計算機アーキテクチャ 142－9
ハイパフォーマンス
コンピューティング 85－9
（2001．3．8）

ワールドワイドメタコンピューティングの試みについて

今村 俊幸[†], 村松 一弘[†], 北端 秀行[†‡], 金子 勇[†*],
山岸 信寛[†], 長谷川 幸弘[†], 武宮 博[†], 平山 俊雄[†]

[†] 日本原子力研究所 計算科学技術推進センター
[‡] 科学技術振興事業団 科学技術特別研究員
*(現) エクス・ツールス株式会社

世界各国の計算機資源のみならず様々なネットワーク上の装置を有機的に結合し，一つの仮想計算機システムを構築する試みとしてメタコンピューティングが提案されている．日本原子力研究所では，これまで所内 LAN での仮想計算機上を構築し数値アプリケーションの実験を行ってきたが，さる 2000 年 11 月アメリカ，ダラスにて開催された国際会議 SC2000 期間中に日独米英 4ヶ国のスパコンを結合して世界規模での実験の試みに成功した．本実験では，放射線情報推定システムを題材として世界 5 機関の並列計算機を利用し最大計 510CPU の仮想計算機の構築並びに，仮想計算機上での計算を行った．また，計算と同時に仮想計算機から大気中に放出された放射性物質の拡散過程を可視化することも実施した．本報告では，世界規模での実アプリケーションの実験の概要とその結果についてまとめる．

Worldwide metacomputing experiment at SC2000

Toshiyuki Imamura[†], Kazuhiro Muramatsu[†], Hideyuki Kitabata[†‡], Isamu Kaneko[†*],
Nobuhiro Yamagishi[†], Yukihiro Hasegawa[†], Hiroshi Takemiya[†], Toshio Hirayama[†]

[†]Center for Promotion of Computational Science and Engineering,
Japan Atomic Energy Research Institute
[‡]Domestic Research Fellow from Japan Science and Technology Cooperation
*ExpressionTools Inc.

Metacomputing, which enables us to construct a virtual computer system with some computer resources or experimental devices via internet connection, was proposed. Japan Atomic Energy Research Institute, JAERI, also continued to carry out several numerical simulations on a virtual computer system even though it was restricted in the JAERI's LAN. At SC2000, we had an opportunity to construct a worldwide virtual supercomputer with help of several supercomputer centers at Germany, US, UK and Japan, and we succeeded to execute a "Quick responsible source estimation system" with 510 processor units on 5 sites. Furthermore we demonstrated a real-time visualization for the dispersion process of radioactive particles released into atmosphere. In this report, we summarized the result of worldwide metacomputing experiment.

1 緒言

計算機能力やネットワーク性能の飛躍的な発展は大規模な問題への強力な武器となってきたことは事実であるが，常に現状の最高性能の計算機以上の能力を要求する問題は多く存在する．ネットワーク上に散在する様々な計算リソースは，その数を単純に数え上げるだけで (スーパコンピュータもそれに含まれるので当然ではあるが)，スーパコンピュータを凌駕するものとなる．分散する並列計算機を接続しより高性能な仮想並列計算機を構築する試みは，大規模科学技術計算を実践するために不可欠なものと考えられてきている．この試みは，メタコンピューティングの一領域であるとと

らえられ，いくつかの計算機センターが国際会議SC'XY期間中にお互いのスーパコンピュータを利用して実験を行ってきた[1].

本稿では，SC2000期間中に原研が行った世界規模のグローバルメタコンピューティングについて報告する.

2 グローバルメタコンピューティング実験

SC2000におけるグローバルメタコンピューティング実験では，各参加機関: シュツッツガルト大学(HLRS)，ピッツバーグ計算機センター(PSC)，サンディア研究所(SNL)，電総研(ETL)，マンチェスター大学(CSAR)，電脳中心(NCHC)，原研(CCSE)，北陸先端大(JAIST)，がそれぞれのセンターで所有する並列計算機をネットワーク上での仮想計算機構築のためのテストベッドとして資源の確保・提供を行うとともに科学技術計算アプリケーションを持ちより，メタコンピューティングの実践を行った. 参加した機関の資源を図1に示した. Web等で公表されている数値によると，CPUの総数は2629，総論理性能は2.78TLOPSに達した. SC2000の期間のみ実施可能ではあったが，世界規模の実験に相応しい環境と言える.

原研は，"Risk management system for environmental crisis" というテーマを打ち立て，放射能放出源推定システムを構築するとともに日独米英4ヶ国の仮想計算機での実験を行った.

