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1回会合資料

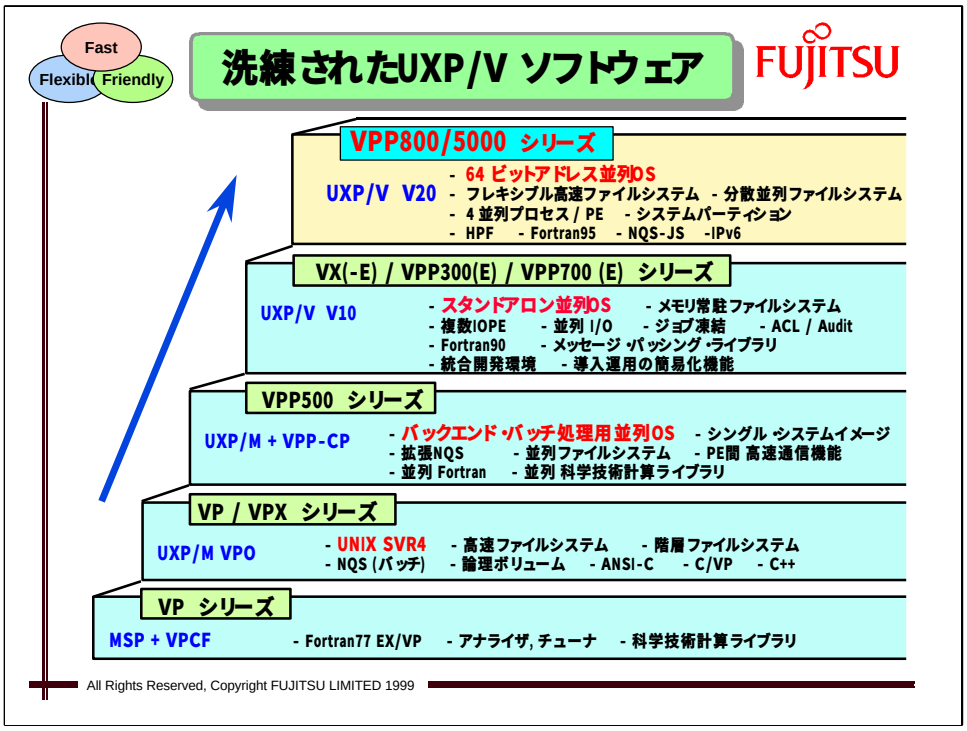
1999年8月20日

VPP800/VPP5000の OS機能について(報告)

富士通株式会社
ソフトウェア事業本部HPCソフトウェア統括部
木村　治夫
MAE03261@nifty.ne.jp

All Rights Reserved, Copyright FUJITSU LIMITED 1999

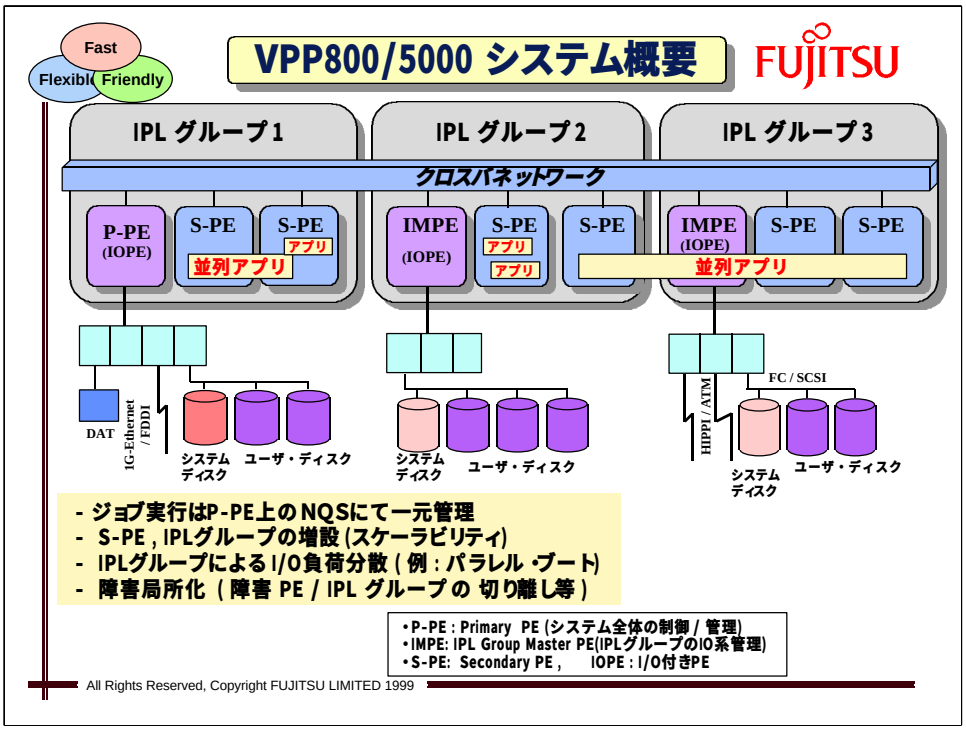
VPP800/5000シリーズのOSであるUXP/V V20について、報告致します。



UXP/V V20は、VPシリーズから始まって、VPXシリーズ、VPP500シリーズ、VX/VPP300/700シリーズでの豊富な経験と実績を生かして開発したOSです。

従来との互換性を維持しつつ、Fast/Flexible/FriendlyのトリプルFコンセプトに基づいて、大規模高速計算・高速入出力・ユーザインタフェースを強化しています。

