次期ベクトル並列型スーパーコンピューターシステム SR11000 のご紹介

（株）日立製作所

１．はじめに

1999 年 3 月に納入したベクトル並列型スーパーコンピューターSR8000（128 ノードモデル）の後継機として、2005 年 3 月に、ベクトル並列型スーパーコンピューターSR11000 を納入する予定です。今回のベクトル並列型スーパーコンピューターシステムは、2 段階の納入となっており、2005 年 3 月の「フェーズ 1」では、SR11000 モデル J1（44 ノードモデル）を納入します。「フェーズ 1」の 2 年後（2007 年）の「フェーズ 2」では、「フェーズ 1」の 44 ノードにノードを増設して 128 ノードモデルにアップグレードします。ここでは、「フェーズ 1」の SR11000 モデル J1（44 ノードモデル）について、ご紹介します。

２．SR8000 と SR11000 の比較

表 1 に、東京大学情報基盤センターの現行ベクトル並列型スーパーコンピューターSR8000 と次期ベクトル並列型スーパーコンピューターSR11000（フェーズ 1）の比較を示します。

表 1．SR8000 と SR11000 の比較

項目	SR8000（現行機種）	SR11000（次機種フェーズ 1）
システム性能	1024GFlops	5350.4GFlops
ノード数	128	44
1 ノード当たりの 主記憶容量	8GB	128GB
ノード性能	8GFlops	121.6GFlops
演算プロセッサー性能	1GFlops	7.6GFlops
1 ノード当たりの 演算プロセッサー数	8	16
ノード間ネットワーク	3 次元クロスバー	3 段クロスバー
ノード間転送性能	1GB/s（単方向）×2	12GB/s（単方向）×2
OS 及び構成	HI-UX/MPP for SR8000 シングルイメージ OS	AIX 5L クラスター構成

３．ハードウェアの特長

SR11000 は、科学技術計算システムに要求される「高性能ノード」、「高スケーラビリティ（高いシステム性能）」、「高信頼性」を追求するコンセプトに基づいて開発しております。SR11000 のハードウェア構成を図 1 に示します。

