

WHITE PAPER

Xeon 5500 (Nehalem EP) 搭載 PRIMERGY サーバの メモリパフォーマンス

バージョン 1.0
2009 年 8 月

ページ数 17

要約

Xeon 5500 番台 (Nehalem EP) 搭載 PRIMERGY デュアルソケットモデルの優れたパフォーマンスを実現した主な原動力は、メインメモリを従来のフロントサイドバス (FSB) ではなく、QuickPath インターコネクト (QPI) で接続するというパラダイム転換によるものです。この新しいアーキテクチャーで、可能な限り強力なシステムを構成しようとする場合、いくつかの新しいパラメータについて検討する必要があります。中心になる考え方は、800、1066、および 1333 MHz という異なるメモリ周波数であり、さらには各プロセッサの 3 つのメモリチャネルにメモリモジュールを可能な限り等分に分散するという点です。このホワイトペーパーでは、これらの要因がパフォーマンスに及ぼす効果を説明し、強力かつ低コストな構成の定義に役立つ情報を提供します。

目次

はじめに.....	2
メモリアーキテクチャー	3
パフォーマンスを考慮した構成	5
メモリパフォーマンスに対する影響	10
ベストプラクティスのルール	16
関連資料.....	17
お問い合わせ先.....	17

はじめに

インテル Xeon 5500 (Nehalem EP) プロセッサとインテル 5520 チップセット搭載の最新のデュアルソケット PRIMERGY ラック、タワー、およびブレードサーバが、飛躍的なパフォーマンスの進歩を遂げています (前バージョンの 2 倍)。進歩を示すテスト結果は、さまざまな負荷プロファイルで確認され、さまざまなベンチマークで実証されています。このようなパフォーマンスの向上は、主に、プロセッサと他のシステムコンポーネント、特にプロセッサとメインメモリ間の接続方法が改善された結果によるものです。システムインターコネクトのリンクは、インテル Pentium Pro プロセッサ (1995 年) 以来、x86 搭載のサーバ内のフロントサイドバス (FSB) を使用して実装されてきました。近年、FSB テクノロジーは、例えばチップセット内で FSB ごとに必要なピン数などの点で、複雑さにおいて限界に達していました。そこで Xeon 5500 では、インテル QuickPath インターコネクト (QPI) という新しいアプローチが導入されています。QPI では、システムアーキテクチャーにおける対称型マルチプロセッシング (SMP) から NUMA (Non-Uniform Memory Access : 非均等型メモリアクセス) へのパラダイム転換が実現されています。

QPI は、プロセッサモデルに応じて、6.4、5.9、または 4.8 GT/s (ギガトランスファー/秒) の片方向のシリアルリンク経由で、プロセッサ同士およびプロセッサと I/O を管理するチップセット間を接続します。Xeon 5500 番台のプロセッサには、メインメモリをリンクするためのメモリコントローラーが内蔵されており、割り当てられたメモリモジュールを各プロセッサが直接制御します。プロセッサは、QPI リンク経由でメモリの内容を近接のプロセッサに提供し、同時に近接のプロセッサからの情報を要求します。

プロセッサとメモリが直接接続されているため、メモリパフォーマンスの向上が可能です。ローカル要求とリモート要求のパフォーマンスが異なるため、このアーキテクチャーは NUMA に分類するのが妥当であるといえます。オペレーティングシステムは、物理メモリの割り当て時と、プロセスのスケジューリング時に、NUMA を考慮します。RAM の総容量は、可能な限り 2 つのプロセッサに等分に分配する必要があります。

このルールを前提とすると、メモリシステム機能から生じる他の多様な項目を考慮する必要があります。これによりメモリは、1333、1066、または 800 MHz のクロック速度を提供するようになります。各構成の実効値は、プロセッサの種類、使用される DIMM の種類、および各プロセッサに 3 つあるメモリチャネルに対するメモリ容量の配分によって決定されます。理想的には、プロセッサごとだけでなく、チャネルごとの DIMM 数についても対称性が求められます。その結果、DIMM の数は、6 の倍数が推奨されます (3 チャネルのプロセッサが 2 個)。このガイドラインに従うと、8、16、32、64、および 128 GB のメモリを構成する従来のマトリクスは実装できません。しかし、お客様がこれらのメモリサイズを指定する場合、パフォーマンスにどのような影響が生じるのでしょうか。

このホワイトペーパーでは、まず Xeon 5500 搭載 PRIMERGY サーバのメモリアーキテクチャーの概要を説明します。第 2 セクションでは、実際のアプローチを取ります。構成を定義するために情報を求めているという場面を想定し、パフォーマンスを考慮した構成の表も用意しています。また、この想定では、システムと CPU タイプが指定され、特定のメモリ容量 (または大まかなメモリ構成) を満たす最適な構成を検討していると仮定しています。多くの場合、ここで示す表を慎重に検討するだけで、十分に適切な構成を作成できます。第 3 セクションでは、STREAM ベンチマークテストの結果に基づいて、推奨される構成の背景情報を説明します。このセクションの説明は、要求されるメモリ容量が 2 セクション目までの表に示されておらず、個別に構成を定義しなければならない場合にお勧めします。最後のセクションでは、**ベストプラクティス**に関するルールをまとめて示します。

この複雑な問題に関しては、次の項目が適用されます。最初は多くの要因によってパフォーマンスが左右されるように見える場合でも、ベスト プラクティスに関するさまざまなルールに従うことにより、強力なシステムを迅速に構成できます。コスト面に基づいてバランスのとれたソリューションを検討してみると、多くの場合でコストを無限にかけることは可能ですが、平均 5 % 未満のわずかなパフォーマンスの向上しか図れません。限りなくコストをかけるべきか、それが常に必要かを検討する場合には、ある程度の注意深さが必要です。同様に、プロジェクトの背景に関する知識も必要になります。本番システムを購入する場合は、ベンチマークテストの扱いが、異なる場合があるので注意してください。

