サイエンティフィック・システム研究会
科学技術計算分科会１回会合資料

FUJITSU

富士通のPCクラスタへの取り組み
～システム～

2001年8月3日
株式会社富士通プライムソフトテクノロジ
第一開発統括部第三開発部
小宮山　　孝
E-Mail:komiyama@pst.fujitsu.com

1

内容構成：

1)ＰＣクラスタの動向と取り組みの背景
2)計算性能の向上
3)ＰＣクラスタの構成と特徴
4)BeowulfとSCoreの比較
5)SCore使用ユーザ
6)ＰＣクラスタへの富士通の取り組み
7)ＰＣクラスタの利用分野
8)クラスタ関連の製品・サービスの提供
　・PCクラスタ向けPRIMERGY
　・クラスタシステム構築支援サービス
　・ＨＰＣユーザプログラムチューニングサービス
　・システムの検証支援
9)PCクラスタコンソーシアム

ＰＣクラスタの動向と取り組みの背景

■ＰＣクラスタの動向
－米国ＮＡＳＡのBeowulfプロジェクト
１９９４年　１６台のＰＣを１０MbpsのEthernetで接続
コモディティな部品の組み合わせでクラスタシステムを構成

－日本RWCPのSCoreプロジェクト
１９９６年　３２台のＰＣをMyrinetで接続
SCoreというＰＣクラスタ用ソフトウエアを開発

－計算サーバとしての性能の向上
ＰＣクラスタが、Top５００(２００１年）の上位にランキング

■ＰＣクラスタへの取り組みの背景
ＰＣの高性能化、ユーザの利用拡大、言語等開発ツール整備
→富士通としてトータルなソリューションを提供

 2

ＰＣクラスタの動向についてご説明します。
－米国ＮＡＳＡのBeowulfプロジェクトは、１９９４年に１６台のＰＣを１０Mbps のEthernetで接続したシステム構成で、MPI等のコモディティな部品の組み合わせで構成したＰＣクラスタシステムです。
－日本RWCPのSCoreプロジェクトは、１９９６年に３２台のＰＣをMyrinetで接続したシステム構成で、SCoreというＰＣクラスタ用ソフトウェアを開発したＰＣクラスタシステムです。
－ ＰＣクラスタは、計算サーバとしての性能が向上しており、２００１年度では計算サーバTop５００の上位にランキングされています。

ＰＣクラスタへの取り組みの背景として、ＰＣの高性能化、研究部門を中心としたユーザの利用拡大、コンパイラ等の言語やデバッガ等の開発ツールが整備されていることがあげられます。
このような背景から、富士通では、トータルなソリューションを提供いたします。

