のような、プロプライエタリな接続チップがないプロセッサを直接結合するシステム設計（いわゆる *グルーレス設計*）の場合、NUMA ノードの総数は 8 つに制限されます。これがつまり、PRIMEQUEST 2000 タイプ 3 シリーズでは COD が利用できない理由です。

メモリ周波数の定義

構成の効率的なメモリ周波数は、メモリパフォーマンスに関する重要なパラメーターで、全般的な条件の範囲によって異なります。Broadwell-EX 搭載サーバの場合は、1866 MHz、1600 MHz、1333 MHz の 3 つの値が問題となります。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。

一般的な条件では、構成されているプロセッサモデル、メモリチャネルの動作モード（ロックステップまたは独立／パフォーマンス）、3DPC 構成の場合はさらに、構成されている DIMM タイプが問題となります。また、PRIMERGY RX4770 M3 では、*Energy optimized* というキーワードで、メモリ周波数を最小値の 1333 MHz に落とすオプションが追加されています。

まず、構成されたプロセッサ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、Broadwell-EX シリーズを次の表にしたがって分類することをお勧めします。この表は、メモリチャネルの動作モードごとの最大メモリ周波数を示しています。Xeon E7 v4 モデルの全リストが表に示されています。PRIMEQUEST 2000 タイプ 3 シリーズおよび PRIMERGY RX4770 M3 のサーバモデルでの可用性については、システムの構成を参照してください。

CPU タイプ	QPI	独立（2:1）		ロックステップ（1:1）		Xeon E7 v3 モデル
		SMI	DDR4	SMI	DDR4	
Advanced	9.6	3200	1600	1866	1866	E7-8890 v4、E7-8880 v4、E7-8870 v4、E7-8860 v4、E7-8891 v4、E7-8893 v4、E7-8867 v4
Standard	8.0	2666	1333	1866	1866	E7-8855 v4、E7-4850 v4、E7-4830 v4
Basic	6.4	2666	1333	1866	1866	E7-4820 v4、E7-4809 v4

メモリ周波数は DDR4 の周波数を意味します。ただし、メモリ接続のアーキテクチャーによると、オンチップメモリコントローラーとオフチップメモリバッファー間の SMI Gen2 リンクの周波数にリンクされています。ロックステップ動作モードの場合、周波数の割合は 1:1 です。独立モードの場合は 2:1 です。これは、独立メモリチャネルで 4 つの SMI Gen2 リンクの周波数と 8 つの DDR4 チャネルの周波数間のバランスをとるためです。これはロックステップモードでは必要ありません。8 つの DDR4 チャネルが 4 つの論理チャネルのペアを形成するために統合されているためです。

通常はメモリ周波数が主な焦点になりますが、これがベースとなっているロジックは SMI トピックが関係する場合に限り明確になります。このため、表には対応する SMI 周波数も記載してあります。上限の 3200 MT/s は、独立モードで最大メモリ周波数 1600 MHz を必要としており、Jordan Creek 2 のメモリバッファーに適用されます。一方、ロックステップモードのメモリ周波数の方が高いのは、独立モードの場合に比べて SMI 周波数が低いことに関係しています。周波数はこのリソースの帯域幅に相当します。SMI 周波数を見るだけで、パフォーマンスレベルがより低いモード（ロックステップ）がより高いメモリ周波数をサポートするという明らかな変則性があることが分かります。

各チャネルに DIMM を 3 つ搭載したフル構成のメモリチャネル（3DPC 構成）では、より一般的な条件として、結果的にメモリ周波数が低下する場合があります。この場合、メモリチャネルの静電負荷が設計によって DIMM のタイプごとに異なるため、DIMM のタイプに依存します。

そこで、特定の構成における有効なメモリ周波数を次に示します。

メモリチャネルのロックステップ動作モード

	8 および 16 GB 1Rx4 RDIMM			16 および 32 GB 2Rx4 RDIMM			64 GB 4Rx4 LRDIMM		
CPU タイプ	**1DPC**	**2DPC**	**3DPC**	**1DPC**	**2DPC**	**3DPC**	**1DPC**	**2DPC**	**3DPC**
Advanced Standard Basic	1866	1866	1600	1866	1866	1333	1866	1866	1600