システム全体は次の4サブシステムから構築されそれぞれ密接に結び付いた関係となっている. 利用者とのGUIに相当するTMEを除く計算システムの関係は図2のようになる.

- 放射能放出源推定サブシステム
- 異機種並列計算期間通信ライブラリ: Stampi
- 実時間可視化サブシステム: PATRAS
- TME (Task Mapping Editor)

以下，システム構築に要した各要素技術に関して説明を行う.

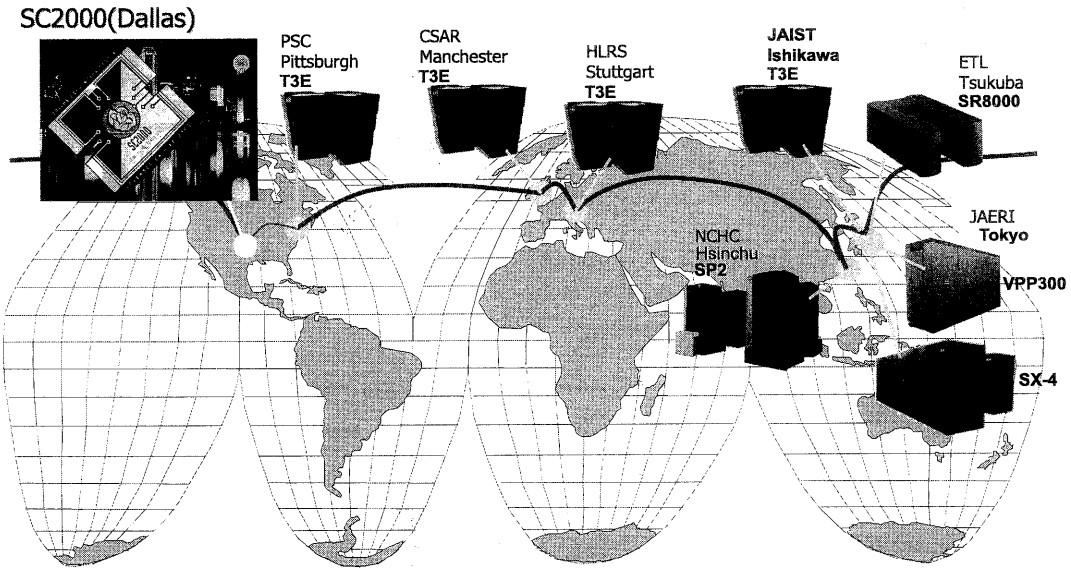

図 1: SC2000における仮想計算機テストベッド

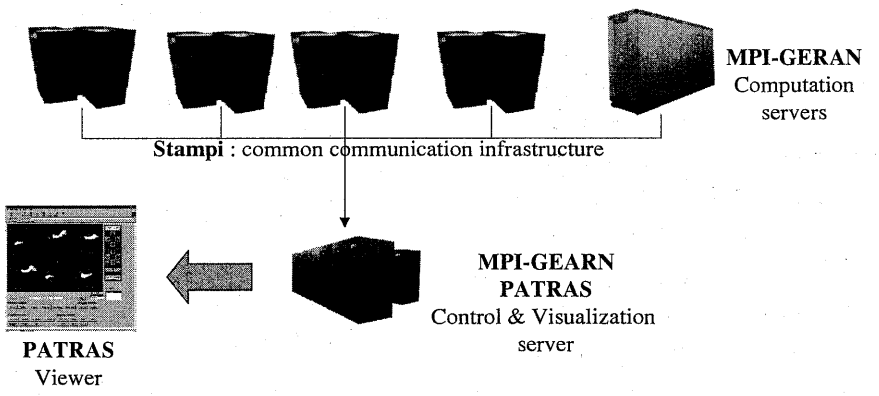

図 2: 各サブシステムの関係

2.1 放射能放出源推定サブシステム, MPI-GEARN

原研では国内 (外) でのモニタリングポイントにおける異常を検出し, モニタリング情報と大気拡散計算からブラックボックスである放出地点や放出量を精度よく迅速に計算できるシステムを開発してきている [2]. 本システムで用いられる推定手法は, 従来の検出地点からの逆流線解析ではなく移流・拡散過程を含んだ順方向の大気拡散計算と統計解析を組み合わせた手法である.

