



## IBM System x™ 3455 のパフォーマンス

Douglas M. Pase および Matthew A. Eckl  
IBM Systems and Technology Group

## 要約

本書では、IBM System x™ 3455 サーバー (x3455) のパフォーマンスについて考察します。この分析には、メモリー帯域幅と待ち時間、浮動小数点ベクトルのパフォーマンス、および SPEC<sup>®</sup> CPU2000 の速度と比率のベンチマークのパフォーマンスが含まれます。本書では、システムのプロセッサと周波数のスケールリング、およびサーバー内で 12 個の DIMM をサポートするためのパフォーマンスへの影響を考察します。すべてのテストで x3455 は優れたパフォーマンスを示しています。調査の結果は、次世代の 2 ソケット AMD Opteron プロセッサ・ベース・サーバーに期待されるものと一致するパフォーマンスを示しています。

## 1. はじめに

IBM System x 3455 は、IBM 製の次世代型 2 ソケット AMD Opteron プロセッサ・ベース・ラック最適化サーバー製品の一つです。このシステムは、1.8 GHz、2.2 GHz、2.4 GHz、2.6 GHz、および 2.8 GHz の速度の新しい Rev. F デュアル・コア・プロセッサをサポートしています。また、667 MHz DDR2 の 12 個の DIMM をサポートしています。メモリーのサイズは、DIMM 当たり 512MB から 4GB<sup>1</sup> までになり、合計のメモリー容量はシステム当たり最大 48GB に達します(下記の図 1 を参照)。また、x3455 は 2 つの I/O スロットもサポートしています。その内の 1 つは、PCI-Express (PCI-E) x16 スロットです。もう 1 つのスロットは、PCI-E x8 スロットまたは Hyper-Transport Expansion (HTX) スロットとして構成可能です。さらに、コールド・スワップ SATA ディスク・ドライブ 2 台、前面と背面に複数の USB ポート、および Gigabit Ethernet ポート 2 つを収容するスペースも確保されています。

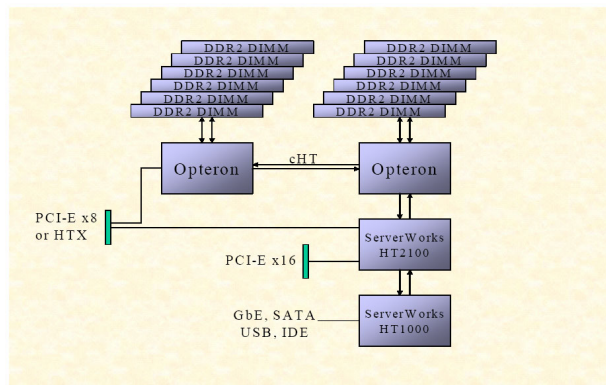


図 1. x3455 のブロック・ダイアグラム

## 2. メモリー・パフォーマンス

x3455 は、あらゆるマルチプロセッサ Opteron 設計と同じように、共有アドレスの Non-Uniform Memory Access (NUMA) 設計です。アドレス・スペースは、システム内のすべてのメモリーにわたるので、すべてのプロセッサがメモリー内の任意の場所にデータを保管したり、任意の場所からデータを検索したりすることができます。ただし、プロセッサに直接接続されているメモリーの方が、他方のプロセッサに接続されているメモリーよりも速く検索できます。参照により L1 と L2 の両方のキャッシュにデータが見つからない場合、メモリー・サブシステムは、キャッシュ内にデータがあるかどうかを他方のプロセッサに尋ねなければなりません。これは、スヌープ要求と呼ばれます。プロセッサはデータで応答するか、要求されたデータがないことを知らせるメッセージで応答する必要があります。これは、スヌープ応答と呼ばれます。ローカル・メモリーの参照では、ローカル・メモリーから収集されるデータを使用する

1 x3455 は、8GB (2x4GB) DIMM オプションが入手可能になれば、このオプションをサポートする予定です。

前に、スヌープ要求とスヌープ応答が必要です。また、リモート・メモリー (他方のプロセッサに接続されているメモリー) の参照では、HyperTransport リンクを経由してデータが送信される待ち時間が追加されます。

