

JAXA統合スーパーコンピュータシステム —概要とシステムの中核「富士通のテクニカルコンピューティングサーバFX1」—

JAXA Supercomputer Systems with Fujitsu FX1 as Core Computer

あらまし

独立行政法人宇宙航空研究開発機構（JAXA : Japan Aerospace Exploration Agency）様は、富士通のテクニカルコンピューティングサーバFX1を中核とするJAXA統合スーパーコンピュータシステム（JSS）を、2009年4月の本運用に向けて導入する。

JAXA統合スーパーコンピュータシステムは、大規模（超）並列計算機システム、ストレージシステム、大規模共有メモリシステムおよび遠隔利用システムの四つの要素から構成され、卓越した並列アプリケーション実効性能、大容量・高速ストレージおよび遠隔利用といった特徴を有する。

本稿では、JAXA統合スーパーコンピュータシステムの概要とその特徴について説明する。

Abstract

In April 2009, the Japan Aerospace Exploration Agency (JAXA) will deploy an integrated supercomputer system called JAXA Supercomputer Systems (JSS) using the Fujitsu FX1 technical computing server as the core computer. JSS consists of a massively parallel supercomputing system, storage system, large-scale shared memory system, and remote access system. It features outstanding application-execution performance, large-capacity/high-speed storage, and remote access.

This paper outlines JSS and describes its features.



阿部孝之（あべ たかゆき）

テクニカルコンピューティング事業
本部計算科学ソリューション統括部
所属
現在、独立行政法人宇宙航空研究開
発機構様のJAXA統合スーパーコン
ピュータシステムの導入、およびシ
ステム運用業務のサポートに従事。



稲荷智英（いなり ともひで）

テクニカルコンピューティング事業
本部計算科学ソリューション統括部
所属
現在、独立行政法人宇宙航空研究開
発機構様のJAXA統合スーパーコン
ピュータシステム向けのプログラム
高速化支援業務に従事。



関 堅（せき けん）

次世代テクニカルコンピューティング
開発本部システム開発統括部 所属
現在、テクニカルコンピューティン
グサーバFX1の開発業務に従事。

ま え が き

独立行政法人宇宙航空研究開発機構（JAXA：Japan Aerospace Exploration Agency）様は、「空へ挑み、宇宙を拓く」というコーポレートメッセージのもと、人類の平和と幸福のために役立てるよう、宇宙・航空が持つ大きな可能性を追求し、様々な研究開発に挑んでいる。その代表的な活動として数値シミュレーション技術の研究開発に取り組んでおり、スーパーコンピュータの高速計算処理能力を利用し、数値流体力学（CFD：Computational Fluid Dynamics）に代表される数値シミュレーション技術の発展と普及を推進している^①

また、宇宙科学研究所（ISAS）・航空宇宙技術研究所（NAL）・宇宙開発事業団（NASDA）の3機関時代からの経緯で、調布事業所、角田事業所、相模原キャンパスの3箇所にそれぞれスパコンを保有してきたが、JAXA統合スーパーコンピュータシステム（JSS：JAXA Supercomputer Systems）は、スパコンによる数値シミュレーション技術を宇宙開発などのJAXA事業に本格的に活用することを企図して、宇宙3機関統合のシンボリックな位置づけで導入される^②

本稿では、この研究活動を支える計算基盤システムとして、富士通が新たに開発したハイエンドテクニカルコンピューティングサーバFX1を中核としたJSSの概要とその特徴について紹介する。

JSSの概要

JSSは、超高速計算エンジンとなる大規模（超）並列計算機システム、大容量・高速ストレージであるストレージシステム、大規模共有メモリシステム、遠隔からの利用性を向上させる遠隔利用システムの四つの要素から構成される。

大規模（超）並列計算機システムは、総計135 TFLOPSの計算性能と総計100 Tバイトのメモリ空間を有する最大級のスカラ型計算エンジンである。

このシステムは、利用者に様々な規模・形態の数値シミュレーション環境を提供するメインシステムと、特定プロジェクトへの柔軟な数値シミュレーション環境を提供するプロジェクトシステムから構成され、JSSの中核となる。

