

ホワイトペーパー

Fujitsu PRIMERGY サーバ

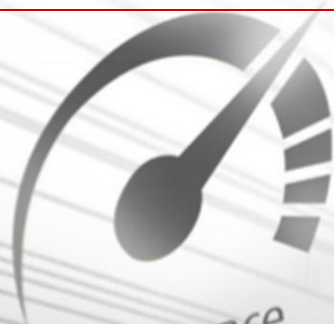
Xeon E5-2600 v2 (Ivy Bridge-EP) 搭載システムのメモリパフォーマンス

Xeon E5-2600 v2 (Ivy Bridge-EP) 搭載 PRIMERGY モデルでも、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で、パフォーマンスが目覚しく向上します。これは 3 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、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

2013-11-12



目次

ドキュメントの履歴	2
はじめに	3
メモリアーキテクチャー	4
DIMM スロット	4
DIMM タイプ	6
メモリ周波数の定義	8
BIOS パラメーター	9
パフォーマンスを考慮したメモリ構成	11
パフォーマンスモード構成	11
独立モード構成	13
対称型メモリ構成	13
メモリパフォーマンスに対する定量的影響	14
測定ツール	15
メモリチャネルへのインターリーブ	16
メモリ周波数	17
メモリランクへのインターリーブ	18
リモートメモリへのアクセス	19
冗長性を考慮した際のメモリパフォーマンス	20
関連資料	21
お問い合わせ先	21

ドキュメントの履歴

バージョン 1.0 (2013 年 9 月 13 日)

初版

バージョン 1.0a (2013 年 11 月 12 日)

マイナー修正

はじめに

Intel Xeon E5-2600 v2 (Ivy Bridge-EP) プロセッサが搭載されたデュアルソケット PRIMERGY サーバの現行機種では、大部分の負荷シナリオにおけるパフォーマンスが旧世代のサーバと比較して 30~40 %向上しています。これは、マイクロアーキテクチャーと Intel C600 (Romley-EP) チップセットはそのまま、32 nm 製造技術から 22 nm 製造技術へと移行したことによります。22 nm に移行したことで、プロセッサあたりのコア数は、これまでの 8 から最大 12 に増えました。この、パフォーマンス向上における主要パラメータは、メモリシステムの最大メモリ周波数を 1600 MHz から 1866 MHz に上げることで支えられています。DDR3 メモリ (800~2133 MHz) 用の範囲において、1866 MHz は最後から 2 つ目のレベルです。

先行世代の Nehalem-EP、Westmere-EP、Sandy Bridge-EP のメモリアーキテクチャーで実証済みの主要機能は維持されています。プロセッサには、オンチップメモリコントローラが搭載されており、各プロセッサはそれぞれに割り当てられているメモリモジュールのグループを制御します。このローカルメモリアクセスのパフォーマンスは、非常に優れています。このプロセッサは、片方向シリアル QPI (QuickPath Interconnect : QuickPath インターコネクト) リンク経由でメモリの内容を隣接プロセッサに提供し、同時に隣接プロセッサにメモリの内容を要求します。リモートアクセスのパフォーマンスは、さほど高くありません。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。

ただし、直前の世代である Sandy Bridge-EP [関連資料 3] のメモリシステム機能について言えば、多くがそのまま引き継がれています。各プロセッサに 4 つのメモリチャネルが存在し、DIMM (Dual Inline Memory Module : デュアルインラインメモリモジュール) スロットは 3 本あります。そのため、プロセッサあたりの最大 DIMM 数は 12 のままです。同じことが、QPI リンクの周波数にも言え、プロセッサモデルに応じて 8.0、7.2、6.4 GT/s (ギガトランスファー/秒) となっています。また、これらのリンクを 2 つ使って両プロセッサを結合することもできます。新しい点は、前述のとおり、最大メモリ周波数が 1600 MHz から 1866 MHz に増えたことです。加えて、メモリコントローラーのスケジューリングと一貫性プロトコルが最適化されています。その結果、最も基本的なメモリパフォーマンスの指標であるメモリ帯域幅が、デュアルソケットサーバで 80 GB/s から 100 GB/s に向上しました。

可能な限り強力なシステムを構成するためには、メモリアーキテクチャーに関する基本的な知識が必要です。このホワイトペーパーでは、この基本情報を提供します。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、両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し、さらに高速化するために、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、プロセッサあたり 4 つのメモリチャネルが横方向に存在します。各プロセッサ上の PRIMERGY コンフィギュレータを同じタイプの 4 枚 1 組の DIMM による「パフォーマンスモード」構成にすることで、この方向への最適なインターリーブを実現しています。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。このための決定的なメモリリソースが、いわゆるランク数です。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。
- パフォーマンスに影響を与えるメモリ周波数は、プロセッサのタイプ、DIMM のタイプと数によって、1866 MHz、1600 MHz、1333 MHz、1066 MHz があります。また、消費電力を削減するために BIOS 設定によって周波数を 800 MHz に下げることができます。メモリ容量が非常に大きい場合、メモリモジュールを低電圧低消費電力モードに設定すると、メモリ周波数が制限されます。そのため、パフォーマンス、容量、消費電力の 3 点を相互に調整する必要があります。

このホワイトペーパーでは、影響を与える要因を取り合げ、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。

測定結果では、プロセッサのパフォーマンスごとの影響をパーセンテージで示します。構成プロセッサモデルが強力であるほど、本書で取り上げているメモリ構成の問題について十分に考慮する必要があります。

ミラーリングやランクスペアリングなど、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、4 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、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。