メモリアーキテクチャー

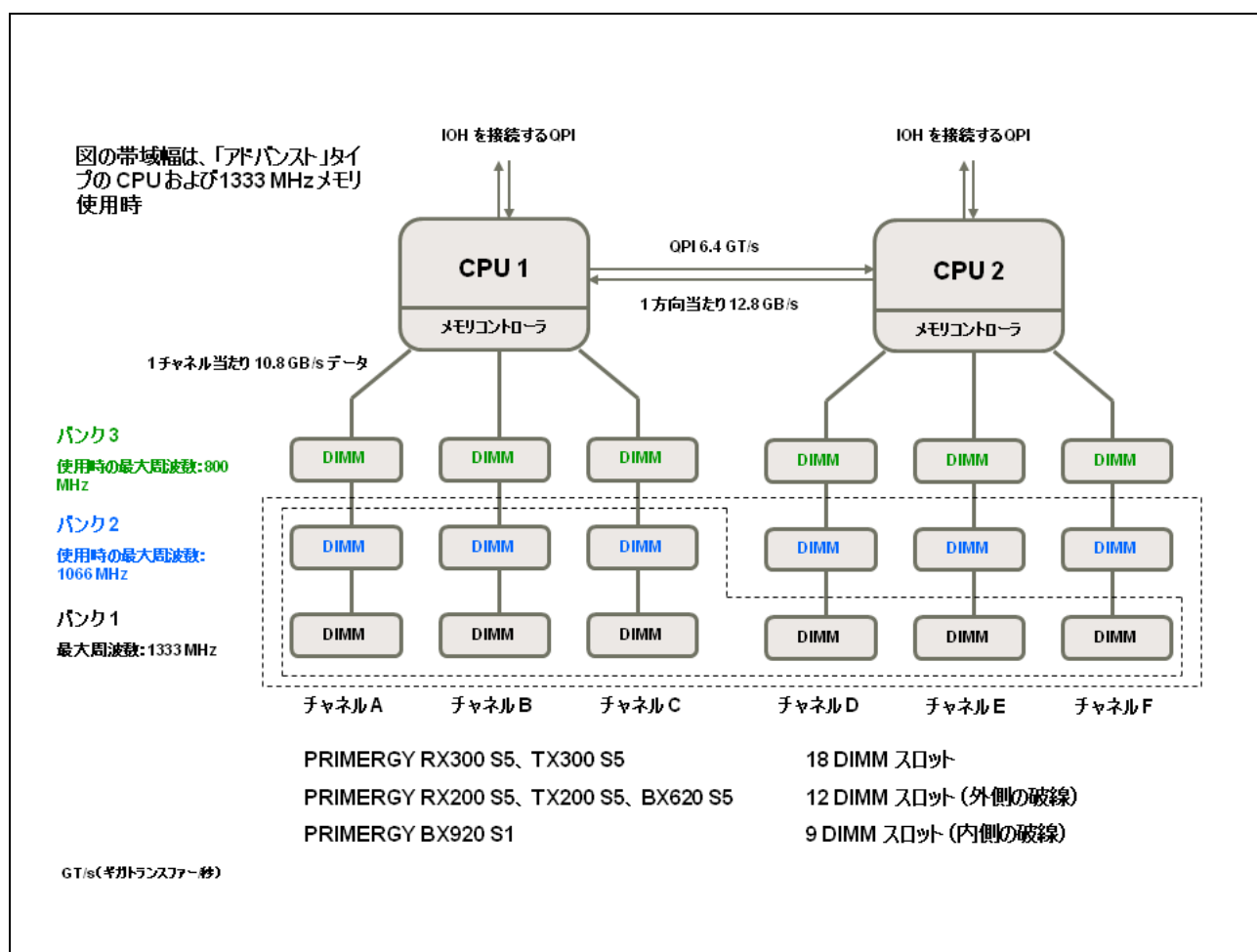
ここでは、3 部構成でメモリシステムの概要を説明します。まずブロック図で、利用可能な DIMM スロットの配置を説明します。2 番目に、4 つのメモリ構成モードについて説明します。これらのモードは、PRIMERGY システム構成図によっても参照されます。3 番目に、利用可能な DIMM のタイプを説明します。

DIMM スロット

次の図は、メモリシステムの構造を示します。DIMM スロットとその配置に関して、次の 3 つのモデルグループがあります。

- グループ 1 (18 スロット) : PRIMERGY RX300 S5 および TX300 S5
- グループ 2 (12 スロット) : PRIMERGY RX200 S5、TX200 S5 および BX620 S5
- グループ 3 (9 スロット) : PRIMERGY BX920 S1

説明は、2009 年 7 月時点でのシステムに基づきます。



1 つのプロセッサには、常に 3 つのメモリチャンネルが存在します。ただし、筐体内のスペースが決定的要因となる場合、チャンネル当たりの装着可能な最大 DIMM 数に応じて、これら 3 つのモデルグループが変更されます。チャンネル当たりの DIMM 数が変わると、メモリ周波数に変化が生じ、ひいてはメモリパフォーマンスに影響を与えます。このサイズは DPC (DIMM per channel : チャンネルあたりの DIMM 数) と呼ばれ、以下でも、この用語を使用します。例えば、PRIMERGY RX300 S5 の 2DPC 構成では、チャンネル当たり 2 個の DIMM、合計 12 個の DIMM が搭載されます。

DPC 値は、システム内のすべてのチャンネルで同じである必要はありません。構成を記述する場合は、例えば次のような省略形を使用します。

2 - 2 - 2 / 1 - 1 - 1

上の例で、1 つ目のプロセッサには各チャンネルに 2 個のメモリモジュールがあり、2 つ目のプロセッサには各チャンネルに 1 個のモジュールがあります。

以下ではまた、「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 3 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、サーバのマザーボード上で、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。プロセッサ当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、最適なインターリーブが得られます。インターリーブは、メモリのパフォーマンスを左右する主な要因です。これについては後ほど説明します。

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。1 つのプロセッサのみで運用する場合、空のソケットに割り当てられたスロットは使用できません。

4 つのメモリ構成モード

メモリ構成を定義する際には、パフォーマンス以外にも考慮すべき側面があります。これらの側面は、RAS (Reliability (信頼性)、Availability (可用性)、Serviceability (サービス性)) という略語で表されます。メモリシステムには、特に高度な RAS 要件を求めるお客様向けのオプションが用意されています。これは、以下に示す 4 つのメモリ構成モードのうち、最初の 2 つに関係しています。これら 2 つのモードは、必要に応じて、BIOS で指定します。指定しない場合は、実際の DIMM 構成によって、パフォーマンスチャンネルモードか独立チャンネルモードかを判定されます。適正な DIMM が適切に配置されていれば、自動的にパフォーマンスモードに設定されます。

- **スペアチャンネルモード** : 各バンクは、空か、3 つの DIMM (同一タイプ、同一容量) で構成します。チャンネル A と B (または D と E) の DIMM のみが使用されます。チャンネル C (または F) には、DIMM に障害が生じた場合のスペアが装着されます。このモードは、BIOS で明示的に指定する必要があります。
- **ミラーチャンネルモード** : 各バンクでチャンネル A と B (または D と E) のみが使用されます。これらのチャンネルは、同じタイプの DIMM で構成する必要があります。チャンネル C (または F) のすべてのスロットは、空のままです。ハードウェアは、オペレーティングシステムやアプリケーションに対して透過的に、メモリの内容をミラーリングします。実効物理メインメモリは、構成された容量のわずか半分になります。DIMM が故障しても、システムのダウンタイムは生じません。このモードは、BIOS で明示的に指定する必要があります。
- **パフォーマンスモード** : 各バンクは、空か、3 つの DIMM (同一タイプ、同一容量) で構成されます。この構成では、3 つのメモリチャンネル経由で、最適なインターリーブが実現されます。
- **独立チャンネルモード** : 上に述べた 3 つ以外のすべての構成が、このカテゴリに分類されます。**Unbuffered** モジュールと **Registered** モジュールが混在しない限り、各スロットには、以下に示す DIMM のどれでも割り当てることができます。

