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　第2回会合資料

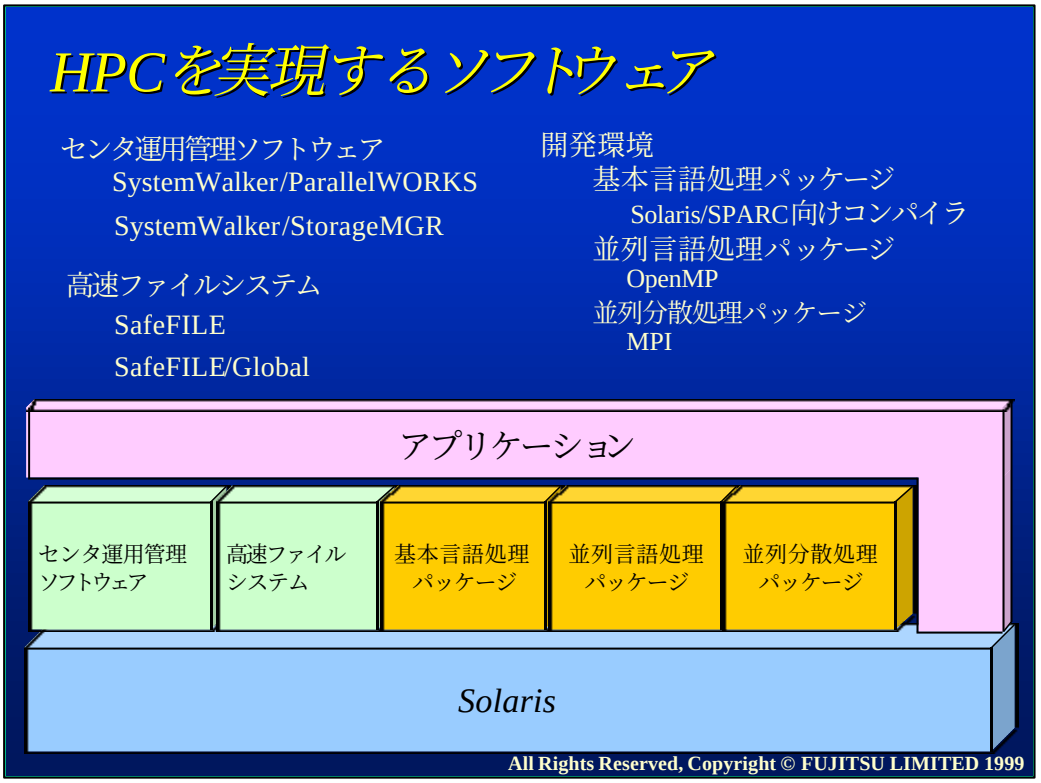
新スカラサーバGP7000F ソフトウェアと利用環境について（報告）

1999年11月5日
富士通（株）ソフトウェア事業本部
第二ソフトウェア事業部
山根　恒美
nae01725@nifty.ne.jp

All Rights Reserved, Copyright © FUJITSU LIMITED 1999

本稿では、新スカラサーバGP7000Fのソフトウェアとその利用環境について報告する。

従来、富士通はR&D分野にスカラサーバとしてSファミリと、Sファミリをノードに採用したAP3000システムを提供してきた。新スカラサーバGP7000Fは自社製のUNIXサーバ機であり、当社のサーバ戦略の中で非常に重要な地位を占めるものである。



このGP7000Fは従来のR&D分野用UNIXサーバと、ビジネス分野用UNIX
サーバの統合を図った汎用、かつオープンなUNIXサーバである。

OSとしてSolaris7の６４ビットOSを採用し、またアプリケーションについてはSolaris 7以降のアプリケーションはバイナリがそのままでGP7000F上で動作するため、ユーザは非常に幅広い選択肢を享受できる。

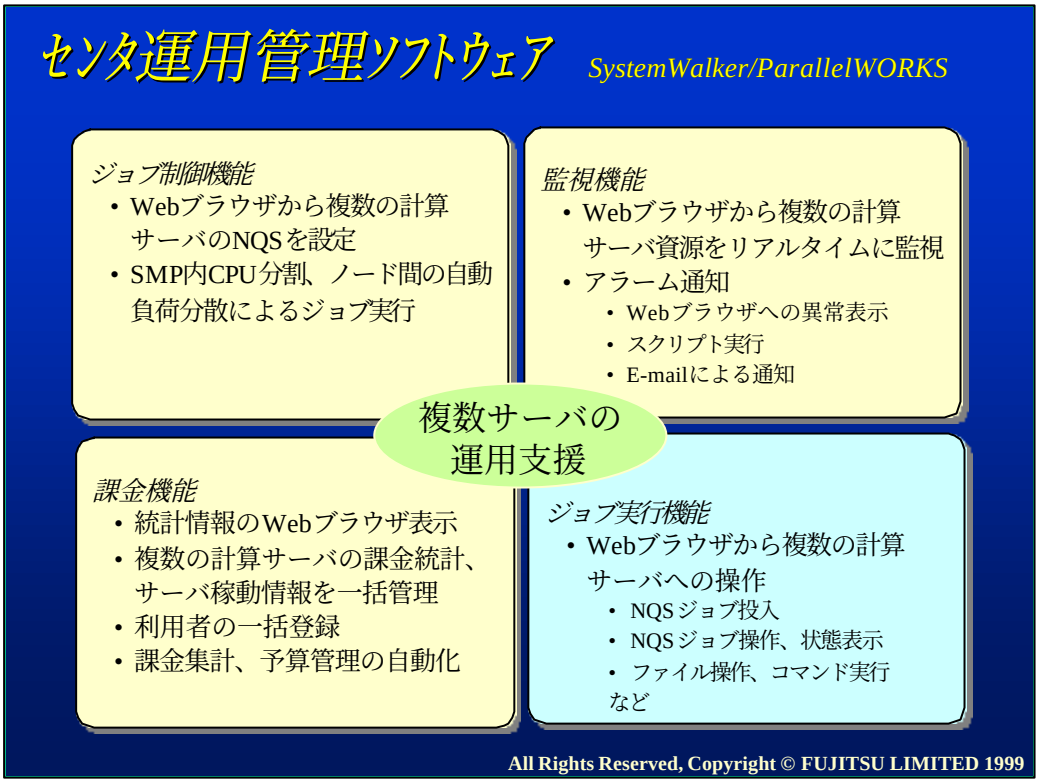
本稿ではHPCへの取り組みを主眼に、OSとアプリケーションの間に位置す
るミドルウェアの領域から、センタの運用管理のためのソフトウェア、ファイル
システムへの取り組み、GP7000F用プログラム開発支援環境とGP7000Fにおけるプログラミングパラダイムについて紹介をする。

センタ用運用管理ソフトウェアを説明する前に、富士通の運用管理ソフトウェアの体系について説明する。

富士通の運用管理ソフトウェアはグローバルサーバ（メインフレーム）からUNIXサーバ、さらにPCサーバにいたるまでSystemWalkerという製品体系にまとめようとしている。ここに幾つか例となる製品を挙げている。