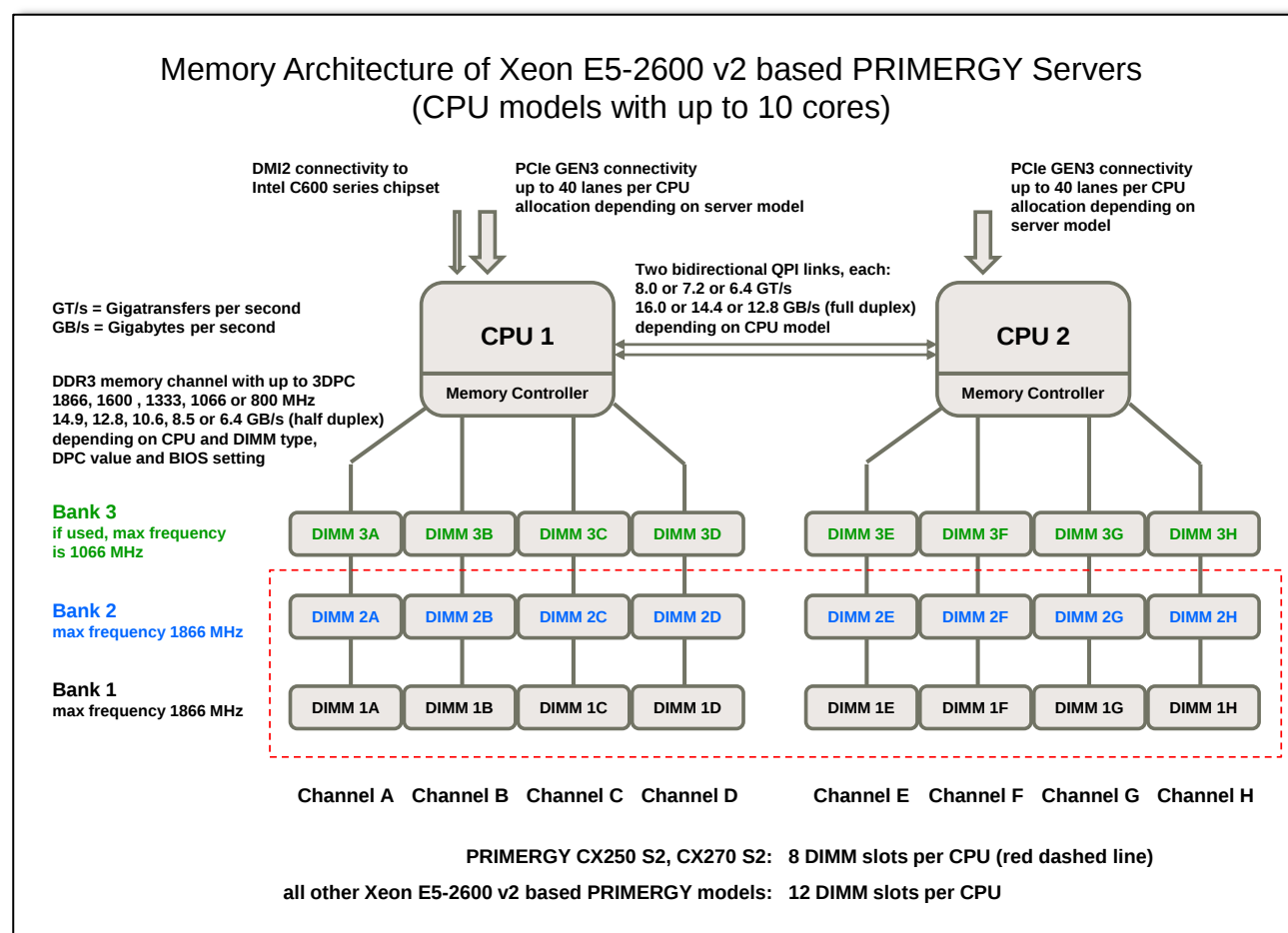
DIMM スロット

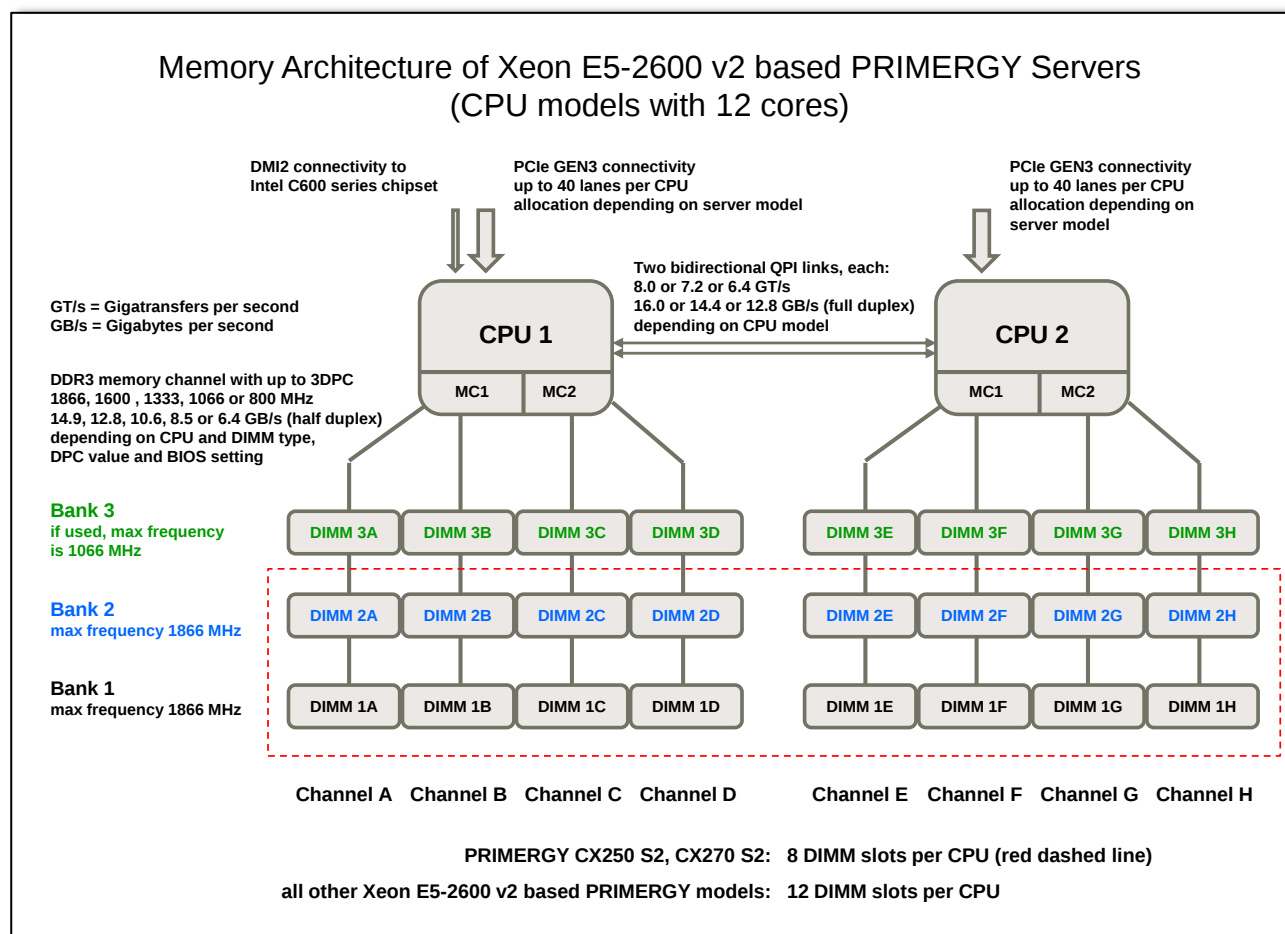
次の図は、メモリシステムの構造を示します。2 つの図の細かな違いについて説明する前に、まず重要な要素が共通している点に注目してみましょう。

Xeon E5-2600 v2 搭載の PRIMERGY サーバは、通常、プロセッサあたり 12 本の DIMM スロットを装備しています。例外は、PRIMERGY CX250 S2 モデルと CX270 S2 モデルで、フォームファクターが高密度であるため、スロット数は 8 となっています。

図でリソースのメモリチャネルと QPI リンクを見ると、周波数と帯域幅の間に接続があり、それぞれのデータパス幅から続いています。DDR3 メモリチャネルには 64 ビット、QPI リンクには 16 ビットがあります。双方向 QPI リンクの場合は、帯域幅が各方向で有効であるため、全二重方式と呼ばれています。メモリチャネルでは、リード/ライトアクセスがデータパスを共有しなければならないため、半二重方式と呼ばれます。

1 つのプロセッサには、常に 4 つのメモリチャネルが存在します。チャネルあたりの DIMM 数が変わると、メモリ周波数に変化が生じ、さらにはメモリパフォーマンスに影響を与えます。この値は、DPC（チャネルあたりの DIMM 数）と呼ばれます。以降、この用語を使用します。チャネルの構成が異なる場合、発生する最大 DPC 値は、メモリ構成が周波数に与える影響の決定的な要因になります。