最初の 2 つのモードは、Xeon 5500 搭載のサーバモデルによってはサポートしていないことがあります。

利用可能なメモリのタイプ

指定された PRIMERGY モデルについて構成を検討する場合、次の表に示す DIMM を使用します。メモリモジュールには、ECC 保護の DDR3 メモリモジュールが使用されます。

表内の最後の 1 列には、相対的な価格差を示します。比較の基準として、2009 年 7 月現在の PRIMERGY RX200 S5 の料金表価格を使用しました。この列では、4 GB のサイズの Registered PC3-8500 DIMM を基準 (1 として強調表示) に、GB 当たりの相対価格を示します。8 GB モジュールが高コストであることと、Unbuffered モジュールが手ごろな価格であることが目を引きます。このように価格差があるので、メモリを構成する際には、コストの問題を考慮に入れる必要があります。例えばパフォーマンスを 2% 改善するために 2 倍のコストをかける対応は、賢明とは言えません。

タイプ	制御	最大周波数 (MHz)	ランク	容量	GB 当たりの相対価格	注記
DDR3-1333 PC3-10600 rg ECC	Registered	1333	1	2 GB	1.1	
DDR3-1333 PC3-10600 rg ECC	Registered	1333	2	4 GB	1.0	
DDR3-1333 PC3-10600 rg ECC	Registered	1333	2	8 GB	3.7	
DDR3-1066 PC3-8500 rg ECC	Registered	1066	1	2 GB	1.1	
DDR3-1066 PC3-8500 rg ECC	Registered	1066	2	4 GB	1	
DDR3-1066 PC3-8500 rg ECC	Registered	1066	2	8 GB	3.4	
DDR3-1066 PC3-8500 ub ECC	Unbuffered	1066	1	1GB	0.8	a)
DDR3-1066 PC3-8500 ub ECC	Unbuffered	1066	2	2 GB	0.7	a)

a) PRIMERGY RX300 S5 および TX300 S5 では使用不可

Unbuffered モジュールは構造が単純なため、Registered モジュールより最大容量が小さく、1DPC または 2DPC 構成でのみ使用可能です。

販売区域によっては、特定の DIMM タイプの使用に制限がある場合があります。

パフォーマンスを考慮した構成

次の表には、パフォーマンスを考慮した場合に適切なすべてのメモリサイズについて、構成例を示します。この表から、最初の表の構成が「理想的」であることがわかります。なぜなら、これらの構成はすべて、システム内のすべてのメモリチャネルにわたって、メモリが均等に分配されているからです。これらの構成は、パフォーマンスモードに対応します。

2 番目の表は、8、16、32 GB など、従来のシステムアーキテクチャの「古典的」構成について示します。これらの構成は、理想的な構成と比較して**慎重に評価すると**、(容量の違い自体が、テスト結果に影響を与えない限りにおいて) パフォーマンス率が若干のマイナス面があります。しかし、このマイナス面は、ほとんどの用途で問題になりません (下で説明するとおり、この差異の原因は、古典的なサイズで 2-way インターリーブが使用されていることにあります。理想的な構成は、3-way インターリーブです)。

最初の 2 つの表は、メモリが両方のソケットに対称に分配されており、NUMA について最適です。非対称メモリ構成については、後半の表に示します。

最大速度である 1333 MHz で動作できる代替構成を赤で示します。このオプションが適用される**アドバンス**トのプロセッサクラスには、Xeon 5500 モデル X5570、X5560、および X5550 があります。赤で示した代替構成を検討してみましょう。Unbuffered DIMM の代わりに Registered DIMM を使用する際の、余分のコストはそれほどでもありません。ただし、8 GB のモジュールを使用すると、相当な差になります。

これらの表の最後の 3 列は、それぞれの構成が可能な PRIMERGY モデルを示します。

ここで扱っていない構成に対するメモリ構成については、「メモリパフォーマンス」のセクションを参照してください。

理想的なメモリサイズ							
総容量	メモリタイプ	モジュール サイズ GB	構成	注記	R/TX200 S5 BX620 S5	BX920 S1	R/TX300 S5
6 GB	PC3-8500 Unbuffered	1	1 - 1 - 1 / 1 - 1 - 1		✓	✓	
12 GB	PC3-8500 Unbuffered	2	1 - 1 - 1 / 1 - 1 - 1	Registered について 価格優位性あり	✓	✓	
	PC3-8500 Registered	2	1 - 1 - 1 / 1 - 1 - 1		✓	✓	✓
	PC3-10600 Registered	2	1 - 1 - 1 / 1 - 1 - 1	アドバンスド CPU と 1333 MHz が必須の場合	✓	✓	✓
18 GB	PC3-8500 Unbuffered	1 および 2	2 - 2 - 2 / 2 - 2 - 2	一番目のバンク 2 GB モジュール 二番目のバンク 1 GB モジュール	✓		
24 GB	PC3-8500 Unbuffered	2	2 - 2 - 2 / 2 - 2 - 2	1DPC、4 GB DIMM より価格的に優位	✓		
	PC3-8500 Registered	4	1 - 1 - 1 / 1 - 1 - 1		✓	✓	✓
	PC3-10600 Registered	4	1 - 1 - 1 / 1 - 1 - 1	アドバンスド CPU と 1333 MHz が必須の場合	✓	✓	✓
36 GB	PC3-8500 Registered	2 および 4	2 - 2 - 2 / 2 - 2 - 2	一番目のバンク 4 GB モジュール 二番目のバンク 2 GB モジュール	✓		✓
48 GB	PC3-8500 Registered	4	2 - 2 - 2 / 2 - 2 - 2		✓		✓
	PC3-8500 Registered	8	1 - 1 - 1 / 1 - 1 - 1	BX920 S1 用	✓	✓	✓
	PC3-10600 Registered	8	1 - 1 - 1 / 1 - 1 - 1	アドバンスド CPU と 1333 MHz が必須の場合	✓	✓	✓
60 GB	PC3-8500 Registered	2 および 8	2 - 2 - 2 / 2 - 2 - 2	一番目のバンク 8 GB モジュール 二番目のバンク 2 GB モジュール	✓		✓
	PC3-8500 Registered	2 および 4	3 - 3 - 3 / 3 - 3 - 3	一番目のバンク 4 GB モジュール 二番目のバンク 4 GB モジュール サードバンク 2 GB モジュール			✓
72 GB	PC3-8500 Registered	4 および 8	2 - 2 - 2 / 2 - 2 - 2	一番目のバンク 8 GB モジュール 二番目のバンク 4 GB モジュール	✓		✓
	PC3-8500 Registered	4	3 - 3 - 3 / 3 - 3 - 3				✓
84 GB	PC3-8500	2、4	3 - 3 - 3 / 3 - 3 - 3	一番目のバンク			✓

