JAXA Supercomputer System(*JSS*)の構成と特徴

藤田直行、高木亮治、松尾裕一
宇宙航空研究開発機構

Configuration and features of JAXA Supercomputer (*JSS*)

by
Naoyuki Fujita, Ryoji Takaki and Yuichi Matsuo

ABSTRACT

JSS is JAXA's newly installed supercomputer system, which consists of three parts. Compute engines part has totally 141TFLOPS computing power which consists of four types of compute engines. *M-System* and *P-System* are 3,008 CPUs (12,032 cores) and 384 CPUs (1,536 cores) distributed memory processor system, *A-System* is scalar CPU shared memory processor system which has 1TB main memory, *V-System* is vector CPU shared memory processor system. Storage part has about 1PB disk space and 10PB LTO4 tape cartridge space. Integrating distributed environment part realizes integrated secure user environment on geographically distributed condition. Some *JSS* features, which are efficiency of parallelization on large number of parallelism (62%-97%), a large amount of main memory (94TB), correspondence of variety of computing needs (four types of computers), distributed user environment (*JSSnet*, *L-System*, and *J-SPACE*), and energy saving approach on cooling, are discussed.

１．はじめに

宇宙航空研究開発機構(JAXA)では、旧 3 機関時代からの経緯で、調布航空宇宙センター、角田宇宙センター及び、相模原キャンパスの 3 ヵ所にスーパーコンピュータ(スパコン)を保有してきたが、調布航空宇宙センターと角田宇宙センターのスパコンがほぼ同時にリースアウトするのを機に、JAXA 次期スーパーコンピュータシステム(*JSS*)として整備を行い、 2009 年 4 月から全システムの稼働を開始した。*JSS* 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等の JAXA 事業に本格的に活用することを企図すると共に、宇宙 3 機関統合のシンボル的な位置付けで導入されたものである。本報では、*JSS* の構成とその特徴について述べる。

JSS は、地理的に分散して存在した 3 個のスパコンが 1 ヵ所に統合され運用が行われることを考慮し、遠隔地からの利用の利便性確保と、多様な計算需要を扱えるシステム構成をとっている。スカラプロセッサとベクトルプロセッサを持つ総演算能力 141TFLOPS の計算エンジン部、RAID5 ディスク 1PB、LTO4 テープ 10PB の容量を持つストレージ部、計算エンジン部のフロントエンド機能の遠隔配置や遠隔ファイルシステム等を用いて、分散した利用場所の統合利用環境を提供する分散環境統合部により構成される。

本報では、2 章で *JSS* のシステム構成を紹介し、3 章で、並列化効率 62～97%を実測した高並列高効率計算、総メインメモリ容量 94TB の大規模メモリ、4 種類の計算機群による多様な計算需要への対応、主要機能の遠隔配置と SINET3 による遠隔地利用環境、計算機室の省エネへの配慮といった *JSS* の特徴について述べる。

２．システム構成

本章では、*JSS* のシステム構成を述べる。最初に、*JSS* の全体構成を述べる。図 1 に *JSS* の全体構成図を示す。*JSS* は、大きく、①計算エンジン部、②ストレージ部、③分散環境統合部から成る。

２．１　計算エンジン部

計算エンジン部は、演算器の種類、演算器と主記憶メモリの接続方式、及び運用方式の違いにより、4 種類のシステムから構成される。この 4 種類のシステムを、*M* システム、*P* システム、*A* システム、*V* システムと呼ぶ。

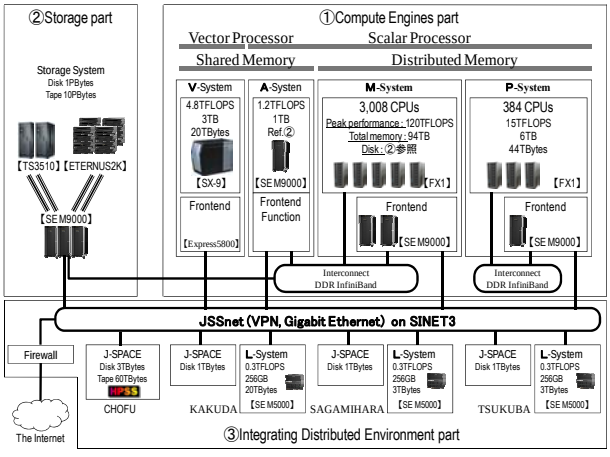

Fig.1 *JSS* Configuration

２．１．１　*M* システム（高並列スカラシステム）

M システムは、富士通(株)製 FX1 で構成され、3,008 個の CPU を持つ分散メモリ計算機(DMP:Distributed Memory Processor)である。CPU は SPARC64VII プロセッサ(1)を採用している。ユーザは、フロントエンド計算機を介して FX1 を使用する。フロントエンド計算機

は、富士通(株)製 SPARC Enterprise M9000(SEM9000)であり、1 ノード当たり 8 個の CPU と 256GB の主記憶メモリを持つ共有メモリ計算機を 2 ノード有し、2 台は冗長構成を組んでいる。

