◇◇◇◇◇◇
解 説
◇◇◇◇◇◇

九州大学研究用計算機システムの利用支援について

林 豊洋[1]

1 はじめに

大量のデータに対する数値計算や，単一の計算に大容量のメモリを必要とするアルゴリズムを適用する場合，個人が所有する計算機では規模や速度の面から限界が生じます．一般的にこのような計算を行う場合は，大規模演算に適した研究用計算機システムを利用します．

今年度情報科学センターでは，大規模演算や高速演算の要求に応えるため，研究用計算機システムの利用支援を行いました．具体的には，研究用計算機システムとして，九州大学情報基盤研究開発センターに導入されているスーパーコンピュータシステム及び高性能アプリケーションサーバ (以下，九大研究システム) を活用します．九大研究システムの利用登録や問い合わせの代行，利用方法に関する講習会などを行い，ユーザの利用を支援するというものです．

本稿では，研究用計算機システムの概要や，情報科学センターが行った九大研究システムの利用支援の詳細について記述します．また，実際に九大研究システムにテスト用のアルゴリズム (簡単な画像処理) を適用し，研究システム向けのプログラムの記述方法やシステムの処理性能について紹介します．

2 研究用計算機システム

2.1 研究用計算機システムとは

科学技術計算の分野では，大量のデータに対する数値計算や大規模な行列等の計算を行い，データの解析を行うことがあります．このような計算では，数億個のデータに対する処理や，数十 GByte オーダーのメモリを要する演算も珍しくはありません．したがって，個人の PC や研究室レベルで所有する小規模な共用計算機では，大規模演算の実行に関して以下の問題が生じます．

演算速度 PC/WS に搭載される CPU の性能は向上していますが，一般的に入手できる計算機で同時に利用できる CPU の個数は 2〜8 個であることが多く，演算速度に不足が生じます．したがって，実行結果を得るためにデータの間引き等を行う必要があり，近似的な数値計算にとどまる可能性があります．

記憶容量 大規模な行列演算等を行う場合は，数 GByte オーダーのメモリが必要なものがあります．また，多くの入力データを持つ場合はストレージも大容量である必要があります．個人レベルで入手できる計算機ではメモリやディスク容量が不足する場合があり，データの分割や小規模化を行って近似的な演算を行う必要があります．

[1]情報科学センター　助教　toyohiro@isc.kyutech.ac.jp

演算のスケジューリング 通常のオペレーティングシステムは複数のジョブをタイムシェアリングによって効率よく動作させることを目的としており，大規模演算向けのスケジューリングは重要視されていません．したがって，複数の数値計算が同時に実行され，単一の計算にリソースを集約できないことや，逆に一つのジョブが長時間リソースを占有し，他の数値計算が長期に亘って実行されない可能性があります．

このような問題に対処するため，効率的な大規模演算の実行に適したシステムが存在します．このシステムは，「大規模計算機システム」「バッチジョブシステム」「研究用計算機システム」などと呼ばれています (本稿では「研究用計算機システム」と呼びます)．

研究用計算機システムは小規模な計算機システムと比較して，下記の特徴を有します．

演算規模 高速で大規模な演算を可能とするため，単一の計算で多くの CPU が利用できるように設計がなされています．多くの CPU が利用できる構成として，一台の計算機に多くの CPU(64 コア等) が搭載された大規模 SMP や，複数の計算機を結合網 (InfiniBand 等) で接続したクラスタなどがあり，数万の CPU が利用可能なシステムも存在します．また，CPU だけでなく，大容量のメモリ (数百 GByte オーダー) やストレージ (数百 TByte オーダー) が利用できるように設計されています．

ジョブシステム 通常のタイムシェアリングシステムと異なり，大規模演算に適したジョブシステムが利用できます．一般的な OS(タイムシェアリングシステム) ではシェルを介した対話的なコマンドの実行を行います．また，多くのジョブがリソースを共有し，見かけ上同時に実行されます．これに対して研究用計算機システムでは,「コマンド (計算) に対して，どの程度 CPU とメモリ (ジョブクラス) を占有するか」を記述したリクエスト[2]に基づき，ジョブがリソースを占有して実行されます．このようなシステムは「バッチシステム」と呼ばれ，リソースの占有が小規模なジョブの優先実行や，長時間リソースを占有したジョブの待避・再実行などを行い，計算機リソースを有効活用しつつ，多くのジョブを実行する仕組みを有しています．また，ユーザが占有したリソースに応じて課金を行う仕組みを有しており，システムの利用時間に比例した負担を求める設備も存在します．