メモリーのパフォーマンスは、メモリー帯域幅またはスループット、および待ち時間の形で示すことができます。システムの合計帯域幅は、システム内のチャンネル数と各チャンネルの速度との積です。これはパフォーマンスの上限を示します。パフォーマンスの上限とは、決して超えないことが保証されると共に、必ずしも達成できるとは限らない量です。フル構成時の x3455 の帯域幅は 21,333 MB/秒です。メモリー・スループットは、実際に達成可能な測定値です。このテストでは、メモリー・スループットの測定に Stream ベンチマーク [1] を使用しました。Stream Scale ベンチマークを使用して 2 プロセッサ構成で 12,803 MB/秒を達成できました。また、メモリー待ち時間は、データの局所性とメモリー・ロードに応じて、48 ns 弱から 105 ns 強までが測定されました。この測定には、pChase と呼ばれるカスタム・ベンチマークを使用しました。

Stream ベンチマークは、メモリー・スループットのテストを目的として設計されました。このベンチマークは、科学および技術コンピューティングからの 4 つの一般的な演算を使用します。4 つの演算とは、Copy、Scale、Add、および Triad です。Copy と Scale は、保管される 64 ビット値ごとに 1 つの 64 ビット浮動小数点値を要求します。Add と Triad はそれぞれ、保管される値ごとに 2 つの値をロードします。4 つの演算はすべて、最大キャッシュよりはるかに大きい一連のメモリー全体で順次にデータを要求し、保管します。この方法で、HPC (high-performance computing) アプリケーションで頻繁に行われる重要な演算の一部を概算します。

pChase ベンチマークは、ポインター追跡ベンチマークです。必要なサイズのメモリーを満たすために、ポインターのチェーンが設定されます。これらのポインターは、メモリーを順次に参照しないように特別に設定されます。キャッシュ・ラインごとに 1 つのポインターだけが使用されます。このチェーンは、経過時間がシステム・クロック・レゾリューションよりも相当長くなるまで繰り返され、正確な測定が行われたことを保証します。プロセッサ・コアごとにたった 1 つのスレッドが実行されます。このベンチマークは、ロードとアンロードの遅延だけでなく、ローカルおよびリモートの遅延も測定できます。pChase と Stream のどちらのベンチマークも、さまざまな条件がメモリーのパフォーマンスに与える影響を判別するのに使用されます。

特に指定のある場合を除いて、本書におけるパフォーマンス結果はすべて、2.8 GHz プロセッサおよび 12GB の CL5、667 MHz DDR2 メモリー (12 個の DIMM) を使用する x3455 で測定されたものです。BIOS では、Node Interleave と ChipKill (どちらもデフォルト設定) を使用不可にするように設定しました。(これらの設定の影響については、「IBM System x 3755 のパフォーマンス」(Douglas M. Pase および Matthew A. Eckl 著[2]) を参照してください。)

## 2.1 プロセッサ・スケーリング

プロセッサ・スケーリングは、パフォーマンス分析の興味深い特徴の 1 つです。プロセッサ・スケーリングは、プロセッサをシステムに追加したり、システムから取り外したりするときにパフォーマンスがどのように変化するかを示します。図 2 は、2 プロセッサ構成のスループットが、シングル・プロセッサ構成のスループットの約 2 倍であることを示しています。その理由は、各 Opteron プロセッサには独自の内蔵メモリー・コントローラーがあり、2 つ目のプロセッサをシステムに追加すると、メモリーにチャンネルが追加されるからです。さらに、リモート・プロセッサに対するデータ変更のスヌープは、ローカル・メモリーが読み取られる前に完了するので、スループットは減少しません。

メモリー待ち時間は、プロセッサ数の影響を比較的受けないように見えますが、2 プロセッサ・システムはシングル・プロセッサ・システムより少し高速であることがやや興味がある点です。これは、現在真相がつかめていない影響です。