計算性能の向上

■クラスタTop500(*1)のベスト5状況（2001年6月）

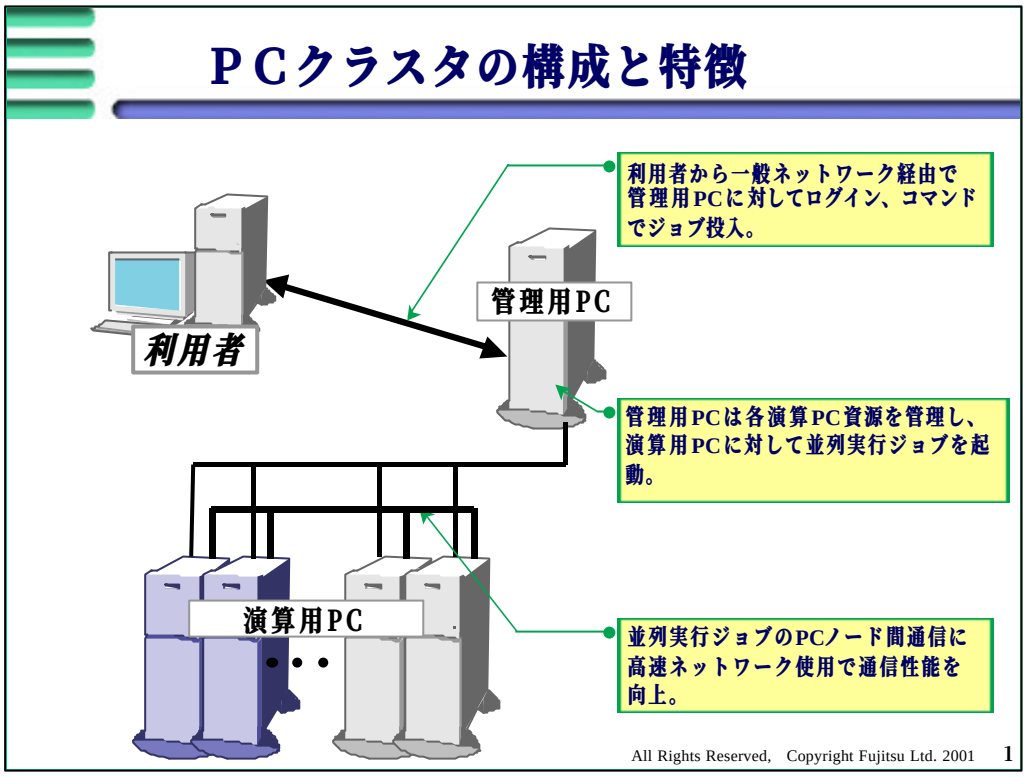
3、4位がSCoreを使用

	構築者	ノード数	CPU数	ピーク性能	OS	PC	NETWORKS
1.	IBM	1030	1038	1037GFlops	Linux	1024MHz	Gig abit Ethenet
2.	IBM	516	1032	1032GFlops	Linux	1000MHz(P3)	Myrinet2000
3.	NEC	520	1040	967GFlops	(SCore)	933MHz(P3)	Myrinet2000
4.	RWCP	512	1024	955GFlops	(SCore)	933MHz(P3)	Myrinet2000
5.	Inp(*2)	500	920	701GFlops	Linux	700MHz(P3)	Fast Ethenet

＊1：クラスタTop500は、Linpackベンチマークプログラム絶対性能による世界のスーパーコンピュータのランク付けリスト

＊2：Inpharmatica Ltd

Linpackベンチマークプログラム絶対性能による世界のスーパーコンピュータのランク付けリスト「Clusters TOP500」では、ベスト5位が全てPCクラスタでした（2001年6月現在）。また、そのうち3,4位ではScoreを使用していました。



PCクラスタの構成についてご説明いたします。上の図は、 複数の利用者PC,通常一台管理用PC、複数の演算用PCのネットワーク接続構成図です。

・利用者PCと管理用PC間は、LANなどの一般ネットワークで接続します。
・管理用PCと各演算用PC間は、Ethernetなどの高速ネットワークで接続します。
・各演算PCノード間は、Myrinetなどの超高速ネットワークで接続します。

また、PCクラスタの特徴として以下があげられます。

・利用者から一般ネットワーク経由で管理用PCに対してログイン、コマンドでジョブ投入できます。
・管理用PCでは、各演算PC資源を管理し、投入されたジョブ処理として演算用PCに対して並列実行ジョブを起動します。
・並列実行ジョブのPCノード間通信には、Myrinetなどの超高速ネットワークを使用することで通信性能の向上を図ります。

BeowulfとSCoreの比較（１）

■Beowulf型
下記に代表される既存ソフトウェアの組み合わせで
システム構築
－MPI（＊1）やPVM（＊2）通信ライブラリ
－PBS（＊3）のバッチシステム
－Fortran C/C++コンパイラ

■SCore
PCクラスタシステムソフトウェアを提供
－運用管理機能、ユーザレベル通信用専用デバイス
－MPI（注1）やPVM（注2）通信ライブラリ
－PBSバッチシステム

注１：MPICH 1.2.1 規格を提供
注２：SCore 4.0 から PVM 3.4 規格を提供

＊1：Message Passing Interface
＊2：Parallel Virtual Machine
＊3：Portable Batch System