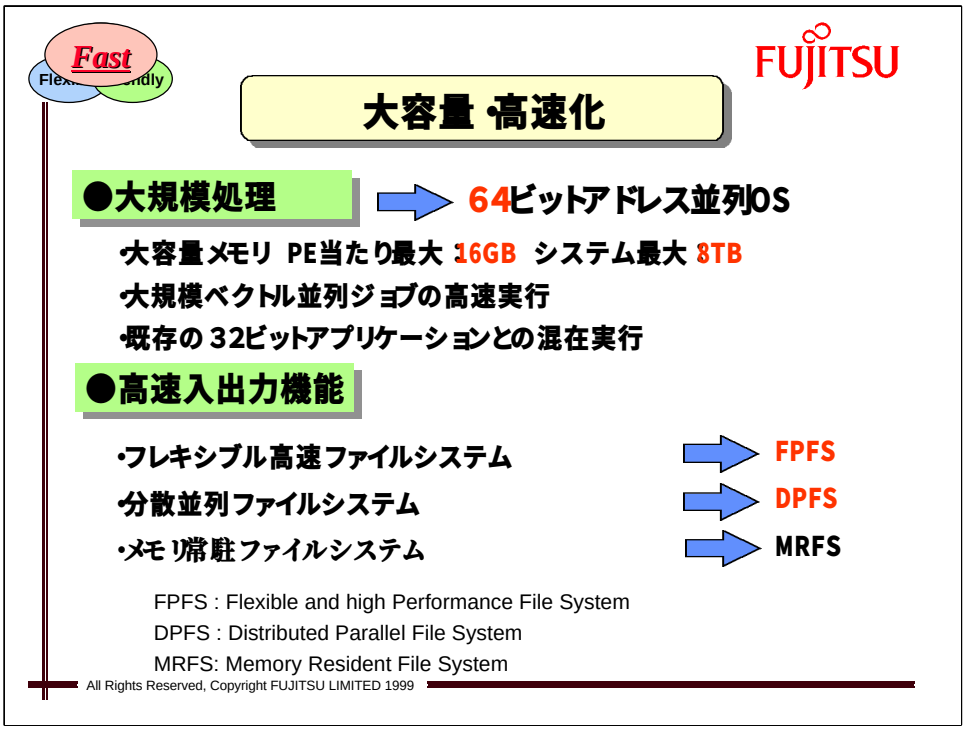
以降、UXP/V V20で新規にサポートした、新ファイルシステムのサポートや運用機能強化などの機能を中心に報告していきます。



VPP800/5000の大規模システムでのシステム概要図です。VPP300/700システムから基本的には変更ありませんが、簡単に説明します。

IO 負荷の分散やIPL時間の短縮のためにIPLグループがあります。S-PEやIPLグループを増設することにより、処理能力をスケーラブルに拡張することができます。また、障害時には障害PEやIPLグループを切り離すことで障害を局所化できます。

また、システム全体を制御するのがP-PEで、1システムビューを実現しています。



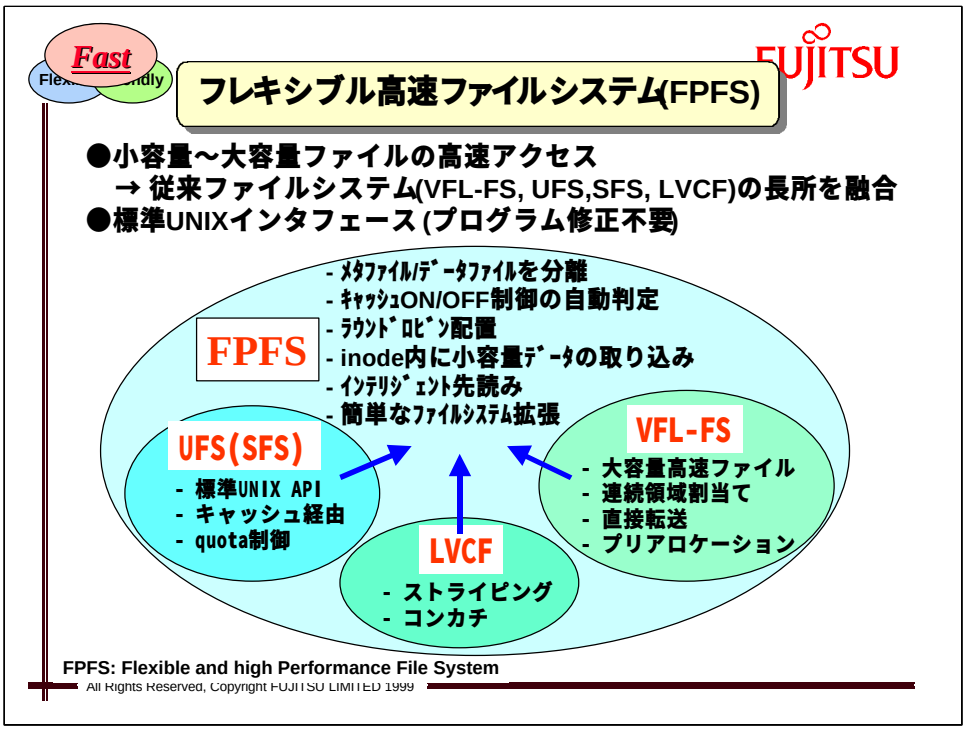
トリプルFコンセプト(Fast/Flexible/Friendly)のFastである大容量・高速化機能として、64ビットアドレスサポートと高速ファイルシステムサポートがあります。

64ビットアドレスをサポートすることにより、PE当たり最大16GBのメモリを有効に利用した大規模ベクトル並列ジョブの高速実行が可能となります。既存の32ビットアプリケーションとの混在実行も可能です。

V20でのファイルシステムの推奨版は、フレキシブル高速ファイルシステムFPFS、分散並列ファイルシステムDPFS、メモリ常駐ファイルシステムMRFSです。

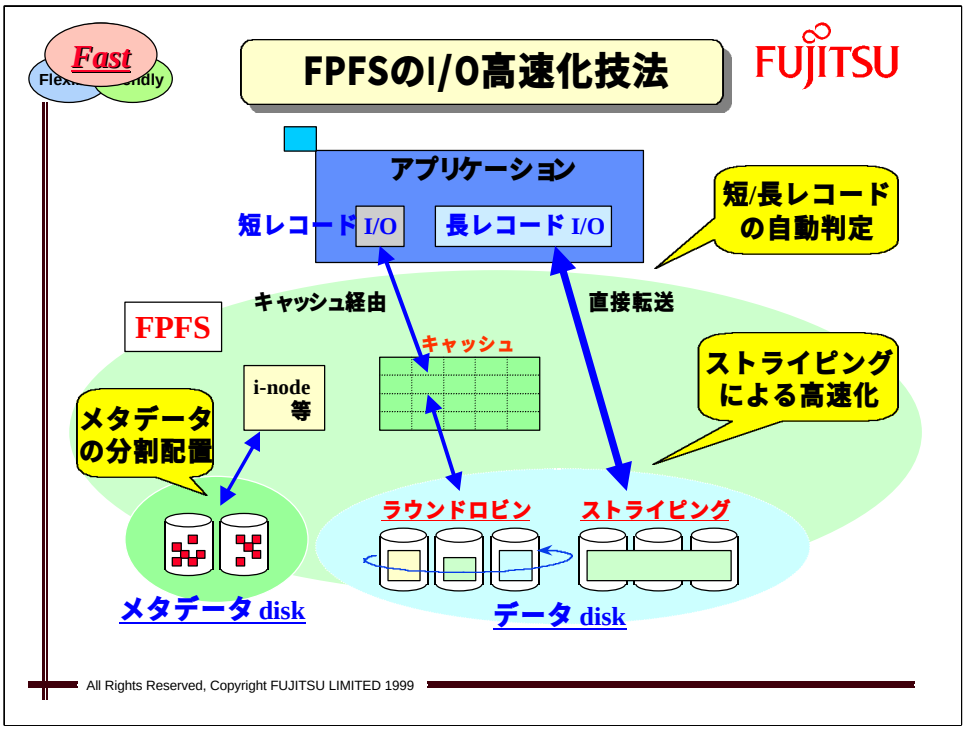
標準的には、FPFSをご利用下さい。DPFSはIOPEが複数存在するシステムで、1つのプログラムからのIOを高速化したい場合に有効です。MRFSは一時的なファイルについて、メモリ上に展開して高速化を図るものです。MRFSは、VX/VPP300/700でもサポートされていますが、メモリが拡張された分V20では有効性が高まります。