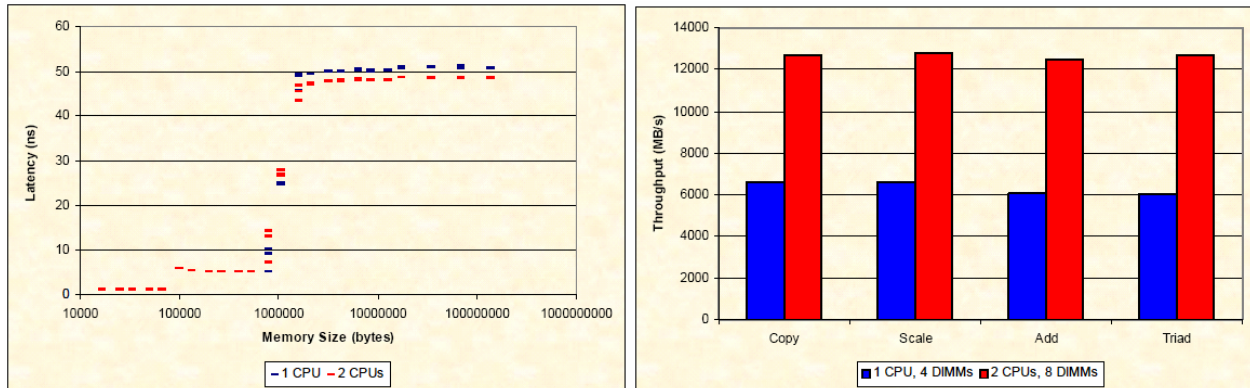


図 2. プロセッサ・スケーリングがメモリー・パフォーマンスに与える影響

## 2.2 装着されている DIMM の数

次に考察を加える項目は、DIMM 数がメモリー・パフォーマンスに与える影響です。DIMM 数が 2 の累乗 (たとえば、2 または 4) である場合、メモリー・コントローラーは、メモリー・アクセスをより高速にする一定の最適化を実行できます。このような最適化の 1 つは、順次アクセスにおける競合を減らすように DIMM 間のアドレスをインターリーブすることです。DIMM 数が 2 の累乗ではない場合、この種のインターリーブ操作はより複雑であり、メモリー・コントローラーの処理は簡単ではありません。

ここで興味深い質問があります。すなわち、システム内の DIMM が 8 個ではなく、12 個ある場合はどうなるでしょうか。メモリー容量を増やすためにパフォーマンスを犠牲にしますか? 図 3 は、この質問の答えを示しています。この図は、興味深く、やや予想外の結果をいくつか示しています。プロセッサ 1 個の場合には、4 個の DIMM を使用する場合、6 個の DIMM と比較すると約 10% スループットが低下し、待ち時間が長くなることを示しています。しかし、プロセッサ 2 個の場合には、8 個の DIMM のスループットが、12 個の DIMM の場合より 10% 強向上しています。ただし、待ち時間は同じです。