	Registered	および 8		8 GB モジュール 二番目のバンク 4 GB モジュール サードバンク 2 GB モジュール			
96 GB	PC3-8500 Registered	8	2 - 2 - 2 / 2 - 2 - 2		✓		✓
108 GB	PC3-8500 Registered	2 および 8	3 - 3 - 3 / 3 - 3 - 3	一番目のバンク 8 GB モジュール 二番目のバンク 8 GB モジュール サードバンク 2 GB モジュール			✓
120 GB	PC3-8500 Registered	4 および 8	3 - 3 - 3 / 3 - 3 - 3	一番目のバンク 8 GB モジュール 二番目のバンク 8 GB モジュール サードバンク 4 GB モジュール			✓
144 GB	PC3-8500 Registered	8	3 - 3 - 3 / 3 - 3 - 3				✓

古典的なメモリサイズ							
総容量	メモリタイプ	モジュール サイズ GB	構成	注記	R/TX200 S5 BX620 S5	BX920 S1	R/TX300 S5
4 GB	PC3-8500 Unbuffered	1	1 - 1 - 0 / 1 - 1 - 0		✓	✓	
8 GB	PC3-8500 Unbuffered	1	2 - 1 - 1 / 2 - 1 - 1	Registered および 3 チャンネルについて 価格優位性あり	✓		
	PC3-8500 Registered	2	1 - 1 - 0 / 1 - 1 - 0		✓	✓	✓
	PC3-10600 Registered	2	1 - 1 - 0 / 1 - 1 - 0	アドバンスド CPU と 1333 MHz が必須の場合	✓	✓	✓
16 GB	PC3-8500 Unbuffered	2	2 - 1 - 1 / 2 - 1 - 1	Registered について 価格優位性あり	✓		
	PC3-8500 Registered	2	2 - 1 - 1 / 2 - 1 - 1	3 チャンネルを使用	✓		✓
	PC3-8500 Registered	4	1 - 1 - 0 / 1 - 1 - 0		✓	✓	✓
	PC3-10600 Registered	4	1 - 1 - 0 / 1 - 1 - 0	アドバンスド CPU と 1333 MHz が必須の場合	✓	✓	✓

32 GB	PC3-8500 Registered	4	2 - 1 - 1 / 2 - 1 - 1		✓		✓
	PC3-8500 Registered	8	1 - 1 - 0 / 1 - 1 - 0	BX920 S1 用	✓	✓	✓
	PC3-10600 Registered	8	1 - 1 - 0 / 1 - 1 - 0	アドバンスト CPU と 1333 MHz が必須の場合	✓	✓	✓
64 GB	PC3-8500 Registered	8	2 - 1 - 1 / 2 - 1 - 1		✓		✓
	PC3-8500 Registered	4	3 - 3 - 2 / 3 - 3 - 2	8 GB モジュールについて価格優位性あり			✓
128 GB	PC3-8500 Registered	8	3 - 3 - 2 / 3 - 3 - 2				✓

赤色で示される状況でのみ、1333 MHz が可能です。3 バンク構成では、あらゆる状況でタイミングが常に 800 MHz です。1 つのチャネルのみが 3DPC だとしても、タイミングは常に 800 MHz です。それ以外の構成では、CPU タイプ Xeon E5520 ~ X5570 の赤以外の状況で 1066 MHz が適用され、E5502 ~ E5506 で 800 MHz が適用されます。

必要なメモリ容量は仮定条件です。例えば I/O 速度など、アプリケーションのパフォーマンスに対するその暗黙の影響は、ここでは無視する必要があります。

非対称型メモリ構成

そのフォームファクタによっては、すべてのシステムが、すべての構成バージョンで対称型メモリ構成を形成できるわけではありません。3 ページの図は、PRIMERGY BX920 S1 の DIMM スロットの非対称型配置を示しています。最初のソケットに 2 つのメモリバンクがあり、2 番目のソケットに 1 つのメモリバンクがあります。この非対称配置は、両方のソケット経由でメモリを対称に分配するようにという NUMA の推奨に関して、推奨構成をリストする際の別の側面を示唆しています。

NUMA 最適化構成

PRIMERGY BX920 S1 は、スロットは非対称ですが、48 GB の容量まで合計メモリ容量を 2 等分することができます。これは、NUMA 最適化構成です。この構成は、先の理想的なメモリサイズと古典的メモリサイズの表内ですでに示されています。

これまでに示した表では、32 GB と 48 GB の容量に、標準の 8 GB モジュールが使用されました。6 つの DIMM スロットを持つ CPU で 4 GB モジュールを使用する場合は、そのコスト面のデメリットが緩和されます。次の表では、高価な 8 GB モジュールが、2 つまたは 3 つのみ必要な、これらのバージョンを示します。これらは 2DPC 構成なので、メモリ周波数 1333 MHz は達成できません。「アドバンスト」プロセッサの場合、これが、このバージョンと前に示したバージョンとのパフォーマンス上の唯一の違いです。他のプロセッサの場合は、コストの違いのみです。

PRIMERGY BX920 S1 NUMA 最適化バージョン				
総容量	メモリタイプ	モジュール サイズ GB	構成	注記
32 GB	PC3-8500 Registered	4 および 8	2 - 1 - 1 / 1 - 1 - 0	左が 4 GB モジュール 右が 8 GB モジュール
48 GB	PC3-8500 Registered	4 および 8	2 - 2 - 2 / 1 - 1 - 1	左が 4 GB モジュール 右が 8 GB モジュール

非対称型構成

次の表では、左側のメモリが右側より多くなっています。超過分は、総容量の 1/5 ~ 1/3 です。最大超過分の半分、つまり 1/10 ~ 1/6 には、QPI リンク経由の「リモート」アクセスが行われます（統計的に確認）。このような軽度の非対称型の場合、対称型と比較して 1 ~ 2 % のパフォーマンスのマイナス分を計算に入れる必要があります。大容量の共有メモリセグメントを持つデータベースなど、いずれにしてもリモートアクセスが避けられない作業負荷の場合、パフォーマンス上のマイナスはありません。これは、PRIMERGY BX920 S1 で Windows Server 2008 と SQL Server 2008 を使用して行われた OLTP2 評価で確認されました。