ストレージシステムは、1 Pバイトの磁気ディスクと、10 Pバイトのテープライブラリを有し、大量の数値シミュレーションデータ保存と、そのデータを高速にアクセスさせるためのシステムである。

大規模共有メモリシステムには、それぞれ1 Tバイトの共有メモリを有する、スカラSMP型計算機のAシステムと、ベクトル型計算機のVシステムがある。

Aシステムは、大規模共有メモリを必要とする数値シミュレーションやデータ処理、およびISVアプリケーションを実行するためのシステムである。Vシステムは、角田事業所および相模原キャンパスで実行されていたベクトル型計算機上のプログラム資産を継承するとともに、ベクトル機用のプログラムを実行するためのシステムである。

遠隔利用システムは、角田事業所、つくば事業所、相模原事業所などの各拠点から調布事業所のJSSを容易に利用するためのシステムである。

JSSのプログラム開発・実行環境は、Parallelnavi^(注1)によって実現される。プログラム開発言語としては、Fortran, C, C++が提供され、並列実行環境としては、データパラレル方式のXPFortran、メッセージパッシング方式のMPI（Message Passing Interface）が提供される。また、Parallelnaviは、CPUなどの資源管理とジョブスケジューリングを柔軟にカスタマイズ可能なインタフェースを提供しており、JAXA独自のジョブ制御機構のきめ細やかなノード割当て方式に対応可能であり、公平かつ効率的な数値シミュレーション環境を実現している。JSSの構成概念を図-1に示す。

次章以降では、大規模（超）並列計算機システムとストレージシステムを中心に紹介する。

大規模（超）並列計算機システム

本章では、大規模（超）並列計算機システムの構成と特徴について説明する。

● 大規模（超）並列計算機システムの構成

大規模（超）並列計算機システムは、演算ノードとして、3392ノードのFX1を有している。各ノード

(注1) FX1, SPARC Enterpriseシステムのハードウェア特性を最大限に引出すためのプログラム開発環境と高速実行環境を提供する富士通製のソフトウェア。

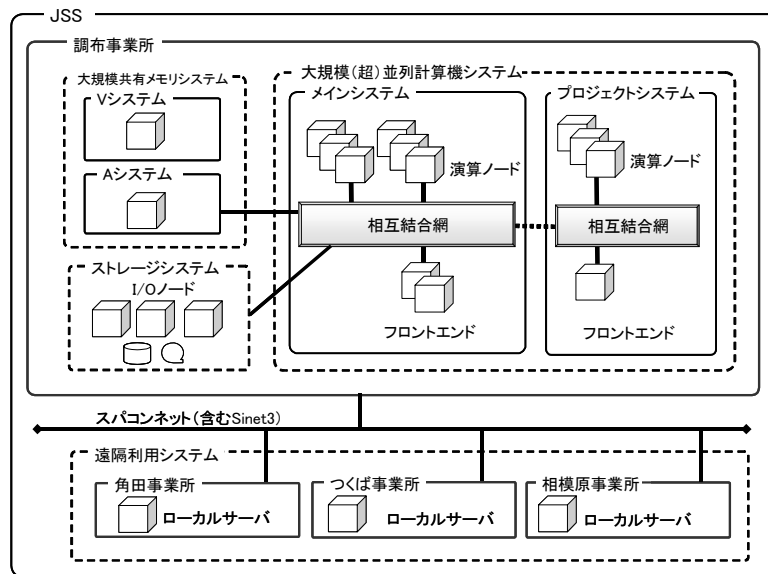


図-1 JSS構成概念
Fig.1-JSS configuration.