以降では、「メモリバンク」という用語も使用します。図では、複数のチャンネルに分配されている 4 つの DIMM のグループが、1 つのバンクを形成しています。図中の色（黒、青、緑）は、サーバのシステムボード上で、構成エラーを防ぐためにバンクに付けられたカラーマークに対応します。プロセッサあたりの利用可能なスロット経由で DIMM を分配する場合、バンク 1 から順に割り当てることにより、チャンネル全体で最適なインターリーブが得られます。インターリーブは、メモリパフォーマンスに影響を与える主要な要因です。

DIMM スロットを使用するためには、対応するプロセッサを搭載する必要があります。最大構成がない場合、空の CPU ソケットに割り当てられたスロットは使用できません。

NehalemEP (2009) で QPI を実装したメモリアーキテクチャーが導入されて以来、デュアルソケットサーバ用に設計された-EP プロセッサには、単一のオンチップメモリコントローラーが搭載されています。Xeon E5-2600 v2 でも、最大 10 コアのプロセッサモデルについてはこれが当てはまります（最初の図）。12 コアである Xeon E5-2695 v2 モデルと E5-2697 v2 モデルの場合、2 つのメモリチャンネルに対し、それぞれ 2 つのコントローラーが用意されています（2 つ目の図）。これは、2 つのメモリコントローラーを搭載した 4 -および 8 ソケットサーバ用の EX プロセッサと似ていて、そのトポロジーにより、12 のコアが 2 つのリングで相互接続されています。コアが 12 未満のモデルには、リングが 1 つしかありません。チャンネルインターリーブがパフォーマンスに及ぼす影響についてのセクションでは、この 12 コアを持つ 2 つのプロセッサモデルの特殊な機能を思い出す必要があります。

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。

どの DIMM タイプでも、データは 64 ビット単位で転送されます。これが、DDR3-SDRAM メモリテクノロジーの特徴です。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 または x8 を参照してください)。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。4 ランクまたは 8 ランクの DIMM のメリットは、最大容量である点にあります。同時に DDR3 の仕様はメモリチャネルあたり最大 8 ランクしかサポートしません。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。

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ド内でバッファされます (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) で構成が可能になります。
- 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仮想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仮想ランクを監視するだけです。この機能は、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。

アンバッファード (UDIMM) DIMM は、Xeon E5-2600 搭載の PRIMERGY サーバでも用意されていました。アドレスおよびデータのバッファリングを行わないこの単純な DIMM タイプは、メモリチャネルの静電負荷が高いため、3 DPC での構成をサポートしません。ただし、UDIMM は、構造が単純なため価格的に優位であり、長期にわたって人気を博しました。この優位性は、現在では消失しています。そのため、Xeon E5-2600 v2 搭載の PRIMERGY サーバに対しては UDIMM が提供されず、8GB 2Rx8 PC3L-12800E UDIMM が特別リリースにおいて提供されるだけになっています。

DIMM タイプ (JEDEC/SystemArchitect)	制御	最大周波数 (MHz)	電圧	ランク数	容量	SDDC	GB あたりの 相対価格
4GB 1Rx4 PC3L-12800R 4GB (1x4GB) 1Rx4 L DDR3-1600 R ECC	Registered	1600	1.5/1.35	1	4 GB	ある	1.4
8GB 1Rx4 PC3L-12800R 8GB (1x8GB) 1Rx4 L DDR3-1600 R ECC	Registered	1600	1.5/1.35	1	8 GB	ある	1.0
16GB 2Rx4 PC3L-12800R 16GB (1x16GB) 2Rx4 L DDR3-1600 R ECC	Registered	1600	1.5/1.35	2	16 GB	ある	0.9
8GB 2Rx8 PC3-14900R 8GB (1x8GB) 2Rx8 DDR3-1866 R ECC ¹	Registered	1866	1.5	2	8 GB	ない	1.3
16GB 2Rx4 PC3-14900R 16GB (1x16GB) 2Rx4 DDR3-1866 R ECC	Registered	1866	1.5	2	16 GB	ある	1.2
32GB 4Rx4 PC3L-12800L 32GB (1x32GB) 4Rx4 L DDR3-1600 LR ECC	load reduced	1600	1.5/1.35	4	32 GB	ある	1.6
64GB 8Rx4 PC3L-10600L 64GB (1x64GB) 8Rx4 L DDR3-1333 LR ECC	load reduced	1333	1.5/1.35	8	64 GB	ある	7.9

¹ 他の DIMM タイプとの併用はできません

DIMM の x4 または x8 構造は、エラー（修正できるものと修正できないものを含めて）の検出度に影響します。このため、表内の 8GB 2Rx8 PC3-14900R RDIMM は他の使用可能な RDIMM（すべて x4）と一緒にできません。SDDC（Single Device Data Correction、表の最後から 2 列目を参照）は、x4 モジュールに限定された拡張 ECC 機能であり、DRAM チップ全体の障害を補うことができます。

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、非常に小さいため、たい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。

- DIMM にコンポーネントを追加したために、DIMM タイプの RDIMM と LRDIMM の複雑さが増し、アクセス遅延がわずかに数ナノ秒ほど長くなります。
- 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、ランク乗算を行うと、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わずかに（RDIMM を使った構成と比べて 5 %未満）減少します。

1866 MHz 対応のものを除き、どの DIMM タイプも、1.5 V または省電力の 1.35 V で稼動します。ただし、1.35 V での動作は、メモリ周波数、さらにはメモリパフォーマンスの低下につながる可能性があります。次のセクションでは、メモリ周波数について取り上げ、こうした相互関係にスポットを当てます。Xeon E5-2600 v2 搭載の PRIMERGY サーバは、新しく、3 DPC 構成における省電力の 1.35 V による稼動をサポートしています。先行世代では、3 DPC での構成は 1.5 V でしか動作しませんでした。

特定の構成の有効な周波数は、一連の影響に左右されます。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。

表の最終列は、各 DIMM の価格を相対比で示しています。それぞれの価格は、2013 年 9 月現在の PRIMERGY RX300 S8 の料金表を使用しています。列は、GB あたりの相対価格を示し、1600 MHz に制限された 8GB の RDIMM（測定値 1 として強調表示）に標準化されています。DDR3-SDRAM メモリモジュールの導入以降、相対価格の状況は絶えず変化しています。現時点では、例えば 1866 MHz 対応の RDIMM タイプや LRDIMM など、メモリ容量が大きいもので価格の上昇が見られることがあります。容量が最大で価格も最も高い 64 GB LRDIMM は、Xeon E5-2600 v2 搭載の PRIMERGY サーバで導入され、その最大ストレージ容量は、先行世代の 768 GB から 2 倍の 1.5 TB に増大しています。

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。

メモリ周波数の定義

メモリの周波数には、1866、1600、1333、1066、または 800 MHz の 5 種類があります。システムに電源が入ると、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ごとに適用されます。まず、定義上、構成プロセッサモデルが非常に重要になります。本書では、Xeon E5-2600 v2 モデルを次の表に従って分類することをお勧めします。パフォーマンスに与える影響に関する下の表でも、モデルシリーズは同様に 4 クラスに分類され、プロセッサクラスに従って個別に影響が検討されます。12 コアで最も強力な 2 つのモデルは、Advanced クラスに割り当てられていないメモリコントローラーが 2 つあります。