図 1 に，研究用計算機システムの一般的な構成を示します．

システムは，利用者用のフロントエンド，計算用のバックエンド，ストレージ，演算専用の結合網等で構成されています．ユーザはフロントエンドシステムにログイン (SSH 等を利用) し，研究用計算機システムで利用可能なプログラムのコンパイルや，実行するジョブに対するリクエストを作成し，バッチシステムにジョブを投入します[3]．バッチシステムは，リクエストに基づき CPU やメモリの割当量を決定し，ジョブをバックエンドサーバにて実行します．ジョブの実行終了後，実行結果がストレージ上に保存され，研究用計算機システムでの計算は終了します．

2.2 研究用計算機システムの提供

上記で述べたとおり，大規模な演算を行うためには研究用計算機システムの利用が効率的であるといえます．しかし，研究用計算機システムは多くのシステムで構成されるため，このようなシステムを研究室レベルで導入

[2]これらを記述したファイルを「バッチリクエストファイル」と呼び，実行するコマンド，演算規模，標準入出力をリダイレクトするファイル名，演算終了連絡用のメールアドレス等を記述します

[3]ジョブの投入後は，フロントエンドからのログアウト後もジョブは実行されます

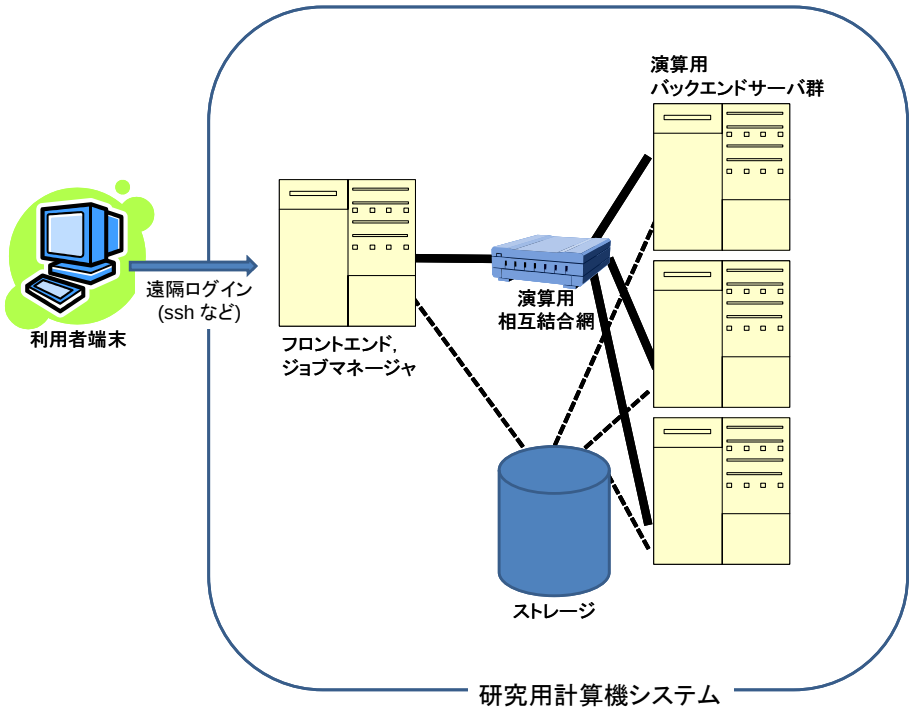

図 1: 研究用計算機システムの構成

する場合，システム維持に多大な労力を要します．したがって，大規模演算の需要に対し，センターが何らかのサービスを提供することが望ましいと考えられます．

今年度情報科学センターでは，九州大学情報基盤研究開発センターが提供する研究用計算機システムを活用し，大規模演算に対するサービスを提供する試みを行いました (サービス提供の詳細は 4 節に示します)．

3　九州大学研究用計算機システム

研究用計算機システムは，多くの研究機関が導入しています．その中でも，「全国共同利用大型計算機センター」には，全国の学術機関が学術研究のために共同で利用できる研究用計算機システムが導入されています．

全国共同利用大型計算機センターは全国で 7 箇所[4]設置されており，九州地区では九州大学情報基盤研究開発センターがサービスを提供しています．九州大学の研究用計算機システム [1] は 2007 年度に機種更新が行われ，2007 年 6 月より最新システムでのサービス提供が開始されました．以下に，新システムの概要とサービス体制について示します．

3.1　スーパーコンピュータシステム

従来，九大のスーパーコンピュータシステムはベクトル型[5]が採用されていましたが，2007 年の 6 月にスカラプロセッサを搭載した 2 種類の計算機システム (スーパーコンピュータシステム A，B) のハイブリッド構成に更

[4]北海道大学，東北大学，東京大学，名古屋大学，京都大学，大阪大学，九州大学に設置されています

[5]複数個のデータに同様の演算を同時実行可能なプロセッサを搭載した計算機．代表的なシステムに，地球シミュレータがあります．

新されました.