以降、V20で新規にサポートされたFPFSとDPFSを中心にその高速化の仕組みなどについて簡単に紹介していきます。



FPFSはVFL-FS、UFS、LVCFなど従来のファイルシステムの長所を融合したファイルシステムで、小容量ファイルから大容量ファイル、短レコードIOから長レコードIOまでを高速化するものです。

特別なインタフェースはなく、プログラムの修正は不要です。



FPFSの高速化技法について説明します。

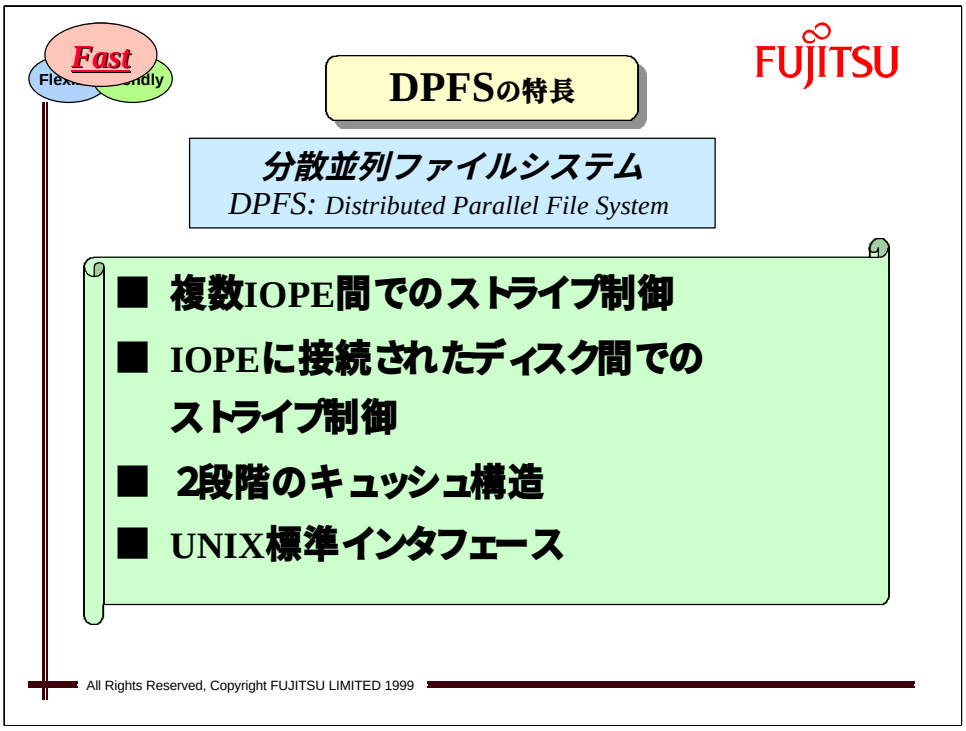
一般に短レコード長のIOはキャッシュを利用してまとまった単位でDISKとIOしたほうが、逆に長レコード長のIOはキャッシュを使用せず直接転送したほうが良いIO性能が得られます。

従来のUFSはキャッシュを利用、VFL-FSはキャッシュを使用せず直接転送しています。短レコード長のIOが多い場合はUFSを、長レコード長のIOが多い場合にはVFL-FSを使用しますが、短レコード長と長レコード長のIOが混在した場合には性能がでないことになります。

FPFSではレコード長によりキャッシュを使用するか直接転送するかを自動判定することにより、従来のUFSとVFL-FSの利点を融合しています。

また、ファイルを複数のDISKに順番に割り当てるラウンドロビン機能によりスループットの向上を図っています。さらに、1つのファイルを複数のDISKにストライプするストライピング機能によりLVCFと同じように1つのファイルに対するIOを並列化することで、アプリからみたファイルアクセス性能を向上できます。

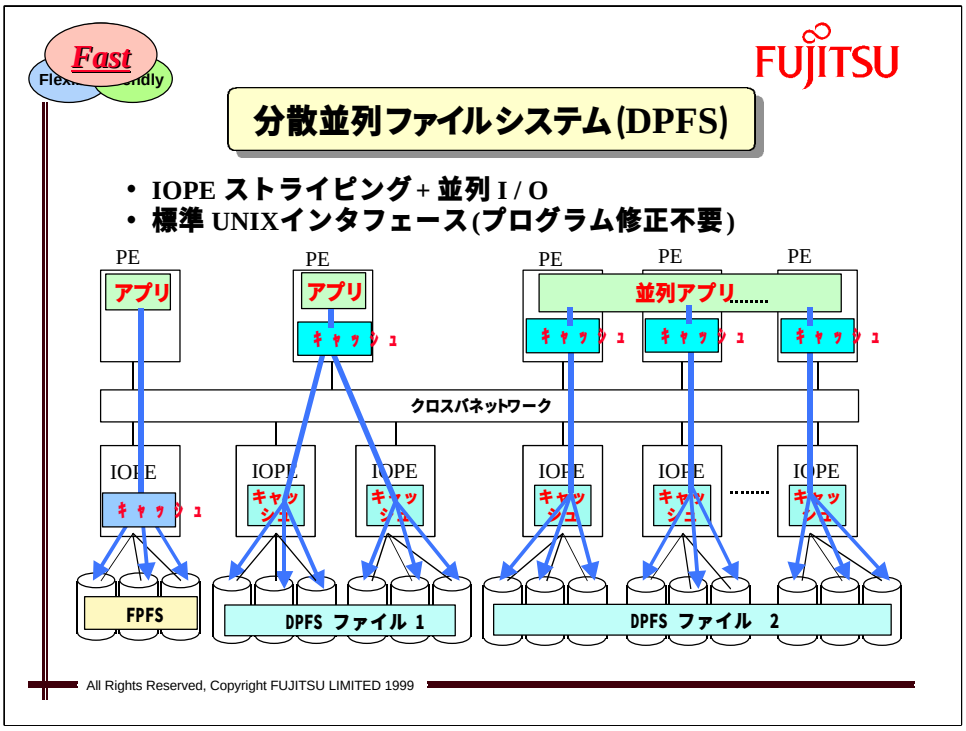
FPFSではファイルのIノードやディレクトリなどの管理情報であるメタデータと実際のデータとを分割して配置します。この分割配置によりfsck処理の高速化も実現しています。