システムの中核となるのは, 大気拡散計算 (MPI-GEARN) と統計解析処理でありその一連の処理過程は図 3のようになる (中核部分を斜線・太枠で示した). 順方向の大気拡散計算は想定する初期条件の個別の計算であるため, 本質的に Embarassingly Parallel の典型でありその初期条件に応じた個数の計算資源 (ここでは CPU) を用意することでその計算時間を短縮できる. 仮想計算機上では, 数百ケースの初期条件を各 CPU に担当させることができるので図 2にもあるようにこの部分を分散化し使用可能な計算機リソースを使うようなシステム構成を取った. 初期条件は, ここでは (放出地点, 放出開始時間, 放出継続時間, 放出量) の 4 パラメータであり, モニタリング情報から経験的にパラメータの候補が想定でき統計解析に必要な十分な精度を与えるサンプル数の初期条件を生成し解析計算を行う.

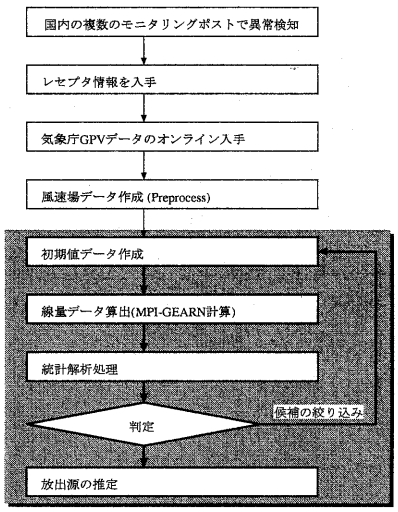

図 3: 放出源推定のフローチャート

仮想計算機上では, 初期条件を管理するとともに担当する CPU に初期条件をブロードキャストするマスタと, 実際の拡散計算を行う複数のスレーブ計算機の構成を取ることができる. マスタは, 各スレーブでの統計結果を集約して測定地点での観測データと, 相対的に最も一致するケースの情報を推定結果とする. マスタは後述する可視化システムのサーバとしても機能するが, 詳細については可視化システムで示すこととする.

2.2 異機種並列計算期間通信ライブラリ: Stampi

Stampi[4] は, ネットワークで接続された複数の並列計算機を利用して並列分散計算を実現するために, 並列計算機相互間の通信を MPI の範囲内で実現可能とするための拡張ライブラリである. ベンダー提供の MPI ライブラリは通常並列計算機内部での通信のみをサポートし当然別ベンダーの計算機への通信はサポートしないが, Stampi は通信相手に応じて TCP ソケットによって通信を実現する. また, MPI に準拠した既存プログラムを修正することなく複数の並列計算機上で実行できるばかりか, MPI-2[5] で盛り込まれた一部機能を実現している. 以下 Stampi の特徴を示す.

1. 並列計算機 [内外] 通信での通信機構の自動選択
2. 間接通信の実現
3. 動的なプロセス管理機能
4. データフォーマットの自動変換機能
5. Stampi/Java の実現

1~4 の特徴は複数の並列計算機を使う上で必要となる内部 / 外部プロトコルの選択やルーティングなどを利用者から隠蔽するものであり, 利用者は複雑な内部機構を意識することなく MPI の通信セマンティックスに従うことで複数の並列計算機を効率的に利用することができる (図 1 に Stampi の概観を示す). 特に動的なプロセス管理が可能であることから, ネットワーク上の資源管理機構を利用してより効率的な処理を実現することが容易である. さらに, Java との連携を可能とすることにより, Applet から並列機を呼び出すシステムを構築することができ本実験においても, 複数の並列計算機間の通信とともに並列計算機と Web ブラウザ上で動作する Applet 間の通信にも利用されている.

2.3 PATRAS

PATRAS[6] は原研が開発した, 計算サーバである分散並列計算機上でのピクセル生成をコンセプトとした可視化システムである. 本システムの特徴としては, 異機種計算環境との親和性, ベクトル並列計算機上でのピクセルデータ生成, ピクセルデータ圧縮技術によるクライアント / サーバ間のネットワーク負荷の軽減, JAVA Applet による Web ベースの GUI などが挙げられる.

グローバルメタコンピューティング実験のベースとなる放射線情報推定サブシステムに対して, PATRAS システムは以下の様に機能する.

1. 異機種計算環境において, 計算サーバ上で MPI-GEARN を実行する.
2. MPI-GEARN の結果の中で, トレーサ移動計算の結果のみを可視化サーバに転送する.
3. 可視化サーバ上で地形データとトレーサの可視化処理を行い, 画像 (ピクセル) データを生成する.
4. 可視化サーバ上で画像データを圧縮し, ユーザ側の PC クライアントに転送する.
5. PC クライアント上で画像データを解凍し, 表示する.