メモリチャネルの独立動作モード

	8 および 16 GB 1Rx4 RDIMM			16 および 32 GB 2Rx4 RDIMM			64 GB 4Rx4 LRDIMM		
CPU タイプ	**1DPC**	**2DPC**	**3DPC**	**1DPC**	**2DPC**	**3DPC**	**1DPC**	**2DPC**	**3DPC**
Advanced	1600	1600	1600	1600	1600	1333	1600	1600	1600
Standard Basic	1333	1333	1333	1333	1333	1333	1333	1333	1333

独立モードはパフォーマンスモードとも呼ばれ、通常はパフォーマンス測定およびベンチマークに適しています。この理由については前述しました。メモリチャネルが独立していることで帯域幅の利点が得られます。ロックステップモードではメモリ周波数が比較的高いため、この利点が低減される可能性がありますが、そのようにはなっていません。それぞれの SMI 周波数は帯域幅比の指標として使うことができます。

PRIMERGY RX4770 M3 の Energy Optimized 設定

BIOS パラメーターに関する前のセクションで説明したように、PRIMERGY RX4770 M3 で *DDR Performance = Energy optimized* を設定すると、通常はメモリ周波数が最小値 1333 MHz まで下がります。このため上の表は使われなくなりました。PRIMEQUEST 2000 タイプ 3 シリーズには、パラメーターは存在しません。

Energy optimized によって節約できる電力はかなり小さいということをもう一度指摘しておく必要があります。メモリモジュールの消費電力は主に電圧によって決まります。Broadwell-EX 搭載サーバでは常に 1.2 V です。

理想的なメモリ容量

ここまで、Broadwell-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、メモリチャネルの動作モードによって制御され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 M3 では、プロセッサごとにメモリボードが 2 枚必要になる点に注意してください。

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）	独立	ロックステップ	ベンチマーク
					最大 MHz	最大 MHz	
128	256	512	1	8GB 1Rx4 RDIMM	1600	1866	
256	512	1024	2	8GB 1Rx4 RDIMM	1600	1866	
			1	16GB 1Rx4 RDIMM	1600	1866	
			1	16GB 2Rx4 RDIMM	1600	1866	
384	768	1536	3	8GB 1Rx4 RDIMM	1600	1600	
512	1024	2048	2	16GB 1Rx4 RDIMM	1600	1866	+
			2	16GB 2Rx4 RDIMM	1600	1866	++
			1	32GB 2Rx4 RDIMM	1600	1866	
768	1536	3072	3	16GB 1Rx4 RDIMM	1600	1600	
			3	16GB 2Rx4 RDIMM	1333	1333	
1024	2048	4096	2	32GB 2Rx4 RDIMM	1600	1866	+
			1	64GB 4Rx4 LRDIMM	1600	1866	
1536	3072	6144	3	32GB 2Rx4 RDIMM	1333	1333	
2048	4096	8192	2	64GB 4Rx4 LRDIMM	1600	1866	
3072	6144	12288	3	64GB 4Rx4 LRDIMM	1600	1600	

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 2 メモリバッファー、DDR4 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、構成について各行で 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。この技術的な背景は、NUMA システムアーキテクチャーでのローカルメモリアクセスとリモートメモリアクセスの違いです。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

この表は、それぞれの構成ごとの最大メモリ周波数を示していますが、メモリチャネルの動作モードについては、すでに説明したケースの違いに注意してください。独立モードの場合は、消費電力がより低いプロセッサモデルで構成すると、結果的に表に示した値より周波数が低くなる場合があります。さらに、BIOS に *Energy optimized* を設定すると周波数が低くなることがあります（PRIMERGY RX4770 M3 の場合で 1333 MHz）。後者は両方の動作モードに当てはまります。

いずれにしても、表のメモリ構成は、RAS（ロックステップ）とパフォーマンスとのトレードオフがどのように決められたかにかかわらず、理想的なチャネルインターリーブの特性を示しています。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。

PRIMEQUEST 2000 タイプ 3 シリーズおよび PRIMERGY RX4770 M3 の標準的な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。最適なメモリパフォーマンスの構成には ++ の印が付いています。