ドは、それぞれ1個のCPU、32 Gバイトまたは16 Gバイトのメモリから構成されている。大規模(超)並列計算機システムは、3008ノード(32 Gバイトメモリ搭載)のメインシステムと、384ノード(16 Gバイトメモリ搭載)のプロジェクトシステムから成る。また、それぞれのノードは、一つのFX1より構成されている。各システムの構成要素を表-1に示す。

各ノードは、往復それぞれ2 Gバイト/sの転送性能を有するFBB (Full Bisectional Bandwidth)^(注2)のファットツリー型相互結合網によって結合されており、ノード間的高速データ転送を実現している。

● 大規模(超)並列計算機システムの特徴

本節では、大規模(超)並列計算機システムの特徴である科学技術計算に特化したFX1、高速かつ高機能な相互結合網、卓越したアプリケーション実行性能について説明する。

【科学技術計算に特化したFX1】

FX1のCPUとして富士通が開発した高性能CPU SPARC64VIIプロセッサ⁽³⁾を採用した。SPARC64VIIは1個のCPUチップに4個のコアを搭載し、クロック2.5 GHz、プリフェッチ、アウトオブオーダー実行、4個の浮動小数点演算の同時実行などの機

(注2) Nノードのネットワークを任意に2等分したとき、その間の理論通信性能の合計(バイセクションバンド幅: 双方向)が、1ノードのバンド幅のN/2倍であることを指す。

表-1 JSS構成要素

システム名	ノード名	ノード数	CPU数	用途
大規模(超)並列計算機システム	メインシステム 演算ノード	3008	3008	ジョブ実行
	メインシステム フロントエンド	2	16	ログイン など
	プロジェクト システム 演算ノード	384	384	ジョブ実行
	プロジェクト システム フロントエンド	1	20	ログインや I/O処理
ストレージシステム	I/Oノード	3	96	I/O処理
大規模共有メモリ計算機システム	Aシステム	1	32	ISVなど
	Vシステム	3	48	ジョブ実行 など
遠隔利用システム	ローカルサーバ	3	24	ログイン など

能を有している。

さらに、マルチコア時代をリードする新たなアーキテクチャであるIntegrated Multicore Parallel Architecture⁽⁴⁾を採用した。Integrated Multicore Parallel Architectureは、マルチコアCPUの性能を引き出す「自動並列コンパイラ」とコア間の同期処理を高速化する「高速コア間バリア機構」および、False Sharing回避に効果的な「コア間共有L2キャッシュ」の性能向上技術の協調動作により、チップ内の複数コア(4コア)を一つの高性能演算ユニットとして利用可能とする技術である。また、高性能コンパイラ

(Parallelnavi), および高いメモリバンド幅を実現する専用チップセット (JSC : Jupiter System Controller) の連携により, 従来スカラ計算機で効率の低かったベクトル向けコードの高速化と, ループ並列化の適用範囲拡大を実現する。さらにノード内はスレッド並列により1プロセスで複数コアを有効利用することができる。FX1の仕様概要を表-2に示す。

CPUは, CPU内にハードバリア (高速スレッド同期) 機構を入れるとともにL2キャッシュをコア間で共有化し, 同期オーバーヘッドやキャッシュ Falseシェアリング (複数スレッド間のキャッシュ競合) を回避することで, スレッド並列^(注3)性能の向上を目指している。また, 浮動少数点演算ユニットのレイテンシ^(注3)と並列動作性を向上することにより, Linpack^(注4)の中核ルーチンであるDGEMM^(注4)で92%の実効効率を達成した。SPARC64VIIとJSCの関係を図-2に示す。

一方, チップセットでは, JSC-CPU間のバスを1.25 GHzで高速動作させるとともに, DDR2 (Double Data Rate 2) メモリのインタフェースを4組持つJSCを二つ使うことで, 理論ピーク40 Gバイト/sのメモリスループットを確保し, 実測におい

表-2 FX1の仕様概要

CPU	プロセッサ	SPARC64VII
	L2キャッシュ	6 Mバイト
ノード	CPU数	1
	メモリ容量	32 Gバイトまたは16 Gバイト
	メモリバンド幅	40 Gバイト/s
	I/O	HDD (73 Gバイト) ×1, InfiniBandHCA (DDR) ×1 1000BASE-T×1
筐体	ノード数	4
	外形寸法	19インチラックマウント5U

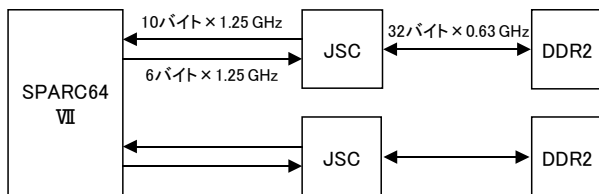


図-2 ノード内LSI間, LSI-メモリ間の関係
Fig.2-LSI interconnections within node.