- CentricMGR ― 分散システム用の統合運用管理ソフトウェア
- PerfMGR ― 1日毎、1週間毎、あるいは1月毎にシステムの稼動状況を把握し、システム資源のどこにボトルネックが生じているかを見出すためのキャパシティプランニングツール
- SoftDelivery ― 数千台規模のPC/WSにソフトウェア資源を一括配布、版数管理を行う

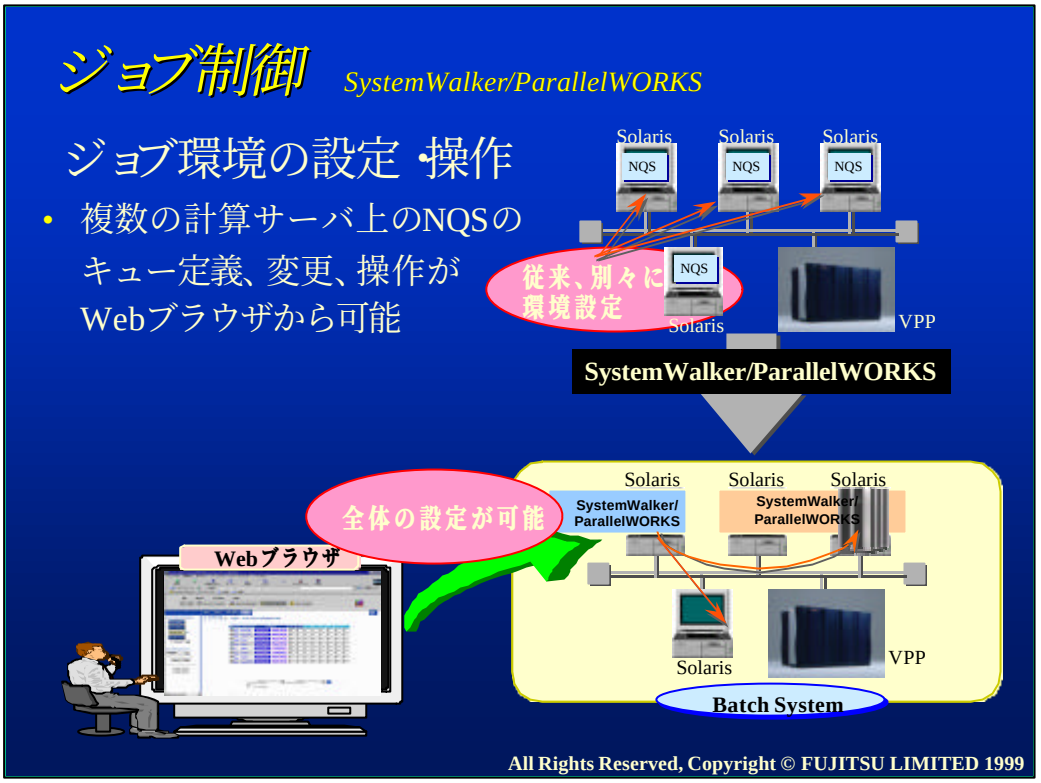
今回、SystemWalker配下に二つの新製品を提供する。一つは新ディスクアレイ装置GR720の瞬時バックアップ機能のサポートを行うStorageMGRであり、これについては後のファイルシステムの一環として説明する。もう一つの新製品が共同利用センタ用運用管理ソフトウェアであるParallelWORKSである。



共同利用センタの特徴は、不特定多数のエンドユーザから任意の時点に非定形のバッチジョブやTSSの要求が投入されることであり、その運用管理ソフトウェアに求められる要件としては、特定ユーザにシステム資源を独占使用させることなく、各ユーザの要求を公平にさばいていく中で、システム資源を最大限効果的に利用するジョブスケジューリングとスケジューリング状況の監視、および課金統計情報を採取出来ることが命題になっている。

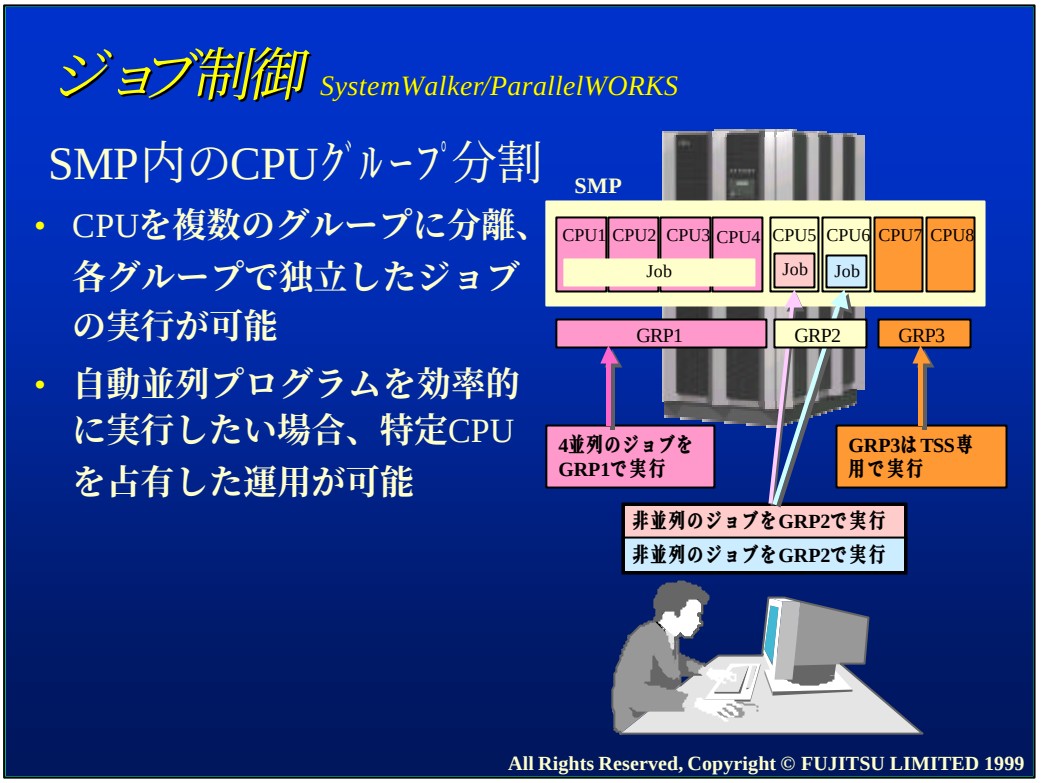
従来このためにWWSiteシリーズ、バッチシステムとしてのNQS、課金のためのUTMS-EXを提供していたが、今回それらを統合、機能エンハンスを行いSystemWalkerの体系の中に組み入れたのがParallelWORKSである。。

ParallelWORKSの機能はジョブ制御、スケジューリング、システム監視、および課金というシステム管理者向け機能と、ジョブの投入・実行というエンドユーザ向けの機能からなるが、これらが管理者にとってもエンドユーザにとっても、Webブラウザから操作できることと、複数の計算サーバに対する操作を一括して行える、いわゆるワン・システム・ビューの提供がParallelWORKSの特徴である。