以下の表は、32GB の対称型の構成に対する、8 GB モジュールなしで機能する低価格の代替構成から順に各構成を一覧したものです。

PRIMERGY BX920 S1 非対称型構成				
総容量	メモリタイプ	モジュール サイズ GB	構成	注記
32 GB	PC3-8500 Registered	4	2 - 2 - 2 / 1 - 1 - 0	
36 GB	PC3-8500 Registered	4	2 - 2 - 2 / 1 - 1 - 1	
40 GB	PC3-8500 Registered	4 および 8	2 - 2 - 2 / 1 - 1 - 0	4 GB モジュール左 8 GB モジュール右
60 GB	PC3-8500 Registered	4 および 8	2 - 2 - 2 / 1 - 1 - 1	一番目のバンクは左右の ソケット共に 8 GB モジュール、 左のソケットの二番目のバンク は 4 GB モジュール
64 GB	PC3-8500 Registered	8	2 - 2 - 2 / 1 - 1 - 0	
72 GB	PC3-8500 Registered	8	2 - 2 - 2 / 1 - 1 - 1	

メモリパフォーマンスに対する影響

ここでは、RAM のパフォーマンスに影響のあるコンポーネントについて説明します。まず、このホワイトペーパーの基準となったテストで、メモリパフォーマンスをどのように測定し、それらのデータをどのように判断したかについて説明します。

測定ツール : STREAM ベンチマーク

STREAM ベンチマーク (開発者 : John McCalpin 氏) [関連資料 3] 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。ベンチマークでは、常に GB/s 単位のスループットの結果を得られます。Triad の値が最も良く引用されます。以下の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、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、複雑な構成空間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。各状況で STREAM は、メモリに影響を与える構成アクションによって生じるパフォーマンスに対する最大の影響 (低下または向上) を示します。以降に示している STREAM ベンチマークに関する値は、パフォーマンスに対する影響の**範囲**とみなすことができます。

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、各アクセスのレイテンシ (遅延) 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。実際のレイテンシは帯域幅の増加とともに増加するため、値は連動します。並列メモリアクセスによってレイテンシが認識されなくなる範囲は、アプリケーションと、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全般的な予測を立てることは非常に困難です。

STREAM ベンチマークと実際のアプリケーションのパフォーマンスの関係は、経験値に基づいてのみ示すことができます。

- 商用アプリケーション (データベース、SAP など) では、STREAM 値の 1/4 の値を使用します。たとえば、STREAM で 20 % の低下が示された場合、経験測によれば、商用アプリケーションで用いた場合 5 % 低下します。
- 科学技術アプリケーションでは、メモリパフォーマンスに対するさらに大きな影響を想定する必要があります。そのようなアプリケーションは、商用アプリケーションよりも STREAM のように大きな効果が出ます。この場合は、STREAM 値の約 1/2 の値を想定します。

商用アプリケーションシナリオの方がより実用的なため、以降のセクションでは、実際のアプリケーションのパフォーマンスに対する影響を見積もる際に前者の経験則を使用します。

これらの経験則は、標準ベンチマークの結果によっても実証されています。たとえば、メモリのクロックが 1066 MHz ではなく 1333 MHz に設定された場合に STREAM では 11% のパフォーマンス向上が示され、SPECint_rate2006 (代表的な商用ワークロード) での向上率は 3 % で、SPECfp_rate2006 (科学技術ワークロード) での向上率は 5 % です。また、SPECjbb2005 (Java パフォーマンス) での向上率も 3 % です。

実際の影響の程度を評価する際には、使用率が 100 % のシステムでのみ影響が直接現れ、それ以外の場合は CPU 負荷の変化に伴って影響が現れることも考慮してください。使用率があまり高くない場合、応答時間に対する顕著な影響は現れません。

パフォーマンスに影響を与える主な要因

ここでは、メモリパフォーマンスに対する 2 つの主な影響要因である、周波数とインターリーブについて説明します。どちらのパラメータもそれぞれ 800、1066、1333 MHz の周波数と 1 ウェイ、2 ウェイ、3 ウェイのインターリーブの 3 つのオプションがあります。メモリ構成を計画する際には、まずこれらのパラメータの計画を立てる必要があります。

メモリの有効な周波数

システムの電源を入れたときに BIOS によって判断される有効な周波数は、次の 3 つの要素によって決まります。

- プロセッサのタイプ。Xeon 5500 番台のモデルは、次の表に従って 3 つのグループに分類されています。インテルでは、これらを Basic (ベーシック)、Standard (スタンダード)、Advanced (アドバンスド)と呼んでいます。Basic グループのプロセッサは 800 MHz のみをサポートしており、Standard グループは 800 と 1066 MHz、Advanced グループは最大 1333 MHz をサポートしています。
- DIMM のタイプ。最大 1066 MHz と最大 1333 MHz までの DIMM を使用できます。
- DPC 値 (チャンネルあたりの DIMM)。周波数 1333 MHz は、1DPC でのみ可能です。2DPC 構成では周波数は 1066 MHz に制限されており、3DPC 構成では 800 MHz に制限されています。6 個のチャンネルを同じ構成にしていない場合は、最も大きい DPC 値によって判断されます。

クラス	Xeon タイプ	コア数	GHz	L3 キャッシュ (MB)	QPI (GT/s)	最大メモリ MHz	TDP (ワット)
Advanced	X5570	4	2.93	8	6.4	1333	95
	X5560	4	2.80	8	6.4	1333	95
	X5550	4	2.67	8	6.4	1333	95
Standard	E5540	4	2.53	8	5.9	1066	80
	E5530	4	2.40	8	5.9	1066	80
	E5520	4	2.27	8	5.9	1066	80
	L5520	4	2.27	8	5.9	1066	60
Basic	E5506	4	2.13	4	4.8	800	80
	L5506	4	2.13	4	4.8	800	60
	E5504	4	2.00	4	4.8	800	80
	E5502	2	1.87	4	4.8	800	80

できるだけ高い周波数を使用したほうが効果的です。ただし、3 つの要素の中で最も低い値が使用されます。つまり、構成の周波数は 3 つの要素の中の最も低い値によって決まります。また、周波数は、プロセッサ単位ではなく、システムの標準値として定義されます。

次の例でこのメカニズムについて説明します。PRIMERGY RX200 S5 が、タイプ X5570 プロセッサと 4 GB の PC3-8500 モジュールを以下のように搭載するとします。