DPFSはディスク間でのストライプに加えて複数IOPE間でもストライプすることにより1つのプログラムからのIOの高性能化を目指したもので、複数IOPEシステムで特に効果を発揮するファイルシステムです。

また、拡張されたPEメモリを有効に利用するためにキャッシュを有効利用して高性能化も図っています。

FPFSと同じように特別なプログラミングは不要であり、プログラムの修正は不要です。

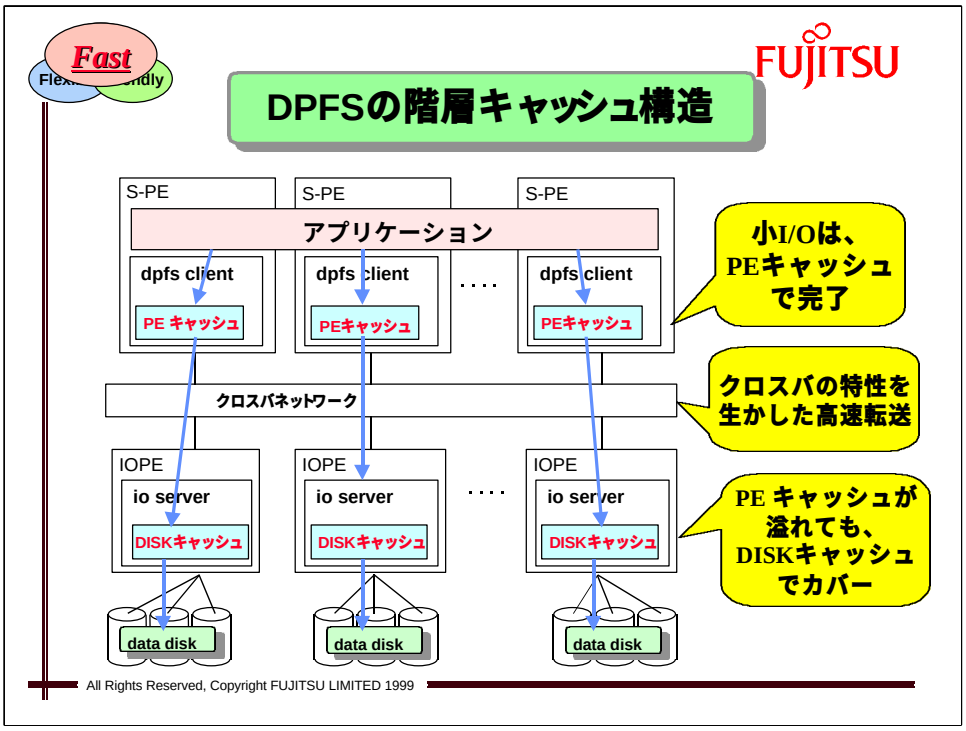


DPFSの特徴の1つであるIOPE間へのストライピング機能について説明します。

左の図はFPFSでDISK のストライピングをした例です。

真ん中の図は、IOを発行するPEは1つだがDPFSによりそのIOをIOPE間にストライピングしている例です。IOPEからDISKへのストライピングにより2段階のストライピングにより、RAID　DISKの場合に、3段階のストライピングにより、高速化が図られています。

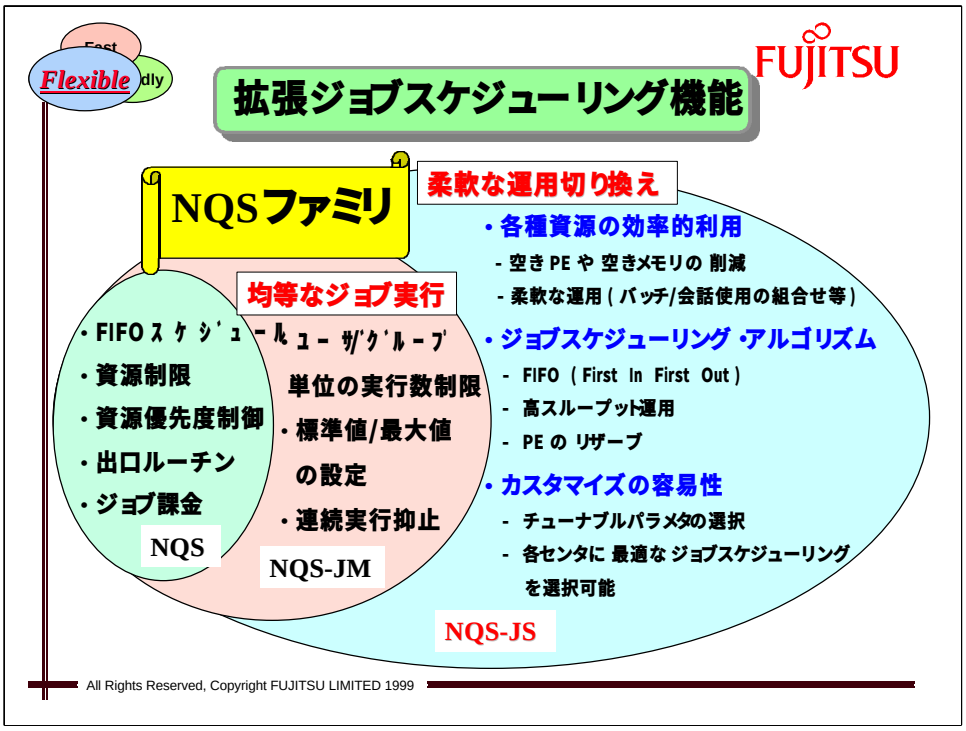
右の図はIOを発行するPEが複数あり、個々のPEからIOEPに対して並列にIO て発行される例です。