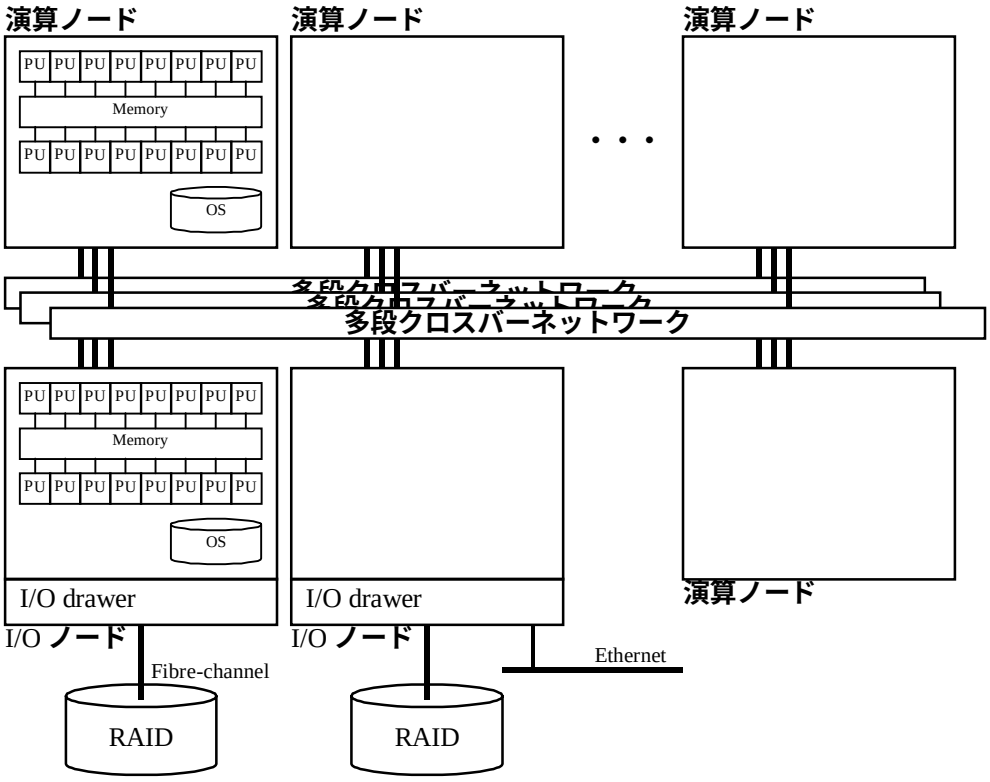

図 1. ハードウェア構成

SR11000 の特長は以下の通りです。

（1）高性能プロセッサーと大容量キャッシュによる高性能ノード

高性能ノードを実現するために、SR11000/J1 のマイクロプロセッサーは、最新の高速 RISC プロセッサーPOWER5 を採用しています。POWER5 は、最新の半導体テクノロジーを採用したものであり、周波数、集積度ともに、世界トップレベルです。SR11000/J1 の演算ノードは、この POWER5 プロセッサー16 台を高いメモリーバンド幅のスイッチで接続した高性能 SMP（Symmetric Multi Processor）構成となっています。また、SR8000 と同様に、日立独自のプロセッサー間高速同期機構を組み込み、ショートベクトル性能を改善しています。

SR11000/J1 では 288MB/ノードの大容量 L3 キャッシュにより、高いキャッシュヒット率が確保でき、実効メモリーレイテンシーの削減により、様々なアプリケーション性能が向上します。また、コンパイルや対話処理等のスカラー処理、I/O 処理性能も向上します。

（2）高度なプリフェッチ技術による高い実効メモリー性能

ハードウェアプリフェッチは、ハードウェアが自動的に主記憶データーアクセスのパターンを検出し、先読みデーターを主記憶メモリーーキャッシュ間で効率的なデーター転送を行い、予めキャッシュにデーターを取り込むことにより、キャッシュヒット率を向上し、実効メモリーレイテンシーの削減と実効メモリースループットの向上を図る機能です。

ソフトウェアプリフェッチは、コンパイル時に生成したプリフェッチ命令により、主記憶からキャッシュメモリーへのデーター転送を、プログラムのループ中でそのデーターを参照

するまでに完了させることにより、実効メモリーレイテンシーを削減し、メモリースループットを向上する機能です。

ハードウェアプリフェッチは、ソフトウェアプリフェッチと異なり、ハードウェアが自動的に検出し、起動されるので、チューニングレベルの低いプログラムや、低い最適化レベルでコンパイルしたプログラム、C 言語のようなポインター型を使用したプログラムにおいても、実効性能を向上することができます。

ハードウェアプリフェッチとソフトウェアプリフェッチの動作例を表 2 に示します。

表 2. ハードウェアプリフェッチとソフトウェアプリフェッチの動作例

アクセスパターン	ハードウェアプリフェッチ	ソフトウェアプリフェッチ
連続アクセス	ハードウェアが自動検出	コンパイラーが自動検出
インデックスロード A(L(I)) （L(I) = 1, 2, 3, …）	ハードウェアが自動検出	コンパイラーの自動検出不可 （ユーザー指示文にて対応）

（3）高いメモリースループット

メモリーコントローラーのプロセッサーチップへの統合、弊社独自の太い専用パスによるプロセッサーチップ間の接続、主記憶への DDR2 SDRAM メモリーの採用、等により、高メモリースループットと低レイテンシを同時に実現します。

（4）高速ノード間ネットワーク：多段クロスバーネットワーク

SR11000 では、ノード間ネットワークとして「多段クロスバーネットワーク」を採用しました。「フェーズ 1」、「フェーズ 2」いずれも 3 段構成です。

「多段クロスバーネットワーク」は、任意のノード間データ転送間での衝突を最小限にし、システム全体の高いスケーラビリティを実現しています。また、衝突の少ないネットワークは、複数ノード JOB を異なるノードグループで実行しても、ほぼ同一の性能を達成できるなどの特長があります。