(注3) リクエストを発してから, 結果が返ってくるまでにかかる遅延時間のこと。
(注4) 線形代数処理で多用される倍精度行列積を計算する。

てもSTREAM (Triad) ベンチ^(注5)で, 13.5 Gバイト/s以上を達成した。HPC2500のSPARC64Vから今回のVIIまでの機能比較を表-3に示す。

【高速かつ高機能な相互結合網】

相互結合網は, InfiniBand^(注6)のDDRインタフェース (理論ピーク2 Gバイト/s全二重) によるFBBのファットツリーを採用し, 120 TFLOPSの計算性能のノード (3008ノード) を, LeafとSpineの2段 (通過LSI段数5段) のInfiniBandスイッチで相互接続した。FBB接続により通信の競合による実効性能の劣化を回避することにより, 通信時間の変化を極小化することをねらっている。

さらに, 並列ジョブの実行効率を高め, 多ノードジョブの性能の揺らぎを抑えるために, ノード間的高速バリアトリダクション演算, OSスケジュールのための同期割込みを実行する高機能スイッチ^(注5)を768ノードごとに追加した。この同期割込み機構は, 768ノードの間でユーザ時間とOS時間の割当て周期を一致させることにより, ユーザジョブのノード間同期動作が, ノードごとに異なるOS動作により長時間阻害され遅延することを防止し, 性能劣化・揺らぎを抑えることをねらっている。

【卓越したアプリケーション実行性能】

ここでは, Integrated Multicore Parallel Architectureを利用した計算性能の指標となる標準ベンチマークテストの測定結果について紹介する。

なお, FX1については, 導入前の評価用システムを使用して性能測定を行った。

表-3 歴代SPARC64のHPC向け機能比較

機能	SPARC64		
	V ^{★1}	VI	VII
(1) CPUコア数	1	2	4
(2) 動作周波数 (GHz)	1.3	2.4	2.5
(3) FP積和演算レイテンシ	12 τ	7 τ	6 τ
(4) FPリネームレジスタ数	32	48	48
(5) FPレジスタライトポート数	2	4	4
(6) コア間L2キャッシュ共有	無	有	有
(7) CPU内ハードバリア	無	無	有
(8) TLBエントリ数 ^{★2}	32	2048	
(9) システムバス幅	16 バイト	20バイト	+12バイト

★1: 130 nm版での数値 ★2: ラージページのオペランド用

(注5) 文部科学省の委託事業「次世代IT基盤構築のための研究開発」の研究開発領域「将来のスーパーコンピューティングのための要素技術の研究開発」の研究開発課題「ペタスケール・システムインターコネクト技術の開発」の成果の一部を活用。

科学技術計算用コンピュータの単体CPU性能とMPI並列性能を測定するためのベンチマークテストであるEuroBen Benchmark[®]を使用して、FX1における演算の素性能を測定した。本ベンチマークテストには多数の評価項目があり、単体CPUの演算性能については、基本的な演算を抽出した31種類の計算ループを使用する。今回はその中から、SPARC64VIIの特長を示す計算ループの例を2種類紹介する。

(1) 逐次性能の測定結果

逐次性能の測定結果を図-3に示す。これは計算ループ14 (9次の多項式計算) について、問題サイズを変化させて逐次実行した場合の演算性能を計測したものであり、グラフの横軸が問題サイズ、縦軸が演算性能を示す。既存クアッドコアCPU (Xeon, Opteron) およびベクトル機 (VPP5000) との性能を比較すると、SPARC64VIIは、問題サイズが小さいところではベクトル機に比べ高い立ち上がり性能、問題サイズが大きいところでもほかのCPU、ベクトル機と遜色ない高い実効性能を示している。