CPU タイプ	メモリ周波数[MHz]	QPI (GT/s)	Xeon E5-2600 v2 モデル
12 Cores	1866	8.0	E5-2697 v2、E5-2695 v2
Advanced	1866	8.0	E5-2690 v2、E5-2680 v2、E5-2670 v2、E5-2660 v2、E5-2650 v2、E5-2667 v2、E5-2643 v2、E5-2637 v2
Standard	1600	7.2	E5-2640 v2、E5-2630 v2、E5-2620 v2、E5-2650L v2、E5-2630L v2
Basic	1333	6.4	E5-2609 v2、E5-2603 v2

DIMM タイプおよびメモリ構成の DPC 値によっても、周波数が制限されます。プロセッサタイプ、DIMM タイプ、および DPC 値はメモリ周波数に大きな影響を及ぼします。これを BIOS で無効にすることはできません。ただし、BIOS パラメーター「DDR Performance」により、パフォーマンスと消費電力を調整することはできます。パフォーマンスを選択した場合、有効なメモリ周波数は次の表のようになります。

DDR Performance = パフォーマンスに最適化 (デフォルト) グレーの網掛け : 1.5V - 網掛けなし : 1.35V												
	RDIMM 1866 MHz			LV RDIMM 1600 MHz			32 GB LV LRDIMM			64 GB LV LRDIMM		
CPU タイプ	1DPC	2DPC	3DPC	1DPC	2DPC	3DPC	1DPC	2DPC	3DPC	1DPC	2DPC	3DPC
12 Cores	1866	1866	1066	1600	1600	1066	1600	1600	1066	1066	1066	1066
Advanced	1866	1866	1066	1600	1600	1066	1600	1600	1066	1066	1066	1066
Standard	1600	1600	1066	1600	1600	1066	1600	1600	1066	1066	1066	1066
Basic	1333	1333	1066	1333	1333	1066	1333	1333	1066	1066	1066	1066

次の表は、低消費電力、1.35 V の低電圧での動作を優先する場合に有効です。

DDR Performance = 低電圧に最適化 グレーの網掛け : 1.5V - 網掛けなし : 1.35V												
	RDIMM 1866 MHz			LV RDIMM 1600 MHz			32 GB LV LRDIMM			64 GB LV LRDIMM		
CPU タイプ	1DPC	2DPC	3DPC	1DPC	2DPC	3DPC	1DPC	2DPC	3DPC	1DPC	2DPC	3DPC
12 Cores	1866	1866	1066	1333	1333	800	1600	1600	1066	1066	1066	1066
Advanced	1866	1866	1066	1333	1333	800	1600	1600	1066	1066	1066	1066
Standard	1600	1600	1066	1333	1333	800	1600	1600	1066	1066	1066	1066
Basic	1333	1333	1066	1333	1333	800	1333	1333	1066	1066	1066	1066

当然ながら、PRIMERGY CX250 S2 と CX270 S2 はスロットがないため、3 DPC の列は該当しません。

1866 MHz で 2 DPC がサポートされるのは、Xeon E5-2600 v2 搭載の PRIMERGY サーバの特長です。Intel の仕様では、2 DPC における最大許容限度は 1600 MHz です。PRIMERGY モデルまたは DIMM タイプによっては、このサポートにより遅延が生じる場合があります。BIOS バージョンの最新のコンフィギュレーターとデータシートが決定的な要素となります。

メモリパフォーマンスは、次の 3 つ目の構成で最も低くなります。

	DDR Performance = 電力に最適化 グレーの網掛け 1.5V – 網掛けなし : 1.35V											
	RDIMM 1866 MHz			LV RDIMM 1600 MHz			32 GB LV LRDIMM			64 GB LV LRDIMM		
CPU タイプ	1DPC	2DPC	3DPC	1DPC	2DPC	3DPC	1DPC	2DPC	3DPC	1DPC	2DPC	3DPC
12 Cores	800	800	800	800	800	800	800	800	800	800	800	800
Advanced	800	800	800	800	800	800	800	800	800	800	800	800
Standard	800	800	800	800	800	800	800	800	800	800	800	800
Basic	800	800	800	800	800	800	800	800	800	800	800	800

低電圧での動作は、低消費電力という点では大きな決定的要因ですが、メモリ周波数はそれほど低下しません。そのため、電力に最適化する設定はそれほど魅力的ではありません。メモリパフォーマンスが確実に低下する一方で、1.35 V での動作による消費電力の削減に加えてさらに消費電力を削減できるかどうかは確実ではないからです。原則として、*[Low-voltage optimized]* (定電圧に最適化) は、パフォーマンスよりエネルギー効率を重視する場合に使用してください。

機能の説明は、その多くがメモリ周波数に関連しています。アプリケーションパフォーマンスに対するメモリ速度の影響に関する数値を後で示します。

BIOS パラメーター

BIOS の [Advanced / Memory] に、次のパラメーターを使用したメモリ構成に関するサブメニューがあります。

- Memory Mode : Independent / Mirroring / Sparing
- NUMA : Enabled / Disabled
- DDR Performance : Low-voltage optimized / Energy optimized / Performance optimized
- DRAM Maintenance : Disabled / Auto

1 つ目のパラメーターは、冗長性機能に関連しています。SystemArchitect で構成中にこれらの機能が要求された場合、工場で適切なデフォルト設定が行われます。それ以外の場合、パラメーターは *Independent* (冗長性なし) に設定されます。冗長性機能がシステムパフォーマンスに与える影響に関する数値を後で示します。

物理アドレス空間をローカルメモリのセグメントから構築するか、またオペレーティングシステムに構造を通知するかを NUMA パラメーターで定義します。デフォルト設定は *Enabled* です。明確な理由がないこれを限り変更しないでください。

3 つ目のパラメーター *DDR Performance* については、すでに前のセクションで詳しく説明しました。

4 つ目のパラメーター *DRAM Maintenance* のデフォルト設定は *Auto* (自動) です。明確な理由がない限りこれを変更しないでください。Xeon E5-2600 v2 搭載の PRIMERGY サーバに対して提供されているメモリタイプの場合は、デフォルトの *Auto* (自動) に設定しておくで最適なメモリパフォーマンスが実現されます。また、これらのメモリタイプについては、デフォルト設定の変更につながる次のような理由が存在しません。

「*DRAM Maintenance = Disabled*」という設定は、Xeon E5-2600 搭載のプロセッサ世代など、旧型のメモリタイプにのみ有効です。これらのメモリタイプでは、アクセスパターンによっては修正可能なメモリエラーが蓄積される「パスゲート効果」が生じるためです。パスゲート効果を排除するために、「*DRAM Maintenance = Auto*」というデフォルト設定が、これらのメモリタイプに対して DRAM のリフレッシュレートを 2x (2 倍)、それ以外に対して 1x (1 倍) に設定します。2 倍でのリフレッシュを行うと、パフォーマンスは 2 %ほど低下します。修正可能なメモリエラーが蓄積される可能性を許容した上で、パフォーマンスの低下を防ぎたい場合は、「*DRAM Maintenance = Disabled*」とすることでリフレッシュレートを 1x に設定できます。

「*DRAM Maintenance = Disabled*」に設定すると、次のサブメニューが表示されます。

- Patrol Scrub : Disabled / Enabled
- Fast Patrol Scrub : Disabled / Enabled
- Refresh Rate Multiplier : 1x / 2x / 3x / 4x

サブメニューが開いた時点では、最後のパラメータは構成のメモリタイプにかかわらず常に 2x となっています。これには、前述したような意味があります。「*DRAM Maintenance = Disabled*」という設定を使用する場合は、ユーザーが責任をもってパラメータを正しく設定しなければなりません。