ParallelWORKSの機能を幾つか例を用いて説明する。まずNQSによるバッチジョブ制御のための環境のセットアップは、従来個々の計算サーバ毎に別々に必要な作業であったが、ParallelWORKSでは一台のPC/WSのWebブラウザから複数の計算サーバのNQS環境の設定・変更、NQSキューの定義・変更が可能となった。また最新のベクトル型スーパコンピュータVPP800とVPP5000についてはParallelWORKS連携という製品を導入すれば、他のSolarisの計算サーバと同様にParallelWORKSのメンバの一員となる。

ジョブ制御機能としては、たとえば複数の計算サーバをグループ化し、共通のNQSキューを用意すると、そのキュー経由で投入されたジョブはグループ内のサーバのCPU能力比（環境設定であらかじめ定義）とジョブ実行開始時の各サーバのCPUの使用状況に応じて、実際に実行すべきサーバを自動選択する、いわゆる自動負荷分散機能を用意している。またバッチシステムを特定のユーザが占有することのないよう、ユーザ毎に一度に実行可能なジョブ数を制限することも可能である。



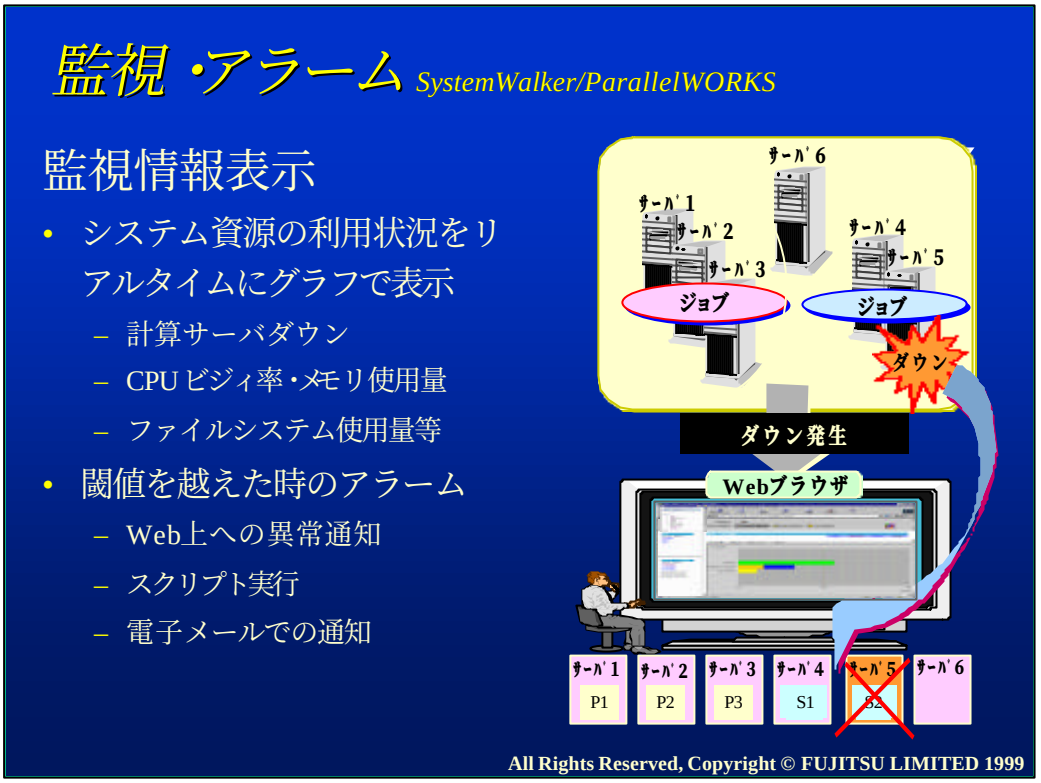
次にParallelWORKSの新機能である、NQSによるSMP内のCPUグループ分割機能について説明する。

CPUのグループ分割自身はSolarisのプロセッサセット機能としてサポートされているが、これをNQSのジョブキューと対応付けることで、各CPUグループ毎に独立した柔軟なジョブ運用が可能となる。NQSキューから投入されたジョブは対応するグループに属するCPUでのみ動作するので、別のグループで動作するジョブからのCPUの干渉を防いだり、一つのジョブがシステム内のすべてのCPUを使用するような事態を防ぐことができる。

本機能の類似機能として、GP7000Fのハードウェアのシステムパーティション機能がある。しかし、NQSによるSMP内のCPUグループ分割とシステムパーティションは以下のように全く異なる機能である。

・システムパーティション ― システムボードを単位として物理的にメモリとCPUを分割。OS(Solaris)は各パーティションで個別に動作。パーティション構成の変更はファームウェアで実施。

・CPUグループ分割 ― CPUは論理的（ソフト的）にグループ分け。メモリは分割されず各CPUで共用する。また、動作OS(Solaris）も全体で一つ。CPUグループの構成変更はSolarisのコマンドで実施し、リブートは必要としない。



ParallelWORKSによる各計算サーバの監視機能として以下がある。

・プロセス状況表示

各サーバで動作中のプロセスをリアルタイムに監視。さらに、ユーザID等の抽出キーを指定し表示データ量を制限したり、CPU時間等をもとにデータを並び替えて表示可能。

・システム資源の利用状況表示

複数計算サーバの負荷状況（CPU、メモリ）をWebブラウザに同時表示。データはリアルタイムに更新されるため、画面を開いておけばシステム資源の状態を簡単に把握可能。また、一定期間のデータをExcel形式で保持し、グラフ表示することも可能。

・アラーム機能

ジョブ実行に必要なファイルシステムやコンソールメッセージ等の状況を監視し、資源の使用量があらかじめ定義された閾値を越えた場合にアラームを上げたり、ユーザが定義したスクリプトを実行。また、コンソールメッセージやアラームログの表示が可能。

本図の左側はシステム管理者用の課金機能、右側はエンドユーザのため
のジョブ投入、実行機能である。

課金機能としては、

•利用者一括登録機能

センタ利用者をWindowsのAccessデータベースに一括登録。登録された利用者情報は複数の計算サーバで有効。

•ジョブ改札制御

利用者登録機能で設定した予算情報と課金集計機能で集計した課金高により、ユーザ単位の残予算を計算。予算不足者の利用を禁止することも可能である。

•サーバの統計情報の一括管理とブラウザ表示

•センタカストマイズ機能の充実

課金コマンドの提供、TSS改札制御のNQS出口ルーチンの提供

等が特徴である。

エンドユーザ用のジョブ実行機能としては、WebブラウザからのNQSジョブ
投入、ジョブ操作、ジョブ状態表示、ファイル操作やコマンド投入、メール操
作等を可能にしている。

次にGP7000Fのファイルシステムである SafeFILEと、SafeFILEをベースにファイル共用機能をサポートしたSafeFILE/Globalについて説明する。