スーパーコンピュータシステム A - 大規模 SMP システム　スーパーコンピュータシステム A は多数の CPU(スカラプロセッサ) と大容量メモリが搭載された，大規模 SMP システムで構成されています．計算機一台 (1 ノード) にデュアルコアの Intel Itanium2 プロセッサが 32 基搭載されており，1 ノードで 64 コアのプロセッサと 128GByte のメモリを有しています．なお，オペレーティングシステムには Linux が採用されていますが，ファイルシステムやバッチジョブシステムには大規模計算を想定したシステムが用いられています.

このような大規模 SMP 構成の計算機が 32 ノード導入されており，各ノードは 4GByte/秒の通信速度を有する結合網で相互接続されています．表 2 にスーパーコンピュータシステム A の仕様を示します．表に示す通り，1

表 1: スーパーコンピュータシステム A の仕様 ([1] に掲載の表を転載)

演算ノード	富士通株式会社 PRIMEQUEST580 Intel Itanium2 1.6GHz（デュアルコア）× 32 プロセッサ (=64 コア) 主記憶容量 128 GB
総ノード数	32 ノード
総プロセッサ（コア）数	1,024 プロセッサ（2,048 コア）
理論演算性能の総和	13.1 TFLOPS
主記憶容量の総和	4 TB
相互結合網	InfiniBand 4x (理論転送性能:片方向 4GByte/s)
オペレーティングシステム	Red Hat Enterprise Linux AS (v.4 for Itanium)
ファイルシステム	Parallelnavi SRFS (Shared Rapid File System) for Linux
バッチジョブ管理システム	Parallelnavi for Linux Advanced Edition
言語処理系	Fortran，C（OpenMP 対応，自動並列化機能有り），C++
メッセージパッシングライブラリ	MPI
数値計算ライブラリ	BLAS，LAPACK，ScaLAPACK，PARDISO 等
科学技術計算アプリケーション	Gaussian，GAMESS，Molpro，AMBER

ノードに限定した場合においても大規模な計算が可能な構成となっています．OpenMP[2][6]のみを用いた場合でも，64 コアのプロセッサと約 90GByte のメモリを用いるプログラムが利用できます.

また，MPI[3][7]を用いて各ノードでメッセージ通信を行った場合，最大で 2048 コアの CPU と 4TByte のメモリを用いたプログラムが動作します．最大の構成でプログラムを動作させた場合の理論性能は 13.1TFlops となります.

スーパーコンピュータシステム B - 大規模 PC クラスタ　スーパーコンピュータシステム B は，384 台の 1U サイズのラックマウントサーバが相互結合された，大規模 PC クラスタで構成されています．計算機一台 (1 ノード) にデュアルコアの Intel Xeon プロセッサが 2 基搭載されており，1 ノードで 4 コアのプロセッサと 8GByte のメモリを有しています．なお，スーパーコンピュータシステム A と同様に，オペレーティングシステムには Linux が採用されています.

単一ノードに搭載された CPU は 4 コアで，メモリも 8GByte であるため，単一ノードでの演算性能には限界がありますが，このようなシステムが 192 ノード ×2 セット導入されており，各ノードは 2GByte/秒の通信速度を有する結合網で相互接続されています．したがって，MPI を用いてメッセージ通信を行った場合，最大で 1536

[6]SMP 向けの並列プログラムを作成する規格およびその実装．OpenMP ではノード間の通信を要するプログラムは作成できません.

[7]各ノードで計算に必要なデータを交換する規格およびその実装

表 2: スーパーコンピュータシステム B の仕様

演算ノード	富士通株式会社 PRIMERGY RX200S3 Intel Xeon 3.0GHz（デュアルコア）× 2 プロセッサ (=4 コア) 主記憶容量 8 GB
総ノード数	192 ノード × 2 セット
総プロセッサ（コア）数	384 プロセッサ（768 コア）× 2 セット
理論演算性能の総和	18.4 TFLOPS
主記憶容量の総和	3 TB
相互結合網	InfiniBand 4x (理論転送性能:片方向 2GByte/s)

コアの CPU と 3TByte のメモリを用いたプログラムが動作します．最大の構成でプログラムを動作させた場合の理論性能は 18.4TFLOPS となり，効率の良いプログラムを作成すれば，大規模 SMP を上回る性能を示すことが可能です．

3.2 高性能アプリケーションサーバ

高性能アプリケーションサーバは，スーパーコンピュータシステムと同様の大規模な演算が可能であることに加え，多くの科学技術計算用のアプリケーションを利用することができるシステムです．高性能アプリケーションサーバも，2007 年の 6 月に機種更新が行われました．スーパーコンピュータシステムと異なり，プロセッサに IBM の POWER アーキテクチャを採用した SMP 計算機が用いられています．