（5）運用機能の充実

・LPAR（Logical PARtitioning）機能によるノードの分割運転
  CPU、メモリー、ノード間ネットワーク、I/O アダプター等の各資源を分割し、一つのノードを二つの論理ノードとして使用することが可能です。
・CSM-MS（Cluster Systems Management-Management Server：管理コンソール端末）からの操作による全ノードもしくは指定したノードグループの電源投入と立上げ／停止と電源断
・自動運転装置との連動

（6）障害処理

・RISC プロセッサーには、データー／アドレス系に ECC（Error Correction Code）またはパリ

ティチェッカーを備えます。
・L1 キャッシュには、パリティチェックによる誤り検出を備えます。L2/L3 キャッシュには、1 ビット誤り訂正、2 ビット誤り検出可能な ECC を備えます。主記憶にはメモリー素子の 1 チップ誤り訂正、2 チップ誤り検出可能な ECC を備えます。
・主記憶には、メモリーラインの 1 ビットエラーが閾値に達した場合に、ラインを予備メモリーチップに交代するメモリービット交代機能を備えます。
・ノード間ネットワークの転送路には CRC（Cyclic Redundancy Check）を設け、1 ビットエラー検出時は、メッセージを再送します。
・ハードウェア障害情報の自動通報機能を備えます。

４．ソフトウェアの特長

SR11000 の OS（オペレーティングシステム）には、米国 IBM 社の AIX 5L を採用しています。コンパイラーや数値計算ライブラリーは、弊社の最適化 FORTRAN や MATRIX/MPP といった製品をご提供し、SR8000 上でご利用いただいているプログラムを変更することなく、原始プログラムを SR11000 上でリコンパイルするだけで移行できます。なお、SR8000 でご提供してまいりました、ノード間通信ライブラリーである RemoteDMA 転送ライブラリーは SR11000 でのご提供ができないため、MPI への移行をお願いいたします。

4.1　OS（AIX 5L）

AIX 5L は、64 ビットアドレッシング対応など多彩な先進テクノロジーを提供する最新の UNIX OS です。Linux とのシームレスな連携や業界標準に対応したオープンな環境を提供します。AIX 5L の特長は以下の通りです。

（1）64 ビットアドレッシング対応
・64 ビットアドレッシングによって、1TB（テラバイト）のファイルや 2GB を超えるメモリーを利用した大規模なプログラムを高速に処理できます。
（2）多彩な機能
・1 つのノードを 2 つのパーティションに分け、パーティションごとに異なるシステム環境を構築できます。両パーティション間で影響を与えることなく、独立した運用ができます。
・LVM（Logical Volume Manager）は、物理ディスクを任意の構成の論理ボリュームとして管理し、柔軟なディスク運用ができるようにします。耐故障性の高いファイル入出力処理を実現するディスク・ミラーリング機能も備えています。
・システムハングアップの検出、自動システム再起動、ネットワークアダプター障害時の経路交代機能、および障害連続多発時のエラーログの抑制など、強力な障害対応機能を提供しています。
・メインフレーム並みの約 2000 種類の情報が採取できる強力なトレース機能を備えています。トレースレポート機能によって、フォーマットファイル（テンプレートファイル）のルー

ルに従って、ログファイルを編集する機能を提供します。

・TCSEC（Trusted Computer Security Evaluation Criteria。通称オレンジブック）の C2 相当のセキュリティサービス（パケット認証、一貫性、アクセス制御など）を提供し、システムの信頼性や安全性を高めることができます。Native Kerberos V5 Network Authentication Service の認証機能を提供しています。

（3）オープン性

・Linux Affinity と呼ばれる Linux 環境を提供しています。これは、Linux のプログラムソースの移植性を高めるもので、アプリケーション開発者の負担を軽減します。具体的には、AIX 上で動作するミドルウェアやアプリケーションを、Linux 上で開発する場合と同じ API で開発できます。これによって、PC Linux を開発機として利用できます。このソース互換機能を使用したミドルウェアやアプリケーションでも、AIX 固有の HA（High Availability）機能などを使用できます。