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。

その次のセクションでは、メモリコントローラーとチャネルにおけるインターリーブについて説明します。また、チャネルのロックステップ動作モードと独立動作モードの違いをトピックで取り上げています。それ以降のセクションでは、メモリ周波数、ランクでのインターリーブ、さまざまな DIMM タイプ固有のその他の影響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、その他の機能をできるだけ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 タイプ 3 シリーズでは、それぞれ 2 つのプロセッサが搭載された 1 枚および 4 枚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を実施しました。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 2800E3	PRIMERGY RX4770 M3
CPU 種類	Xeon E7-8890 v4	Xeon E7-8890 v4
メモリタイプ	16GB (2x8GB) 1Rx4 DDR4-2400 R ECC 32GB (2x16GB) 2Rx4 DDR4-2400 R ECC	32GB (2x16GB) 2Rx4 DDR4-2400 R ECC 128GB (2x64GB) 4Rx4 DDR4-2133 LR ECC
ディスクサブシステム	1 × RAID Ctrl SAS 6G 1 GB 1 × HD SAS 6G 300 GB 15K HOT PL 2.5” EP	1 × RAID Ctrl SAS 6G 1 GB 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 16043 （BIOS、BMC、MMB）	BIOS R1.0.0, BMC 8.13F
オペレーティングシステム	Red Hat Enterprise Linux Server release 6.7	Red Hat Enterprise Linux Server release 6.7

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 2800E3 のパフォーマンスレポート［関連資料 6］および PRIMERGY RX4770 M3 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 v4 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。

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

インターリーブは、同じタイプの複数メモリリソース間で変更することによる、物理アドレス領域のセットアップです。まず、Broadwell-EX 搭載サーバの場合は、メモリコントローラーが 2 つのプロセッサが適しています。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、以下で説明します。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。

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。メモリコントローラーおよびメモリチャネルにわた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。

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 タイプ 3 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 1600 MHz	**100 %**	58 %	29 %
	ロックステップ 1866 MHz	70 %	36 %	18 %
商用アプリケーションパフォーマンス（SPECint_rate_base2006）	独立 1600 MHz	**100 %**	93 %	77 %
	ロックステップ 1866 MHz	96 %	82 %	62 %

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 SPECint_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。

独立動作モードとロックステップ動作モード間のケースの違いに関して、表の元になっている測定値と同様、メモリ周波数も通常異なっていることに注意してください。つまり、チャネルインターリーブの主な影響とは別に、これらの比較は、さまざまなメモリ周波数の二次的な影響も組み込んでいます。

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。

前の表は PRIMEQUEST 2000 タイプ 3 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 2 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 M3 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 M3 のコンフィギュレータに関する知識は、次の表を理解するうえでの前提条件です。

PRIMERGY RX4770 M3 のチャネルインターリーブ					
	動作モード	**CPU あたり：2 枚のメモリボード 全体で 8 枚の DIMM を分散** **理想的な容量**	**CPU あたり：2 枚のメモリボード 全体で 4 枚の DIMM を分散**	**CPU あたり：1 枚のメモリボード 全体で 4 枚の DIMM を分散**	**CPU あたり：1 枚のメモリボード 全体で 2 枚の DIMM を分散** **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 1600 MHz	**100 %**	65 %	51 %	33 %
	ロックステップ 1866 MHz	69 %		35 %	
商用アプリケーションパフォーマンス（SPECint_rate_base2006）	独立 1600 MHz	**100 %**	95 %	92 %	79 %
	ロックステップ 1866 MHz	96 %		82 %	

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影響（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次に示します。ベンチマークでは常に品質 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。80 %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。60 %を少し超えるケースの場合は、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 のケースも、この表の 4 枚の DIMM の場合に相当するメモリパフォーマンスになる可能性があります。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。

メモリ周波数の影響

コントローラーとチャネルインターリーブがメモリパフォーマンスに与える二次的な影響は、メモリ周波数の影響です。

Broadwell-EX 搭載サーバに関する限り、この影響が問題となる典型的な状況は、3DPC 構成に関連して周波数が低下する場合です。この関係については、メモリ周波数の定義について説明したセクションですでに説明しています。3DPC 構成は大容量のメモリに必要です。つまり、パフォーマンスと容量間のトレードオフの問題です。

次の表に、メモリチャネルの動作モードと DIMM のタイプによって、3DPC で周波数が低下するケースを示します。

- ロックステップモードでは必ず周波数の低下が発生します。
- 独立モードつまりパフォーマンスモードでは、16 GB および 32 GB 2Rx4 RDIMM のみに適合します。独立モードでの周波数 1600 MHz は、強力な *Advanced* プロセッサモデルでのみ可能であることにも留意してください。周波数のトピックは、パフォーマンスレベルが低いモデルには適していません。