表 3 に高性能アプリケーションサーバの仕様を示します．

表 3: 高性能アプリケーションサーバの仕様

演算ノード	日立製作所 SR11000 モデル J1，モデル K2 IBM POWER5 1.9GHz（デュアルコア）× 8 プロセッサ (J1) IBM POWER5+ 2.3GHz（デュアルコア）× 8 プロセッサ (K2) 主記憶容量 128 GB
総ノード数	23 ノード
総プロセッサ（コア）数	184 プロセッサ（368 コア）
理論演算性能の総和	3 TFLOPS
主記憶容量の総和	2.9 TB
相互結合網	専用クロスバーネットワーク (転送性能:片方向 4GByte/s)
オペレーティングシステム	AIX 5L 5.3
ファイルシステム	General Parallel File System
バッチジョブ管理システム	LoadLeveler
言語処理系	Fortran，XL C/C++（OpenMP 対応，自動並列化機能有り），C++
メッセージパッシングライブラリ	MPI
数値計算ライブラリ	BLAS，LAPACK，ScaLAPACK，PARDISO MATRIX/MPP，MATRIX/MPP/SSS，MSL2
科学技術計算アプリケーション	Gaussian，GAMESS，Molpro，AMBER，CHARMM，TINKER， VASP，PHASE，MSC.Marc/Mentat，MSC.Nastran，MSC.Patran， CONFLEX，CFX，IDL

3.3 サービス体制

スーパーコンピュータシステムおよび高性能アプリケーションサーバの機種更新に合わせて，システムの利用負担金に関して，大きな変更がなされました．従来の研究用計算機システムでは，利用者がシステムを利用したCPU 時間やストレージの容量に応じた利用負担金が発生していました．2007 年度からはこれが改められ，利用する計算機の機種と規模 (ノード数および占有か共有か[8]) に応じた 1 年単位の定額料金となりました．

また，九州大学情報基盤研究開発センターでは，各種システムやアプリケーションの利用講習会を九大の各キャンパスで開催していますが，この講習会に準じる内容を他大学で行うサービスも行われています．

4 システムの利用支援

3 節にて述べたとおり，九州大学情報基盤研究開発センターには最新のスーパーコンピュータシステム及び多くの科学技術計算が可能なアプリケーションサーバが導入されました．また，利用負担金も 1 年単位の定額料金に改められました．このような料金体系となったことで，情報科学センターが研究用計算機システムの一部分を一旦借り上げ，希望するユーザにアカウントを発行することにより，学内の大規模演算に関する需要に応えることが可能であると判断しました．

非常に高性能なシステムの提供が可能な状態となりましたが，以下の 2 点を考慮する必要がありました．

システムの提供方針 スーパーコンピュータシステムがベクトル型から一般的なスカラー型に改められ，システムが大規模 SMP，PC クラスタ，IBM POWER アーキテクチャの 3 種類が存在する状態となりました．各システムの特性が大きく異なるため，ユーザに対してどのシステムを提供すべきか判断する必要があります．

システム利用講習会の開催 研究用計算機システムの利用を円滑に行うため，大規模な演算が可能なプログラムの作成法やバッチジョブの投入法を解説する利用講習会を開催する必要があります．

上記の内容に関してはセンターの職員のみでは構成が決定できないことや，研究用計算機システムが更新された初年度であることから，大規模演算の活用を行っているユーザの意見を反映し方針を決定する形式を採りました．したがって，体系的な研究用計算機の提供やサポートを行うものではなく，次年度に提供すべき研究用計算機システムの方向性を確定する意味合いが強いため，今年度は「研究用計算機システム利用支援」という名称でサービスを開始しました．

4.1 今年度の利用支援について

本節では，今年度センターが行った研究用計算機システム利用支援の内容に関して示します．

4.1.1 旧研究用計算機システム利用者へのヒアリング

前述の通り，機種更新により最新のシステムが 3 種類存在する状態となりました．また，各システムの利用体系も，ノードを一括で借り上げることが可能な占有タイプから，1 ノード内の一部の CPU のみが利用可能な小規模な共有タイプまで，多くの選択肢が存在します．

[8] 自組織でノードを利用するユーザを管理できる体型が占有タイプで，他研究機関とノードを共有する体型が共有タイプという位置づけがなされました．

したがって，どのシステムを提供すべきかを判断するため，平成 17 年度から 19 年度の期間に九大の旧研究用計算機システムの利用登録を行っていた本学のユーザ (20 名弱) に対し，適用する大規模演算の分野や，必要とするライブラリ等に関してヒアリングを行いました．ヒアリングの結果得られた回答の一部を以下に示します．

- 近年は，研究室にて導入した計算機で対応できるアルゴリズムを適用することが多い．しかし，時々配列が確保できない場合があるため，大容量のメモリが利用できるシステムは使いやすいと思う．
- ベクトル化が難しく並列化も困難な手法であるため，単一ノードが大規模なシステムがふさわしい．
- 現在は PC クラスタで計算を行っているが，大規模ノードで計算を行うことにも興味はある．
- どの構成が演算性能が出しやすいのかわからないため，様々な試行を行いたい．