言うまでもなく、システムの性能や処理能力は、単にCPU性能だけではなく、ファイルシステムの性能にも大きく依存する。

Solarisの標準ファイルシステムであるUFSには

・スペース不足を来した場合にはファイルシステムの再作成が必要

・離散ブロックや間接ブロックにより、運用中にI/O性能が劣化

・fsckがファイルシステム全体をスキャンする方式であるため、ファイルシステムの拡大に伴うfsckによる復旧時間の増大

等の課題がある。これらの課題を克服し、さらにHPC環境にふさわしい高速化、大規模化を実現するファイルシステムとして新たにSafeFILEの開発を決定した。開発に際しては、APIについてはUFS互換とし、UFS上の資産の単純移行（ファイルのコピーのみ）を可能としている。

SafeFILE/Globalは近年増加しつつあるSAN（Storage Area Network）技術を実現したDiskに適合した、複数サーバでファイル共用を可能とするソフトウェアである。

SafeFILE　高速ﾌｧｲﾙｼｽﾃﾑ

- ﾒﾀﾃﾞｰﾀ、ﾃﾞｰﾀ、ﾛｸﾞﾃﾞｰﾀの分散配置による高速ﾌｧｲﾙｼｽﾃﾑ
- ﾏﾙﾁﾎﾞﾘｭｰﾑｻﾎﾟｰﾄとﾌｧｲﾙｼｽﾃﾑの動的拡張
- 連続域割り当てによる巨大ﾌｧｲﾙの性能改善
- ｼﾞｬｰﾅﾙ（ﾛｸﾞ）による高速 fsck
- 管理ﾌｧｲﾙ数最大１２８M個
- ﾌｧｲﾙｼｽﾃﾑでの大規模並列ｱｸｾｽｻﾎﾟｰﾄ　（2000年度下期提供予定）
  - DISK性能を引き出すI/Oサイズ拡張(1MB)
  - 複数DISKを並列アクセスして高速化するファイルストライプ
    (最大16ストライプ)
  - 大規模ファイルのサポート
    最大ファイルサイズ16TB、最大ファイルシステムサイズ512TB

All Rights Reserved, Copyright © FUJITSU LIMITED 1999

SafeFILEでは性能改善として

・メタデータ（管理データ）、データ、ログデータを別ボリュームに分散配置可能

・ブロックの連続域割り当て

・更新ログ情報による高速fsck(ファイルシステムの大きさに無依存)

等を図った。（なお、fsckについてはSolaris7のUFSでは同様の手段により改善が図られている。）

運用性改善として

・マルチボリュームサポートによるファイルシステムの動的拡張

・管理ファイル数128M個

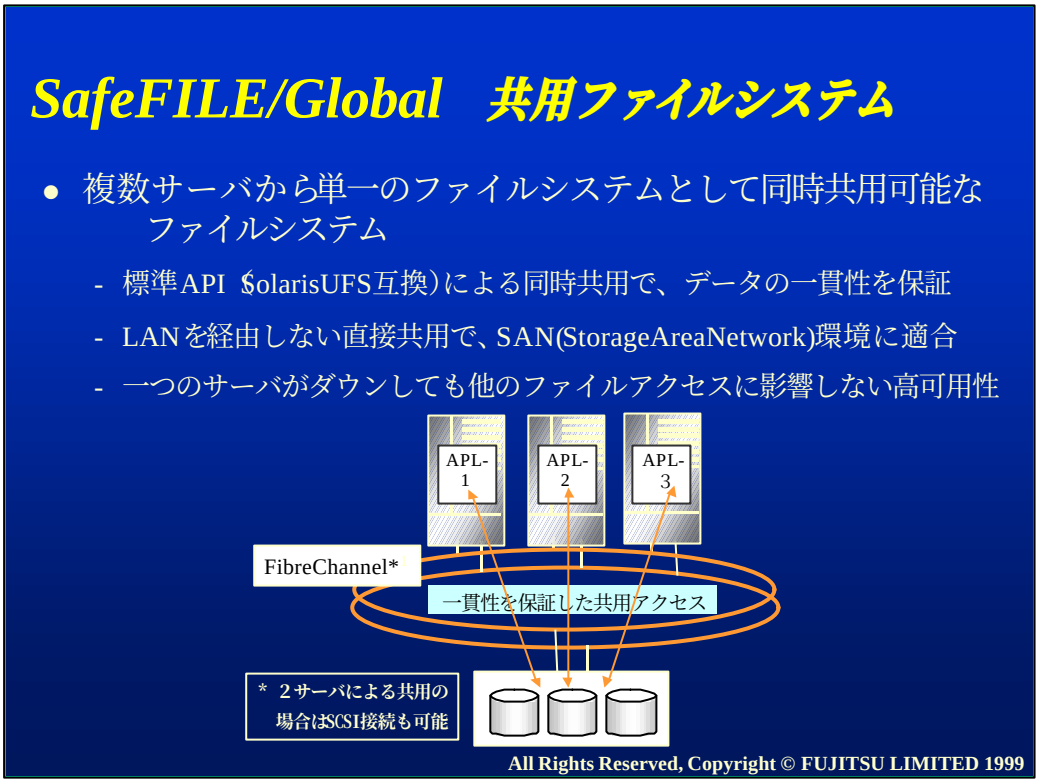
・スーパブロックと呼ぶスペース管理用データや、パーティション構成情報の二重化によるRAS強化

を実現した。さらに将来のHPC向け機能強化として大規模並列アクセスサポートのために

・I/OサイズをSolarisの最大値（Solaris7では1MB）まで拡張

・最大16ストライプのサポート

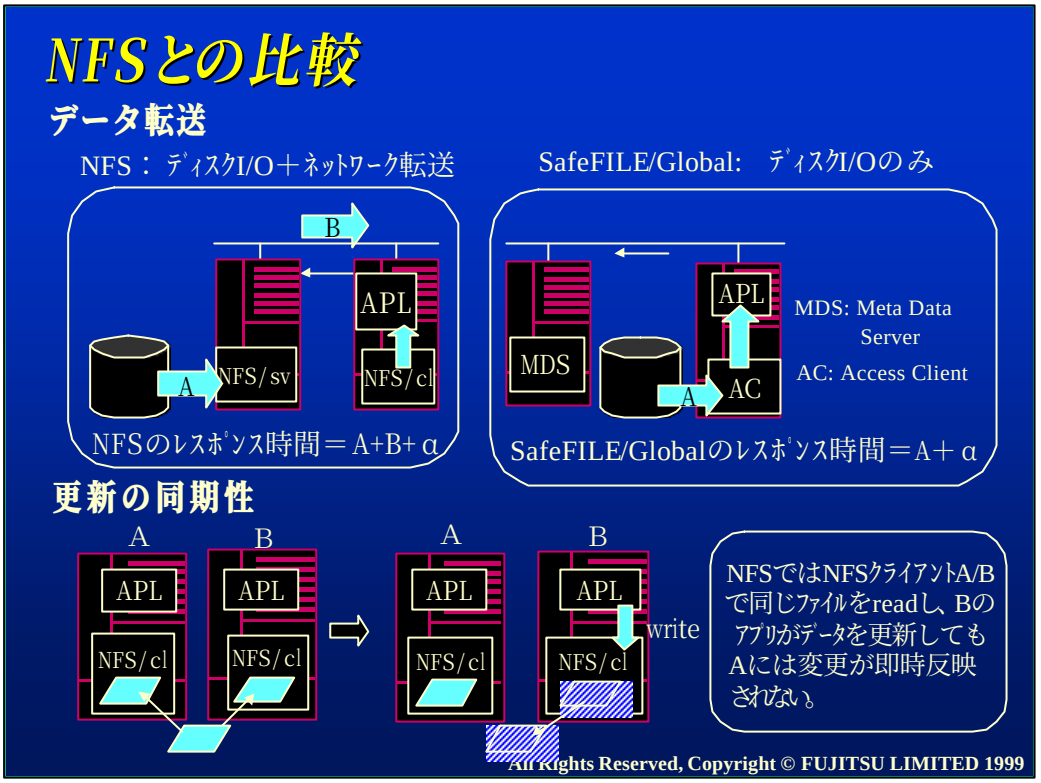
・現在1TBのファイルサイズ、ファイルシステムサイズをそれぞれ16TB、512TBまでの拡張を計画している。