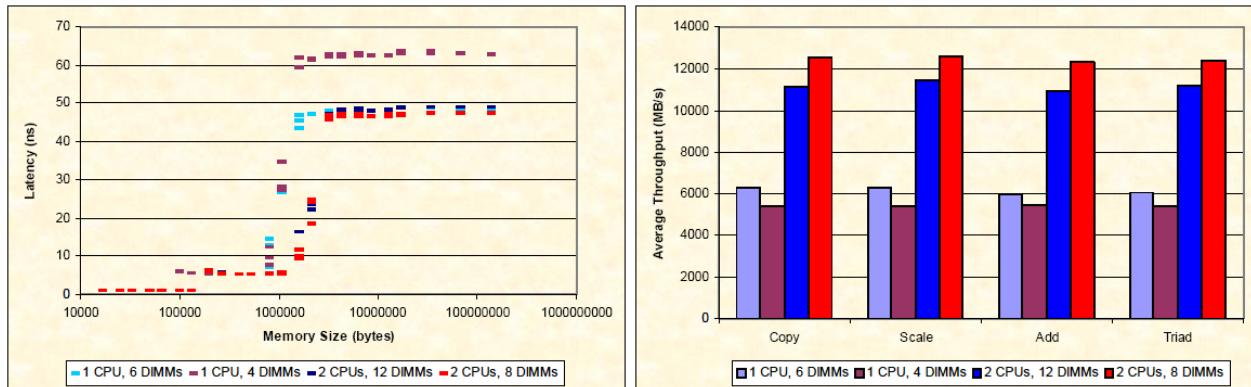


図 3. 装着された DIMM 数がメモリー・パフォーマンスに与える影響

### 2.3 リモート待ち時間

どの NUMA システムにも、ローカル・メモリーとリモート・メモリーがあります。実際に、NUMA の大きな利点は、フラット・メモリーの方が処理が簡単ではあるものの、いくつかのプロセッサで他のプロセッサよりも近くにメモリーを配置すると、少ない費用でパフォーマンスが向上する可能性があることです。これは局所性の利用に重点を置いたものですが、リモート参照は引き続き行われるので、対処が必要です。図 4 は、リモート待ち時間がローカル待ち時間よりも約 30 ns 長いことを示しています。

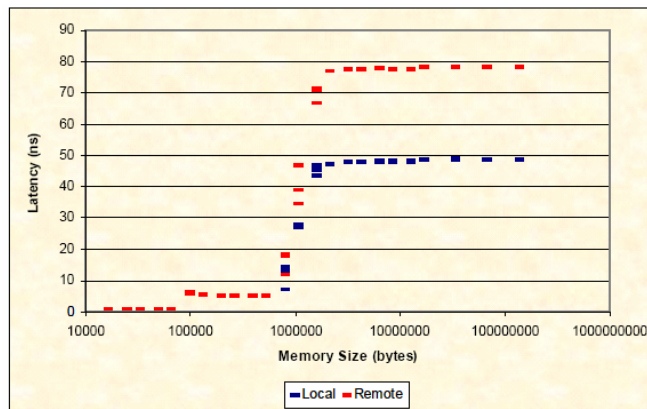


図 4. メモリー待ち時間に対するリモート参照の影響

### 2.4 メモリー・ロード

メモリーの負荷とは、ある時点でメモリーを同時に参照する数がどのくらいあるかであらわされます。サーバーはシステムのスループットを最大化するように設計されていますが、当然メモリー・スループットも含まれます。複数の独立したメモリー参照を同時に実行することはよくあることです。Stream ベンチマークは Add および Triad 演算を使用するときにこの負荷をかけてます。各反復は、保管操作が行われるごとに 2 つのオペランドをメモリーに要求します。また、このベンチマークは、複数のくり返しオペレーションによりメモリーを参照できるようにループを使用しています。このように Stream ベンチマークはメモリーに対し大きな負荷をかけることができます。

pChase ベンチマークもメモリー・ロードを測定しますが、もっと制御された方法で行います。Stream の場合とは異なり、未処理の並行参照数を直接指定できます。1 つの参照のみが処理中である場合、それは、アンロード待ち時間と呼ばれます。複数の参照が並行して実行される場合、ロード待ち時間と見なされます。次の実験では、複数のプロセッサ上での 1 つの参照と 2 つの並行参照とでロードが変化しました。図 5 にその結果が表示されています。