2 - 2 - 2 / 2 - 2 - 2

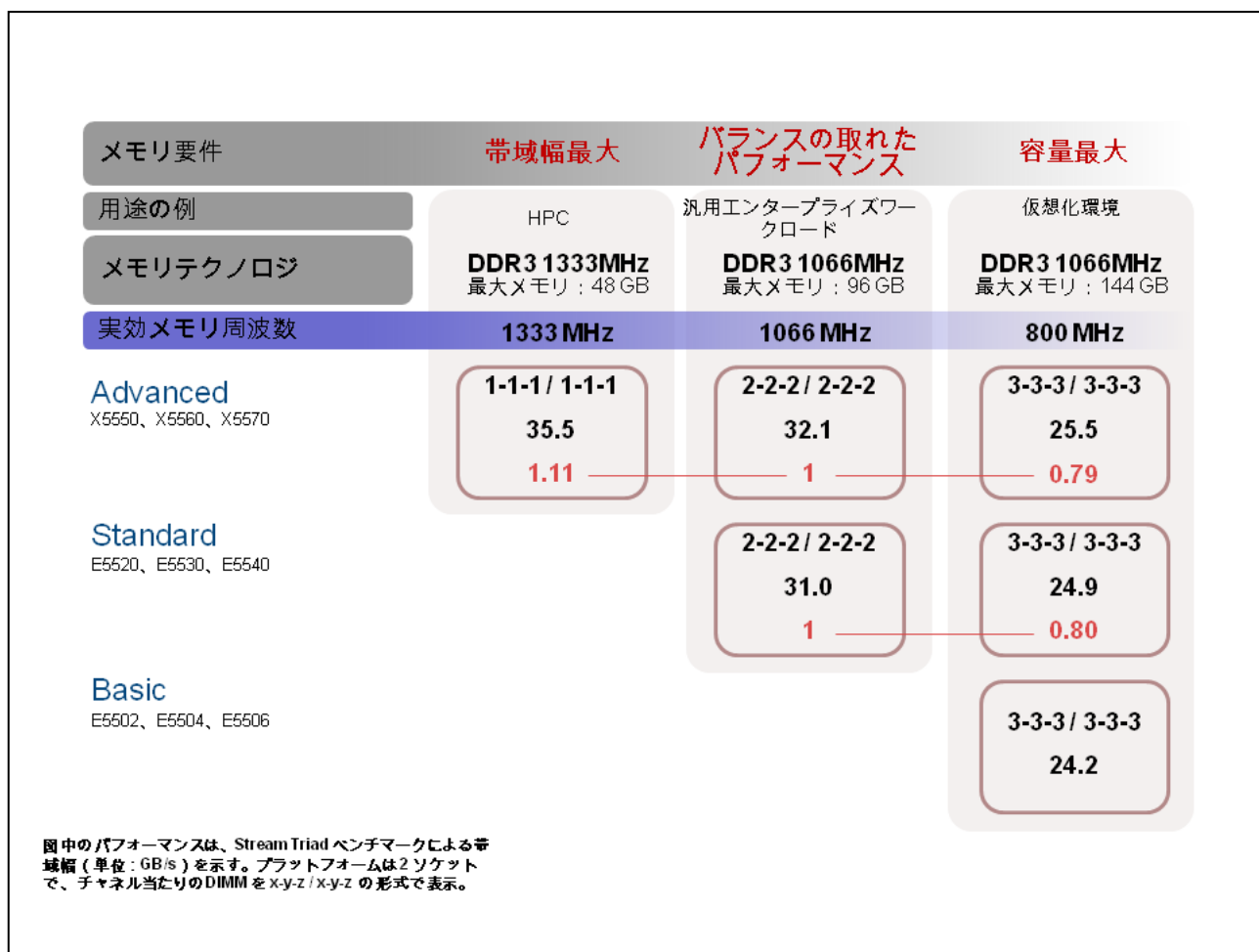
この構成では、メモリ容量は 48 GB、周波数は 1066 MHz になります。プロセッサは 1333 MHz をサポートしていますが、PC3-8500 メモリと 2DPC 構成により、周波数は 1066 MHz に制限されます。2DPC による制限が 1066 MHz なので、1333 MHz 対応の PC3-10600 4 GB モジュールを構成しても意味はありません。また、8 GB PC3-10600 モジュールを以下のように搭載して 48 GB の容量にすることもできます。

1 - 1 - 1 / 1 - 1 - 1

この構成では、1333 MHz の最高周波数にすることができます。

こちらの構成の方は高性能で、前述の構成の方は低価格です。次の図に、各種周波数によるパフォーマンスに対する最大の影響を示します。図では、プロセッサグループとメモリテクノロジーの合理的な関係をマトリクスで示しています。1333 MHz 対応 PC3-10600 メモリは、Advanced グループのプロセッサ (タイプ ID

の先頭が X のプロセッサ) でのみ検出されます。Basic プロセッサでは、メモリの周波数に関して何も考慮する必要はありません。有効周波数は常に 800 MHz です。



図では、STREAM Triad GB/s と「平均的な」周波数 1066 MHz に対する割合 (赤) を示しています。周波数 1333 MHz で 8 GB モジュールを搭載した上図の RX200 S5 の例では、最初の構成と比べて、パフォーマンスは最大約 10 % 向上し、実際には商用アプリケーションで約 3 % 向上します。メモリの価格性能比は、同時に 3 倍向上します。HPC 環境 (High Performance Computing : ハイパフォーマンスコンピューティング) プロジェクトなどの特別な状況に対応する必要がない場合は、通常、4 GB モジュールと周波数 1066 MHz を使用します。

図では、各 DPC の構成の STREAM 値を比較しています。これは、目的の周波数は DPC 値を変化させることでしか得られないためです。総容量が異なる構成を比較することは、STREAM での測定では意味がありません。これはベンチマークでは、常に 1 GB の使用可能な物理メモリのみが使用されるためです。他の理由により周波数が同一な場合、DPC が異なる構成の最大メモリ帯域幅は同一になります。以下は、CPU E5506 と 4 GB のデュアルランク PC3-8500 モジュールの構成の 800 MHz の結果です。

1 - 1 - 1 / 1 - 1 - 1 24.5 GB/s
2 - 2 - 2 / 2 - 2 - 2 24.1 GB/s
3 - 3 - 3 / 3 - 3 - 3 24.2 GB/s

これには、チャンネルあたりのランクの総数が奇数であるという条件があります。これについては、後述の「ランク数」セクションで説明します。ただし、デュアルランクモジュールを構成した場合にはこの状況にはなりません。

RAM クロック周波数についてまとめると次のようになります。周波数を比較した場合、周波数 1333 MHz のパフォーマンスは、1066 MHz と比べて最大 10 % 向上し、実際には商用アプリケーションで約 3 % 向上します。周波数 800 MHz のパフォーマンスは、最大 20 % 低下し、商用アプリケーションで約 5 % 低下します。商用アプリケーションに関するこの情報は一般的な指標であり、特定のアプリケーションでの絶対的な予測値ではありません。

