事例

並列分散科学技術計算を支援するソフトウェア・システム（STA）の構築

武宮　博　今村　俊幸　小出　洋
日本原子力研究所 計算科学技術推進センター
並列処理基本システム開発グループ

日本原子力研究所（原研）計算科学技術推進センターでは，並列処理にかかわる共通基盤技術研究開発の一環として，科学技術計算環境STA（Seamless Thinking Aid）を構築している．STAは，並列分散科学技術計算と呼ばれる新しい形態の科学技術計算を対象とし，プログラム開発から実行，結果解析に至る一連の作業の円滑化，消費される時間の低減を実現することで，利用者の途切れのない思考を支援する（Seamless Thinking Aid）環境である．また，センターではSTA上にいくつかの並列分散アプリケーションを構築し，それらの実用性評価を行っている．本稿では，STAおよびSTA上に構築された並列分散アプリケーションについて紹介する．

■並列分散科学技術計算

実験と理論に並ぶ“第3の科学”と称される計算科学は，実験的再現や理論的解析が困難な現象を理解する新しい科学技術の研究手法である．ベクトル計算機に端を発するスーパーコンピュータの出現により，材料物性，レーザと物質の相互作用，プラズマ挙動など種々の科学技術分野において，この手法が実用に供されつつある．計算科学において行われる処理，すなわち科学技術計算は，これまで単一のスーパーコンピュータ上で行われることが一般的であった．しかしながら多様な計算機の普及とともに，ネットワークに接続された複数の計算機を特性に応じて使い分けたり，互いに連携させたりすることにより，科学技術計算を効率的に実行することが可能となってきた．このような新しい形態の科学技術計算を並列分散科学技術計算と呼ぶ．並列分散科学技術計算の例としては，並列計算機と可視化サーバを互いに連携させ，実時間可視化や実行制御を行うトラッキング，ステアリング処理が挙げられる．また，複数の並列計算機を連携し，計算を分散させることで，メモリ量や計算能力の制約から単一の並列計算機では困難であった大規模シミュレーションを可能とする分散スーパーコンピューティングも並列分散科学技術計算の典型的な例である．

■並列分散科学技術計算の支援

STAでは，以下の2つの項目を実現することで並列分散科学技術計算における利用者の途切れのない思考を支援する．

(1) ネットワーク上に散在する計算機の途切れのない利用の実現．

計算機の使い分け，組合せを容易に実現するためには，利用者に対して複数の計算機を意識させることなく，仮想的に単一計算機を利用しているかのような環境を提供することが重要である．

(2) ツール群の途切れのない利用の実現．

科学技術計算はプログラム開発から計算実行，結果解析に至るいくつものフェーズから構成される．これら一連の作業を効率的に遂行するためには，フェーズ間の移行を円滑にすることが重要である．そのために，各フェーズでの作業を支援するツール群を統合するとともに，個々のツールの保持する情報を組み合わせ，利用者に認識しやすい形式で提供することが必要である．

上記(1)，(2)の課題を並列分散科学技術計算作業全体の流れから捉えてみれば，(1)の課題は作業の空間的な連続性を支援するための要件であり，(2)の課題は作業の時間的な連続性を支援するための要件であると見なすことができる．したがって，上記(1)，(2)を同時に満足させるソフトウェア環境を構築することで，初めて時間的にも空間的にも円滑な並列分散科学技術計算を実現することができる．