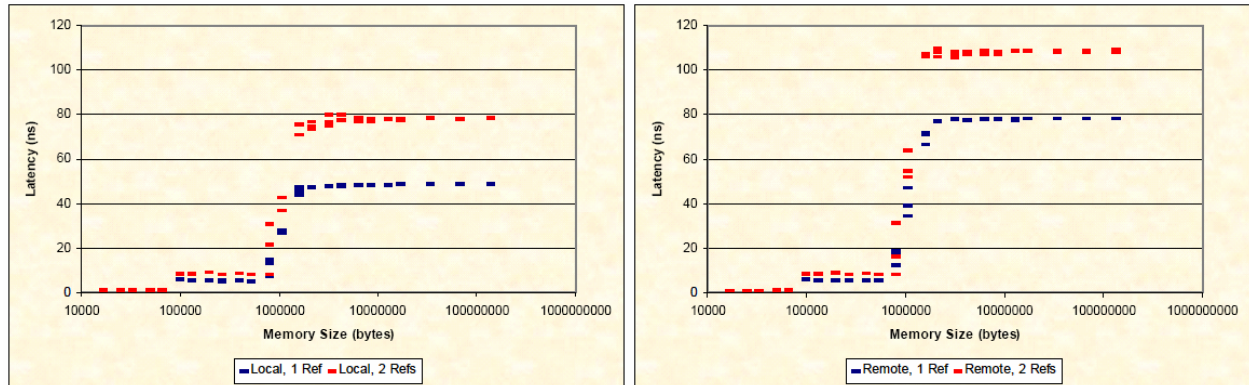


図 5. ローカル・メモリー待ち時間に対するロードの影響

予想されたように、ロードは、メモリー演算の待ち時間を明らかに増やします。参照がローカルであるか、リモートであるかにかかわらず、その影響はほぼ同じです。ローカルの場合、1 つの未処理参照と 2 つの未処理参照との間の待ち時間の差は約 30 ns です。つまり、最初の参照には 48 ns かかり、2 番目の参照には約 78 ns かかります。

### 3.64 ビット浮動小数点パフォーマンス

プロセッサのパフォーマンスは、技術の改善以上の変化は期待できない分野です。2.8 GHz デュアル・コア・プロセッサは、予想どおりのパフォーマンスを示します。このタイプのパフォーマンスを測定するために、High-Performance Linpack (HPL) ベンチマークを使用しています。

Linpack ベンチマークは、(使用可能なキャッシュ内に収まる) 簡単に使用できる long ベクトルを使用して、64 ビット浮動小数点ベクトル・パフォーマンスを測定します[3]。これは、メモリー・パフォーマンスが因数でないことを意味します。また、Linpack パフォーマンスは、プロセッサ・クロック周波数およびシステム内のベクトル (SSE2) ユニット数に比例してスケーラブルであることも意味します。パフォーマンスは、Problem size と使用可能なシステム・メモリーの関数でもあります。HPL は、MPI と呼ばれるインターフェースを介してメッセージを送信することによって情報を交換する複数のプロセスを使用するベンチマークです[4]。Problem size が大きくなると、ベンチマークは浮動小数点演算による支配が高まり、通信動作による影響が少なくなります。しかし、システムが漸近的パフォーマンス限界に近づくと改善の速度が低下します。12 GB のメモリーを使用するこのテストでは、システムは 19.6 GF/秒に達しました。これは、システムの最高パフォーマンスの 87.5% です。(2.8 GHz の 2 つのデュアル・コア・プロセッサを使用した場合、ハードウェアの最高パフォーマンスは 22.4 GF/秒です。) メモリーを増やすと、パフォーマンスはさらに上昇しますが、その量はごくわずかにすぎません。このシステムにおける HPL のパフォーマンス・プロファイルが図 6 に表示されています。