・Solaris の SVR4（System V Release4）のコマンドを提供して、System V 系の経験者に対する親和性も高めています。

・AIX 5L は、次の業界標準や国際標準に対応しています。
UNIX98、XPG4、POSIX1003.1-1996（1003.1c）、POSIX1003.2-1992、
Motif2.1、X Window System Version 11 Release 6.3、IPv6、sendmail8.11.0

4.2　Hitachi Cluster Shared Extended Storage for AIX

SR11000 システムの各ノードにおいて、主記憶を利用した拡張記憶機能をサポートします。この機能により、S-3000 シリーズや SR8000 シリーズの拡張記憶装置を利用したプログラムを容易に SR11000 に移行することができます。

この拡張記憶機能では、主記憶上に構成される仮想的な空間上にディスク装置イメージを持つファイルシステムを作成し、高速ストレージ機能を実現します。ジョブ間のデーター引継ぎなどの中間ファイルとして使用することによりスループットの向上を図るとともに、拡張記憶機能によって確保した空間を擬似デバイスドライバとしてサポートすることで raw I/O によるアクセスにも対応することができます。さらに、ディスクへの入出力を伴わないファイルシステムとして利用することも可能です。この拡張記憶機能は、複数のノードの主記憶を統合して提供できるため、ノード数に応じて容量を拡張することが可能です。

4.3　最適化 FORTRAN77、最適化 FORTRAN90

最適化 FORTRAN77 は、米国標準規格 ANSI X3.9-1978 及び、JIS X 3001-1982（FORTRAN77）上位水準に準拠した FORTRAN コンパイラーです。最適化 FORTRAN90 は、ISO 国際規格 ISO1539;1991、米国標準規格 ANSI X3.198-1992 及び、JIS X 3001-1994（Fortran90）/JIS X3001-1:1998（Fortran95）規格に準拠した Fortran コンパイラーです。最適化 FORTRAN90 は、OpenMP 2.0 仕様をフルサポートします。

最適化 FORTRAN77/90 は、ホストや WS で実現した強力な最適化機能に加えて、SR11000

のハードウェアの性能を最大限に引き出すために、各種最適化機能を提供します。プロセッサー間高速同期処理とコンパイラーの自動並列化技術により、ベクトル機としての利用を実現します。また、単体プロセッサー性能を引き出すためのソフトウェアプリフェッチ機能も合わせて提供します。

最適化 FORTRAN77/90 は以下の特長を有します。

（1）SMP 並列化機能

以下に示す SMP 並列化機能を有します。

・高度な自動並列化機能
・高精度なプログラム解析能力
・自動プライベート化、リダクション並列化
・ループ構造変換による並列化、パイプライン並列化等の先進的並列化
・自動並列化支援のための各種指示文を用意
・高速並列処理方式と同期削減最適化により、高い並列化効率を実現

（2）最適化機能

以下に示す最適化機能を有します。

・3 つの最適化レベル（レベル 0、レベル 3、レベル 4）と、2 つの統合オプション（-Os、-Oss）、CPU オプションにもとづき最適化を実施
・ハードウェア性能を最大限引き出すための豊富な最適化
・ループ構造変換最適化
・命令レベル最適化
・広域自動手続き自動インライン機能
・プロファイル最適化
・その他、一般的最適化の殆どすべてを実装
・各種最適化指示文を用意

4.4　最適化 C

最適化Ｃは、国際標準規格 ISO/IEC 9899:1990 及び米国標準規格 ANSI X3.159-1989 に準拠する C コンパイラーです。また、コンパイルオプションの指定にしたがって旧言語仕様（K&R 仕様）に対応した互換仕様を利用できます。

最適化 C は、ホストや WS で実現した強力な最適化機能に加えて、SR11000 のハードウェアの性能を最大限に引き出す自動ノード内並列化機能を提供します。また、単体プロセッサー性能を引き出すためのソフトウェアプリフェッチ機能も合わせて提供します。

4.5　最適化標準 C++

最適化標準 C++は、国際標準規格 ISO/IEC 14882:1998 をサポートし、高度な最適化機能を持つ C++言語のコンパイラーです。また、コンパイルオプションの指定により、旧言語仕様（ARM 仕様）に対応した互換仕様を利用することもできます。