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ンネルに、2 番目のブロックは 2 番目のチャンネルにという具合に、プロセッサ単位で 3 つのメモリチャン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、局所性原理に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ンネルに分散されます。これは並列化による性能向上です。さらに、アクティブ(開いている)メモリページを変更する際に DRAM メモリの物理特性に従って発生する遅延も目立たなくなります。

次の図に、前述のメモリ周波数よりも優れた効果をもたらすインターリーブの影響を示します。理想的な状況は 3 ウェイインターリーブで、3 つのすべてのチャンネルを同じに構成した場合にこのようになります。メモリ構成オプションのパフォーマンスモードは、このシナリオに基づいています。また、この図は、「推奨」構成 (3 ウェイインターリーブ) であるにもかかわらずこの理想的な状況にならない理由も示しています (スピアチャンネルモードでの構成を除きます。スピアチャンネルモードは図の真ん中の列に対応します)。従来のメモリ構成では 2 のべき乗にしていました。

メモリ要件	最高のパフォーマンス		従来の容量など
容量の例	12、24、48、96、144 GB	8、16、32、64、128 GB	
評価	最高	許容範囲	低
メモリインターリーブ	3-way	2-way	1-way
Advanced X5550、X5560、X5570 ■ 実効周波数 : 1333 MHz	1-1-1 / 1-1-1 35.5 1.23	1-1-0 / 1-1-0 28.9 1	1-0-0 / 1-0-0 15.6 0.54
Standard E5520、E5530、E5540 ■ 実効周波数 : 1066 MHz	1-1-1 / 1-1-1 29.5 1.22	1-1-0 / 1-1-0 24.1 1	1-0-0 / 1-0-0 12.8 0.53
Basic E5502、E5504、E5506 ■ 実効周波数 : 800 MHz	1-1-1 / 1-1-1 24.5 1.28	1-1-0 / 1-1-0 19.1 1	1-0-0 / 1-0-0 10.1 0.53

図中のパフォーマンスは、Stream Triad ベンチマークによる帯域幅 (単位: GB/s) を示す。プラットフォームは 2 ソケットで、チャンネル当たりの DIMM を x-y-z / x-y-z の形式で表示。

図に示している値が極端な値であることを考えれば、2 ウェイインターリーブ構成はまさに合理的であるように思われます。1 ウェイインターリーブ構成は使用しないでください。これは実際には非インターリーブ構成で、分類上 1 ウェイと呼ばれているだけです。パフォーマンスの低下を想定する必要があります。これは、通常の状況下では、プロセッサのパフォーマンス能力に密接に関係しません。この構成は、ミラーチャ

ネルモードでのシステムの信頼性が、パフォーマンスに関するすべての要件に勝る決定的な評価基準である場合のみ意味があります。

インターリーブは、周波数と同様、システムの電源を入れたときに BIOS によって定義されます。チャンネルあたりの GB 数が同じ場合、3 つの構成済みのチャンネルで 3 ウェイインターリーブが可能です。また、1 つのチャンネルが空なら 2 つのチャンネルで 2 ウェイインターリーブが可能です。DPC 値が異なっているにもかかわらず、異なるサイズの DIMM を使用することで、インターリーブに最適な状況を作ることができます。チャンネルあたりの総 GB が決定的な要素です。

チャンネルあたりの総 GB が異なる場合、物理メモリは異なるインターリーブの領域に分割されます。これは、1 ウェイインターリーブの領域を回避することを目的としています。このため BIOS は、次の同一の 4 GB DIMM (たとえば、32 GB の総容量にするのに必要) を

2 - 1 - 1 / 2 - 1 - 1

次のように 2 つの 2 ウェイに分割します。

1 - 1 - 0 / 1 - 1 - 0 (メモリ容量の 50 %) 2 ウェイインターリーブ

1 - 0 - 1 / 1 - 0 - 1 (50 %) 2 ウェイインターリーブ

以下のような不均一な分割はしません。

1 - 1 - 1 / 1 - 1 - 1 (75 %) 3 ウェイインターリーブ

1 - 0 - 0 / 1 - 0 - 0 (25 %) 1 ウェイインターリーブ

インターリーブについてまとめると次のようになります。理想的な 3 ウェイインターリーブでは 2 ウェイよりもパフォーマンスが最大 20 % 向上し、現実的には商用アプリケーションで 5 % 向上します。2 ウェイインターリーブは、従来どおりの 2 のべき乗 (8、16、32 GB など) であるメモリ容量とスピアチャンネルモードで可能です。1 ウェイインターリーブ構成は、パフォーマンスの低下が非常に大きいため、推奨されません。チャンネルあたりの GB 数が異なる場合、異なるメモリ領域でのインターリーブになります。

パフォーマンスの二次的影響

ここまで説明したトピックでは、測定を注意深く行った場合に、アプリケーションパフォーマンスでこれらの影響が認識できるようになると想定しています。以降のトピックでは、測定ツールを使用してパフォーマンスへの影響を実証していきます。ただし、実際のアプリケーションパフォーマンスに影響を与えるとは限りません。

UDIMM と RDIMM

次の表に従えば、PRIMERGY RX200 S5、TX200 S5、BX620 S5、および BX920 S1 では、Registered DIMM (RDIMM) とは別に、Unbuffered DIMM (UDIMM) モジュールも利用できます。より単純な UDIMM 構成は、安価で、エネルギー使用量も多少少なく済みます。必要なメモリ容量に対応できる場合は、これらの理由により好まれる構成です。

タイプ	制御	最大 MHz	ランク	容量	備考
DDR3-1333 PC3-10600 rg ECC	Registered	1333	1	2 GB	
DDR3-1333 PC3-10600 rg ECC	Registered	1333	2	4 GB	
DDR3-1333 PC3-10600 rg ECC	Registered	1333	2	8 GB	
DDR3-1066 PC3-8500 rg ECC	Registered	1066	1	2 GB	
DDR3-1066 PC3-8500 rg ECC	Registered	1066	2	4 GB	
DDR3-1066 PC3-8500 rg ECC	Registered	1066	2	8 GB	
DDR3-1066 PC3-8500 ub ECC	Unbuffered	1066	1	1 GB	a)
DDR3-1066 PC3-8500 ub ECC	Unbuffered	1066	2	2 GB	a)

a) PRIMERGY RX300 S5 と TX300 S5 では使用不可

RDIMM と UDIMM を組み合わせることはできません。

RDIMM では、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内でバッファされます（これが名前の由来です）。これにより、メモリチャネルが解放され、UDIMM では不可能な 3DPC が可能になります。その逆も同様で、UDIMM での 2DPC 構成では負荷が大きくなり（1DPC と比較して）、1N ではなく、2N のタイミングでの DIMM アドレッシングが必要になります。制御コマンドは、メモリチャネルの 2 クロックごとに発行可能です。この結果、UDIMM での 2DPC 構成の最大メモリ帯域幅は、RDIMM よりも 5 %程度低下します。