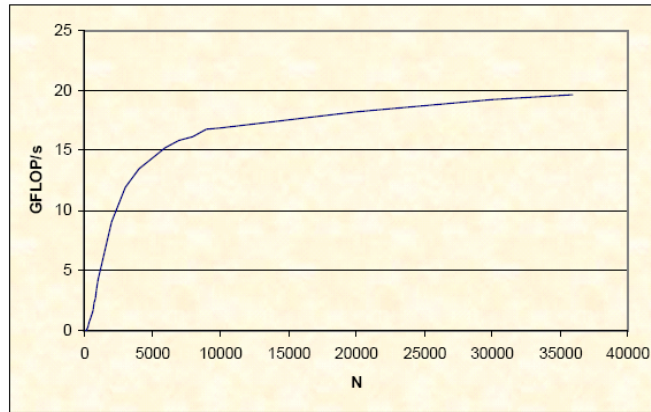


図 6.2 つの 2.8 GHz デュアル・コア・プロセッサ使用時の Linpack パフォーマンス

#### 4. SPEC CPU2000 パフォーマンス

SPEC CPU2000 ベンチマークは、実際には、8 つのベンチマークを 1 つにパッケージしたものです [5]。3 つの独立した特質から構成され、これらの特質を一緒に使用して各ベンチマークが識別されます。これらの特質は、演算タイプ (整数または浮動小数点)、実行モード (速度または比率)、およびコンパイル・モード (基本またはピーク) です。ベンチマークの実行ごとに、そのベンチマークの目的を反映する、コンピューター業界から選ばれたアプリケーションを使用します。各アプリケーションは 3 回実行され、3 回の実行の中間の実行時間に基づいてスコアが与えられます。次に、アプリケーションのスコアをすべて組み合わせて、幾何平均<sup>1</sup>を生成します。これがベンチマークのスコアになります。8 つの異なるベンチマーク・スコアが可能です。すなわち、整数基本速度、整数基本比率、整数ピーク速度、整数ピーク比率、浮動小数点基本速度などです。通常、基本結果とピーク結果はペアで使用されます。したがって、一般に整数速度ベンチマークの場合は SPEC CINT2000 と呼ばれ、整数比率ベンチマークの場合は SPEC CINT2000 Rates と呼ばれます。同様に、浮動小数点速度と浮動小数点比率のベンチマークは CFP2000 および CFP2000 Rates と呼ばれます。

整数ベンチマークは、14 個の整数アプリケーションを使用します。大部分のアプリケーションはキャッシュ可能であるので、全体的に、メモリー・パフォーマンスはベンチマーク・パフォーマンスにほとんど影響を与えません。浮動小数点ベンチマークは、一般にメモリー・パフォーマンスと浮動小数点パフォーマンスの両方への依存度が強い 12 個の浮動小数点アプリケーションを使用します。ただし、大型のキャッシュも、結果に影響を与える可能性があります。

SPEC CPU2000 は、速度ベンチマークまたは比率ベンチマークとして実行可能です。速度ベンチマークは、単一プロセッサ・コアを使用して各アプリケーションの単一コピーを逐次実行し、結果を報告します。これらの結果から、単一タスクを実行可能な速度が概算されます。比率ベンチマークは、各ベンチマークの複数コピーを並行して実行します (通常、プロセッサ・コア当たり 1 つのコピー)。これにより、フルロード時のシステム・スループットが概算されます。サーバーは多数のタスクを同時に実行する (場合によっては個々のタスクの速度を犠牲にして) ように設計されているので、一般に、比率ベンチマークの方がサーバー・パフォーマンスに適した測定です。

1 k 個の一連の数値の幾何平均は、これらの数値の積の k 乗根です。たとえば、x1、x2、x3、x4、および x5 の幾何平均は  $(x1 \times x2 \times x3 \times x4 \times x5)^{1/5}$  になります。

多くの場合、コンパイル・モードは、SPEC CPU2000 ベンチマークで最も誤解される特質です。基本ベンチマークとピーク・ベンチマークの相違点は、アプリケーションのコンパイルに使用できるコンパイラ最適化フラグのみです。基本ベンチマークを正しく実行するには、使用する最適化フラグは 4 つ以下でなければなりません。また、すべてのアプリケーションが同じ最適化フラグを使用する必要があります。ピーク結果では、任意の数のフラグを使用でき、アプリケーションごとにフラグが異なってもかまいません。基本結果から、コード開発などの環境が概算されます。コード開発では、開発者は、有効に機能するフラグの組み合わせを必要としますが、アプリケーションのコンパイルの微調整を好みません。我々の見解では、基本結果は、システムのパフォーマンスをあまりよく反映しません。むしろ、コンパイラが最適化フラグをどのようにうまくパッケージしたか、および汎用最適化フラグ (-O3 など) をどのようにうまく実装したかを反映します。

ピーク結果は、最適化フラグの数や選択で制約を受けないので、コンパイラは、サーバーの設計上の特徴を自由に利用することができます。その結果、ピーク結果の方が、システムのパフォーマンスをより正確に反映します。このレポートでは、ピーク結果のみが使用されます。

#### 4.1 整数結果

図 7 は、プロセッサ周波数別の整数ベンチマークの結果を示しています。このベンチマークは非常にキャッシュへの依存度が高いため、速度結果と比率結果の両方が、プロセッサ周波数に合わせてほぼ完全なスケラビリティを示しています。これらの結果はここに示されていませんが、このベンチマークはプロセッサ数と非常に密接に対応すると推定しても差し支えありません。