SafeFILE/Globalは複数の計算サーバから、一つのファイルの同時共用を
可能とする共用ファイルシステムである。SafeFILEを開発のベースとし、
SolarisのUFS互換のAPIによる複数サーバの同時共用で、データの一貫性
はファイルシステムが保証する。各サーバからのデータアクセスはSAN環境
に適合しており、LANを経由せずFC経由で直接にDiskをアクセスする。
ファイルシステムを共用している一つのサーバがダウンしても、他のサーバによるファイルアクセスには影響を与えない高可用性を備えている。

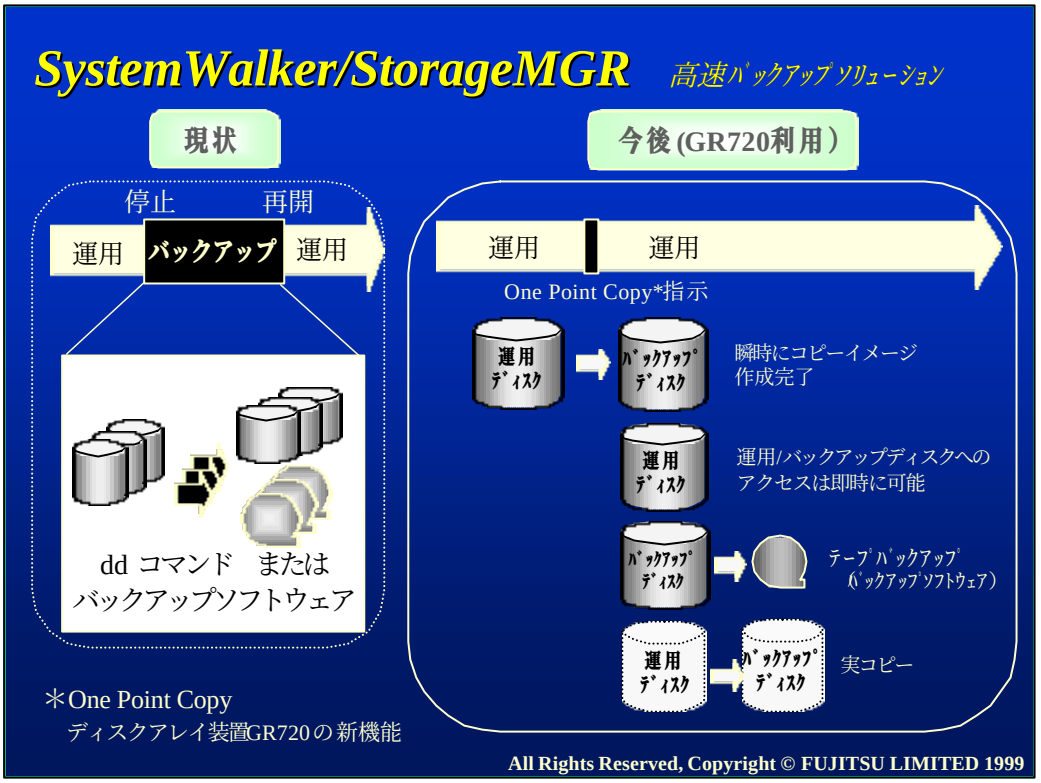
SafeFILEとSafeFILE/Globalでは、データ部の構造は共通であり、SafeFILEからの移行はメタデータを変換するだけであり、そのためのコマンドを提供し、容易な移行を実現する計画である。

SafeFILE/Globalは論理的には最大32のサーバによる共用を可能としているが、実用上の限界はシステム構成と使い方に大いに依存する。これについては今後実績と経験を積んで情報を提供していきたい。



SafeFILE/GlobalとNFSとの相違を図解している。データ転送については、NFSではNFSサーバによるDiskアクセス時間とサーバからクライアントへのネットワーク転送時間を必要とする。一方、SafeFILE/GlobalではDiskへのI/O時間を必要とするだけである。しかし、メタデータは一つのサーバが管理
する方式を採用しており、排他制御のためのメタデータアクセスのネットワー
ク上のトラフィックが実際に共用できるサーバ数を決定するが、これは各サーバ上のアプリによる共用ファイルの使い方に依存する。

他の相違点は更新の同期性に関するものである。SafeFILE/Globalでは、一つのサーバによるデータ更新はデータを共有している他サーバに即時に
反映される。



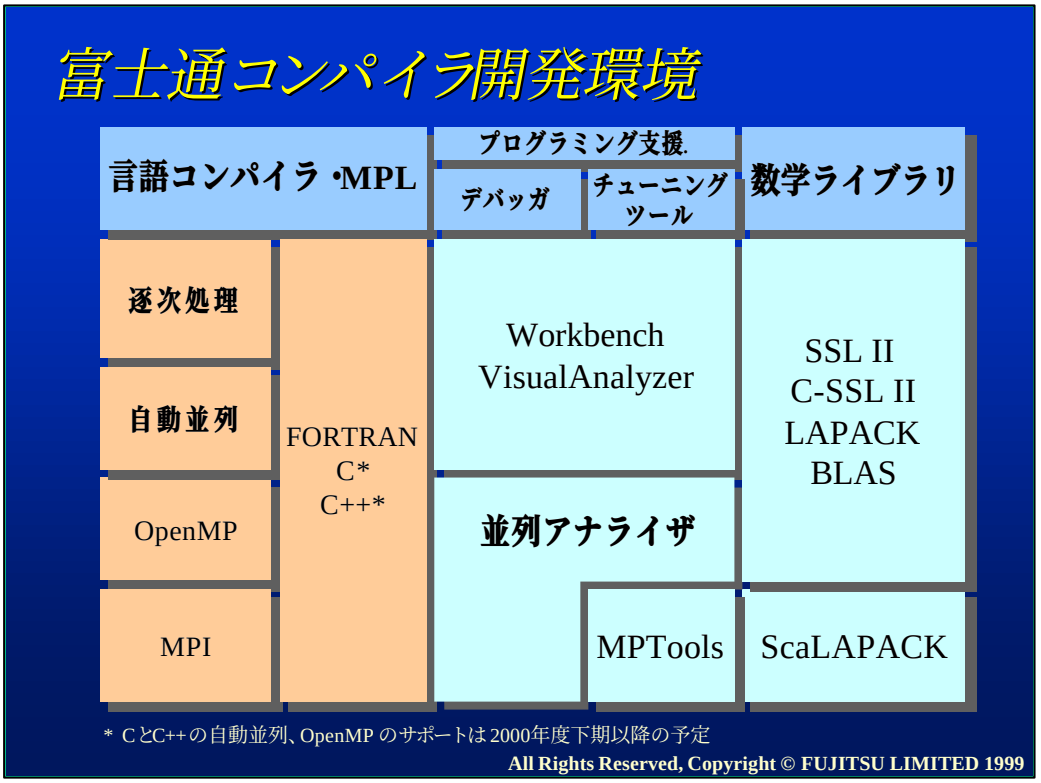
次にディスクアレイ装置GR720の新機能であるOPC(One Point Copy)機能と、OPCをサポートするSystemWalker/StorageMGRによる高速バックアップソリューションを説明する。バックアップの手順は以下のとおりである。なお、バックアップ単位は論理ボリュームまたはディスクパーティションである。