このように，単一ノードで大規模な演算を行っているユーザ，既に PC クラスタを活用しているユーザ，実際に様々な構成を試すことにより演算性能を判断すべきであると考えるユーザなど，様々なシステムの利用が想定されることがわかりました．また，多くのユーザは既に研究室で比較的大規模な演算を行っており，これらのシステムを超える構成でなければ，研究用計算機システムを利用するメリットは薄いと考えられます．

4.1.2 利用を支援するシステム，利用負担等の決定

ヒアリングの結果より，ユーザは様々なシステムを利用することが想定されるため，今年度はスーパーコンピュータシステム (A および B) と高性能アプリケーションサーバの大規模共有タイプを九大より借り受けました．各システムにおいて，ユーザが利用可能なリソースは表 4 の通りです．

表 4: 大規模共有タイプリソース一覧

	スーパーコンピュータ A	スーパーコンピュータ B	大規模アプリケーションサーバ
CPU	Itanium2 1.6GHz	Xeon 3.0GHz	POWER5+ 2.3GHz
ノード数	1	32	4
プロセッサ数	64	128	96
メモリ容量 (GByte)	89.6	179.2	96

どのシステムが高い演算性能を示すかは，適用する数値計算の内容に依存します．ユーザも実際に全てのシステムで演算を行うことにより，自分に適したシステムを知ることが可能であると考えられます．従って，今年度はこれら 3 システムを全て利用するユーザを募集することを決定しました．また，各システムへ適用した演算の概要や使い心地等を記述した利用レポートの提出と，九州大学情報基盤研究開発センターとの手続きや連絡は全て情報科学センターを経由することを条件に，今年度は利用負担金をセンターが全額補助することとしました．なお，利用負担金の全額補助を行う募集ユーザ数は 15 人としました．

4.1.3 利用者の募集，アカウントの配布

上記で述べたシステムの利用補助に関して，2007 年 6 月に発行した ISC News No.210 にて「九大 研究用計算機システム利用支援について 」として学内に告知を行いました．利用補助を希望するユーザに対して，アカウントを配布するまでの流れを以下に示します．

解説

1. ユーザから情報科学センターに利用希望の連絡が届く
2. 情報科学センターはユーザに，利用申込書を返送する
3. ユーザから必要事項が記入された利用申込書が届く
4. 利用負担金の支払いや連絡先の追記を行い，九州大学情報基盤研究開発センターに利用申込書を発送する
5. 情報科学センターにアカウント情報が届く
6. アカウント情報の控えを保存し，ユーザにアカウント情報を転送する

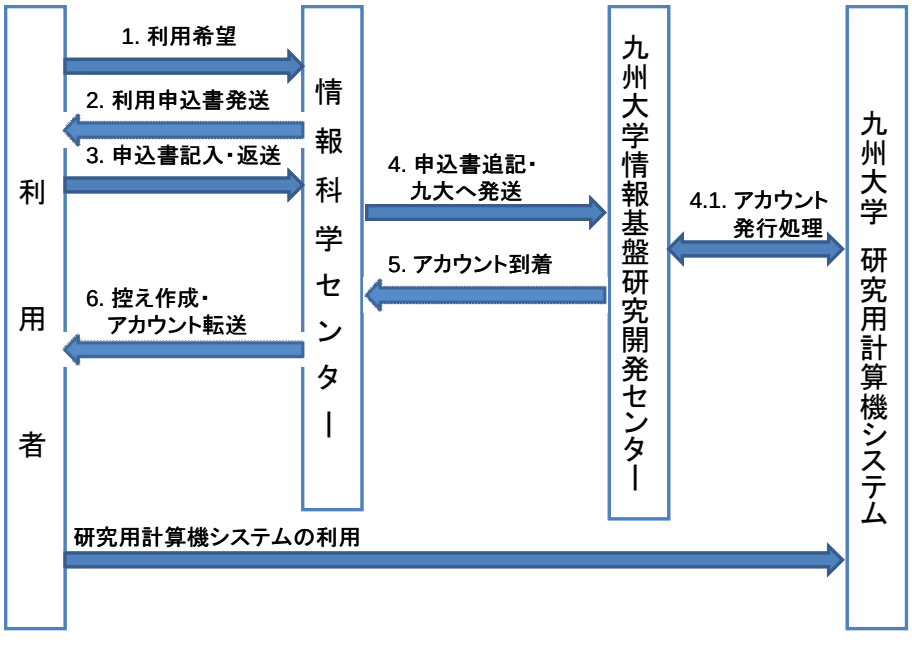

図 2: 研究用計算機システム利用支援の流れ

上記の手順により，ユーザは九大との連絡や利用負担金に関する会計上の手続きを一切行う必要がありません．またこの手順では，利用希望からアカウントの配布が完了するまで，2 週間程度を要します．