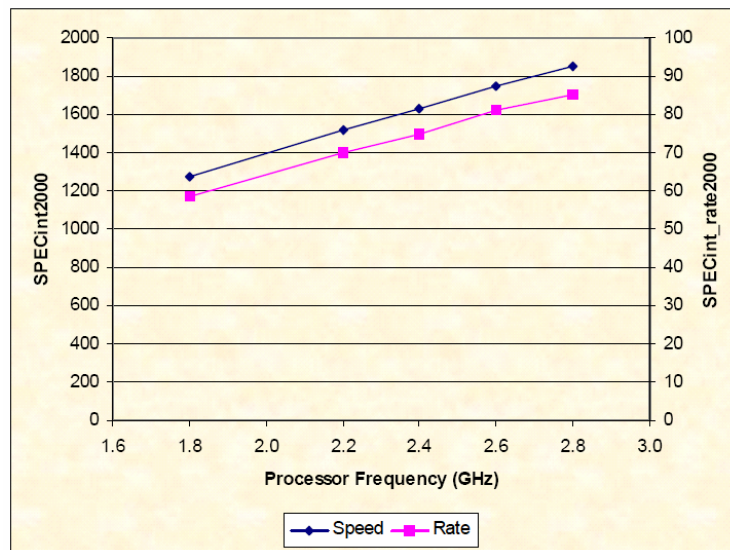


図 7. プロセッサ周波数別の SPEC CINT2000 速度結果と比率結果

#### 4.2 浮動小数点結果

図 8 は、プロセッサ周波数別の浮動小数点ベンチマークの結果を示しています。この図表から、パフォーマンスはプロセッサの周波数に対して非常に直線的にスケラリングすることが分かります。その理由は、システム内に十分なメモリー・スループットがあるため、メモリーのパフォーマンスが制限要因にならないからです。

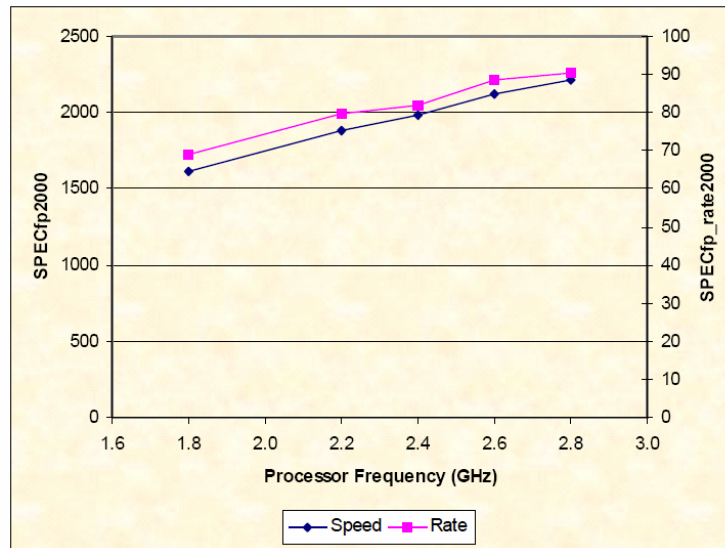


図 8. プロセッサ周波数別の SPEC CFP2000 速度結果と比率結果

このセクションですでに述べたように、このベンチマーク・スイートのパフォーマンスは、メモリーのパフォーマンスに大きく依存しています。これは、Opteron プロセッサ設計の 2 つの特徴に関係します。どちらの特徴も、内蔵メモリー・コントローラーと関係があります。第一の特徴は、プロセッサ周波数の増加に合わせてメモリー・パフォーマンスが向上することです。プロセッサ周波数が上がると、メモリー・コントローラーへの要求が高速化します。これは、図 8 の結果の直線性に大きく寄与しています。

内蔵メモリー・コントローラーの 2 番目の特徴は、システムにプロセッサが追加されるとメモリー帯域幅が増えることです。これらの結果を示していませんが、システムにプロセッサが追加されるときに、比率ベンチマークのパフォーマンス上昇が直線形に非常に近いと予想するのが妥当です。1 プロセッサの比率パフォーマンスは、2 プロセッサのパフォーマンスの半分であるか、おそらく半分よりやや上であると予想されます。

## 5. 結論

x3455 は、1 つまたは 2 つの AMD Opteron プロセッサをサポートする新規 IBM サーバーです。システム当たり 12 個の DIMM、すなわち最大 48GB のシステム・メモリーをサポートします。本書で提示した調査結果に基づいて出した結論は次のとおりです。