STAでは，(1)の課題を解決するためにネットワーク上に散在する計算機群の通信を支援する通信基盤SCEを構築し，その上に並列分散科学技術計算

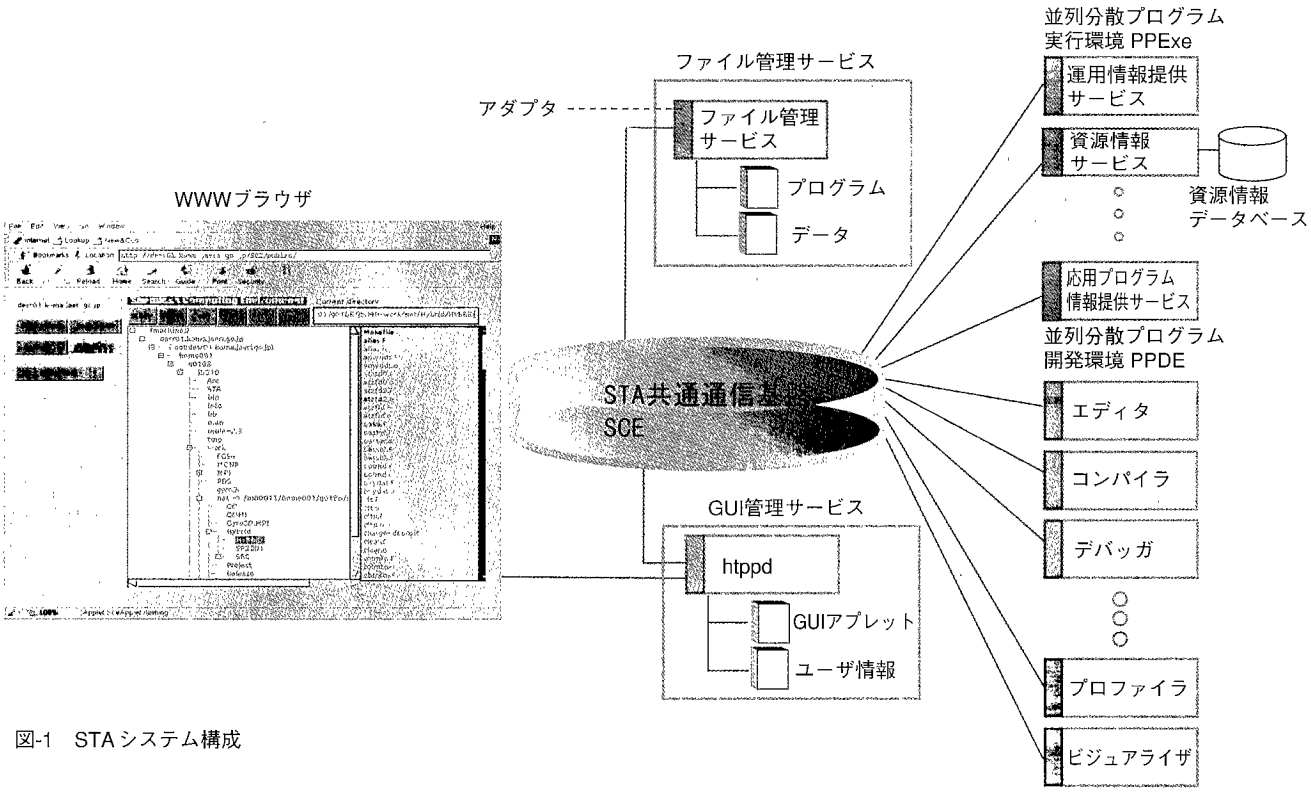

図-1　STA システム構成

を支援する各種ツール群を統合したプログラム開発環境PPDEおよびプログラム実行環境PPExeを構築することで(2)の解決を図っている(図-1)．

以下に，STAを構成するサブシステムの概要を述べる．

■STAを構成するサブシステムの概要

◎通信基盤：SCE

SCEは，複数の並列計算機上に構築された各種ツール間，あるいは利用者プログラム間の通信を支援する通信基盤である．並列分散科学技術計算は従来の分散処理的な側面と並列処理的な側面を併せ持っている．SCEでは，分散処理を支援する遠隔関数呼び出し型の通信ライブラリStarpc，および並列処理を支援するメッセージパッシング型の通信ライブラリStampiを提供している．StarpcはPPDEおよびPPExeを構成するツール間の通信を支援する．また，StampiはMPI-2ライブラリの仕様に基づいており，複数の並列計算機上で動作する並列アプリケーションの通信を支援する．

Starpc，Stampiライブラリの設計にあたっては，大規模科学技術計算を対象としていることから効率的な通信の実現に留意した．

StarpcではArgonne研究所において開発されたNexus通信ライブラリ[1]を利用することで通信の効率化を図っている．Nexusは，TCP/IP，AAL5，MPL等多様な通信プロトコルへの対応を考慮して開発された通信ライブラリで，利用する通信プロトコルを動的に切り換えられる．StarpcはNexusの機能を利用し，そのネットワークに最も適した通信プロトコルを採用することで効率的な通信を行うことができる．