1~5 が, 計算のタイムステップごとに実時間で実行される.

文献 [6] などでは, PATRAS の設計思想では計算サーバと可視化サーバを同一計算機で行わせることで物理データの移動を削減させることにあったが, 今回の実験では使用する並列計算機の使用可能メモリ容量の制限があったため, 計算サーバと可視化サーバを分離する方式をとった. ハードウェアの構成上, 計算サーバと可視化サーバの途中が帯域の狭い WAN での結合となるため, 可視化対象はデータ量の多い密度分布ではなく粒子の軌跡追跡に限られることとなった.

PC クライアント上の画像データ表示 Viewer (GUI) は, Java Applet(JDK1.1+swing) で作成されており, GUI と可視化サーバ間の通信は Stampi によって制御されてる. GUI の主な機能は, 可視化サーバへの起動 · 一時停止 · 終了などの制御, 描画範囲の変更, 視点の変化, 描画オブジェクトの色変更, コンター図などの範囲指定などの様々な制御が実時間可視化と並行して指定可能である.

2.4 TME(Task Mapping Editor)

TME(Task Mapping Editor) は, 視覚的にアプリケーションの関係を定義する一種のプログラミングツールであると共に実行制御ツールでもある. TME は原研で開発された STA(Seamless Thinking Aid) の一機能であり, AVS 等各種ビジュアルプログラミング環境で採用されているデータフローをウィンドゥ上で定義する方式を取っており, 分散する資源をアイコンとして取り扱い, それらを結び付けることによりそのデータフローと制御関係を定義するインターフェイスを提供する. 利用者は, TME 上でアプリケーションの関係を定義するとともに, 実行に関しても TME 上の START ボタンをクリックすることで必要なアプリケーションの起動をかけることができる. TME からのバックサイド計算機群への通信並びにプロセス起動に関しては STA の機能を使用するため詳細は文献 [3] を参照して頂きたい.

TME で定義される分散アプリケーションの結合パターンとして, 2 種類の関係が存在する. 1 つは, 入出力を定義するものであり, 例えば, マシン A の a というファイルを使って, マシン B 上の b という FORTRAN アプリケーションで装置番号 1 に結合させて実行させるなどの関係を表す. もう一つは, MPI-2 のプロセス生成関数に基づいた MPI アプリケーションの結合関係の定義するものである.

SC2000 の実験では各機関に分散する MPI-GEARN をクライアント – サーバの形式で TME 上で定義し, 全アプリケーションの起動・監視を行った. 図 4に TME 画面の一例を示す.

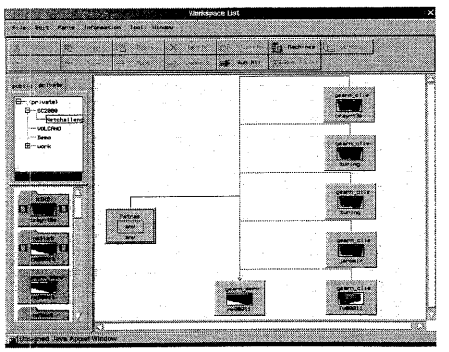

図 4: TME 画面の一例

3 放射能放出源推定実験と考察

SC2000 の期間中に実施した計算では 1998 年 5 月にスペインのアルヘシラスで起こったセシウム 137 汚染事故についての放出源推定計算を試みた. 並列計算機は, HLRS, PSC, CSAR, JAIST の CRAY T3E4 台と CCSE の富士通 VPP300 と NEC SX-4 の計 6 台を使用した.

SX-4 は拡散計算 (MPI-GEARN) のマスタならびに可視化サーバとして, T3E と VPP300 は拡散計算スレーブならびに可視化クライアントとして機能させた. 各サイト合わせて 510CPU 上 (使用 CPU の内訳は HLRS 100, PSC 100, CSAR 250, JAIST 45, CCSE 15) で初期条件の異なる大気拡散シミュレーション 510 ケースを実行し, その中から特徴的な 6 ケースを選択して CCSE の SX-4(6CPU) の各 CPU でそれぞれのケースの可視化処理に成功した. 計算時間は, 他のグループと占有時間 (2 時間) を均等に分けたため約 20 分間使用することができた. 図 5に SC2000 実験時の撮影した実時間可視化システムの様子を掲載する. (2 色刷りの予稿集では粒子の飛散が確認しにくいが, イベリ