なお，募集期間が 4 日間と短かったため，告知した期間内での応募は 7 件にとどまりました．しかし，その後も応募が継続的にあり，現時点で 13 名がシステムの利用を行っています (表 5)．全てのキャンパスに利用者がいることから，広範な分野で利用が行われていることが考えられます．

表 5: 研究用計算機システム応募者内訳

応募時期	07/06 月	07/07 月	07/08 月	07/09 月以降	合計
戸畑	3	0	0	0	3
飯塚	4	2	2	0	8
若松	0	0	0	2	2

4.1.4 利用講習会の企画

研究用計算機システムの利用者およびシステムに興味を持つ職員や学生を対象に，システムの利用講習会を計画しました.

講習会にて取り扱う内容に関して，研究用計算機システムの応募者に対してヒアリングを行い，プログラムの編成を行うことを計画しました．ヒアリングの結果,「プログラムの実行方法」「OpenMP/MPI による並列プログラミング方法」の講習に関する意見が多数を占めました．また,「応用アプリケーションでどのような処理が可能であるか知りたい」といった意見もありました.

この結果に基づき，下記の編成での講習会を検討しました.

1. 3 時間程度の講習会を開催し，システムが最低限利用できるトピックを取り扱う．具体的には,「システムの概要紹介」「バッチジョブの作成および投入方法」「OpenMP/MPI による並列プログラミング講座」「応用アプリケーションの紹介」に関して取り扱う.
2. 講義形式に加えて，実習形式を含んだ講習とする.
3. 試行として，飯塚キャンパスのみで開催を行う.
4. 九州大学情報基盤研究開発センターより講師を招き開催する.

上記の内容に関して，九州大学情報基盤研究開発センターに講習会が開催可能であるか打診したところ,「3 時間という時間上の制約から，並列プログラミング講座に関しては，OpenMP と MPI の何れかに限定すれば可能である」との回答を頂きました．この回答に基づき，並列プログラミング講座に関しては MPI を取り扱う編成として，利用講習会の開催が決定しました.

最終的に決定した利用講習会の概要は以下の通りです.

講座名 九大スーパーコンピュータシステム利用講習会

日時・会場 2007 年 10 月 25 日 14:00〜17:00 , 情報科学センター (飯塚) 端末演習室 I

講師 九大より招聘

定員 20 名

プログラム
1. 九大研究システムの概要
2. システム利用の基本操作 (プログラムの作成方法，ジョブの投入方法について)
3. MPI 概論 (MPI を用いた並列計算が可能な最小限の書式や関数について)
4. 応用ソフトェア (Gaussian) 解説

4.1.5 利用講習会の開催

上記で述べた利用講習会に関して，2007 年 10 月に発行した ISC News No.213 にて「九大スーパーコンピュータシステム利用講習会について 」として学内に告知を行いました.

講習会は，飯塚キャンパスの情報科学センター端末演習室 I にて開催しました．募集人員 20 名に対して，応募締め切り日までに 16 名の応募があり，当日参加や資料のみの希望者を含めると 18 名が講習を受けました．学生を募集対象に含めたため，学生からの応募が多く，11 名の学生が講習を希望しました．講習会を受講した本学教職員・学生の内訳を表 6 に示します.

表 6: 九大スーパーコンピュータシステム利用講習会応募者内訳

	教育職員	技術職員	学生	合計
戸畑	1	0	0	1
飯塚	4	1	11	16
若松	1	0	0	1

当日の講習会の進行は，九州大学情報基盤研究開発センターの南里准教授，渡部准教授，上田技術職員により行われました．プログラムはほぼ予定通り進行し，研究用計算機システムの概要に関する紹介，システムへのログインの方法，バッチジョブの作成および投入方法に続き，MPI によるプログラムの作成法に関する講習が 3 時間にわたって行われました．

講習は講義形式のみではなく，配布されたテキストに沿ってプログラムを実行する演習形式も含まれ，研究用計算機システムの活用につながる有用なものであったと思います．また，配布されたテキストには MPI で用いる関数のリファレンスや実装上の注意事項等が含まれており，利用者にとって便利な資料になるのではと思います．

4.1.6 今後の予定・課題

今後の予定として，年度末にユーザからの利用レポートを集計し，来年度に借り受けるシステムの選定作業が挙げられます．今年度はユーザからの希望に応じて共有タイプのシステムを借り受けていましたが，来年度は何れかのシステムに限定し，占有タイプを借り受けることを検討しています．これに合わせて，ユーザへの利用負担方法も検討する必要があります．

また，講習会の開催方法を再検討する必要があります．講習会は飯塚キャンパスのみでの開催であったため，飯塚キャンパスからの参加が多かったものの，戸畑，若松からの参加は難しい状況でした．教職員からのヒアリング時にも，各キャンパスで開催して欲しいという意見があり，開催地を考慮する必要があります．加えて，講習会の回数や取り扱うトピックも検討する必要があると考えています．