(2) 自動並列性能の測定結果

自動並列性能の測定結果を図-4に示す。これは計算ループ8 (ベクトルの定数倍とベクトルの和) について、問題サイズを変化させて演算性能を計測したものである。VPP5000以外は自動並列化を使用して、1CPU内4コアに4スレッドを割り付けて並列実行している。SPARC64VIIは問題サイズが小さい区間においてもほかのシステムより実効性能が高く、従来の自動並列化で効果が出ないような細粒度並列化にも適用できることを示している。これらの結果から、FX1は従来にくらべ広範囲のアプリケーションに対して有効な高速化機構を備えていると言える。

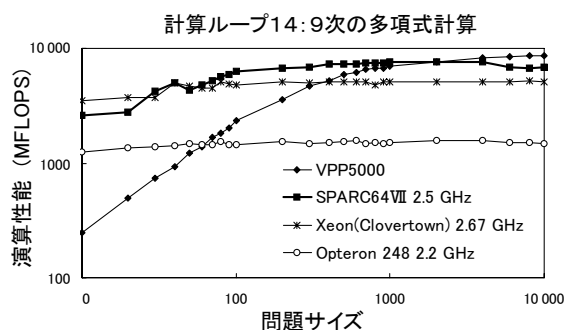


図-3 EuroBen Benchmark逐次性能
Fig.3-Single-core performance measured by the EuroBen Benchmark.

大容量・高速ストレージシステム

JSSの大容量・高速ストレージシステムは、SPARC Enterprise M9000 (3台)、1 PバイトのRAID5ディスク (ETERNUS2000 90台) と総容量10 Pバイトのテープライブラリ (IBM製 TS3500 LTOドライブ48台) から構成され、高いI/O性能を有する。3台のSPARC Enterprise M9000とディスク、テープライブラリは、ETERNUS SN200のSANスイッチ^(注6)により、ファイバチャネルで接続されている。この構成のねらいは、単体I/Oノード構成によるI/O性能の限界やI/Oノードの障害が全系停止につながるという課題を、I/OノードのスケールアウトによるI/O性能向上と、冗長構成による可用性の向上で解決することである。

JSSのストレージシステム構成を図-5に示す。

ストレージシステムには、ファイルシステムが二つ存在している。一つは、ストレージシステムに接続されているハードディスクやテープ装置、ストレージデバイスを制御するローカルファイルシステムであり、もう一つは、多数の計算ノードから一つの大規模ストレージシステムを共用するための制御を行う共有ファイルシステムである。

ローカルファイルシステムには、ディスクの高速アクセス機能とテープの階層管理 (HSM : Hierarchical Storage Management) 機能を持つ

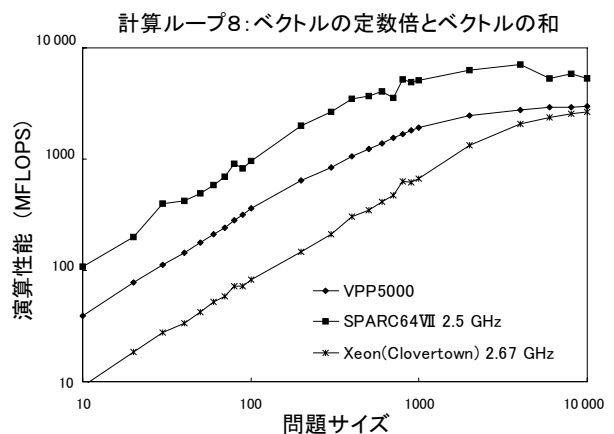


図-4 EuroBen Benchmark並列性能
Fig.4-Parallel performance measured by the EuroBen Benchmark.