BeowulfとSCoreの比較をいたします。

Beowulf型は、既存ソフトウェアの組み合わせでシステムを構築します。

SCoreは、 PCクラスタシステムソフトウェアを提供します。 特に、独自開発されたクラスタ運用管理機能とMyrinetなどのユーザレベル通信用専用デバイスが提供される点が異なります。また、MPIは、MPICH 1.2.1 規格を提供、PVMは、SCore 4.0 から PVM 3.4 規格を提供します。

BeowulfとSCoreの比較（２）

■Beowulf型
通信層は、TCP/IP（＊1）プロトコル

■SCore
通信層は、UDP（＊2）プロトコルで高性能。
同一アプリケーションバイナリは実行時にオプションで
下記のネットワーク等を選択可能
－Myrinetギガビットネットワーク
－10/100/1000 Mbpsイーサーネット

＊1：Transmission Control Protocol/Internet Protocol
＊2：User Datagram Protocol

 6

また、通信層の比較をいたしますと、Beowulf型は、TCP/IP プロトコルを使用しています。

SCoreは、UDP プロトコル使用により高性能となっております。
SCore上で開発された同一アプリケーションバイナリは、実行の時にオプションで、異なるネットワークを選択することが可能です。

BeowulfとSCoreの比較（３）

■Beowulf型
クラスタ運用管理機能は、 PBS(Portable Batch System)
DQS(Distributed Queueing System)など

■SCore
クラスタ運用管理機能は,PBSとSCore-Dオペレーティング
システムでホスト、ネットワーク、ディスク等を一括管理
－3つのジョブスケジューリングモード提供
－チェックポイント・リスタート機能（＊1）
－ジョブマイグレーション機能（＊1）

＊1：ジョブが扱うファイル情報復元不可等の制限事項あり、
CPUバンドジョブに限定。

クラスタ運用管理機能の比較をいたしますと、Beowulf型は、 PBSやDQSなどです。

SCoreは、PBSとSCore-Dオペレーティングシステムにより、ホスト、ネットワーク、ディスク等を一括管理することができます。

・シングルユーザモード、マルチユーザモード、およびバッチ処理モードの
３つのジョブスケジュールモードが選択可能です。
・長時間ジョブのために、チェックポイント・リスタート機能（＊１）や
・ジョブマイグレーション機能（＊１）が提供されます。

＊1：ジョブが扱うファイル情報復元不可等制限事項あり、CPUバンドジョブに限定されます。

SCore使用ユーザ

■日本
- 科学技術振興事業団(JST)
- 三菱産業システム研究所
- 独立行政法人通信総合研究所
- 先端医療センタ
- 理化学研究所、東京大学、東京工業大学、大阪大学、、、

■米国
- Los Alamos国立研究所

■イギリス
- オックスフォード大学スーパーコンピューティングセンタ

■ドイツ
- ボン大学
- ティエビンゲン大学
- ハイデルベルグ大学

■フランス
- パリ南大学

上の図は、Ｓｃoreを使用されているユーザの一覧です。ご覧の通り、多くのユーザに利用されています。

ＰＣクラスタへの富士通の取り組み

■ＰＣクラスタの位置付け
価格性能比のよい、中小規模向け、分散並列計算サーバ
（数百台規模の大規模サーバは、特定用途向け）

■サポートするPCクラスタシステム
機能、性能、実績などから以下の２つを採用
－ＳCore　：RWCP(新情報処理開発機構)で開発されたフリーソフトウェア含むPCクラスタシステムをサポート
－Beowulf：米国ＮＡＳＡで開発されたPCクラスタシステム構築をサポート。*Beowulfは今年後半よりサポート予定

■サポートの特徴
製品に加えサービス、検証支援のトータルなサポートを提供
－製品：PCハード／ネットワークハードと言語処理システム
－サービス：管理者向けクラスタシステム構築支援サービスとエンドユーザ向けプログラムチューニングサービス
－検証支援：検証センターでのシステム検証支援