表の元になっている測定値は、プロセッサあたり 8、16、または 24 枚の DIMM 構成で、理想的なチャネルインターリーブで測定したものです。

		独立モード（Advanced CPU）	ロックステップモード		
		RDIMM 2Rx4	RDIMM 1Rx4	RDIMM 2Rx4	LRDIMM 4Rx4
2DPC	メモリ帯域幅 (STREAM)	**100 % （1600）**	**100 % （1866）**	**100 % （1866）**	**100 % （1866）**
	商用アプリケーションパフォーマンス (SPECint_rate_base2006)	**100 % （1600）**	**100 % （1866）**	**100 % （1866）**	**100 % （1866）**
3DPC	メモリ帯域幅 (STREAM)	87 % （1333）	83 % （1600）	74 % （1333）	93 % （1600）
	商用アプリケーションパフォーマンス (SPECint_rate_base2006)	96 % （1333）	97 % （1600）	93 % （1333）	96 % （1600）

表で比較のベースにしているのはそれぞれ 2DPC 構成で、最大周波数を許可します。影響を受けるメモリ周波数をパフォーマンスのパーセンテージ値の下に括弧で示しています（MHz 単位）。

比較している 4 つのそれぞれの構成におけるパフォーマンスの低下は、周波数の違いが主な原因です。ただし、この影響には、次のセクションで説明する DIMM 設計にかかわる影響が加味されています。1Rx4 RDIMM および 4Rx4 LRDIMM で影響が異なるのはこのためですが、両方のケースで 3DPC のメモリ周波数が、ロックステップモードで 1866 MHz から 1600 MHz に低下しています。

PRIMERGY RX4770 M3（*DDR Performance = Energy optimized*）の場合に周波数が全般的に 1333 MHz に低下するという影響は、2Rx4 RDIMM で 2DPC から 3DPC に移行する場合とほぼ同じです。つまり、商用アプリケーションパフォーマンスは、動作モードによって 4 %（独立モード）から 7 %（ロックステップモードモード）低下します。

独立モードに比べてロックステップモードモードの方がパフォーマンス低下が大きいのは、ロックステップでメモリ帯域幅が低いことが原因です。より厳しい条件下では、周波数低下のような影響が加わります。

DDR Performance = Energy optimized オプションによる特別なケースは例外ですが、周波数については、PRIMEQUEST 2000 タイプ 3 シリーズと PRIMERGY RX4770 M3 間で区別する必要はありません。

ランクでのインターリーブと DIMM タイプの影響

次の表では、同じメモリ周波数の DIMM 構成のケースを比較しています。ここでも動作モード別に示しています。最大周波数は、独立モードで 1600 MHz、ロックステップモードで 1866 MHz です。一連の測定は理想的なチャネルインターリーブ下（プロセッサあたり 8、16、または 24 個の DIMM）で行いました。構成を同一にして、メモリパフォーマンスでの 2 つの主な影響、つまりチャネルインターリーブとメモリ周波数を比較しています。

相対的なパフォーマンスの説明は、絶対的にベストなケース（太字の 100 %で強調表示）に関連していることが分かります。両動作モードとも、2Rx4 RDIMM での 2DPC 構成が最高のメモリパフォーマンスを示しています。プロセッサあたりのメモリ容量が十分ある場合に、ベンチマークに適しているのはこのためです。

ただし、実稼働環境の商用アプリケーションパフォーマンスが 1〜2 %低下しますが、通常は無視してかまいません。このセクションで示したパフォーマンスの違いは、ベンチマーク時に主に考慮に入れた微妙な差異です。

		独立モード 1600 MHz			ロックステップモード 1866 MHz		
		RDIMM 1Rx4	RDIMM 2Rx4	LRDIMM 4Rx4	RDIMM 1Rx4	RDIMM 2Rx4	LRDIMM 4Rx4
1DPC	メモリ帯域幅 (STREAM)	92 %	100 %	98 %	87 %	98 %	97 %
	商用アプリケーションパフォーマンス (SPECint_rate_base2006)	99 %	100 %	99 %	98 %	100 %	99 %
2DPC	メモリ帯域幅 (STREAM)	97 %	**100 %**	91 %	96 %	**100 %**	89 %
	商用アプリケーションパフォーマンス (SPECint_rate_base2006)	99 %	**100 %**	98 %	99 %	**100 %**	97 %
3DPC	メモリ帯域幅 (STREAM)	92 %		92 %			
	商用アプリケーションパフォーマンス (SPECint_rate_base2006)	99 %		95 %			