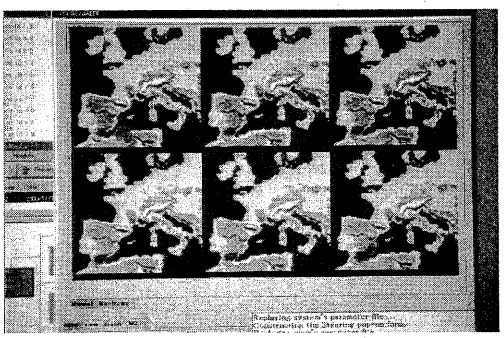

図 5: SC2000 時のメタコンピューティングの様子 (PATRAS GUI 画面)

ア半島の先端から粒子が放出され6画面異る飛散状態が表示されている.)

今回のメタコンピューティング実験では，拡散計算に関してはマスター－スレーブ方式で個々のCPUは全く相互通信を必要としない方式を取った．一方可視化システムではサーバに粒子データ (1000粒子，単精度で各々3成分で12KByte/step) を送る構成としたが，CCSEの6CPUの可視化サーバがWANと接続する部分 (FireWall) で通信が一本にまとめられるため，クライアントが如何に並列にデータを送信してもその効果は得られないハードウェア構成であった．

また，本システムでは受信側で可視化を行うとともに，PCクライアントへピクセルデータを送信する構成を取るため，可視化サーバの描画能力，計算サーバの処理能力，PCクライアントの能力，また途中経路のネットワークの状態によってその性能が変化する．理想は全ての処理時間とデータ転送時間が等しいときに最大限の性能を発揮する (パイプラインの性質と同様である)．実験中は，PC上のJVMのリソース消費のため各種サーバがPCクライアントに待たされるという状態が生じた．したがって今後，長時間にわたる実験並びに負荷変動の大きいネットワークを利用する場合には，構成要素 (可視化サーバや計算サーバ等) の能力バランスを考慮した動的なデータ転送の手法やバッファの利用方法を用いるべきであろう．

さらに，今回の実験では1CPUに1ケースの初期条件を割り当てる極めて単純なタスクの割り当てを行ったが，複数のCPUで並列計算を行ったり，領域分割と分散処理を負荷を考慮しながら行う手法の検討など今後の課題が多く残った．

4 まとめ

本報告では，ワールドワイドなメタコンピューティングの試みとしてSC2000で行った実験について示した．実験には日独米英4ヶ国の6台の並列計算機を用いて，実アプリケーションである放射能放出源推定システムの実験に成功した．同時に実時間可視化を行い，全シミュレーションを世界規模で行うにはネットワークの帯域の問題が残るもののコア部分のみであれば，可視化を含めて現状のネットワーク (公衆回線を含む) でも実行可能であることが示された．また，今回の実験で多くの問題点が明らかになったが今後の課題として (SC2001で行う機会があるかどうかわからないが) 取り組んでいきたい．

最後に，本実験において計算機の利用を快諾して頂いた，シュツッツガルツ大学 Michael Resch氏，ピッツバーグ計算機センター Sergiu Sanielevici氏，マンチェスター大学 John Brooke氏，北陸先端大学院大学 松澤 昭男，井口 寧 両先生にはこの場を借りて深謝致します．

参考文献

[1] Pickles, S. et al. ; The problems and the solutions of the metacomputing experiment in SC99, HPCN'2000, 2000.

[2] H. Kitabata and M. Chino: Development of Source Term Estimation Method During Nuclear Emergency, in Proc. of Mathematics and Computing 1999, 1999.

[3] H. Takemiya, et al.: Software Environment for Local Area Metacomputing, in Proc. of 4th Conf. Supercomputing in Nuclear Applications, 2000.

[4] T. Imamura, et al.: An Architecture of Stampi: MPI Library on a Cluster of Parallel Computers, in J.Dongarra etc. eds. Recent Adavances in Parallel Virtual Machine and Message Passing Interface, LNSC 1908, pp.200-207, Springer, 2000.

[5] Message Passing Interface Forum: Extensions to the Message Passing Interface, University of Tennesee, 1997.

[6] K. Muramatsu, et al.: Parallel Real-Time Visualization System for Large-scale Simulation — Application to WSPEEDI—, in Proc. of 4th Conf. Supercomputing in Nuclear Applications, 2000.