Patrol Scrub パラメータは、あらかじめ *Enabled* に設定されています。メインメモリでは、24 時間周期で修正可能なメモリエラーが検索されます。これは、DRAM の標準機能であり、パスゲート効果の有無に依存しません。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。

Fast Patrol Scrub はあらかじめ *Disabled* に設定されています。この場合、*Enabled* という設定により、パスゲート効果のさらなる防止策として検索周期がおおよそ 30 分に短縮されます。ただし、PRIMERGY サーバを徹底的にテストした結果、2 倍を超えるリフレッシュの必要性が確認されなかったため、デフォルト設定は *Disabled* となっています。

Xeon E5-2600 v2 搭載の PRIMERGY サーバの BIOS パラメーターのうち、パフォーマンスに関連するものについては、[関連資料 6] に説明があります。

パフォーマンスを考慮したメモリ構成

ここで説明する 3 つのメモリ構成は、PRIMERGY コンフィギュレータの用語に基づいています。1 つ目の構成は、理想的な方法でメモリシステムのトポロジーを利用し、最高のメモリパフォーマンスを提供する場合に適用します。コンフィギュレータは、これをパフォーマンスモード構成として参照します。

パフォーマンスモード構成

このモードでの構成は、同タイプの 4 枚 1 組の DIMM によるバンク単位で構成されており、プロセッサの 4 つのメモリチャネルがすべて同等に扱われます。メモリアクセスは、メモリシステムのこれらのリソースに均等に分散されます。技術的に言えば、メモリチャネル経由で最適な 4-WAY インターリーブが実現します。

Xeon E5-2600 v2 搭載の PRIMERGY サーバのパフォーマンスモード構成							
2 CPU システム	DIMM タイプ	DIMM 容量 GB バンク 1	DIMM 容量 GB バンク 2	DIMM 容量 GB バンク 3	パフォーマンスに最適化した場合の最大 MHz	低電圧に最適化した場合の最大 MHz	注
32 GB	PC3L-12800R	4			1600	1333	1WAY ランクインターリーブ (-)
64 GB	PC3L-12800R	4	4		1600	1333	2WAY ランクインターリーブ (+)
	PC3L-12800R	8			1600	1333	1WAY ランクインターリーブ (-)
	PC3-14900R	8			1866		1866 MHz (+) 2WAY ランクインターリーブ (+)
96 GB	PC3L-12800R	4	4	4	1066	800	3 DPC が MHz (-) を減少 1WAY ランクインターリーブ (-)
	PC3L-12800R	8	4		1600	1333	1WAY ランクインターリーブ (-)
128 GB	PC3L-12800R	8	8		1600	1333	2WAY ランクインターリーブ (+)
	PC3L-12800R	16			1600	1333	2WAY ランクインターリーブ (+)
	PC3-14900R	8	8		1866		1866 MHz (+) 4WAY ランクインターリーブ (++)
	PC3-14900R	16			1866		1866 MHz (+) 2WAY ランクインターリーブ (+)
192 GB	PC3L-12800R	16	8		1600	1333	1WAY ランクインターリーブ (-)
256 GB	PC3L-12800R	16	16		1600	1333	4WAY ランクインターリーブ (-)
	PC3-14900R	16	16		1866		1866 MHz (+) 4WAY ランクインターリーブ (++)
	PC3L-12800L	32			1600	1600	低電圧、1600 MHz (+)
384 GB	PC3L-12800R	16	16	16	1066	800	PC3-14900R のメリットなし (3 DPC の周波数は同じになる)
512 GB	PC3L-12800L	32	32		1600	1600	低電圧、1600 MHz (+) 4WAY ランクインターリーブ (++)
768 GB	PC3L-12800L	32	32	32	1066	1066	ランク乗算 (-)
1024 GB	PC3L-10600L	64	64		1066	1066	ランク乗算 (-)
1536 GB	PC3L-10600L	64	64	64	1066	1066	ランク乗算 (-)

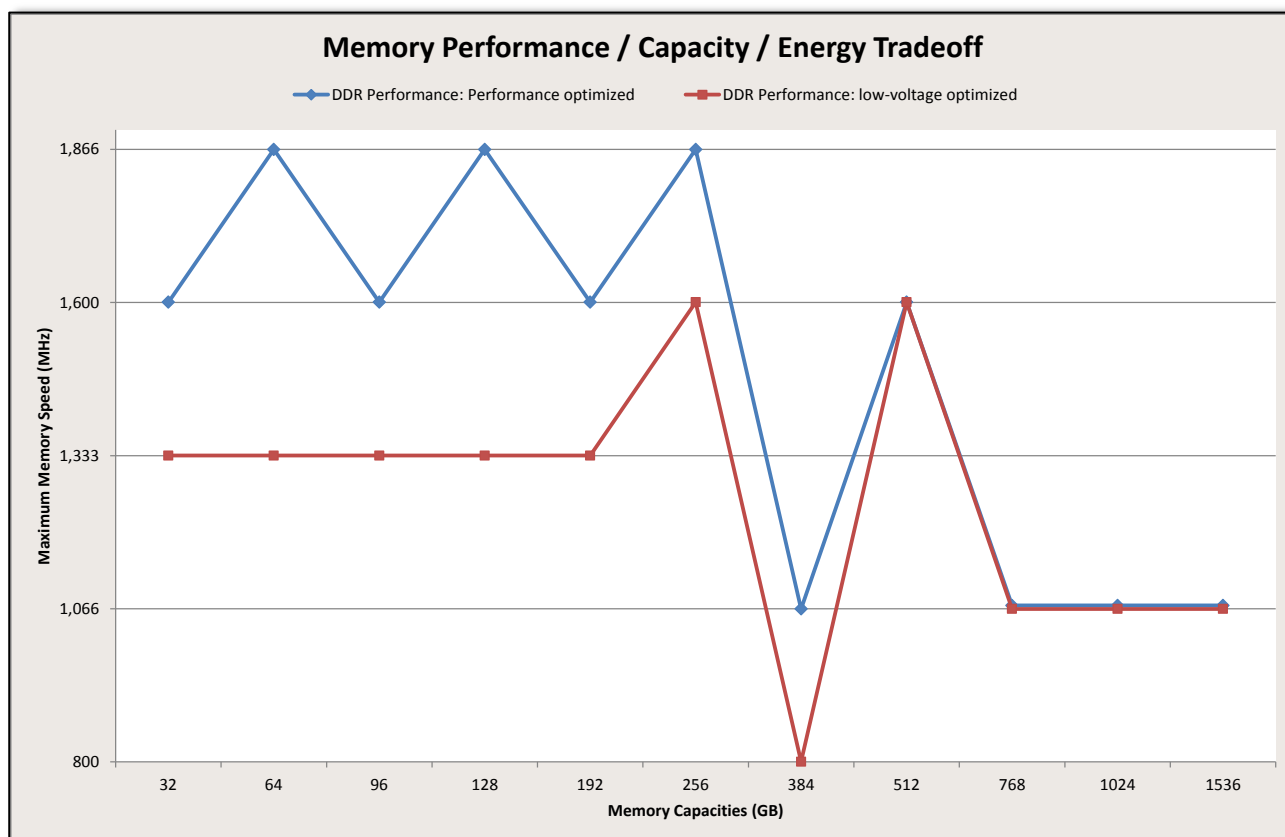
この表は、パフォーマンスモードにおける構成を、代表的なメモリ容量に基づいて一覧にしたものです。表では、2つのプロセッサが、同じメモリ構成で構成されていると想定しています。メモリバンクがそれぞれプロセッサごとに同色の DIMM スロット 4 本で構成されていることから、例えば両プロセッサの最初のバンク（8 本の DIMM スロット）が 4 GB の DIMM で構成されていると、合計は 32 GB となります。DIMM タイプについては、Xeon E5-2600 v2 搭載の PRIMERGY サーバ用にリリースされている表を参照してください。DIMM タイプと容量の組み合わせは、個々のケースによって異なります。