Stampiでは，通信の効率化のために，以下の2点を考慮した．

(1) 並列計算機間通信，並列計算機内通信で用いられる通信プロトコルの切り換え．

複数の並列計算機を用いて処理を行う場合，計算機間と計算機内の2種類の通信を取り扱わなければならない．計算機間通信では，互いに利用可能な通信プロトコル(TCP/IP，UDP/IP等)を用いる必要がある．これらの通信プロトコルは計算機内通信に対しても利用可能であるが，個々の並列計算機で提供されている専用通信プロトコルを利用した通信と比較して一般に通信性能が低い．Stampiでは，相互通信可能性と通信効率を共に満足させるために，並列計算機内通信に対しては専用通信プロトコル，並列計算機間通信に対して共通通信プロトコルを各々利用し，通信対象に合わせてそれらを自動的に切り換えている．

(2) 並列計算機間通信における直接通信，間接通信方式の切り換え．

Stampiでは図-2に示すように並列計算機間の通信を専門に行う通信中継プロセスが存在し，その数を0から2個まで変更できる(図-2は2個の通信中継プロセスを設置した場合の構成を表している)．通信中継プロセスが介在する通信方式を間接通信方式，介在しない方式を直接通信方式と呼ぶ．

間接通信方式では，利用者プログラムから通信中継プロセスまで並列計算機内の高速ネットワークを利用してデータが転送され，その先の通信コストの高い計算機間ネットワークを利用した通信は，通信中継プロセスが担当する．したがって，もし通信処理とは独立に実行できる処理があれば，Stampiの間接通信機能を用いて並列計算機間の通信をうまく隠蔽することができる．一方，そのような処理がなければ，メッセージ中継処理が間に入らない直接通信方式が望ましい．