DPFSでは、キャッシュ性能を高速化するために、S-PEとIOPEにキャッシュを階層的に配置して2レベルのキャッシュ構成としています。

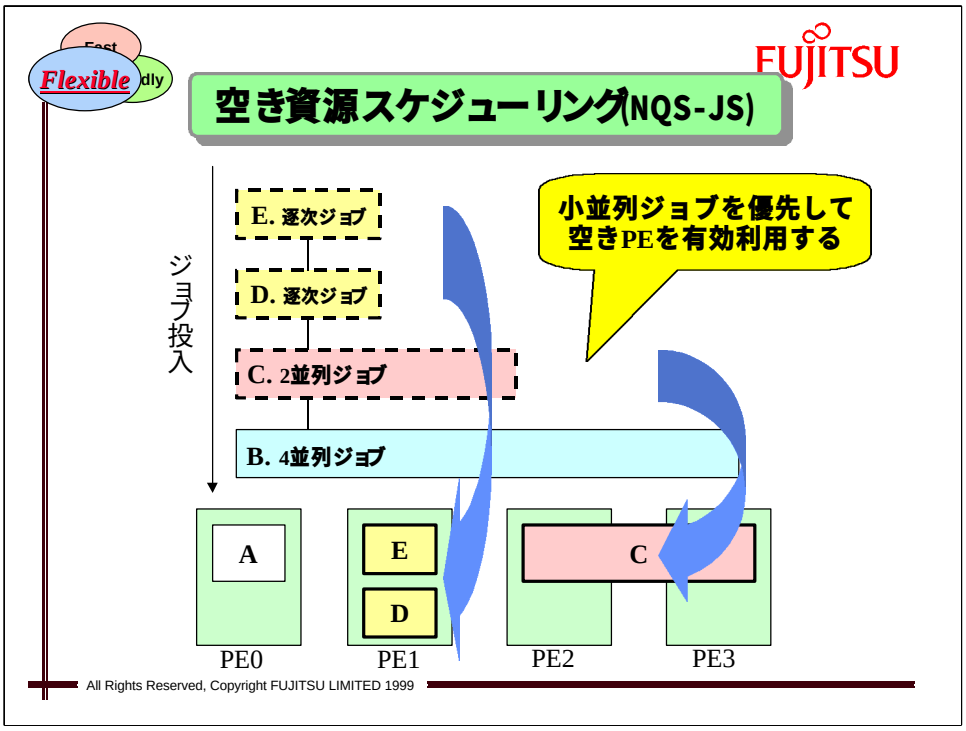
まず、S-PE上のPEキャッシュを使用します。PEキャッシュが溢れた場合にはIOPE上のDISKキャッシュを使用します。S-PEとIOPE間はクロスバで高速転送します。

このように、VPP5000で拡張されたメモリを有効に利用して高速化を図っています。



Flexibleな機能として、まずジョブスケジューリング機能の拡張について説明します。

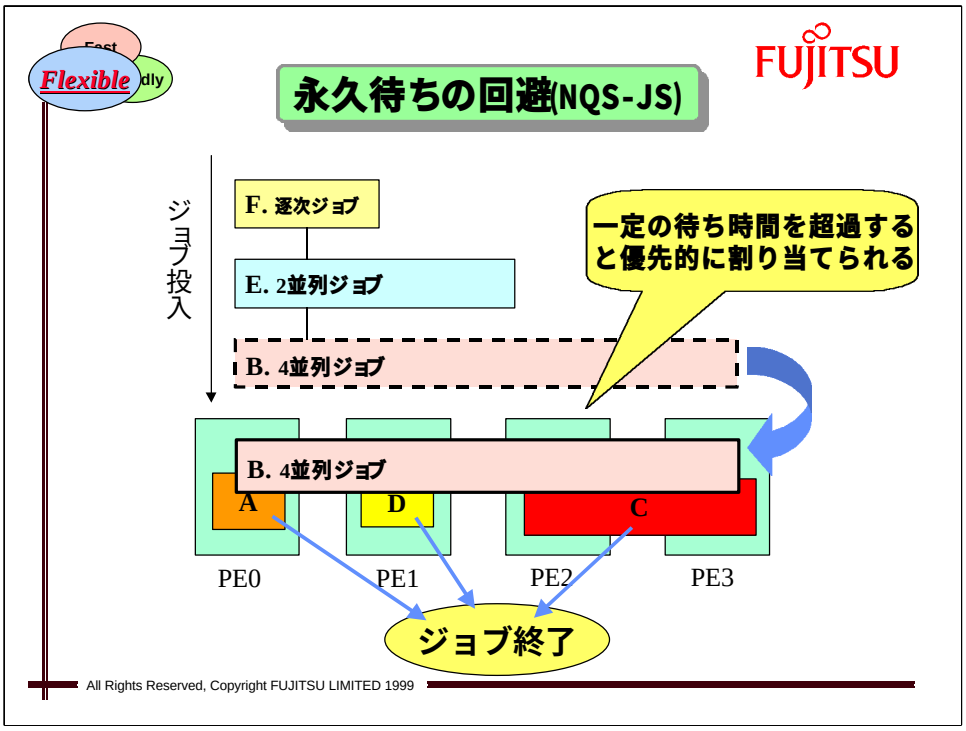
従来からのNQS、NQS-JMに加えて、V20ではスループット向上やカストマイズを容易にする機能としてNQS-JSをサポートしています。



NQS-JSにより、空き資源を有効利用してシステム全体のスループットを向上することができます。

例えば、この図のようにジョブAが実行中でPE1からPE3の3PEが空いている場合、次に実行されるジョブBが4並列のため、Aが終了するまで3PEは遊んでしまいます。これはジョブ実行のFIFOを保証しているからです。