パフォーマンスに違いがあるのは、インターリーブの形式が別ということが主な原因です。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、すでに説明したコントローラーとメモリ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。

パフォーマンスの低下（特に帯域幅）は、1Rx4 RDIMM の 1DPC および 3DPC のランク数が奇数ということで説明できます。ただし、DRAM チップあたりの最大オープンページ数が 8 から 16 の倍になるため、DDR4 の場合にこの悪化はかなり低減されます。DDR3 のケースでは、ランクインターリーブがないことで結果的に帯域幅が 80 %に低下しました。

メモリパフォーマンスへのさらなる影響が LRDIMM のランクインターリーブに加わっています。RDIMM に比べると、まず、データバッファリング用の DIMM コンポーネントが原因で一定のオーバーヘッドが発生します。また、メモリチャネルに 4 つを超えるランクがあるため、DRAM をリフレッシュするために実行されるランクごとのオーバーヘッドが、否定的な意味で目立つようになります。このリフレッシュは、すべてのランクで共有される、メモリチャネルのアドレス行ごとの一定の基本負荷を表します。最後に、メモリチャネルの 8 つを超える物理ランクで、ランク乗算のオーバーヘッドが発生します。

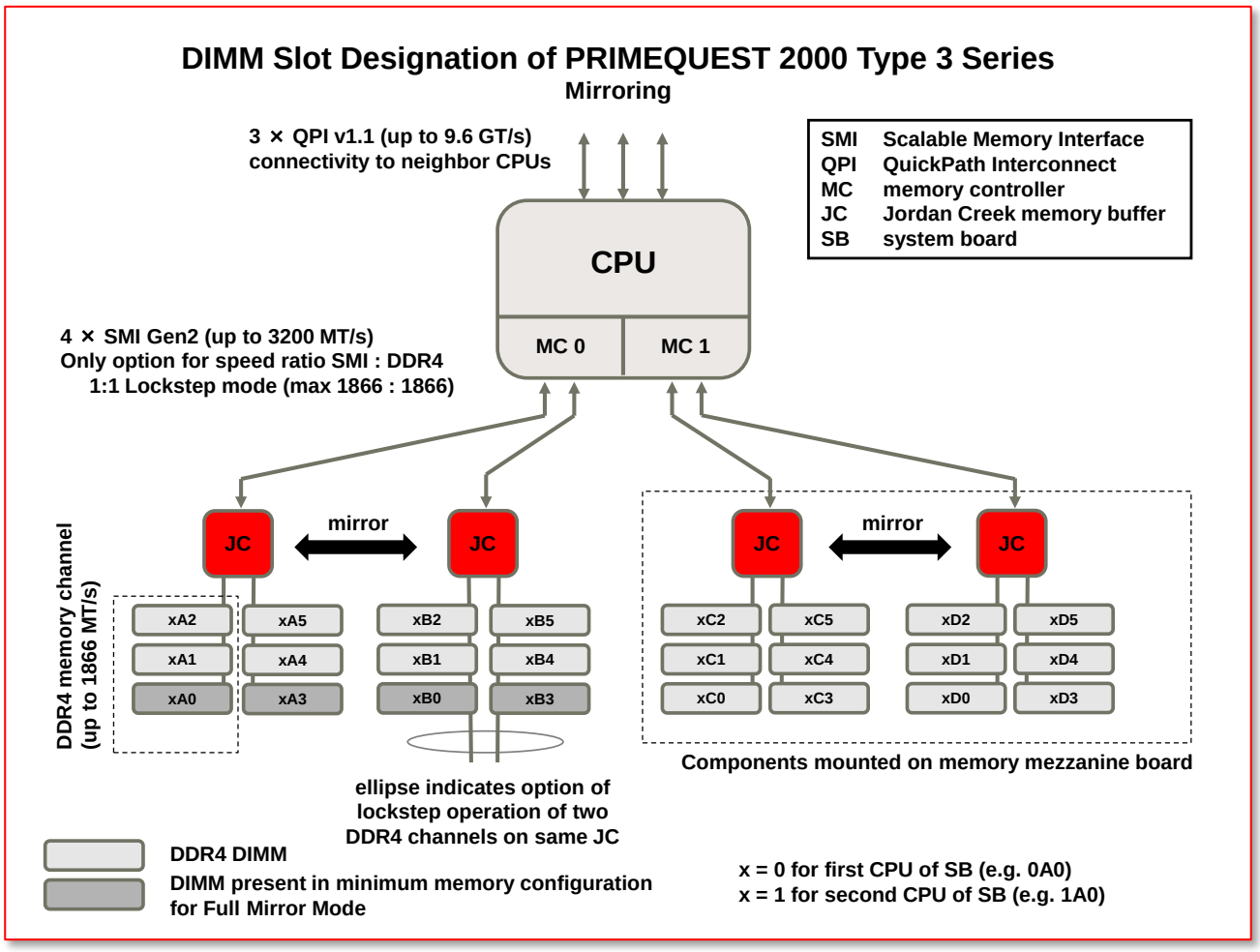
最大メモリ構成を実現するために最適化された DIMM タイプのパフォーマンスがいくぶん低下するという傾向は、このような影響によって説明できます。プロセッサの各世代のメモリコントローラーのさらなる改善と、DDR テクノロジーのさらなる改善の結果、サーバの世代間で影響に若干の変化があります。