(注6) SAN (Storage Area Network) を構築する際に必要な中継装置。複数のサーバやストレージ製品を接続できる。

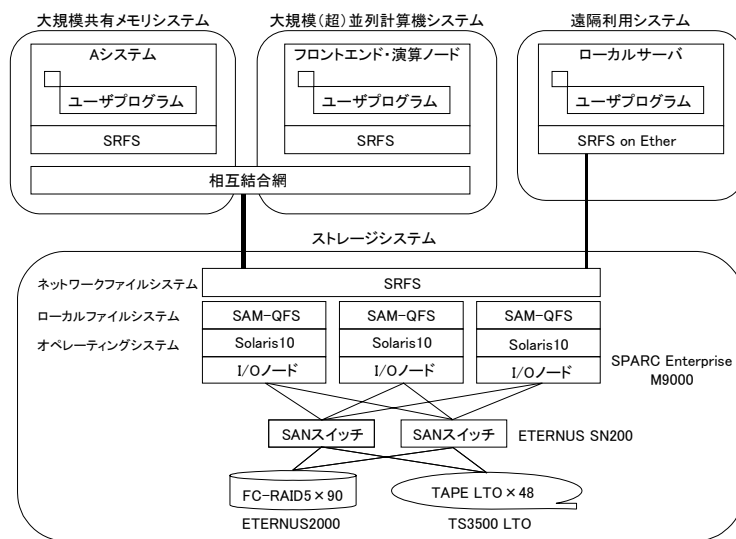


図-5 JSS ストレージシステム構成
Fig.5-JSS Storage system configuration.

SAM-QFS^(注7)を採用した。また、共有ファイルシステムでは、相互結合網を利用した高速アクセスを実現するSRFS (Shared Rapid File System)^(注8)を採用した。

● 透過的かつ高速なストレージの実現

SAM-QFSは、ソフトウェアによるディスクトライピング機能を持ち、高速なアクセス環境を提供するとともに、テープ媒体ではデータを2重に持つことにより、耐故障性とデータのバックアップ機能を提供する。SAM-QFSのHSM機能により、ユーザは自分のファイルがディスクかテープのどちらに存在するか意識することなく、ファイルアクセスが可能である。また、JAXA独自のジョブ制御機構と連携し、ジョブ実行待ちの時間帯に、必要に応じてファイルをテープからディスクへ自動的にステージングする機構により、常に高速なディスクへのアクセスを可能としている。

ストレージシステムのローカルファイルシステム基礎性能であるI/O性能を測定した。ローカルファイルシステムでは、最大33 Gバイト/sのデータ読み込み性能を実現している。I/O性能実測値を図-6に示す。

SRFSは、相互結合網で接続されたシステム(ノード)上で動作し、数千ノード規模でのファイ

First Read 性能 (15stripes x 8groups x 3filesystems)

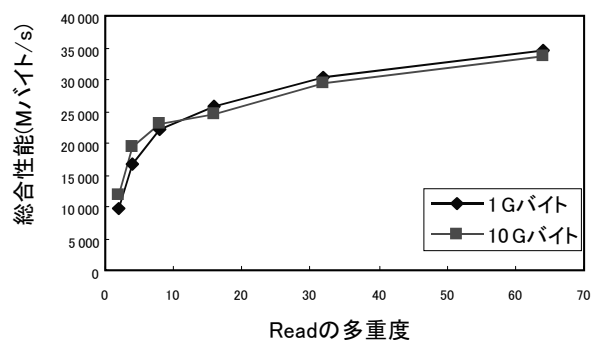


図-6 I/O性能測定
Fig.6-I/O performance values.

ルの共有や大容量データの高速入出力を実現するとともに、複数ノードからの更新に対するデータの一貫性・整合性を保証する。また、SRFSはサーバキャッシュ機能を有しており、超高並列実行でのI/Oや高スループットを実現する。例えば、SRFSクライアントからSRFSサーバに対して、小規模多数ファイルの処理が短時間に集中した場合、ローカルファイルシステムの処理能力不足が懸念される。これに対し、ローカルファイルシステムに集中するI/OをSRFSサーバのメモリにキャッシュすることにより、I/Oのターンアラウンドの向上をねらっている。

● 可用性の実現

今までのJAXAのストレージシステムでは、ストレージシステムのI/Oサーバとディスクやテープ装

(注7) 高速・大容量、かつ階層管理 (HSM) が可能なファイルシステム。Sun Microsystems社製のソフトウェア。
(注8) FX1, SPARC Enterpriseシステム上で動作する分散ファイルシステム。富士通製ソフトウェア。