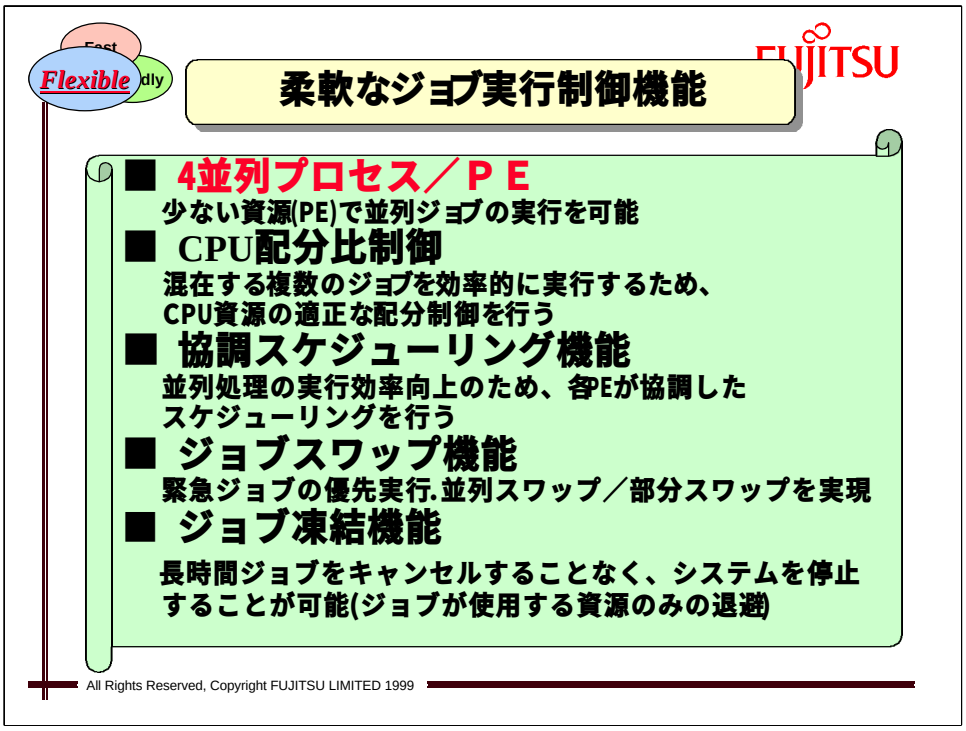
NQS-JSではこのような場合に、ジョブBを追い越して並列度の低いジョブCやDやEを実行させることにより、PEの遊びをなくすことができます。つまり、FIFO保証よりも資源の使用効率向上を狙ったスケジューリングをすることができます。



NQS-JSではシステム全体のスループットを向上させるため、並列度の高いジョブを追い越して並列度の低いジョブ実行させます。その結果、並列度の高いジョブがいつまでたっても実行できないことが起こる可能性があります。

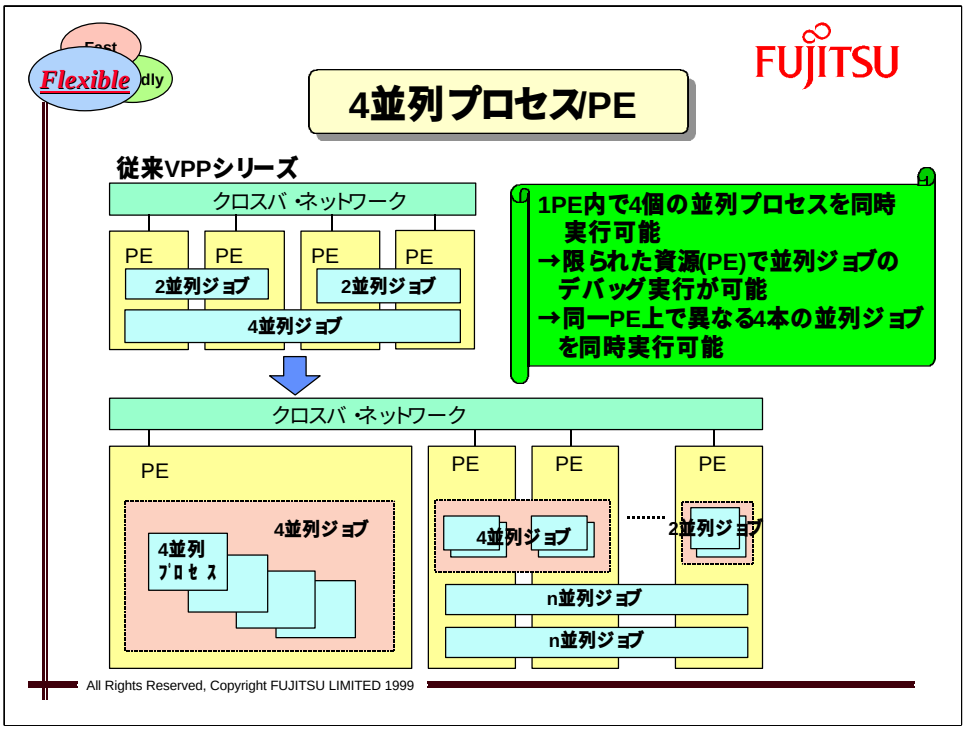
このような問題を解決するために考えられたのが、一定の待ち時間を超過すると優先的にそのジョブを実行させる機能です。

例えば、この図でジョブBが一定時間以上待った場合にはジョブEやFを実行させずに、4PEが空き状態となってジョブBが実行できるのを待ちます。



今まで述べたNQS関連の機能はジョブを実行するまでの機能ですが、実行されたジョブを制御するジョブ実行制御機能として、OHPで示すようないろいろな機能が用意されています。