 9

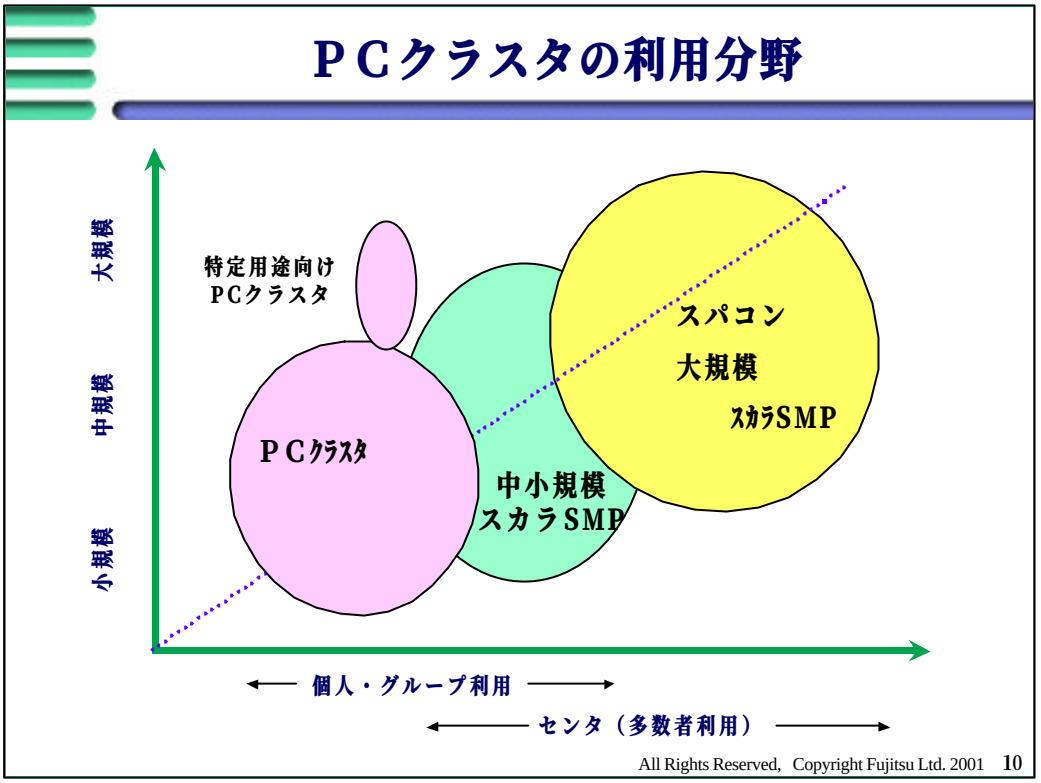
ＰＣクラスタへの富士通の取り組みについてご説明します。

ＰＣクラスタについては、価格性能比のよい、中小規模向け、分散並列計算サーバとして提供します。また、数百台規模の大規模サーバは、特定用途向けとして提供します。

サポートするPCクラスタシステムは、機能、性能、実績などから以下の２つを採用しました。

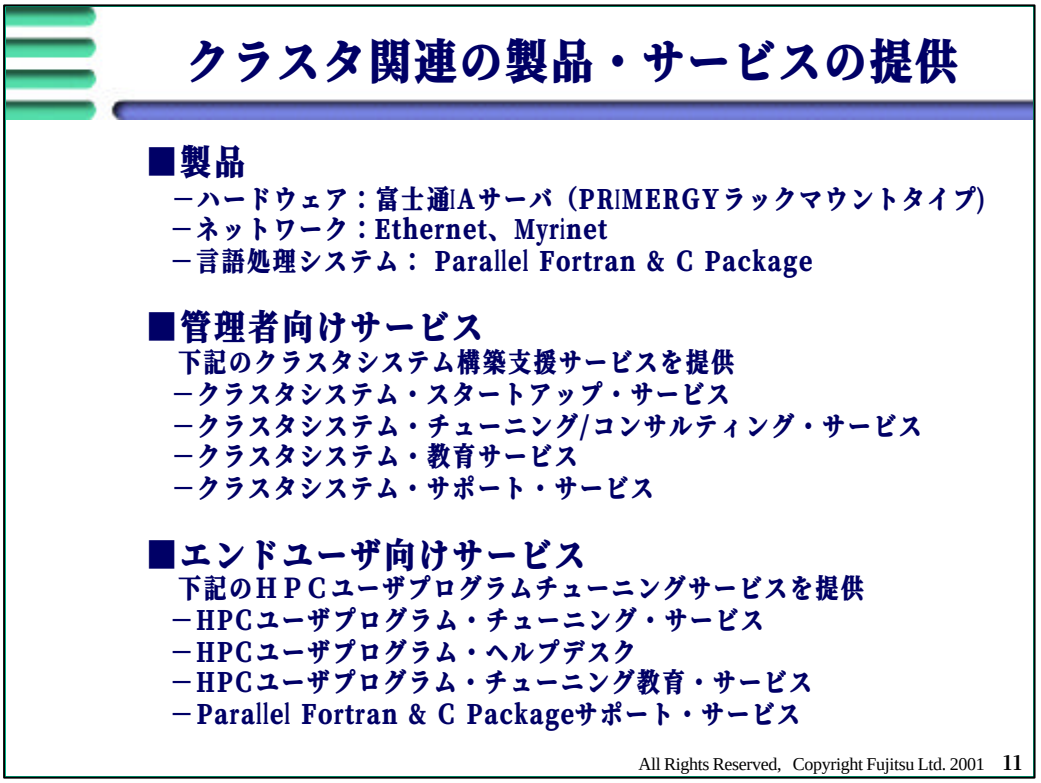
・ＳCoreは、 RWCP(新情報処理開発機構)で開発されたフリーソフトウェアを
含むPCクラスタシステムをサポートします。
・Beowulf は、米国ＮＡＳＡで開発されたPCクラスタシステム構築をサポートします。Beowulfは今年(２００１年)後半よりサポート予定です。

サポートの特徴は、製品に加え、管理者やエンドユーザに対するサービスと検証支援のトータルなサポートを提供します。



上の図は、ＰＣクラスタの利用分野として、縦軸に規模、横軸に利用者数を表すものです。

・超大規模の多数利用は、スパコン、大規模スカラＳＭＰの分野です。
・中規模多数利用は、中小規模スカラSMPの分野です。
・小規模の個人またはグループ利用分野がPCクラスタの分野です。
　また、数百台規模の大規模サーバも特定用途向けとして提供します。



富士通は、クラスタ関連の製品として、以下を提供します。

・ハードウェア：富士通IAサーバ（PRIMERGYラックマウントタイプ）
・ネットワーク：Ethernet、Myrinet
・言語処理システム： Parallel Fortran & C Package

また、管理者向けサービスとして、４つのクラスタシステム構築支援サービスを提供します。

・クラスタシステム・スタートアップ・サービス
・クラスタシステム・チューニング/コンサルティング・サービス
・クラスタシステム・教育サービス
・クラスタシステム・サポート・サービス

エンドユーザ向けサービスとして、４つのＨＰＣユーザプログラムチューニングサービスを提供します。

・HPCユーザプログラム・チューニング・サービス
・HPCユーザプログラム・ヘルプデスク
・HPCユーザプログラム・チューニング教育・サービス
・Parallel Fortran & C Packageサポート・サービス