表の値は、1.5 V および省電力 1.35 V での動作時に、その設定で達成可能な最大メモリ周波数を示しています。ただし、これらの最大周波数が実際に達成されるかどうかは、プロセッサモデルによって異なります。例えば、1866 MHz という周波数は、12 Cores（12 コア）または Advanced（アドバンスド）のモデルクラスのプロセッサを使用した場合にしか実現されません。

表の最後の列は、メモリパフォーマンスの仕上がりという観点から見た相対的な評価を示します。前述の機能の詳細と、影響を表す数値（通常は 5 パーセント以下）については、以下を参照してください。「++」という評価は、メモリサイズが 128、256、512 GB である 3 つの構成に与えられています。これらの構成は、Xeon E5-2600 v2 搭載の PRIMERGY サーバの標準的なベンチマークで使用されます。

次の図は、メモリ容量、低消費電力、最大限のメモリパフォーマンスのトレードオフをメモリ周波数ごとにグラフにしたものです。この図から、容量と低消費電力のために、メモリパフォーマンスがある程度犠牲になることがわかります。ただし、メモリ周波数のアクセシビリティは、構成するプロセッサのタイプによって左右されることも忘れないでください。

青線がジグザグに上下しているのは、1866 MHz 対応の PC3-14900R DIMM では 4 GB というサイズがサポートされていないことと、8 GB および 16 GB の PC3-14900R DIMM が併用可能であることによります。256~512 GB という大容量のメモリ構成において、小容量の構成より高いメモリ周波数で省電力の低電圧（LV）運用が可能であるという事実は、これらの構成で使用する *ロードリデュースド*（LRDIMM）の特性に起因します。小容量の構成では、RDIMM が使用されます。



独立モード構成

これには、パフォーマンスモードや冗長モードには含まれない構成がすべて含まれます。RDIMM と LRDIMM、

- タイプ x4 と x8 の RDIMM
- を混在させることはできない

というルールを除けば、特に制限はありません。

プロセッサあたりの DIMM が 4 つ未満の構成、つまりパフォーマンスモード構成に必要な最小数未満の構成にも特に注意する必要があります。メモリ容量が非常に小さいというだけでなく、さらなる電力の削減を考慮して、このように構成される場合があります。低消費電力は、単に 1.35 V での動作や特定のメモリ構成の周波数の低減だけでなく、DIMM 数を最小限に抑えることでも実現します。この後で紹介する、3 つ以下のメモリチャネルを使った構成がシステムパフォーマンスに及ぼす影響を示した定量的評価から、次のような事項が推奨されます。

- *Advanced* (アドバンスド)、*Standard* (スタンダード)、*Basic* (ベーシック) のプロセッサクラスの場合、プロセッサあたりの DIMM 数を 1 つとした動作は推奨されません。プロセッサあたりの DIMM 数を 2 または 3 とすると、パフォーマンスと電力消費量のバランスが取れた結果になります。
- 12 Cores (12 コア) のプロセッサクラスの場合、プロセッサあたりの DIMM 数を 1 または 3 とした動作は推奨されません。プロセッサあたりの DIMM 数を 2 とすると、パフォーマンスと電力消費量のバランスが取れた結果になります。

推奨されない構成は、全体 (プロセッサあたり 1 DIMM) または部分 (12 Cores のプロセッサの場合、プロセッサあたり 3 DIMM) ということになります。以下のとおり、メモリチャネルを介した 1-WAY インターリーブには、商用アプリケーションのパフォーマンスが最大 30 % も低下するデメリットがあります。12 Cores のプロセッサの特長は、2 つのメモリコントローラーを使った構成で、それぞれのコントローラーに 2 つのメモリチャネルがあることです。

対称型メモリ構成

最後のこのセクションでは、すべてのプロセッサのメモリを可能な限り同等に構成し、BIOS の *NUMA = enabled* のデフォルト設定を確たる理由なく変更するべきではないということに再度焦点を当てます。このように考慮されるのは、システムの QPI を実装したマイクロアーキテクチャーのみです。

工場でのプレインストールでは、このような状況が当然考慮されています。要求されたメモリモジュールは、プロセッサ全体に可能な限り均等に分散されます。

こうした手法と、関連するオペレーティングシステムによって、ローカルのハイパフォーマンスメモリで可能な限りアプリケーションを実行する前提条件が整備されます。プロセッサコアのメモリアクセスは、通常、各プロセッサに直接割り当てられた DIMM モジュールに対して行われます。これにどのようなパフォーマンス上のメリットがあるのかを見積もるため、2-WAY サーバのメモリが対称型に構成されているものの、BIOS オプションが *NUMA = disabled* に設定されている場合の測定結果を以下のリストに示します。統計上、各 2 次メモリアクセスは、リモートメモリに対して行われます。アプリケーションが 100 % リモートメモリによって実行される非対称型メモリ構成、または片側メモリ構成では、50 %/50 % の場合の 2 倍パフォーマンスが低下するものとして見積もる必要があります。

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、メモリ構成に関係するパフォーマンスの向上と低下について説明します。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の使用する 2 つのベンチマークについて説明します。

その後、すでに説明した特徴であるメモリチャネルのインターリーブ、メモリ周波数、およびランクのインターリーブについて、その影響の大きさの順に説明します。最後に、NUMA = disabled で、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測定します。

定量的テストは、プロセッサクラスの 12 Cores (12 コア)、Advanced (アドバンスド)、Standard (スタンダード)、Basic (ベーシック) のそれぞれに対して個別に実施しています。測定は、Linux オペレーティングシステムの 2 つのプロセッサを搭載した PRIMERGY BX924 S4 で行われました。次の表は、測定対象となった構成、特に 4 つのプロセッサクラスの代表的な構成を示します。

SUT (System Under Test : テスト対象システム)	
ハードウェア	
モデル	PRIMERGY BX924 S4
プロセッサ	Xeon E5-2695 v2 (12 Cores) × 2 Xeon E5-2670 v2 (Advanced) × 2 Xeon E5-2630 v2 (Standard) × 2 Xeon E5-2603 v2 (Basic) × 2
メモリタイプ	4GB (1x4GB) 1Rx4 L DDR3-1600 R ECC 8GB (1x8GB) 2Rx8 L DDR3-1866 R ECC 16GB (1x16GB) 2Rx4 L DDR3-1600 R ECC
ディスクサブシステム	HD SATA 6G 1TB 5.4Krpm ×1 (SATA/SAS 用オンボードコントローラーを介して)
ソフトウェア	
BIOS	0.90
オペレーティングシステム	Red Hat Enterprise Linux Server release 6.4

次の表は、相対的なパフォーマンスを示します。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、次の表で 1.0 の値に相当しますが、Xeon E5-2600 v2 搭載 PRIMERGY サーバの個別のパフォーマンスレポートに記載されています。