このように，処理の特質により効率的な並列計算機間通信方式は変化することから，Stampiでは利用者の要求に合わせて両通信方式を切り換えられる機能を実現している．

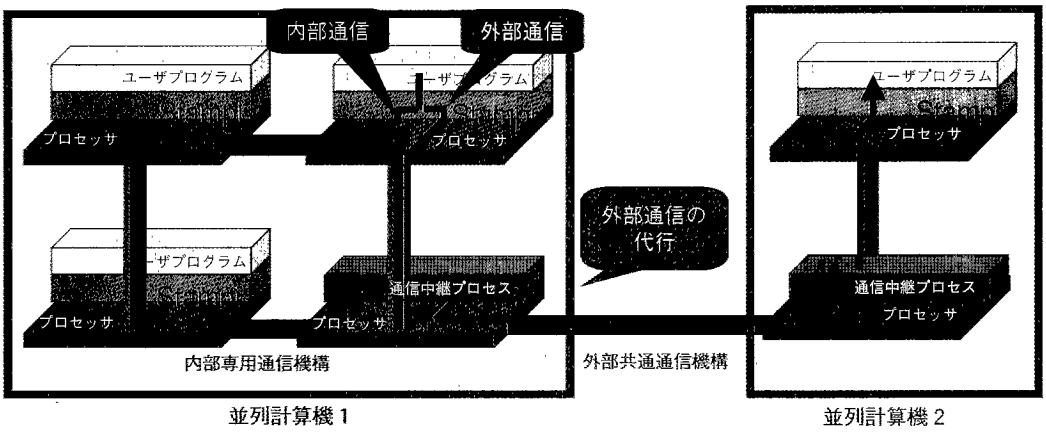

図-2　Stampi ソフトウェアアーキテクチャ

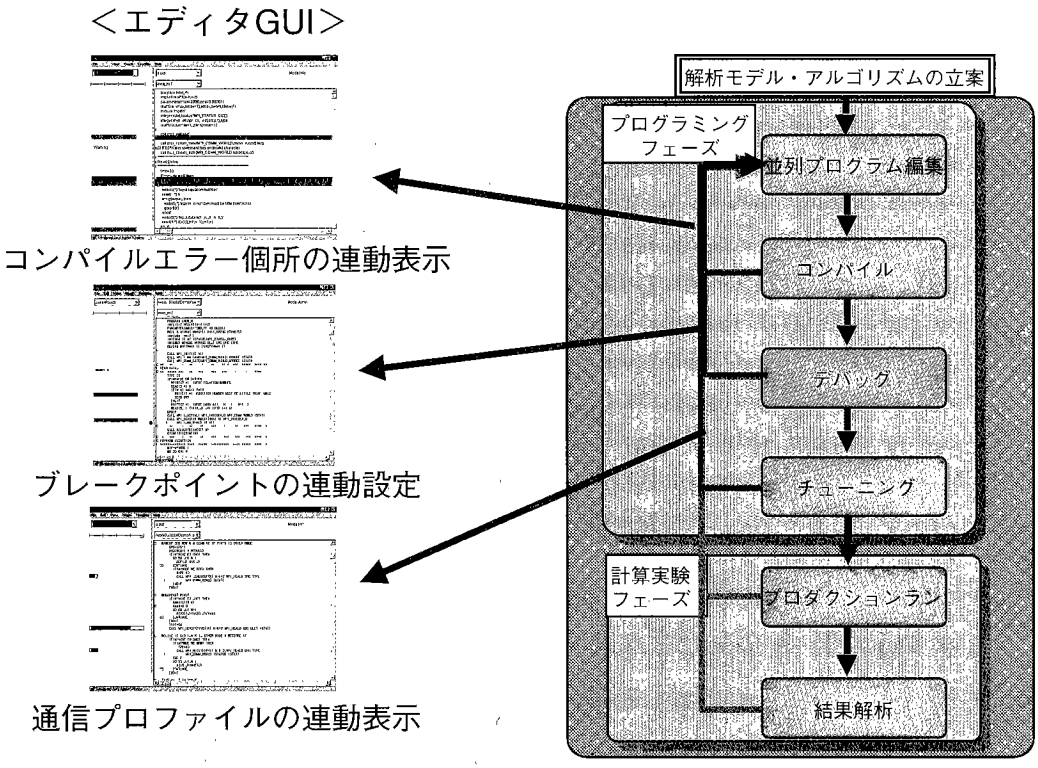

図-3　PPDE におけるツール連携例

◎プログラム開発環境：PPDE

PPDE は，並列分散プログラムの開発を支援する各種ツールの連携を実現した統合プログラミング環境である．並列分散科学技術計算では複数の計算機を組み合わせて利用するため，種々の計算機上でプログラムを開発する必要がある．また，プログラム編集は手元の計算機上で，コンパイル・デバッグは小規模な並列計算機上で，実行性能向上は大規模並列計算機上で各々実行するなど，プログラム開発作業自体でも計算機の使い分けをすることが考えられる．

PPDE は，これらの作業を支援するために，以下の2点を考慮して構築されている．

(1) 複数の計算機を用いたプログラム開発作業の流れの円滑化．

科学技術計算は，一般にプログラム作成，コンパイル，デバッグ，実行性能向上，というフェーズを経て本格的な計算実行に移行する．しかし，この流れは直線的に進行するわけではなく，各フェーズにおいてプログラム編集作業への手戻りが発生する．PPDE では，この流れに合わせて，コンパイラ，デバッガ，性能解析ツール等各種ツールからの情報をエディタ上のプログラムに対応づけて表示することにより，フェーズ間の円滑な移行を実現している（**図-3**）．これらツールの連携は SCE の通信機能を利用することにより，異なる計算機間でも可能である．

(2) 計算機に依存しないツールの統一的な操作感の提供．

エディタ，コンパイラ，デバッガ，性能評価ツール等プログラム開発支援ツールは各種計算機メーカから提供されているが，それらの操作性はメーカごとに異なっている．このことが，

種々の計算機の使い分けを阻害する要因となっている．我々は操作性の統一を図るために，まず各種並列計算機上の既存ツールのインタフェースを調査し，利用者に対し計算機によらず共通なツールGUIの仕様を定めた．また，共通ツールGUIと既存ツールの間にアダプタと呼ばれるソフトウェアを介在させ，共通ツールGUIのインタフェースと既存ツールのインタフェース間の差異を吸収させた．これにより，アダプタのみを機種ごとに構築するだけで共通ツールGUIを介して既存ツールを利用できるようになった．