PCクラスタ向けPRIMERGY

メモリ
最大
4GB

メモリ
最大
16GB

PentiumⅢ 1BGHz/933MHz
最大メモリ　4GB
最大DISK　81.8GB/109.2GB
1U(厚さ4.4cm)タイプ

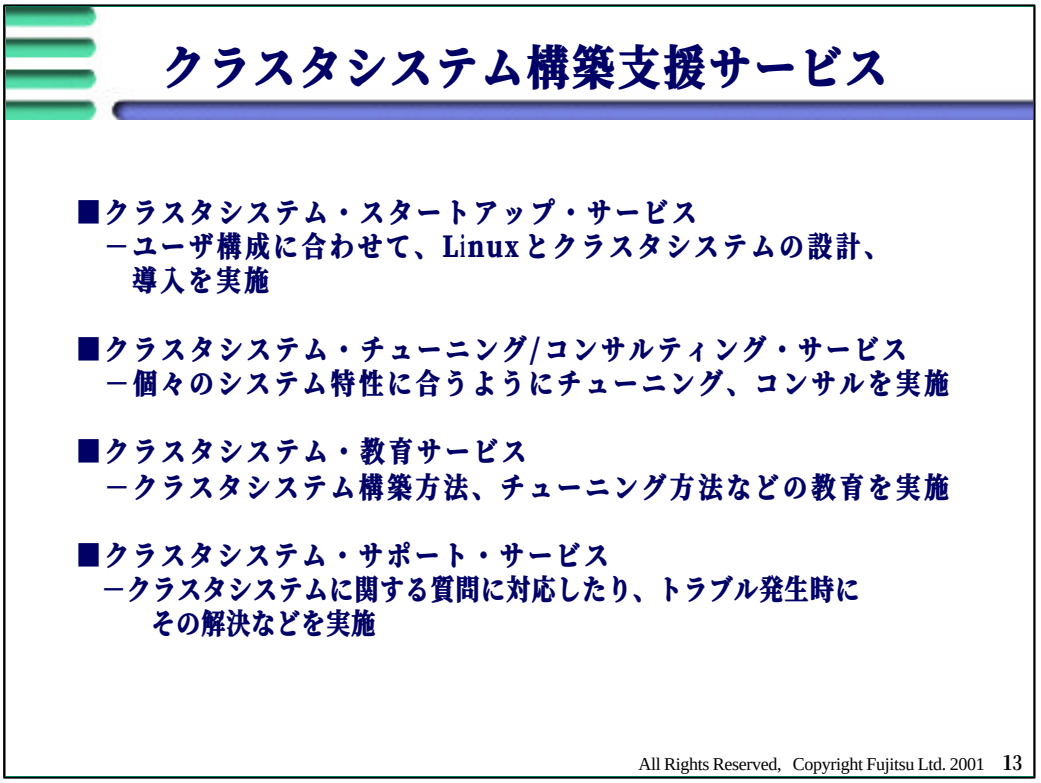
DISK
最大 182GB
/146GB
4Uタイプ

DISK
最大 291GB
7Uタイプ

 12

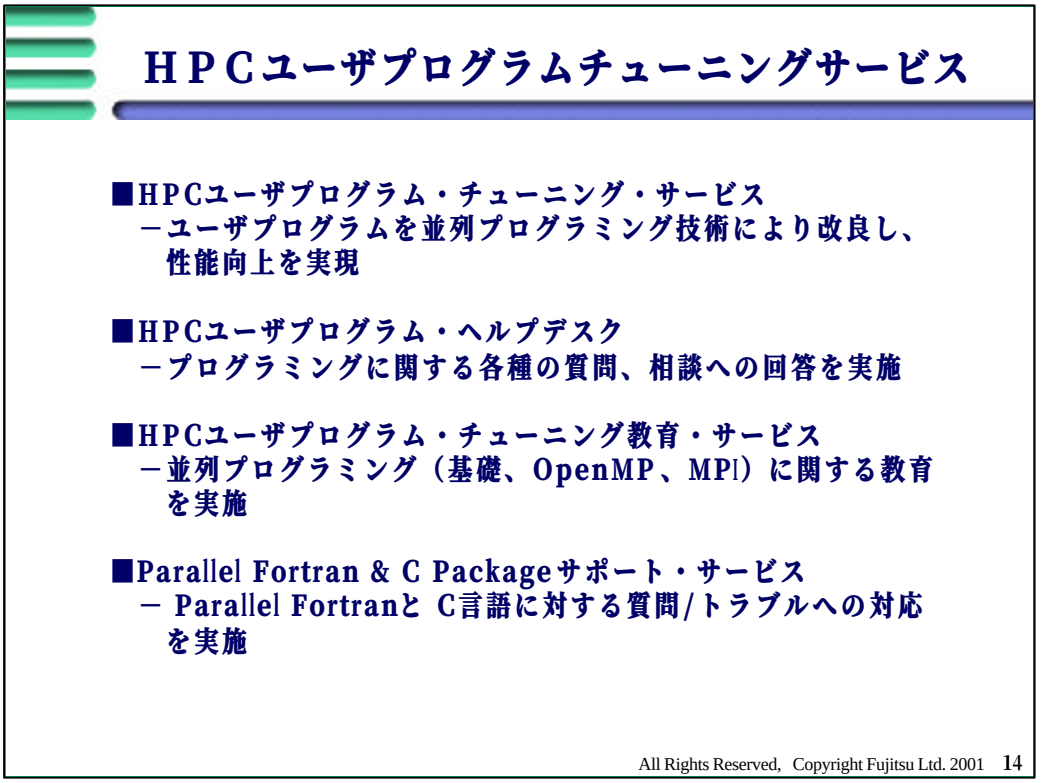
PCクラスタ向けPRIMERGYとして、演算用PC向けに、1Uタイプとしてでは、TS125とTS225の２機種を用意しています。また、1.7GHzのPentium4を搭載したラックマウントタイプを個別に用意しています。