5 九州大学研究用計算機システムの利用例

本節では，今年度ユーザへの利用補助を行った研究用計算機システムを実際に利用し，プログラムの作成方法やシステムの演算性能について例を示します．

今回は，PC クラスタを利用した並列プログラムを想定し，スーパーコンピュータシステム B にて MPI プログラムを適用します．また，適用するアルゴリズムは C 言語で記述を行い，簡単な画像処理である「カラー画像を輝度画像に変換する処理」を実装します．

5.1 MPI プログラムの例

以下に，MPI を利用した並列プログラム例 (輝度画像への変換) を示します．MPI を用いることにより，各プロセスが割り当てられる CPU 番号と利用する CPU の個数を得ることができます．これらの情報を利用することにより，各プロセスが処理を担当するデータの範囲を定義し，並列に処理を行うことが可能となります．

MPI を利用したプログラム例

```
int main(int argc,char **argv)
{
  int ppm_width=0,ppm_height=0;
  unsigned char *ppm_buf=0,*pgm_buf=0,*local_ppm_buf=0,*local_pgm_buf=0;

  int myrank,p;
  MPI_Status status;
  MPI_Init(&argc,&argv);
  MPI_Comm_rank(MPI_COMM_WORLD,&myrank); //CPUl
  MPI_Comm_size(MPI_COMM_WORLD, &p); //pCPUl

  if(myrank == 0)
    {
        ppm_buf = ppm_read("./img/001.ppm",&ppm_width,&ppm_height);
        pgm_buf = (unsigned char*)malloc(ppm_width*ppm_height);
     }
  else
    {
        ppm_read_header("./img/001.ppm",&ppm_width,&ppm_height);
    }

  local_ppm_buf = (unsigned char*)malloc(ppm_width*ppm_height*3/p);
  local_pgm_buf = (unsigned char*)malloc(ppm_width*ppm_height/p);

  // eCPUM
  if(myrank == 0)
    {
      for(int i=0;i<p;i++)
        {
          int o = ppm_width*ppm_height*3/p;
          MPI_Send(ppm_buf+i*o,o,
                   MPI_UNSIGNED_CHAR,i,99,MPI_COMM_WORLD);
        }
    }

  int o = ppm_width*ppm_height*3/p;
@MPI_Recv(local_ppm_buf,o,MPI_UNSIGNED_CHAR,0,99,MPI_COMM_WORLD,&status);
```

MPI を利用したプログラム例 (続き)

```
  //
  for(int i=0;i<ppm_width*ppm_height/p;i++)
    {
      local_pgm_buf[i] =
        local_ppm_buf[i*3+0]*0.3 +
        local_ppm_buf[i*3+1]*0.6 +
        local_ppm_buf[i*3+2]*0.1;
    }

  //
  MPI_Send(local_pgm_buf,ppm_width*ppm_height/p,
           MPI_UNSIGNED_CHAR,0,
           100,MPI_COMM_WORLD);
  if(myrank == 0)
    {
      for(int i=0;i<p;i++)
        {
          int o = ppm_width*ppm_height/p;
            MPI_Recv(pgm_buf+i*o,o,
                   MPI_UNSIGNED_CHAR,i,
                   100,MPI_COMM_WORLD,&status);
        }
    }
  MPI_Finalize();
  return 0;
}
```

5.2 MPI プログラムの実行

スーパーコンピュータシステム B で MPI プログラムをコンパイルする場合は，以下のコマンドを実行します (例のソースファイル名は vision.c とします)．なお，作業は全てスーパーコンピュータシステムのフロントエンドサーバ (tatara.cc.kyushu-u.ac.jp) 上で行います[9].

MPI プログラムのコンパイル

```
% mpifcc -Kfast -pg vision.c
```

-pg オプションは，Intel Xeon プロセッサで動作する実行コードの生成を指定するものであり，スーパーコンピュータシステム B で動作させるために必要となります.

作成したプログラムを動作させるためには，下記のようなバッチリクエストファイル (例のバッチリクエストファイル名は vision.sh とします) を記述します．下記は，大規模共有タイプで利用できる最大のリソースを指定したリクエストとなります.

[9]プログラムの作成に関しては個人の PC で行い，ソースファイルを scp コマンドを用いてフロントエンドに転送することが可能です.