◎並列分散科学技術計算実行環境：PPExe

PPExeは並列分散科学技術計算の実行を支援する統合環境である．並列分散科学技術計算では，冒頭で述べたような実時間可視化解析や分散スーパーコンピューティングが行われる．また，データサーバ上のデータを並列計算機上のアプリケーションの入力データとしてシミュレーションを行い，さらにその結果を手元のワークステーションで解析するといった計算機の使い分けも考えられる．

複数の計算機を用いた一連の処理を効率的に実行するためには，利用する計算機やネットワークの負荷が軽微であることが望ましい．しかし，一般には複数のユーザが同時にそれらを利用するため，特定の計算機やネットワークに利用が集中し，負荷の不均衡が生じやすい．その結果，一連の処理を効率的に行うことは困難であった．

PPExeでは，

(1) 一連の処理の流れをデータフローに基づいて定義できるタスクマッピングエディタ(TME)
(2) 一連の処理の最短実行を目的として個々の処理の計算機割付けを決定するメタスケジューラ
(3) ネットワーク，計算機等資源の稼働状況データ(資源情報)を収集する資源情報モニタ(RIM)
(4) 対象となる計算機上にデータやプログラムを準備し，一連の処理を正しい手順で実行する分散資源管理サーバ(RMS)

の4サブシステムを連携させることにより実行の効率化を図っている．

以下，これら4サブシステムについて説明する．

■TME(タスクマッピングエディタ)

TMEは一連の処理を構成するプログラム間のデータ依存関係や実行制御関係を視覚的に定義するためのエディタである(図-4右上)．利用者は一連の処理をTME上で定義した後，RIMから提供される資源情報を参照しながら実行対象計算機を決定する．あるいはメタスケジューラに実行対象計算機の決定を依頼することもできる．TMEを用いて定義された一連の処理の実行に必要な情報は，TMEからメタスケジューラに送付され，実行対象計算機が決定される．

■メタスケジューラ

メタスケジューラは，

(1) TMEによって利用者が定義した一連の処理を構成するプログラム，
(2) 自動並列化コンパイラにより逐次プログラムから生成された並列実行タスク

の実行対象計算機を決定する．決定は，RIMから提供される資源情報に基づき行われる．

(2)における逐次プログラムからのタスク生成は，笠原らによって提案されたマルチグレイン並列化手法[2]に基づいて行われる．生成されたタスク(マクロタスクと呼ぶ)は，拡張CP/ETF/MISFスケジューリング手法(メタスケジューリング手法)[3]を用いて計算機に割り付けられる．

■RIM(資源情報モニタ)

RIMは，資源情報を常時収集・解析し，必要に応じてメタスケジューラや利用者に提供する．RIMは，以下の3種類の情報，すなわち，

(1) 各計算機の演算性能，アーキテクチャ情報等の静的情報，
(2) 各計算機の計算負荷情報，バッチキュー情報，ネットワークのデータ転送性能情報等の動的情報，
(3) 運用管理情報，

を提供する．RIMは一定時間間隔でこれらの情報を収集するほか，過去の収集情報に基づく統計情報の生成，将来の変動に関する予想も行う．

■RMS(分散資源管理サーバ)

RMSは，TMEを用いて定義された並列分散処理の実行に必要な入力データの準備，プログラムの起動，終了の監視，出力データの管理を行う．RMSによって収集される並列分散処理進行状況に関する情報はTMEに伝えられ，TMEのGUIを介して利用者に提供される．

■並列分散アプリケーション構築事例

計算科学技術推進センターでは，STA上に並列分散科学技術計算アプリケーションを構築し，STAの実用性評価を行っている．以下に，現在開発しているいくつかの並列分散アプリケーションについて述べる．