置は、直接接続されていたため、I/Oサーバの障害がストレージシステム全体の障害につながり、システム全体停止を引き起こす問題があった。JSSでは、I/Oサーバとディスクやテープ装置をSANスイッチで接続することにより、3台以上のI/Oサーバからディスクとテープ装置を共有することが可能となった。さらに、ローカルファイルシステム（SAM-QFS）の共用ファイルシステム機能と連携することにより、ストレージシステムの1台のI/Oサーバがダウンした場合でも、システム全体停止を回避することができる。継続のイメージを図-7に示す。

遠隔利用システム

遠隔利用システムは、SPARC Enterprise M5000, ETERNUS2000から構成される。ユーザは、ローカルサーバでプログラム開発を実施し、数値シミュレーションには調布事業所の計算機システムを利用する。この利用形態では、ユーザは、遠隔利用システムとストレージシステムのファイル保管場所を意識しなければならないという課題があったが、JAXA独自の制御機構と連携し、遠隔利用システムからのジョブ投入に連動したデータ転送を実現する。このために、SRFSをEthernet上で動作させ、ストレージシステムとローカルサーバ間でのSRFSによるファイルシステム共有を提供する。これにより、遠隔地にいながらファイル保管場所を意識することなくJSSを利用することが可能となる。

む す び

本稿ではFX1を中核としたJSSの概要とその特徴について紹介した。

宇宙航空研究開発機構様で今回導入するシステムは、宇宙3機関統合後、JAXAとして初めて調達するスパコンシステムであり、航空分野での利用をこれまで同様に推進するとともに、ロケットエンジン解析、ロケットプルーム音響解析や宇宙機の概念設計への適用を通じて、宇宙開発・宇宙科学や惑星探査分野での本格利用と開発への実質的貢献が期待されている。

今後、著者らはFX1を中核としたJSSの運用・利用操作性の向上に努め、日本の航空宇宙分野の研究活動の推進と発展に寄与していきたいと考える。

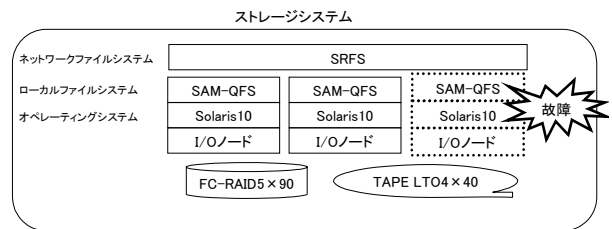


図-7 ストレージの切替え概要
Fig.7-Storage switching to maintain availability.

本稿執筆に当たり、JSSでの計算結果など、数多くの情報を宇宙航空研究開発機構 松尾裕一様にご提供いただきました。この場をお借りして深く感謝いたします。

参考文献

- (1) 富士通：JAXA様の新スーパーコンピュータシステムを受注。プレスリリース（2008年2月19日）。
<http://pr.fujitsu.com/jp/news/2008/02/19.html>
- (2) 松尾裕一：スーパーコンピュータの更新について。PLAINセンターニュース，第173号，2008。
http://www.isas.jaxa.jp/docs/PLAINnews/173_contents/173_contents.html
- (3) 富士通：ハイエンドテクニカルコンピューティングサーバFX1。
<http://jp.fujitsu.com/solutions/hpc/products/fx1.html>
- (4) 富士通：次世代テクニカルコンピューティングサーバFX1の特徴・仕様。
<http://pr.fujitsu.com/jp/news/2008/02/19a.pdf>
- (5) 田中稔ほか：PRIMEPOWER向け並列プログラム開発・運用環境：Parallelnavi。FUJITSU, Vol.52, No.1, p.94-99 (2001)。
<http://img.jp.fujitsu.com/downloads/jp/jmag/vol52-1/paper20.pdf>
- (6) Linpack Benchmark。
<http://www.netlib.org/linpack/>
- (7) STREAM Benchmark。
<http://www.cs.virginia.edu/stream/>
- (8) InfiniBand Trade Association。
<http://www.infinibandta.org/home>
- (9) The EuroBen Benchmark。
<http://www.euroben.nl/>