4並列プロセス/PE機能については、V20で新規にサポートされました。



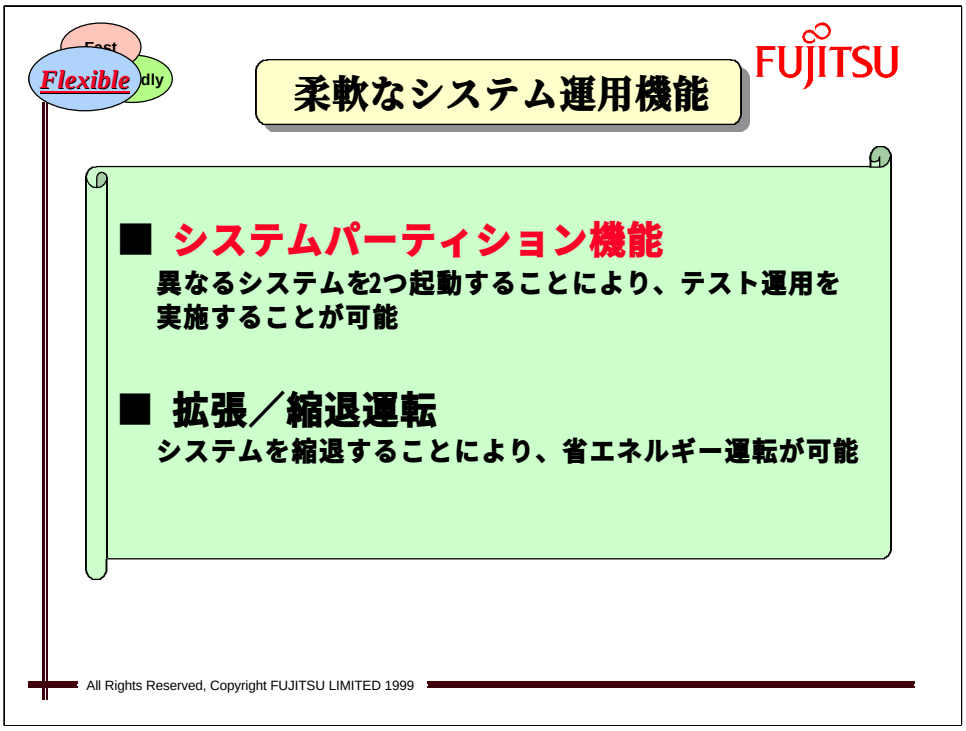
V20で新規にサポートされた4並列プロセス/PE機能について説明します

。

従来のVPPシリーズでは1つのPE上では並列ジョブは2多重までしか実行できませんでした。

UXP/V V20では、1つのPE上で4多重までの並列ジョブを実行できるようになりました。

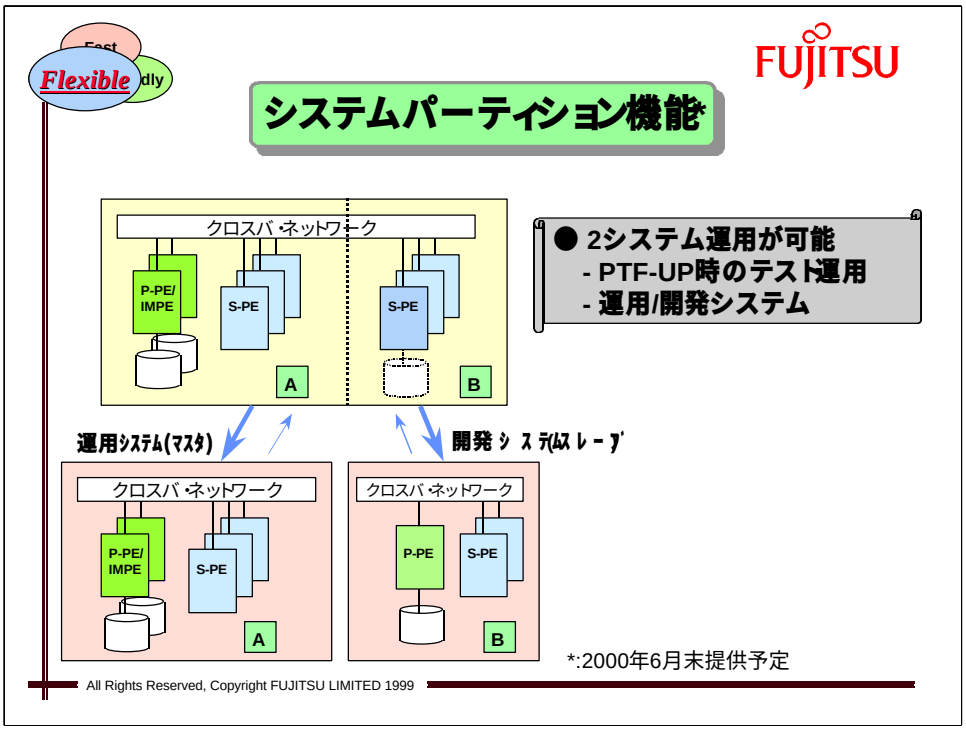
更に、折り返しをサポートすることにより、1つのPE上で最大4並列のジョブを実行することが可能となり、限られたPE上での並列ジョブのデバグが可能となりました。