◎分散スーパーコンピューティング事例：流体・構造連成計算コード

流体・構造連成計算コードは，境界条件を介して流体計算と構造計算をカップリングさせることにより，航空機翼の変形問題等を解析するコードである．流体計算部分からは，時間ステップごとに翼に伝わる荷重分布が構造計算部分に伝えられる．一方，構造計算部分からは，翼の変位データが流体計算部分に伝えられる．伝えられた変位データに基づき流体計算用格子が形成され，その格子を用いて再び流体計算が行われる．この手続きを繰り返すことによって，連成計算(異なる物理モデルを組み合わせて，より詳細に複雑な現象を数値解析すること)が行われる．

本コードをベクトル並列計算機で実行し性能評価を行ったところ，流体計算および格子生成部分は高性能実行されるが，構造計算部分では十分な計算性能の実現が困難であることが判明した．

このことから，流体計算部分をベクトル並列計算機，構造計算部分をスカラ並列計算機で実行し，Stampiを用い

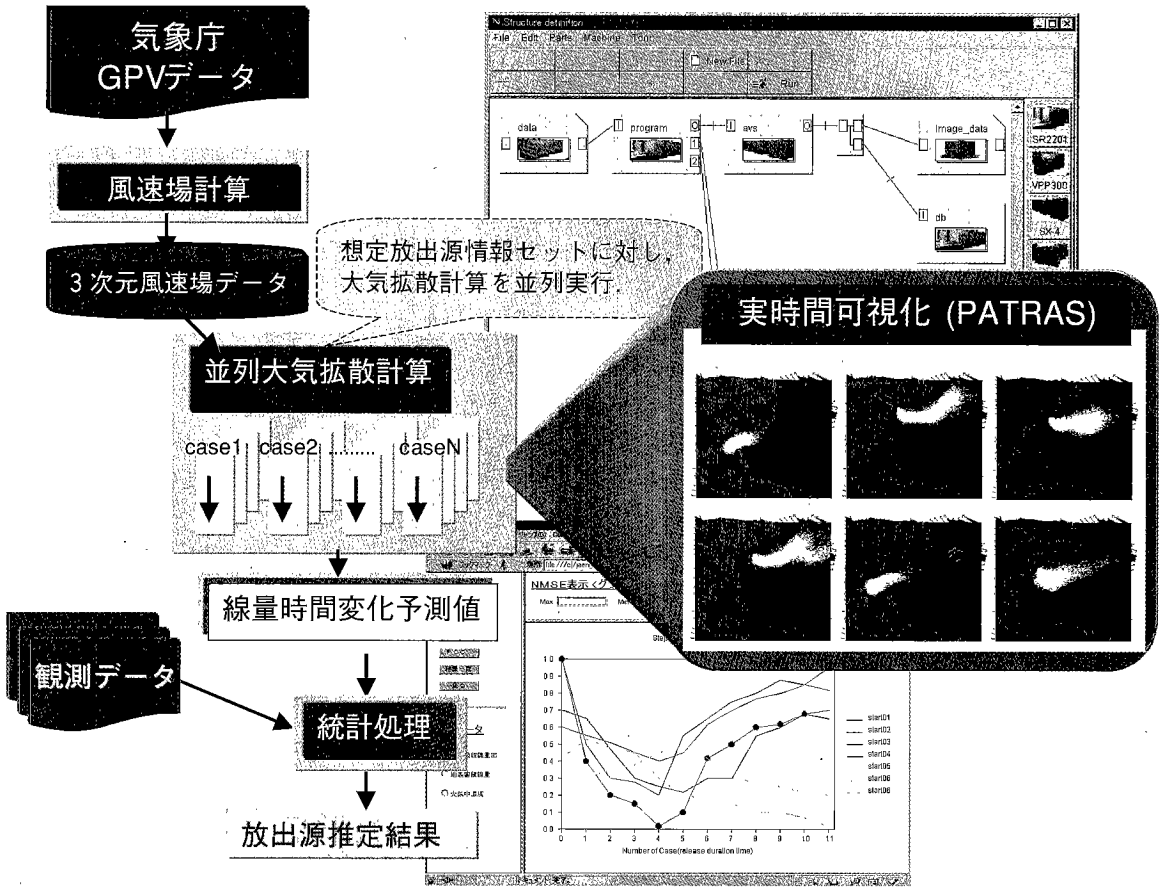

図-4　緊急時放射線源推定システムにおける処理の流れ

てそれらを連携させることにした[4]．

実験には，16プロセッサ構成の富士通VPP300（ベクトル並列計算機），および64プロセッサ構成の日立SR2201（スカラ並列計算機）を用いた．シミュレーションは，流体計算に対して100×100×100格子点，構造計算に対して4500節点を用いて行った．その結果，2台の並列計算機に処理を分散させた場合，単一並列計算機ですべての計算を行った場合より約30％から70％処理時間を短縮できた．