a) StorageMGRにより、あらかじめGR720に対し、運用ディスクとそれに対応
するバックアップ用ディスクを定義。

b) バックアップの前に、OSでキャッシュしているデータの内容を運用ディスクに反映するためにunmountを行い、StorageMGRがOPCコマンドを発行。

c) 運用を継続する場合は再mountを行う。

以上の手順によりOPCが発行された時点で、対象の論理ボリューム、あるいはディスクパーティションのバックアップが保証される。実際にはOPCが
発行された時点からGR720がそのファームウェアにより、運用ディスクの内容をあらかじめ定義されたバックアップディスクに空き時間を利用して順次コピーを行う。運用アプリケーションから運用ディスクに対してデータ更新の要求があれば、GR720はOPC発行時点の該当データをバックアップ用ディスクにコピーしてから、運用ディスクのデータ更新を行う。これにより、OPC発
行後一定時間経過後に実際のバックアップが完成するが、論理的にはOPC
発行時点でバックアップは完了している。



SMPにおけるプログラム開発支援環境について、コンパイラおよびツール、
数学ライブラリの位置付けを図示した。

OpenMPはSMP用並列言語として業界標準となる可能性が最も高い。

当初は自動並列とOpenMPはFortranのみのサポートであるが、2000年度下期以降をターゲットとして、CおよびC++についてもサポート予定である。

基本言語処理パッケージ

- Fortranコンパイラ
  - Fortran95/Fortran90およびFORTRAN77言語仕様
  - スーパーコンピュータ、メインフレーム、UNIXサーバ、PCまで一貫した言語仕様を実現
  - スーパーコンピュータで実現したベクトル化技術を応用した自動並列化機能
- C コンパイラ
  - ANSI及びK&R仕様サポート
- C++コンパイラ
  - ISO規格(ISO/ISE 14882:1998), AT&T互換モード
- 開発支援ツール
  - Workbench :デバッガ, サンプラ
  - Visual Analyzer :ソースアナライザ, ノードカバレッジ
  - SSL II :数値計算ライブラリ(Multi-thread unsafe)

All Rights Reserved, Copyright © FUJITSU LIMITED 1999

前ページのコンパイラ、ツール類のうち、逐次処理コンパイラと開発支援ツール（Workbench, VisualAnalyzer)、およびSSL IIを「基本言語処理パッケ
ージ」として提供する。なお、Fortranの自動並列化機能は本パッケージに含
まれている。

並列言語処理パッケージ

- コンパイラ（OpenMP Fortran）
  - OpenMP Fortran規格 1.0仕様に完全準拠
  - 豊富な機能
    - DOループ並列，タスク並列，リダクション演算，同期/排他制御/逐次化，データスコープの指定，条件的翻訳行など
- ツール（並列アナライザ）
  - GUI統合開発環境（マネージャ）
    - 並列プログラム（OpenMP Fortran, MPI）の翻訳，編集，実行支援
    - 各種ツール（デバッガ，サンプラ，MPTools など）の起動
  - デバッグ機能（デバッガ）
    - OpenMP Fortran言語, MPIプログラムのデバッグ
  - 性能チューニング機能（サンプラ）
    - 手続き単位および1行ごとコスト解析
- 汎用数値計算ライブラリ
  - Multi-thread safeな逐次ライブラリを提供
  - SSL II, LAPACK, BLAS, C-SSL II

All Rights Reserved, Copyright © FUJITSU LIMITED 1999

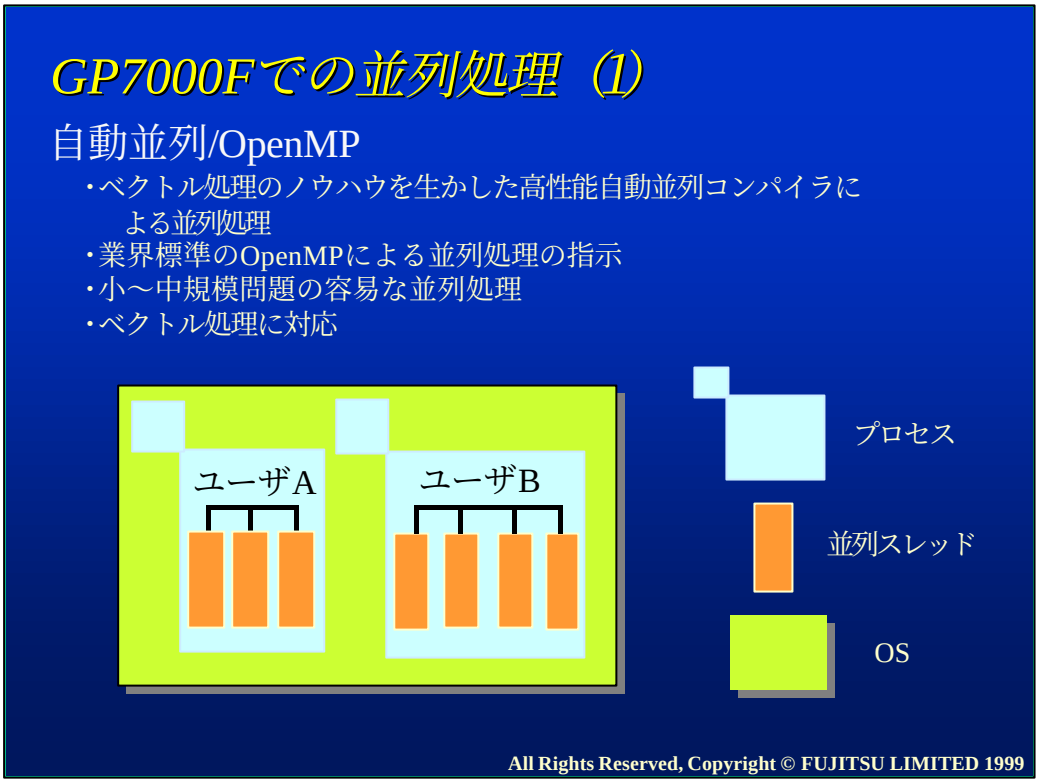
OpenMP Fortran、並列アナライザ、および MT safe(Multi-thread safe)な逐次数値計算ライブラリを「並列言語処理パッケージ」として提供する。