1. BIOS で Node Interleave 設定を使用可能にすると、大幅なパフォーマンス損失が生じる可能性があります。
2. BIOS で ChipKill を使用可能にすると、メモリー・システムの頑強性が向上し、メモリー・パフォーマンスは減少しません。
3. x3455 は、内蔵メモリー・コントローラーを搭載しているので、1 プロセッサと 2 プロセッサとの間でほぼ完全な直線型のプロセッサ・スケーリングを示します。
4. 1 プロセッサから 3 プロセッサまでのローカル (アンロード) メモリー待ち時間が短い。48 ns を測定しました。リモート待ち時間はやや長く、約 78 ns です。
5. 2 つの並行参照でのメモリー・ロードでは、ランダム・アクセスの場合のメモリー待ち時間に 30 ns が加算されます。
6. 64 ビット浮動小数点パフォーマンスは非常に高く、High-Performance Linpack で 19.6 ギガフロップス (87.5%) を実現します。
7. SPEC CPU2000 パフォーマンスも、予想と一致します。

## 6. 参考文献

- [1] Memory Bandwidth: Stream Performance Results, <http://www.cs.virginia.edu/stream/>.
- [2] Douglas M. Pase and Matthew Eckl, 「IBM System x 3755 のパフォーマンス」、[ftp://ftp.soft-ware.ibm.com/eserver/benchmarks/wp\\_x3755\\_081506.pdf](ftp://ftp.soft-ware.ibm.com/eserver/benchmarks/wp_x3755_081506.pdf), IBM, August 2006.
- [3] Douglas M. Pase, “Linpack HPL Performance on IBM eServer 326 and xSeries 336 Servers,” [ftp://ftp.software.ibm.com/eserver/benchmarks/wp\\_Linpack\\_072905.pdf](ftp://ftp.software.ibm.com/eserver/benchmarks/wp_Linpack_072905.pdf), IBM, July 2005.
- [4] Marc Snir, Steve Otto, Steven Huss-Lederman, David Walker, and Jack Dongarra, MPI—The Complete Reference, Vol. 1, 2nd Ed., MIT Press, 1996.
- [5] SPEC CPU2000, <http://www.spec.org/cpu2000/>.



© IBM Corporation 2006

IBM Systems and Technology Group

Department MX5A

Research Triangle Park NC 27709

Produced in the USA.

08-06

All rights reserved.

IBM、IBM ロゴ、BladeCenter、System x、eServer、および xSeries は、IBM Corporation の商標です。

AMD および Opteron は、Advanced Micro Devices, Inc. の商標または登録商標です。

SPEC、SPECint、および SPECfp は、Standard Performance Evaluation Corporation の商標です。

他の会社名、製品名およびサービス名等はそれぞれ各社の商標です。

製品の仕様その他の製品情報は、IBM により予告なしに変更されることがあります。本書に記載の製品、プログラム、またはサービスが日本においては提供されていない場合があります。日本で利用可能な製品、プログラム、またはサービスについては、日本 IBM の営業担当員にお尋ねください。IBM およびその直接または間接の子会社は、本書を特定物として現存するままの状態を提供し、商品性の保証、特定目的適合性の保証および法律上の瑕疵担保責任を含むすべての明示もしくは黙示の保証責任を負わないものとします。国または地域によっては、法律の強行規定により、保証責任の制限が禁じられる場合、強行規定の制限を受けるものとします。

本書には、IBM が制御または保守を行わない第三者のサイトへのリンクを含む場合があります。このような第三者のサイトは、お客様自身の責任でアクセスしてください。IBM は、これらのサイトの情報、データ、意見、助言、または記述の正確性または信頼性について責任を負いません。IBM は、これらのリンクを便宜のため記載しただけであり、決してこれらのリンク先を推奨するものではありません。

すべてのパフォーマンス情報は、管理環境下で決定されたものです。実際の結果は、異なる可能性があります。パフォーマンス情報は、明示的または黙示的な保証なしに、現存のままの状態を提供されます。購入の際には、購入を検討しているシステムのパフォーマンスを評価するシステム・ベンチマークなど、他の情報も参考にしてください。