管理用PC向けには、4UタイプとしてES320とN400の２機種を用意、また、7UタイプとしてH200とMS610の２機種を用意しています。



クラスタシステム構築支援サービスとして、以下の４つを提供いたします。

・クラスタシステム・スタートアップ・サービス
ユーザ構成に合わせて、Linuxとクラスタシステムの設計から、顧客先でのシステム環境設定を含めた導入までを実施します。
・クラスタシステム・チューニング/コンサルティング・サービス
個々のシステム特性に合うようにシステムをチューニング、コンサルティングを実施します。
・クラスタシステム・教育サービス
クラスタシステム構築方法、チューニング方法などの教育を実施します。
・クラスタシステム・サポート・サービス
Linux-OSおよびSCore含めたPCクラスタシステムに関する質問に対応したり、トラブル発生時にその解決などを実施します。



ＨＰＣユーザプログラムチューニングサービスとしては、以下の4つを提供します。

・HPCユーザプログラム・チューニング・サービス

ユーザプログラムを 借用して、当社専門スタッフが並列プログラミング技術により改良し、性能向上を実現します。

・HPCユーザプログラム・ヘルプデスク

プログラミングに関する各種の質問、相談への回答を実施するサービスです。

・HPCユーザプログラム・チューニング教育・サービス

並列プログラミング（基礎、OpenMP、MPI）に関する教育を実施します。

・Parallel Fortran & C Packageサポート・サービス

Parallel FortranとC言語に対する質問/トラブルへの対応を実施します。

システムの検証支援

■インテル株式会社との協調による
「富士通ＩＡソリューションセンター」

- お客様の業務システム構築・性能検証などを支援
- １００台規模のPRIMERGYサーバを常設

FUJITSU IA Solution Center

 15

システムの検証支援として、インテル株式会社との協調により、「富士通ＩＡソリューションセンター」を西新宿に開設しています。

お客様の業務システム構築・性能検証などを支援するため、１００台規模のPRIMERGYサーバを常設しています。

PCクラスタコンソーシアム

■PCクラスタコンソーシアム
－10月設立予定
－SCoreの維持、拡張、配布、普及を中心に活動予定

■コンソーシアムへの富士通の貢献
－正会員として加入し、以下のメンバとして活動を予定
・理事会、専門部会メンバ
－以下の活動が重要であり貢献していきたい
・SCoreの維持、拡張、配布
・ISVアプリケーションの充実
・各種の情報収集と公開、普及活動

PCクラスタコンソーシアムは、今年10月設立予定で、SCoreの維持、拡張、配布、普及を中心に活動予定しています。

富士通は、PCクラスタコンソーシアムに正会員として加入し、理事会、専門部会メンバとして活動を予定しています。　また、SCoreの維持、拡張、配布、ISVアプリケーションの充実、各種の情報収集と公開、普及活動により、今後のPCクラスタ発展に貢献していきたいです。