並列分散処理パッケージ

- メッセージパッシングライブラリ（MPI）
  - 最新のMPI規格（MPI2.0）に準拠
  - SMPノード内（共有メモリ通信），ノード間（AP-Net, TCP/IP通信）双方に対応した並列プロセス間通信
  - サポート機能
    - 1対1通信，集団通信，派生データ型，コミュニケータ，プロセス・トポロジー，動的プロセス生成/管理，片側通信，並列入出力など
- ツール（MPTools）
  - MPIのプログラム解析支援
    - 各プロセスの情報，ジョブの収集した解析情報の集計表示など
    - GUIによる容易な操作性
- 並列線形計算ライブラリ（ScaLAPACK）
  - 逐次数学ライブラリLAPACKの分散メモリ対応版
  - 提供ライブラリ
    - 線形方程式，最小二乗問題，固有値問題，特異値分解など

All Rights Reserved, Copyright © FUJITSU LIMITED 1999

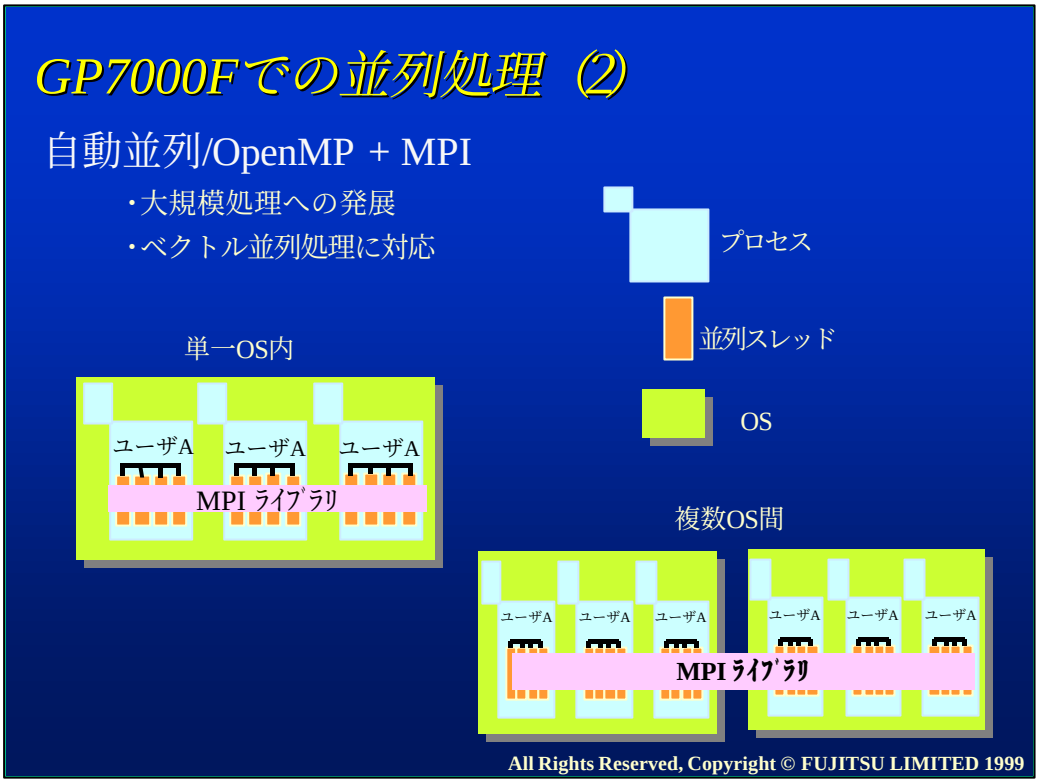
MPI、MPTools、および線形計算ライブラリLAPACKの分散メモリ対応版であるScaLAPACKを「並列分散処理パッケージ」として提供する。



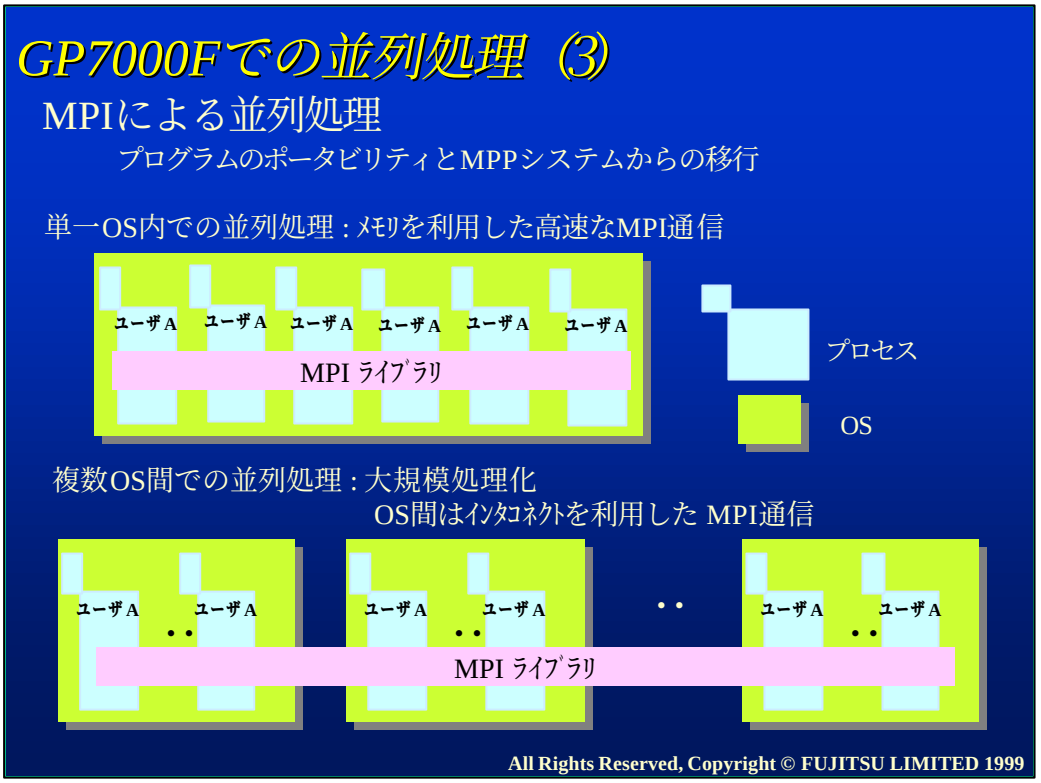
ここでSMPでの並列処理について考察してみたい。

最も広範に使用される並列化手法はコンパイラによる自動並列、またはOpenMPによる並列処理の指示である。この場合コンパイラは、スレッドと呼ぶ並列処理単位を生成するオブジェクトコードを生成する。プログラム実行時には、一つのプロセス配下で複数のスレッドが同時（並列）実行される。一つのスレッドが通常1CPU下で実行される。並列度（スレッド数=動作CPU数）は実行時の環境変数として定義される。