バッチリクエストファイルの例

```
#!/bin/csh
#@$-q c128
#@$-lP 128
#@$-lp 1
#@$-eo
#@$-oi
cd $QSUB_WORKDIR
mpiexec -n 128 ./a.out
```

最後に，バッチリクエストファイルを用いて，スーパーコンピュータシステムにプログラムを投入します．

バッチリクエストファイルの例

```
% qsub ./vision.sh
```

一般的なタイムシェアリングシステムと異なり，命令は一旦待ち行列に保存され，計算機のリソースが確保された時点で演算が行われます．また，フロントエンドサーバからログアウトを行った場合でも，演算は中断されずに行われます．バッチリクエストファイルに連絡用のメールアドレスを記入すれば，ジョブの終了を通知するメールを受信することもできます．

5.3 演算性能

構築した輝度画像への変換プログラムに関して，割り当てる CPU 数を 1 から 128 まで変化させた場合の，実行時間の変化を図 3 に示します．入力画像のサイズは 640 × 480 ピクセルに設定し，変換処理を 100 回繰り返し実行しています．

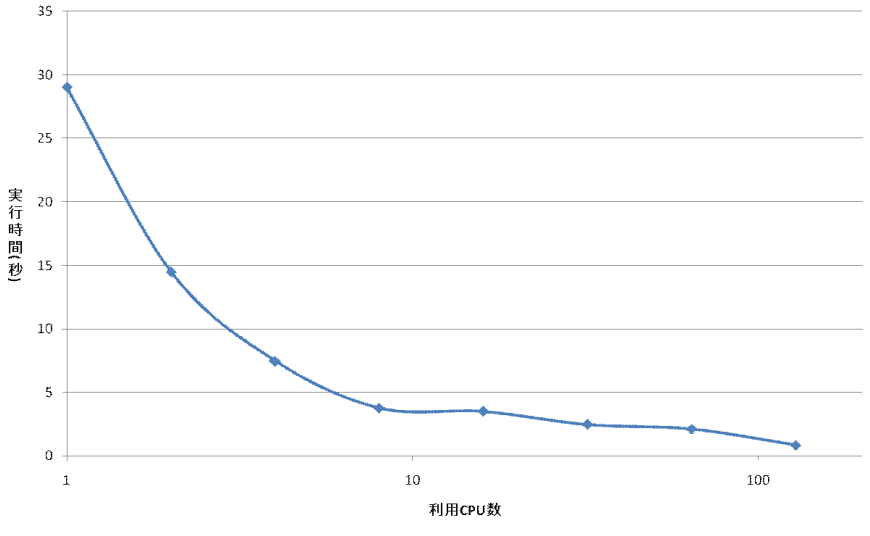

図 3: 並列プログラムの演算性能 (実行時間)

結果は，CPU 数が 8 個以下の場合は，ほぼ理論通りの速度向上率を示しています．16 個を超える場合速度の向上率は鈍くなりますが，CPU1 個による実行と比較し，128 個による実行は 30 倍近い高速化が行えており，プログラムの並列化は有用であるといえます．また，今回は小規模なメモリを利用した演算を行っていますが，MPI

を用いることにより多数の計算機にデータを分散させることが可能であるため，大規模な演算を簡便なプログラムで行うことが可能であるといえます.

6　まとめ

本稿では，本年度情報科学センターが行った研究用計算機システムの利用支援について解説を行いました．研究用計算機システムとして，九州大学情報基盤研究開発センターに導入されているスーパーコンピュータシステム及び高性能アプリケーションサーバを活用し，利用者へのサービス提供を行いました．具体的には，九大研究システムの利用登録や問い合わせの代行，利用負担の補助を行い，13 名のユーザにサービスを提供しました．教職員と学生を対象としたシステムの利用講習会には当日参加を含め 18 名の応募があり，研究用計算機システムの活用につながるポイントを提供できたと考えています.

九大研究用計算機システムは 2007 年に導入された最新のシステムであり，高い計算能力を有するシステムです．来年度も，提供システムや利用負担等の再検討を行い，利用支援を継続する計画です．大規模な演算を行いたい，100 コアを超える PC クラスタシステムを利用したい方は，是非とも利用していただきたいと思います．また，研究用計算機システムに関する問い合わせ先を res-system@isc.kyutech.ac.jp として開設していますので，お気軽にお問い合わせください.

謝辞

利用者登録等に関してお世話になっております九州大学情報基盤研究開発センター共同利用係の皆様に感謝申し上げます．また，システムの借り受けに関するアドバイスを頂きました九州大学の天野准教授，九大スーパーコンピュータシステム利用講習会の準備および当日の進行を頂きました九州大学の南里准教授，渡部准教授，上田技術職員に感謝申し上げます.

本解説記事の執筆にあたり，九州大学情報基盤研究開発センターの研究用計算機システムを利用しました.

参考文献

[1] 九州大学情報基盤研究開発センター 研究用計算機システム，http://www.cc.kyushu-u.ac.jp/scp/.

[2] ”The OpenMP specification for parallel programming”，OpenMP Architecture Review Board，http://www.openmp.org/.

[3] ”Message Passing Interface Forum”，MPI Forum，http://www.mpi-forum.org/.