このテストから得られた 1 つの重要な結果を、最初に明らかにしておきましょう。使用するプロセッサモデルが強力であるほど、パフォーマンスへの影響が顕著になり、構成の詳細をより慎重に検討する必要があります。Advanced (アドバンスド) クラスの最も強力で高価なプロセッサについて不可欠な考慮事項も、Basic (ベーシック) クラスについては、多くの場合無視できます。

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。以降、STREAM のベンチマークの測定値は、Triad アクセスでの値であり、単位は GB/s 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。メモリ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一般的な予測を立てることは非常に困難です。

SPECint_rate_base2006

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。

メモリチャネルへのインターリーブ

インターリーブは、最初のブロックは最初のチャネルに、2 番目のブロックは 2 番目のチャネルにという具合に、プロセッサ単位で 4 つのメモリチャネルを交互に利用するように物理アドレス領域を設定する手法です。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、結果としてすべてのチャネルに分散されます。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。チャネルインターリーブのブロックサイズは、64 バイトというキャッシュラインサイズに基づいています。キャッシュラインサイズは、プロセッサの観点におけるメモリアクセスの単位です。

次の表は、パフォーマンスモードでのメモリ構成によって実現される理想的な 4-WAY インターリーブを行わない場合のパフォーマンスの低下を示しています。この表では、すでに取り上げた、プロセッサが強力になるほどパフォーマンスへの影響が顕著になるという事実が示されています。

ベンチマーク	CPU 種類	4-way	3-way	2-way	1-way
STREAM	12 Cores	1.00		0.52	0.26
	Advanced	1.00	0.78	0.54	0.27
	Standard	1.00	0.80	0.55	0.28
	Basic	1.00	0.91	0.75	0.39
SPECint_rate_base2006	12 Cores	1.00		0.89	0.69
	Advanced	1.00	0.98	0.93	0.76
	Standard	1.00	0.99	0.96	0.84
	Basic	1.00	1.00	0.99	0.94

SPECint_rate_base2006 に関する評価は、商用アプリケーションのパフォーマンスに関するものです。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。

2-WAY および 3-WAY インターリーブのパフォーマンスの低下が穏やかなのは、それなりの理由があります。必要となるメモリ容量が少ないか、または低消費電力のために DIMM 数が最小限に抑えられるためです。1-WAY インターリーブは推奨できません。これは厳密に言うとインターリーブではなく、分類上そのようにと呼ばれているだけです。この場合、プロセッサとメモリシステムの潜在的なパフォーマンスのバランスが取れていません。

12 Cores クラスのプロセッサは、2 つのメモリコントローラーを使った構成により、3-WAY インターリーブはサポートしていません。そのため、これらのプロセッサモデルでプロセッサあたり 3 つの DIM を構成した場合にどうなるか、がポイントとなります。

これは、物理アドレス領域を、インターリーブが異なる領域に分割する必要があるケースの一例です。分割が必要なその他の例としては、メモリチャネルあたりの部分容量 (チャネルあたりの GB) が異なる構成が挙げられます。このような構成は、サイズの異なる DIMM を使った場合や、同サイズの DIMM を 5 つ以上使った場合に生じることがあります。すべての例に共通するのは、メモリチャネル間を切り替える方法で標準化アドレス領域セグメントを設定することができない点です。切り替えは常に「均等に」行われます。このようなケースでは、既存の DIMM をグループ化することにより、可能なかぎり高いインターリーブでセグメントが生成されます。次の表は、その例を 2 つ示しています。

アプリケーションのメモリパフォーマンスは、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。パフォーマンスが特に重要なケースでは、分割の必要な構成は避けた方がいい場合もあります。

分割が必要な DIMM 構成の例 (CPU あたり)	アドレス領域セグメント	サイズ/インターリーブ
1-1-1-0 (12 Cores CPU)	1-0-1-0	アドレス領域の 66 %/2WAY
	0-1-0-0	アドレス領域の 33%/1WAY
2-1-1-1	1-1-1-1	アドレス領域の 80%/4WAY
	1-0-0-0	アドレス領域の 20%/1WAY

メモリ周波数

有効なメモリ周波数への影響は、これまでのセクションで詳しく説明しています。省電力 (BIOS パラメータ *DDR Performance* によって制御) と大型メモリ (3DPC 構成、最大 1600 または 1333 MHz の LRDIMMS を使用) が原因で、有効な周波数がプロセッサタイプによってサポートできるレベルより低くなる場合があります。

次の表は、影響を比較し、バランスを取る際に役立ちます。最初の表の数値は、一連のすべての測定に共通の最小メモリ周波数 800 MHz に基づきます。2 番目の表は、同じ情報を異なる観点から捉えたものです。数値は、理想的なケース、つまりそのプロセッサクラスで最大限の周波数に基づいています。

ベンチマーク	CPU 種類	800 MHz	1066 MHz	1333 MHz	1600 MHz	1866 MHz
STREAM	12 Cores	1.00	1.33	1.63	1.92	2.17
	Advanced	1.00	1.32	1.62	1.90	2.15
	Standard	1.00	1.32	1.59	1.84	
	Basic	1.00	1.13	1.22		
SPECint_rate_base2006	12 Cores	1.00	1.07	1.12	1.17	1.19
	Advanced	1.00	1.07	1.11	1.15	1.17
	Standard	1.00	1.05	1.08	1.11	
	Basic	1.00	1.03	1.05		

ベンチマーク	CPU 種類	800 MHz	1066 MHz	1333 MHz	1600 MHz	1866 MHz
STREAM	12 Cores	0.46	0.61	0.75	0.88	1.00
	Advanced	0.47	0.61	0.75	0.88	1.00
	Standard	0.54	0.72	0.86	1.00	
	Basic	0.82	0.93	1.00		
SPECint_rate_base2006	12 Cores	0.84	0.90	0.94	0.98	1.00
	Advanced	0.85	0.91	0.95	0.98	1.00
	Standard	0.90	0.94	0.97	1.00	
	Basic	0.95	0.98	1.00		

BIOS で「DDR Performance = *Energy optimized* (電力に最適化)」と設定すると、周波数は常に 800 MHz になります。ただし、「DDR Performance = *Low-voltage optimized* (低電圧に最適化)」に設定したときよりも消費電力をさらに削減できる可能性は非常に低くなっています。それは、電力の削減が、メモリ周波数を下げるのではなく、電圧を 1.35 V にすることで主に達成されるためです。これが、*Energy optimized* (電力に最適化) という設定をお勧めしない理由です。*Low-voltage optimized* (低電圧に最適化) という設定では、周波数は 800~1600 MHz になります。

メモリ周波数の低下がメモリ容量に関係している場合、万全を期すために 1 つの問題について説明する必要があります。例えば、I/O 速度など、メモリ容量はアプリケーションのパフォーマンスに対して暗黙の影響を及ぼします。こういった影響は、当然ながらこのセクションで基にしているテストでは考慮していません。表で比較すると、メモリ周波数の違いだけがパフォーマンスに影響を及ぼしています。

メモリランクへのインターリーブ

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、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。

ランクインターリーブは、アドレスビットを経由して直接制御されます。3-WAY ケースの確立のためにチャネルインターリーブで実行されるビット演算は行われません。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、常に 1-WAY インターリーブとなりますが、これは分類上そう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、チャネルでのインターリーブより大きくなります。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。

この表は、4-WAY インターリーブに関するものです。このケースは、PRIMERGY サーバの最も標準的なベンチマークによるものです。大型 RDIMM の 2DPC 構成は、通常、メモリ容量とパフォーマンスが最適なバランスになります。8-WAY インターリーブは、LRDIMM でのみ可能ですが、4-WAY インターリーブと比較しても測定可能な向上が見られないため、省略しています。