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、RAS 機能のミラーリングとランクスペアリングについて、少し説明します。

PRIMEQUEST 2000 タイプ 3 シリーズの完全ミラーモード

ミラーリングは、2 つの Jordan Creek 2 バッファーと、バッファーあたり 2 つの DDR4 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 2 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 2 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。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 1600 MHz	**100 %**	58 %
	通常モード（ロックステップ） 1866 MHz	70 %	36 %
	完全ミラーモード 1866 MHz	50 %	25 %
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 1600 MHz	**100 %**	93 %
	通常モード（ロックステップ） 1866 MHz	96 %	82 %
	完全ミラーモード 1866 MHz	90 %	72 %

[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 M3 の完全ミラーモード

PRIMEQUEST 2000 タイプ 3 シリーズとは異なる DIMM 構成ルールが PRIMERGY RX4770 M3 のケースに適用されます。もう一度コンフィギュレータ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。

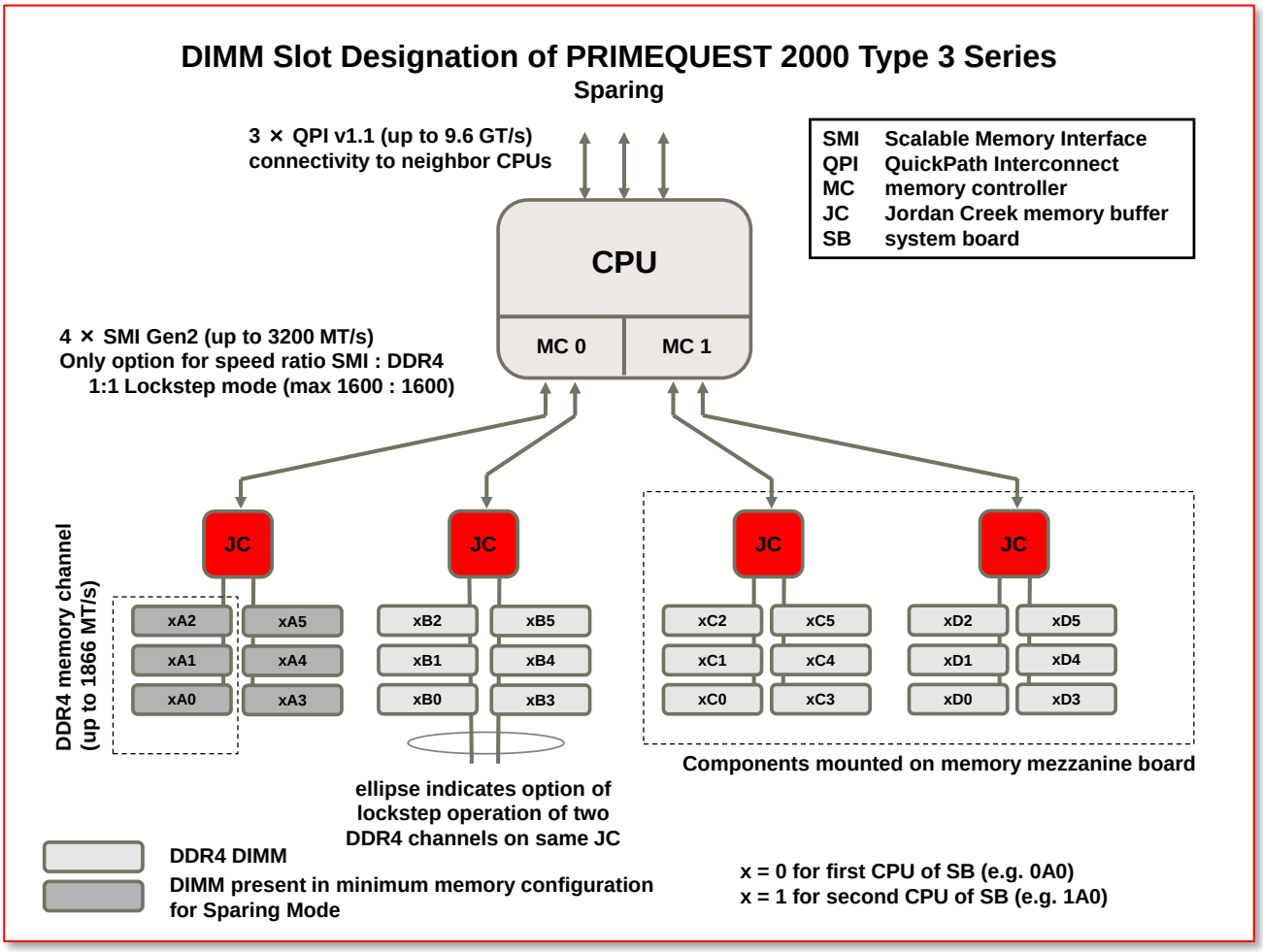
その他の違いとしては、PRIMEQUEST 2000 タイプ 3 シリーズと同様にミラーリングをロックステップ動作モードに追加できるだけでなく、独立動作モードにも追加できる点です。この違いは、次の表に示されています。