現在，センターでは，流体・構造連成計算以外にも分子軌道シミュレーションにおけるHartree-Fock行列計算，大気，海洋，地表間の物質循環を計算する地球環境シミュレーション，原子炉の遮蔽計算，等の分散スーパーコンピューティングに着手している．

◎メタスケジューリング適用事例：粒子・流体ハイブリッドコード

粒子・流体ハイブリッドコードは，トカマク型核融合炉におけるプラズマの挙動解析を目的として開発されたコードである．このコードを対象として，メタスケジューリング手法の性能評価実験を行った[3]．

まず最初にコードの依存性解析を行い，計10個のマクロタスクを得た．次に，SGI Onyx，スカラ並列計算機日立SR2201，SR2201コンパクトモデル（各3プロセッサを使用），ベクトル並列計算機NEC SX-4（1プロセッサを使用）を用いて，得られたマクロタスクのメタスケジューリングを行った．このうち，SGI Onyx上でメタスケジューラを動作させ，他の3つの並列計算機上でマクロタスクを実行させた．

マクロタスクを実行する3つの並列計算機のうち，SX-4はマクロタスクの処理速度が他の並列計算機より3～4倍程度速い．その結果，SX-4が無負荷の状態で行った実験では，すべてのマクロタスクが自動的にSX-4で実行された．一方，人為的にSX-4に4倍程度の負荷をかけた実験では，自動的にSR2201とSR2201コンパクトモデルの2台でマクロタスクが実行された．

また，上記3台の計算機を利用してメタスケジューリングを行った場合と，単一の計算機上ですべてのマクロタスクを実行した場合の実行時間の比較を行った．その結果，単一計算機上で実行した場合，1時間ステップあたり各々41.0秒（4倍程度の負荷をかけたSX-4），36.0秒（SR2201），68.4秒（SR2201コンパクトモデル）要していた計算が，メタスケジューリングを用いると31.7秒で終了した．

以上の実験より，

(1) メタスケジューリングを用いると，各計算機の負荷状況に基づき処理能力の高い計算機を判断し，その計算機で多くのマクロタスクを実行している，

(2) マクロタスクを複数の計算機で実行することにより，1台で実行した場合よりも実行時間の短縮が可能である，

という知見が得られた．

◎PPExe適用事例：緊急時放射線源推定システム

緊急時放射線源推定システムは，原子炉事故等で大気中に放出された物質について，全国に散在するモニタリ

ングポストから随時送付される観測データに基づきその放出源を決定するシステムである[5].

本システムの放出源推定手法は，以下の通りである．まず想定し得る複数の放出源情報（放出地点，放出開始時刻，放出継続時間）の組合せに対し，その時点における気象データに基づいて大気拡散シミュレーションを行う．放出源情報の組合せは数百に及ぶのが一般的であるため，並列処理を行って処理の効率化を図る．その後，シミュレーション結果とモニタリングポストから送付される観測データとを比較し，最も近いものに放出源情報を絞りこむ．絞り込みには，計算結果と観測データの偏差を解析する定量的な絞り込みのほか，拡散計算の計算時間を削減するため，計算データを実時間に可視化し，観測データと大きく異なる場合に実行を途中で打ち切る定性的な絞り込みの2つの手法を併用する．処理の流れを図-4に示す．本システムの性質上，迅速な放出源推定が要求されるため，システム構築においては以下の点に留意する必要がある．

(1) 観測データや気象データ等の管理，数値シミュレーション，実時間可視化等，特性の異なる処理を効率的に実行する必要がある．そのためには，種々の計算機を連携して動作させる必要がある．