商用アプリケーションのパフォーマンスでは、この影響は無視できます。

ランク数

最後の表に、1 ランクまたは 2 ランクが可能なメモリモジュールを示します。これは、64 ビット幅のメモリ領域を同時に読み取りまたは書き込む DRAM チップのグループが 1 つだけの DIMM が存在することを示します。各チップは 8 ビットを受け持ちます。または、そのようなグループが 2 つあります。ただし、DIMM アドレスとデータラインは、両方のグループで共通です。つまり、1 つのグループのみが一度にアクティブになることができます。デュアルランク DIMM のメリットは、表に示しているように、まず容量の大きさが挙げられます。

デュアルランクモジュールの 2 番目のメリットは、前述した物理的な理由です。メモリセルは、2 次元に配置されています。ライン（ページとも呼ばれる）が開かれ、このラインで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値を読み取ることができます。このレイテンシの違いが、メモリコントローラーによる最適化のメリットであり、「開いている」メモリページに関する保留中の命令を再割り当てします。デュアルランクモジュールでは、開いているページにアクセスする確率が高まります。

これは、STREAM でメモリ帯域幅を測定した次の表で確認できます。

CPU	RAM				帯域幅 (GB/s)
	タイプ	容量	ランク数	構成	
X5570	PC3-10600 Registered	4 GB	2	1 - 1 - 1 / 1 - 1 - 1	35.5
X5570	PC3-10600 Registered	2 GB	1	1 - 1 - 1 / 1 - 1 - 1	32.3

DPC 値が高い構成では、チャネルあたりのランク数が奇数の場合に、同様の効果が得られます。デュアルランクモジュールを使用している場合は、この状況にはなりません。2 GB モジュールを使用した構成で、チャネルあたりのランク数が奇数の場合に実際のパフォーマンスが 2~3 %低下することも、シングルランクの RDIMM モジュールよりもデュアルランクの UDIMM モジュールが好まれる理由の 1 つです。

ベストプラクティスのルール

パフォーマンスおよびコスト効果の高いメモリ構成を選択する際には、次の 7 つの簡単なルールを考慮してください。ルールは、重要度の順に示しています。

NUMA	可能な場合は、両方のソケットで対称的にメモリを分散し、BIOS で NUMA モードを有効にします (デフォルト値)。 (BX920 S1 : 非対称スロットであるにもかかわらず、最大 48 GB の NUMA 構成が可能。大規模な構成の中程度の非対称性による影響は、通常ほとんど認められず、これは無視することができます)
インターリーブ	バンク単位で CPU ソケットを構成します。可能な場合は、3 つのすべてのチャネルを含めます。 1 ウェイ (非インターリーブ) 領域を回避します。6~9 ページの表のすべての構成は、3 ウェイまたは 2 ウェイであり、優れたパフォーマンスを提供します。
1333 MHz	「Advanced」CPU と最大 24 GB (低価格モジュール) または 48 GB (高価な 8 GB モジュール) という比較的厳しい前提条件でのみ利用可能です。ベンチマークや特別な条件 (HPC) のプロジェクトに適しています。
800 と 1066 MHz	「Basic」CPU は、3DPC 構成と同様、常に 800 MHz に制限されます。
DIMM 価格の下落	UDIMM モジュールは RDIMM よりも安価です。1 GB モジュールと 4 GB モジュール (UDIMM と RDIMM の両方) では価格は大幅に異なりますが、4 GB モジュールと 8 GB モジュール (RDIMM のみ) の価格差はそれ以上です。
UDIMM と RDIMM	可能な場合は (6 ページ ~ 9 ページの表を比較のこと)、価格およびエネルギー消費量の観点から UDIMM モジュールが推奨されます。
ランク数	チャネルあたりのランク数が偶数であることは、パフォーマンス上あまりメリットはありません。ベンチマークでは効果がありますが、実稼動環境にはそれほどメリットはありません。

商用アプリケーション環境においては以下の数値に相当します。

- 周波数 1333 MHz では、1066 MHz よりもパフォーマンスが約 3 % 向上します。
- 周波数 800 MHz では、1066 MHz よりもパフォーマンスが約 5 % 低下します。
- 3 ウェイインターリーブでは、2 ウェイインターリーブよりもパフォーマンスが約 5 % 向上します。
- 無視しても良い要素は、以下のとおりです。
 - 2DPC 構成での Unbuffered DIMM の 2N タイミング
 - チャネルあたりの奇数のランク数

関連資料

[L1] PRIMERGY systems
http://ts.fujitsu.com/primergy
[L2] PRIMERGY performance
http://ts.fujitsu.com/products/standard_servers/primergy_bov.html
[L3] STREAM benchmark
http://www.cs.virginia.edu/stream/
[L4] PRIMERGY RX200 S5 data sheet and performance report
http://docs.ts.fujitsu.com/dl.aspx?id=9d1f923e-8be1-4a17-a6d6-b00736f5d765
http://docs.ts.fujitsu.com/dl.aspx?id=c4917140-8852-463d-8121-3db977b8f9e1
[L5] PRIMERGY RX300 S5 data sheet and performance report
http://docs.ts.fujitsu.com/dl.aspx?id=c245c1d4a-047c-4217-a1aa-6d5a333d856b
http://docs.ts.fujitsu.com/dl.aspx?id=c058ec307-e8b2-49d7-8ba3-d352cfce6945
[L6] PRIMERGY TX300 S5 data sheet and performance report
http://docs.ts.fujitsu.com/dl.aspx?id=c731a6c-7240-4432-b516-b7c211a829b3
http://docs.ts.fujitsu.com/dl.aspx?id=c05df4b52-64e4-4396-8bae-b7b1a0f4ae3f
[L7] PRIMERGY BX620 S5 data sheet and performance report
http://docs.ts.fujitsu.com/dl.aspx?id=c0703-df6d-4989-b0c9-ada3435f78b2
http://docs.ts.fujitsu.com/dl.aspx?id=d34f4a29-02f5-44ca-baff-63d830701418
[L8] PRIMERGY BX920 S1 data sheet and performance report
http://docs.ts.fujitsu.com/dl.aspx?id=c0703-df6d-4989-b0c9-ada3435f78b2
http://docs.ts.fujitsu.com/dl.aspx?id=fe9adc7a-23ef-471c-8cd5-78835b4deffc
[L9] PC サーバ PRIMERGY (プライマジー)
http://primeserver.fujitsu.com/primergy/

お問い合わせ先

PRIMERGY パフォーマンスとベンチマーク

<mailto:primergy.benchmark@ts.fujitsu.com>