	動作モード	CPU あたり： 2 枚のメモリボード 全体で 8 枚の DIMM を分散 理想的な容量	CPU あたり： 2 枚のメモリボード 全体で 4 枚の DIMM を分散	CPU あたり： 1 枚のメモリボード 全体で 4 枚の DIMM を分散	CPU あたり： 1 枚のメモリボード 全体で 2 枚の DIMM を分散 最小構成
メモリ帯域幅（STREAM）	独立 1600 MHz	**100 %**	65 %	51 %	33 %
	独立 + ミラー 1600 MHz	69 %	45 %	35 %	22 %
	ロックステップ 1866 MHz	69 %		35 %	
	ロックステップ + ミラー 1866 MHz	49 %		26 %	
商用アプリケーションパフォーマンス（SPECint_rate_base 2006）	独立 1600 MHz	**100 %**	95 %	92 %	79 %
	独立 + ミラー 1600 MHz	97 %	87 %	82 %	64 %
	ロックステップ 1866 MHz	96 %		82 %	
	ロックステップ + ミラー 1866 MHz	89 %		73 %	

スペアモード

RAS 機能のランクスペアリングでパフォーマンスに新しい影響はありませんが、これまで見てきた影響が新たな観点でリンクしています。スペアリングというトピックについてさらに測定する必要はありません。その代わり、すでに説明した詳細から、スペアリングを有効化した状態でメモリパフォーマンスをどのように実現するかについて、例を使って説明します。

PRIMEQUEST 2000 タイプ 3 シリーズのケースでは、再度変更した最小構成と、メモリ容量を増やすための構成ルールで、別の見方をすることができます。次の図は最小構成を示しています。ミラーリングの場合と同様、スペアリングはロックステップ動作の PRIMEQUEST 2000 タイプ 3 シリーズのみで可能です。



すべてのメモリチャネルでできるだけ広範に DIMM を配分するというやり方が別にありますが、これとは反対に、スペアリングモードのメモリチャネルを 6 つずつのグループに分けて 3DPC 構成を形成します。最小構成の次は 2 番目のグループの xC0～xC5、次に xB0～xB5、最後に xD0～xD5 です。これはこのプロセスの経済性を優先したものです。メモリチャネルごとに未使用の予備として 1 つまたは 2 つのランクが保持されるため、メモリチャネルごとに最大数のランクを利用できる場合に、正味の割り当て容量が最大となります。