(2) シミュレーションの実行可能な計算機資源は運用，障害等により動的に変化するため，実行対象計算機を柔軟に変更できる機構が必要である．

このことから，PPExeの有する機能を利用したシステム構築を図ることにした．その結果，

(1) 東海研究所設置のデータサーバから東京の計算科学技術推進センター内の並列計算機への観測データおよび気象データの転送，放出源情報の作成，大気拡散シミュレーションの実行，実時間可視化システムPATRAS[6]や統計処理プログラムを利用した放出源情報の絞り込みなどの一連の作業を効率的に実行できるようになった．

東京中目黒	VPP300/16	○	○
	T94/4	○	○
	SX-4/2C×3	○	○
	SR2201/64	○	○
	SP2/48	○	○
東海	VPP500/42	○	○
	AP3000/32	○	○
那珂	Paragon XP/S15-256	○	△（一部ツールのみ）
	SP2/4	○	△（一部ツールのみ）
関西	Paragon XP/S75-MP834	×	×
	VPP300/12	×	×
導入済台数		9台	7台

表-1　STA実装状況一覧

(2) TMEの機能を利用することで，富士通VPP300，NEC SX-4上で並列実行可能な大気拡散シミュレーションプログラムの実行対象計算機を柔軟に変更可能となった．また，RIMのGUIを利用することで，計算機の稼働状況を容易に認識可能となった．

上記のように，現在，大気拡散シミュレーションは2種類の並列計算機上で実行可能である．しかし，本計算は典型的なパラメータサーベイであり，個々のシミュレーションは独立に実行されるため，その時点でネットワーク上に存在する可能な限りの計算機資源を活用することにより一層の計算時間の短縮が可能である．そのために，

(1) Stampiを利用し，大気拡散シミュレーションを複数の並列計算機上で実行可能とする，

(2) メタスケジューラを利用し，利用可能なすべての計算機を用いてシミュレーションを実行可能とする，

という作業を現在行っている．

■システム構築状況

現在，STAは5台の並列計算機から構成される当センター複合並列計算機システムCOMPACS（COMplex PArallel Computer System）を中心として，東海研究所，那珂研究所に設置された並列計算機群に導入されており利用可能となっている（**表-1**）．また，国内外7研究機関にもすでに導入されている（平成11年8月現在）．StmapiおよびPPDEの一部のツールに関しては，当センターホームページ（http://guide.tokai.jaeri.go.jp/program/index.html）を介して広く一般に公開されている．

なお，本稿では触れることができなかったが，並列処理基盤技術開発の一環として当センターで開発された並列数値計算ライブラリPARCEL，構造格子生成システムGRID3DST，PATRAS，プログラム実行性能解析ツールKMtool等も上記ホームページにおいて公開されている．興味を持たれた方はアクセスしていただきたい．

参考文献

1) Foster, I., Geisler, J., Kesselman, C. and Tuecke, S.: Managing Multiple Communication Methods in High-performance Networked Computing Systems, Journal of Parallel and Distributed Computing, Vol.40, pp.35-48 (1997).

2) Kasahara, H. and Yoshida, A.: A Data-Localization Compilation Scheme Using Partial Static Task Assignment for Fortran Coarse Grain Parallel Processing, Parallel Computing, Vol.24, No.5, pp.579-596 (1998).

3) Koide, H., Hirayama, T., Murasugi, A., Hayashi, T. and Kasahara, H.: Meta-scheduling for a Cluster of Supercomputers, in Proc. of International Conference on Supercomputing Workshop, Scheduling Algorithm for Parallel/Distributed Computing (1999).

4) Kimura, T. and Takemiya, H.: Distributed Parallel Computing for Fluid-Structure Coupled Simulations on a Heterogeneous Parallel Computer Cluster, The International Journal of High Performance Computing Applications, Vol.13, No.4, pp.320-333 (1999).

5) Kitabata, H. and Chino, M.: Development of Source Term Estimation Method During Nuclear Emergency, in Proc. of International Conference on Mathematics and Computing (1999).

6) 村松一弘，松本秀樹，武井利文，土肥　俊：並列計算機を利用した実時間可視化システム，第58回全国大会論文集，Vol.1, pp.405-406 (1999).

（平成11年10月5日受付）