柔軟なシステム運用機能としては、システムパーティション機能と拡張/縮退運転機能とがあります。

システムパーティション機能については V20で新しくサポートされました。

拡張/縮退運転機能については、従来からサポートされています。



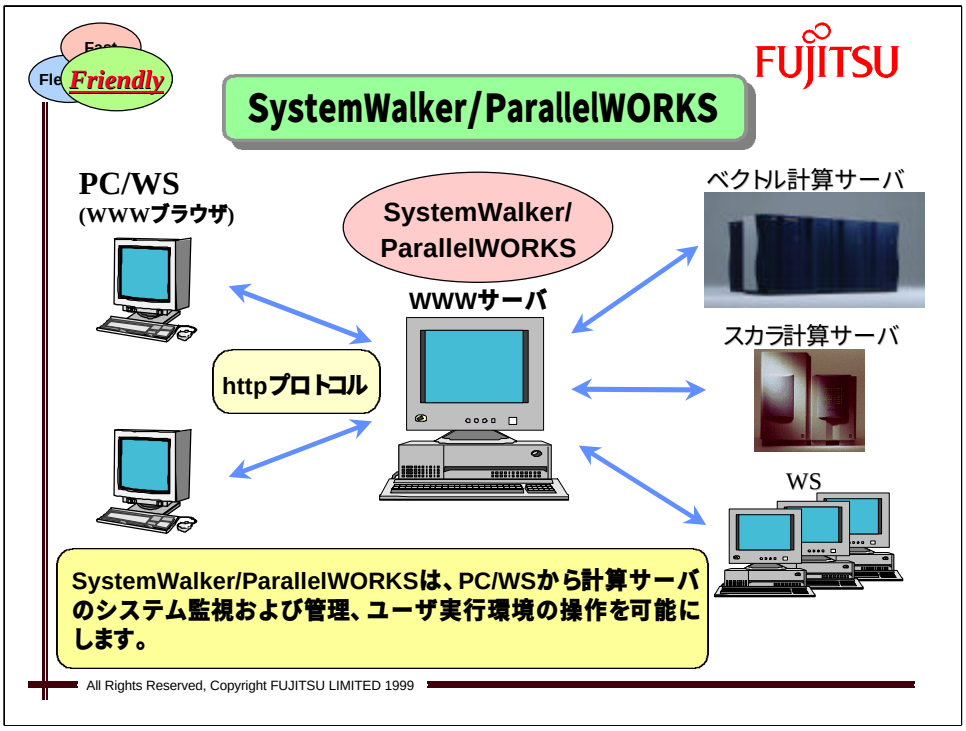
V20で新しくサポートされたシステムパーティション機能について、説明します。

システムパーティション機能はこの図のように、通常は1システムで動作しているシステムを2つに分けて、例えば運用システムと開発システムとして運用できる機能です。

この場合、開発システムは起動・停止が必要ですが、運用システムの起動・停止はなしに運用を切り替えることができます。

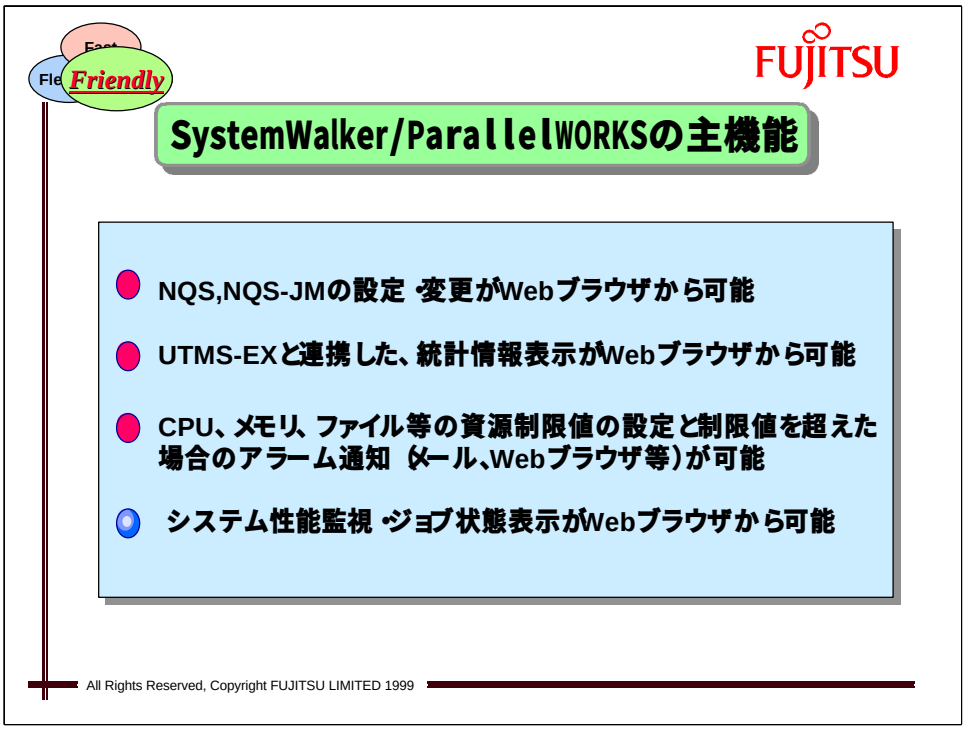
このように、PTF-UP時のテストシステムでの確認など、運用を柔軟に行うことができます。

なお、この機能は2000年の6月に提供予定です。



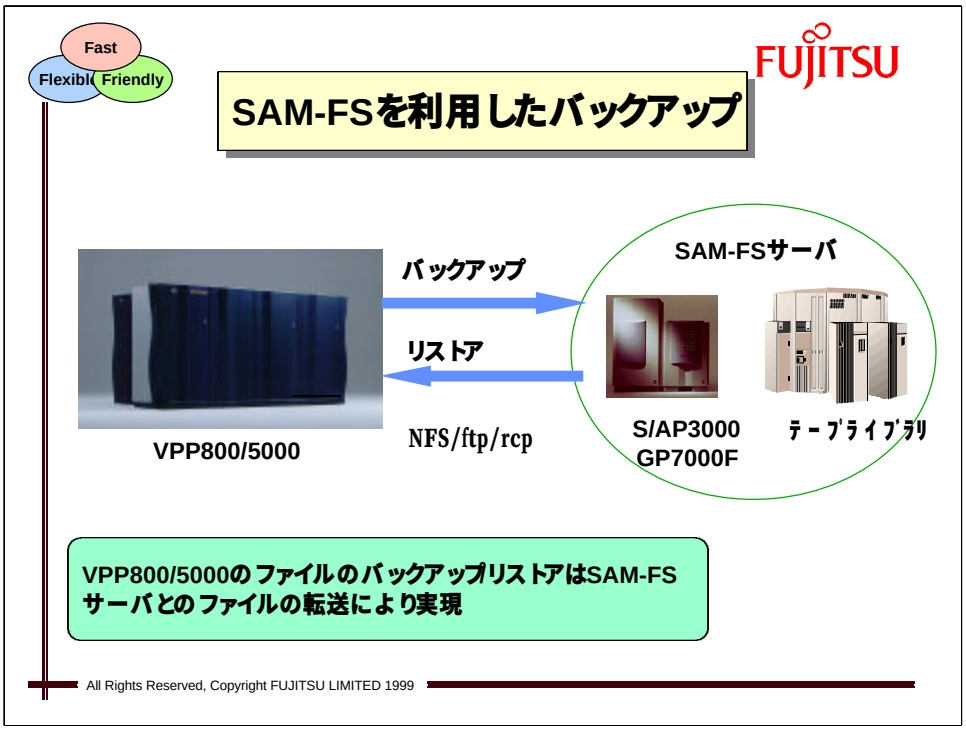
SystemWalker/parallelWORKSは、PC/WS上からWEBビューでVPP800/5000シリーズなどのベクトル計算サーバやスカラ計算サーバのシステム監視や管理を一括して行うことができる機能です。

また、WEBビューでの計算サーバへのジョブ操作やファイル操作などのエンドユーザ向け機能もあります。



SystemWalker/parallelWORKSは、V10で提供依拠していたWW siteをベースに開発した機能です。

WW siteにたいして、赤丸で示した機能が追加されています。



SAM-FSはテープライブラリを使用した階層記憶管理(HSM)機能であり、Solaris、AP3000、GP7000F上で動作します。

VPP800/5000上でのファイルのバックアップ/リストアは、SAM-FSの動作しているファイルサーバに対して、NFSやftp/rcpなどネットワーク経由でアクセスすることにより、大容量のバックアップ環境を利用することができます。

以上でＯＳ機能の報告を終わります。