最適化標準 C++は、各種最適化技術に加え、SR11000 のハードウェアの性能を最大限に引き出す自動ノード内並列化機能を提供します。また、単体プロセッサー性能を引き出すためのソフトウェアプリフェッチ機能も合わせて提供します。

4.6　数値計算副プログラムライブラリー MSL2

MSL2（Mathematical Subprogram Library 2）では、数値計算をする上で必要となる代表的な数値計算上の手法を提供します。利用者が作成した FORTRAN 言語または C 言語で作成された主プログラムから呼び出すことによって、手軽に利用できます。MSL2 には、次の特長があります。

・信頼性の高い手法を集めています。また、同一の目的に対して数種の手法による副プログラムを用意し問題への適応性を高めています。
・通常のデーターだけでなく、性質の悪いデーターについても配慮して設計しています。引数の内容については、特に厳重にエラーチェックをしています。
・引数の名称および並び順を統一し、使いやすくしています。
・エラーコードは、ライブラリー全体で統一されていて、副プログラム実行後、このコードを調べることによって、利用者のプログラム内で適切な処置ができるようになっています。また、エラーのレベルに応じてメッセージのリスト出力を制御できます。

4.7　行列計算副プログラムライブラリー MATRIX/MPP

MATRIX/MPP（MATRIX calculation subprogram library / Massively Parallel Processors）は、基本配列演算、連立一次方程式、大規模疎行列を係数行列とする連立一次方程式、逆行列、固有値・固有ベクトル、高速フーリエ変換、擬似乱数といった技術計算の分野でよく使われる機能において並列計算機の性能を充分に引き出せる工夫をした副プログラムライブラリーです。FORTRAN 言語または C 言語で作成したユーザープログラムから呼び出して利用することができます。MATRIX/MPP には、次の特長があります。

・科学技術計算分野で広く利用される差分法や有限要素法による離散化処理で現れる、大次元の疎行列を係数行列とする連立一次方程式を効果的に解くための、収束性の良い反復解法をサポートしています。
・従来のベクトル計算機上や WS 上で、MATRIX 関連製品を利用して数値計算プログラムを開発したユーザーに対して、同一のインターフェースを提供します。これにより、ユーザーはベクトル計算機、WS 等から SR11000 への移行の際、プログラムをリコンパイルするだけで実行することができます。

4.8　行列計算副プログラムライブラリー MATRIX/MPP/SSS

MATRIX/MPP/SSS（MATRIX calculation subprogram library/MPP/Skyline Sparse matrix Solver）は、大次元疎行列に対するライブラリーです。分散メモリー型並列計算機向きに最適化しており、並列計算機の特長を充分に引き出すことができます。これは、FORTRAN 言語または C

言語で作成したユーザープログラムから利用することができます。MATRIX/MPP/SSS には、次の特長があります。

・スカイライン法の一般的な手法のほかに、複数の要素を塊としてブロック形式のように処理することによって、より高速なセル形式スカイライン法をサポートしています。
・非ゼロ要素の構造に着目し、ゼロ要素との演算を抑止することで、スカイライン法に比べ、演算量を大幅に削減したスパースソルバー機能をサポートしています。
・スパースソルバーもスカイライン法同様、一般的な手法のほかに、複数の要素を塊としてブロック形式のように処理することによって、より高速なセル形式スパースソルバーをサポートしています。
・スパースソルバーの前処理において、ゼロ要素が計算過程で非ゼロ要素となる Fill-in 数を少なくするオーダリング機能をサポートしています。
・既存のスカイライン法ソルバーまたはスパースソルバーとの置き換えが可能です。この置き換えによって、今まで使用していた構造解析プログラムの処理時間を大幅に短縮することができます。
・従来のベクトル計算機上や WS 上で、MATRIX 関連製品を利用して数値計算プログラムを開発したユーザーに対して、同一のインターフェースを提供します。これにより、ユーザーはベクトル計算機、WS 等から SR11000 への移行の際、プログラムをリコンパイルするだけで実行することができます。

以上