ベンチマーク	CPU 種類	4-way	2-way	1-way
STREAM	12 Cores	1.00	1.00	0.90
	Advanced	1.00	0.99	0.91
	Standard	1.00	1.00	0.92
	Basic	1.00	0.99	0.95
SPECint_rate_base2006	12 Cores	1.00	0.99	0.96
	Advanced	1.00	0.99	0.97
	Standard	1.00	1.00	0.98
	Basic	1.00	1.00	0.99

2-WAY および 4-WAY ランクインターリーブは、非常に優れたメモリパフォーマンスを実現します。最終的なパフォーマンスを考えた場合、4-WAY インターリーブにのみわずかなメリットがありますが、通常は無視できる程度です。ただし、1-WAY のケースは、単一ランク 4 GB または 8 GB RDIMM の 1 DPC 構成時に発生します。この場合のパフォーマンスの低下については十分考慮してください。特にパフォーマンスが重視され、強力なプロセッサモデルを使用している状況では、これを回避する必要があります。

Xeon E5-2600 v2 プロセッサの DDR3 メモリコントローラーは、メモリチャネルあたり最大で 8 つのランクをサポートします。物理ランクが 9 以上ある LRDIMM を使った構成では、この DIMM タイプのランク乗算機能が、物理ランクを仮想ランクにマップします。減衰因子が 2 または 4 に設定され、最大仮想ランク数の 6 が達成されます。メモリコントローラーがその仮想ランクを制御して、ランクインターリーブが行われます。

リモートメモリへのアクセス

前述の STREAM および SPECint_rate_base2006 ベンチマークを使ったテストでは、ローカルメモリのみが対象になっていました（プロセッサが自身のメモリチャネルの DIMM モジュールにアクセスする）。隣接するプロセッサのモジュールはまったくアクセスされないか、まれに QPI リンクを経由してアクセスされるのみです。実際のアプリケーションにおいて、オペレーティングシステムやシステムソフトウェアの NUMA サポートによってアクセスできるメモリの大半がローカルメモリである限り、この状況は代表的なものであると言えます。

次の表は、プロセッサあたりの最大メモリ周波数を下回る RDIMM での 4-WAY ランクインターリーブパフォーマンスモードという理想的なメモリ構成でありながら、BIOS 設定が「NUMA = disabled」に設定されている場合の影響を示しています。各 2 次メモリアクセスがリモート DIMM、つまり隣接プロセッサに割り当てられた DIMM に対して行われ、データが QPI リンク経由で迂回する必要があるために、パフォーマンスの低下が発生します。

NUMA = disabled では、プロセッサの詳細メッシュ切り替えによって物理アドレス空間を設定しています。この切り替えは、両プロセッサが同一のメモリ容量であることを前提にしています。こうした一般的な状態が存在しない場合、アドレス空間はソケット間インターリーブが許可される主要部分と、プロセッサ - ローカルの残りの部分に分割されます。

システムソフトウェアまたはシステム関連ソフトウェアで NUMA がサポートされていないか、または十分でないため、NUMA = disabled に関する測定は、設定が推奨される例外的なケースとして、狭い範囲で実施しました。上記の測定はすべて、大部分あるいはすべてのアクセスがリモートメモリに対して行われる場合の影響を見積もる場合に役立ちます。プロセッサがメモリなしで構成されている場合、またはプロセッサあたりに構成されたメモリ容量が大幅に異なる場合に、こうしたケースが発生する可能性があります。ローカルアクセスと比較したパフォーマンスの低下は、表に示した低下分の最大 2 倍になることがあります。

この表では、特に STREAM で測定したメモリ帯域幅において、低下率がこれまでの表のように規則的ではありません。プロセッサが強力であればあるほど低下率が高くなる、という経験則は、ここでは当てはまりません。これは、プロセッサモデルの QPI 周波数が 6.4~8.0 GT/s と異なることの影響に起因します。QPI リンクは、NUMA サポートを無効にした場合の帯域幅測定のボトルネックです。

ベンチマーク	CPU 種類	NUMA = enabled	NUMA = disabled
STREAM	12 Cores	1.00	0.81
	Advanced	1.00	0.84
	Standard	1.00	0.79
	Basic	1.00	0.77
SPECint_rate_base2006	12 Cores	1.00	0.94
	Advanced	1.00	0.94
	Standard	1.00	0.94
	Basic	1.00	0.95

冗長性を考慮した際のメモリパフォーマンス

Xeon E5-2600 v2 搭載 PRIMERGY サーバには、冗長オプションが 2 つあります。ミラーリングの場合、プロセッサの 4 つのすべてのメモリチャネルが構成されますが、2 つのチャネルが他の 2 つのチャネルをミラーリングします。オペレーティングシステムは、実際に構成されているメモリの 50 %を利用できます。スペアリング、厳密に言うとランクスペアリングの場合、メモリチャネルあたり 1 つのランクは未使用で、蓄積されるメモリエラーによってアクティブランクが失敗するケースに備えて予約されています。オペレーティングシステムが利用できる実質メモリ容量は、この場合 DIMM タイプおよび DPC 値によって異なります。次の表では、それぞれのケースで最大メモリ周波数を下回る RDIMM での 4-WAY ランクインターリーブパフォーマンスモードという理想的なメモリ構成でありながら、冗長オプションが有効化されている場合の影響を示しています。

前述の「メモリランクへのインターリーブ」で示したとおり、スペアリングの列は、1-WAY ランクインターリーブと同じです。これは、予約ランクは通常、奇数のアクティブランク数になるためです。一方、上の「メモリチャネルへのインターリーブ」のセクションで示したように、ミラーリング列は 2-WAY インターリーブと同じではありません。これは、ミラーの半分が両方とも読み取りアクセスに利用されるためです。低下は、2-WAY チャネルインターリーブの場合より少なくなります。

ベンチマーク	CPU 種類	冗長性なし	ミラーリング	スペアリング
STREAM	12 Cores	1.00	0.71	0.90
	Advanced	1.00	0.74	0.91
	Standard	1.00	0.76	0.92
	Basic	1.00	0.93	0.95
SPECint_rate_base2006	12 Cores	1.00	0.97	0.96
	Advanced	1.00	0.98	0.97
	Standard	1.00	0.99	0.98
	Basic	1.00	1.00	0.99

関連資料

PRIMERGY システム

[関連資料 1] <http://jp.fujitsu.com/primergy>

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=a344b05e-2e9d-481b-8c9b-c6542defd839>

 <http://docs.ts.fujitsu.com/dl.aspx?id=9c2f6975-753c-47b8-821a-b54caf030979>

 <http://docs.ts.fujitsu.com/dl.aspx?id=43d136df-46f6-443f-9f79-56466dadd1d>

[関連資料 3] Xeon E5-2600 (Sandy 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=0e5b6403-e2f0-4ead-b1d6-ab3e1d15c11b>

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>

BIOS 設定

[関連資料 6] Xeon E5-2600 v2 搭載システムのための BIOS 最適化

<http://docs.ts.fujitsu.com/dl.aspx?id=c9c496f2-3cc9-4d8d-8e34-64d997bd0a86>

PRIMERGY のパフォーマンス

[関連資料 7] <http://jp.fujitsu.com/platform/server/primergy/performance/>

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>