したがって、（ミラーリングは別にして）ランクスペアリングでのメモリパフォーマンスで前に説明したすべての影響が関連しているのは明らかです。

- 構成シーケンスの修正に起因する、チャネルインターリーブの頻繁な減少。
- すべての許可された構成の 3DPC 機能に起因するメモリ周波数の低下。
- 未使用のランクに起因する、ランクインターリーブの変更。

例として、次の 2 つの構成間のパフォーマンスの違いを予測します。この観点で、両方の構成でオペレーティングシステムにプロセッサあたり 128 GB の正味メモリ容量を割り当てているため、この比較は妥当といえます。両方とも、パフォーマンスの最大化ではなく RAS に重点を置いています。

- A：ロックステップ動作、プロセッサあたりタイプ 16 GB 2Rx4 RDIMM の DIMM x 8 枚。DIMM は xA0、xA3、xB0、xB3、xC0、xC3、xD0、xD3 の各位置に装着。
- B：スペア動作、*Memory Sparing Mode = 2Rank* および同じタイプの DIMM x 12 枚。DIMM は xA0～xA5 および xC0～xC5 の各位置に装着。

商用アプリケーションパフォーマンスの場合、セクション「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」の表に従えば、B のチャネルインターリーブが半分になると、A の場合と比較して低下はおよそ 15 %になります。ロックステップのケースでは、DIMM が 8 枚の場合と 4 枚の場合を比較する必要があります。4 枚の場合に、4 つのメモリチャネルに DIMM が装着される構成 B の場合と同じチャネルインターリーブとなるためです。

（A の 1866 MHz の代わりに）周波数が 1333 MHz の B の 3DPC 構成では、低下がおよそ 7 %加わります（セクション「メモリ周波数の影響」の表参照）。

一方、奇数の数のランクが発生していないため、ランクインターリーブの影響は無視してかまいません。このようなケースの場合、例えば、同じ例を設定 *Memory Sparing Mode = 1Rank* で実行した場合（ただしこのケースでは A と B のメモリ容量は同一ではなくなる）、さらに 1 – 2 %の低下が発生します。

全体としては、構成 B のパフォーマンスレベルは A に比べて、ゆうに 20 %は低くなります。ただし、この大きさの低下が明確に表れるのは、システムが完全負荷状態にある場合に限られます。この低下の大きさは、RAS 要件が高いケースでプロセッサリソースの次元設定が十分であることを示しているとみなす必要があります。

また、PRIMERGY RX4770 M3 のスペアモードにはさまざまな構成ルールがあります。このシステムでは、2DPC 構成および 3DPC 構成が問題になります。さらに、PRIMEQUEST 2000 タイプ 3 シリーズとは対照的に、独立モードおよびロックステップモードともにスペアモードと組み合わせることができるため、構成ルールはさらに分化されます。また、メモリボードを 1 枚使うか 2 枚使うかによって動作に違いが生じます。非常に包括的な構成ルールの再現は、本書では取り扱いません。

DIMM 構成が既知の場合、スペアモードのメモリパフォーマンスは、PRIMEQUEST 2000 タイプ 3 シリーズの場合と同様、PRIMERGY RX4770 M3 のそれぞれのパフォーマンスの影響を示した表からも取得できます。

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=7bd26a0c-a46c-4717-be6d-78abebba56b2
:http://docs.ts.fujitsu.com/dl.aspx?id=5569306e-5346-4393-9c9b-44c398c32d86
:http://docs.ts.fujitsu.com/dl.aspx?id=0410aac0-ccd0-4730-9db8-eba50cfbaad7

［関連資料 3］Xeon E5-2600 v4（Broadwell-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=3ce313c6-6713-4350-880c-16959489a510

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 2800E3
http://docs.ts.fujitsu.com/dl.aspx?id=f436ef81-faf5-4a47-831e-53dc912f3c04

［関連資料 7］パフォーマンスレポート PRIMERGY RX4770 M3
http://docs.ts.fujitsu.com/dl.aspx?id=7984ec7d-3891-4a43-8769-71820068ec18

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

2016-07-29 WW JA