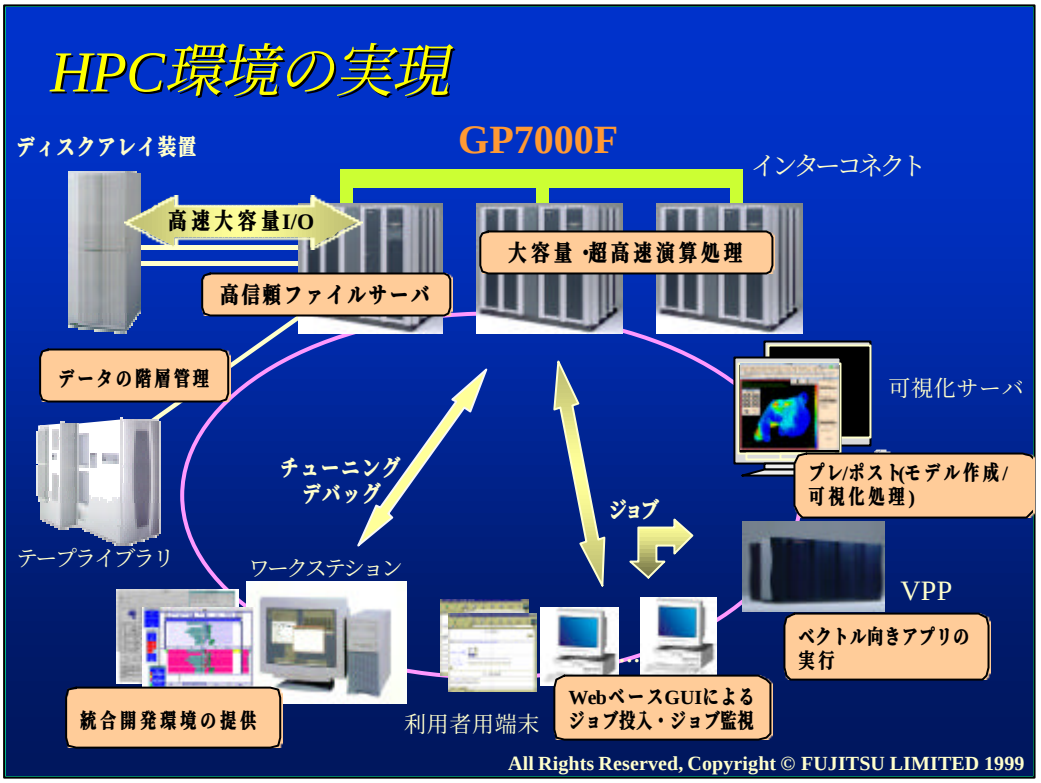
このコンパイラによる自動並列化/OpenMPは小～中規模問題を容易に並列処理したい場合に用いられる手法であり、従来のベクトル機のコンパイラを利用した自動ベクトル化に相当するものである。



問題規模が大きくなった場合には、自動並列/OpenMPとMPIを組み合わせたプログラミングモデルが考えられる。これはコンパイラにより生成されたスレッド並列を行うプロセス間で、さらにMPI通信を用いて並列処理を行う手
法であり、単一OS(一台のSMPマシン)内でも、あるいは問題規模がさらに大きい場合には複数のOS(複数台のSMPマシン)をまたがってMPI通信を利用できる。本手法はベクトル並列機(VPP)におけるベクトル並列処理に相
当するものである。

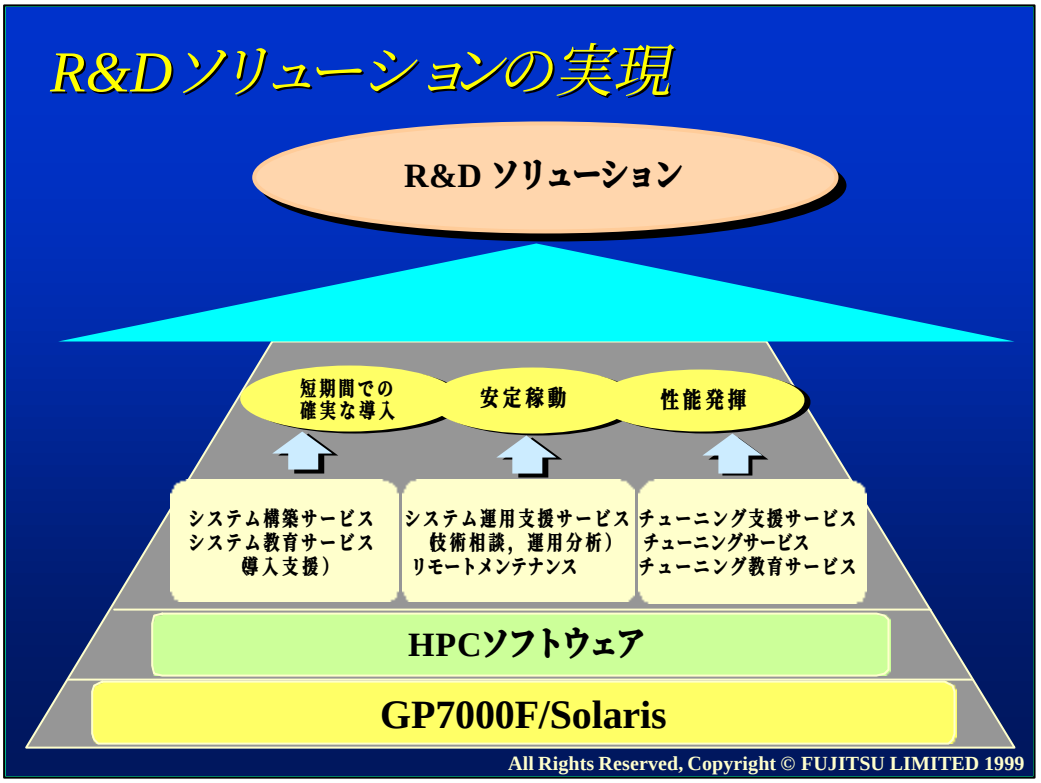


プログラムのポータビリティおよびAP3000システムのようなMPPシステムか
らの移行形態として、MPI通信のみによる並列処理が考えられる。単一OS(一台のSMP機）内のMPI通信はメモリ上の高速データ転送として実現され、複数OS間をまたがるMPI通信は、実際には複数台のGP7000Fを、例えばAP-Netのような高速インタコネクトで接続した分散並列環境として利用する
ことができる。



以上、GP7000FのHPC向けソフトウェアについて説明してきた。これらのソフトウェアはいずれもGP7000S(従来のSファミリ) でも動作可能である。

今日のHPC環境は本図に示したように、ネットワークを介してすべてのサーバとワークステーションが有機的に接続された形態で実現されている。今後とも各サーバのより一層の性能向上（大容量化、高速化）と信頼性の向
上、およびエンドユーザからのHPC環境の使い勝手の向上を図っていく。



他社に類を見ない大規模SMP機であるGP7000 Fのために、PROPOSEサービスとして導入支援、運用支援、さらにチューニングに関するサービスメニューを揃えている。

本機が科学技術計算の分野でも大いに活用されることを願って本報告を終了する。