２．１．２ *P* システム（並列スカラシステムのセキュア運用）

P システムは、*M* システムと基本的に同じ演算ノードを持つが、その総数が 384 個であること、1 ノード当たりのメインメモリが 16GB であるところが異なる。*M* システムと同様にユーザはフロントエンド計算機を介して *P* システムを利用する。また、セキュア運用をする点から、後述のストレージ部とは接続せず、*P* システム内部の専用 RAID 装置を用いてファイルシステムを構成している。*M* システムで採用している DMP 方式、マルチコアプログラミングモデル、高速同期機構は同様に備えている。機密性の高い計算をシステムを占有して実行する運用を提供するシステムである。

２．１．３ *A* システム（大容量主記憶システム）

A システムは、SEM9000 で構成され、4 コア 32CPU の共有メモリ計算機(SMP:Shared Memory Processor)で、主記憶メモリは 1TB である。市販アプリケーションソフトウェア等、大容量のメモリを必要とする計算に使用する。現在、市販アプリケーションとして、NASTRAN、FLUENT、ANSYS、FIELDVIEW、Gridgen を実行できる。ファイルシステムは、後述のストレージ部を用いて構成してあるため、*M* システムとのファイル共有が可能となっており、*M* システム上の自作研究プログラムでの計算結果ファイルを、*A* システムで読み込み可視化等の解析ができるようになっている。

２．１．４ *V* システム（ベクトルシステム）

V システムは、日本電気(株)製 SX-9 で構成され、3 個の演算ノードを持つ。CPU は 1 台 0.1TFOLPS のベクトルプロセッサである。ユーザは、フロントエンド計算機 Express5800 を介して SX-9 を使用する。1 ノード当たり 16 個のベクトルプロセッサと 1TB の主記憶メモリを持つ SMP である。スカラプロセッサでの高速化が望めないプログラムや、ベクトルプロセッサチューニングされた既存プログラムの効率的実行に用いる。*V* システムの演算能力が *M* システム等スカラプロセッサのシステムに比較して少ないため、運用面においては、ベクトル化率の高いジョブを選別して *V* システム上で実行させる予定である。

このように、*JSS* は、高並列計算(*M* システム)、機密性の高い計算(*P* システム)、大容量単一メモリ空間を要する計算(*A* システム)、ベクトル計算(*V* システム)と、多様な需要に応える複数の計算機を総合的に運用する構成になっている。表 1 に、*JSS* の計算エンジン部の 4 つのシステムの一覧を示す。太枠は各システムの特徴を示す仕様である。

Tbl.1 Compute Engine Classification

System name	M-System	P-System	A-System	V-System
Processor type	Scalar	Scalar	Scalar	**Vector**
Processor-memory connection type	Distributed	Distributed	Shared	Shard
Usage	General	**Special/Secure**	General	General
# of nodes	3,008	384	1	3
# of CPUs	3,008	384	32	48
# of cores	12,032	1,536	128	48
Peak TFLOPS	**120**	15	1.2	4.8
Total main memory	94TB	6TB	1TB	3TB
Memory per node	32GB	16GB	**1TB**	1TB
Brand	Fujitsu FX1	Fujitsu FX1	Fujitsu SEM9000	NEC SX-9

Bold Cell: Specialty of each system

２．２ ストレージ部

ストレージ部は、実効容量 1PB、総実効転送性能 25GB/s の RAID5 装置と、総容量 10PB、LTO4 ドライブ 40 台、LTO3 ドライブ 8 台の LTO ライブラリ装置から構成される。図 2 に、ストレージ部の構成詳細を示す。ストレージ部は、ネットワークファイルシステム、ローカルファイルシステムと階層ストレージ管理、OS のデバイスドライバ、ストレージエリアネットワーク、ストレージデバイスから構成される。ネットワークファイルシステムはネットワークへのインターフェースとして

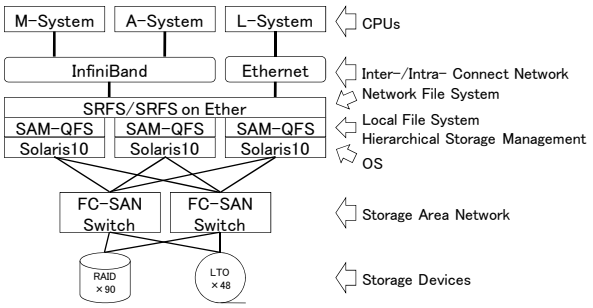

Fig.2 Storage Part Configuration

InfiniBand と Ethernet を持つ。InfiniBand インターフェースにより *M* システムと *A* システムを、Ethernet インターフェースにより遠隔地システムを収容できる。ローカルファイルシステムからストレージデバイスまでは冗長構成をとっており、各部分の一部に故障が発生しても、迂回経路やバックアップ機能を用いてユーザへのファイルシステムサービスを継続できるようになっている。ストレージ部では、数千並列という高並列度演算においても、総実効転送性能 25GB/s を確保しつつ、ファイルシステムが自動的に排他制御を実現するシステムを構築することにより、ユーザのプログラム開発の時間を

確保するよう努めた。また、物理的なストレージ媒体のエラー等によるデータ消失に備えるため、合計 11PB の容量を持つストレージ装置では、ディスク装置は RAID5 構成をとり、テープ装置へのアーカイブ時には同時に２つのテープ媒体に書き込みを行っている。

２．３　分散環境統合部

分散環境統合部は、SINET3[2]上に仮想専用線ネットワーク(VPN)技術を用いて作成したバックボーンを持つ *JSSnet* に、JAXA 事業所内外から *JSS* を利用するためのシステム(*L* システム、*J-SPACE* 等) を接続し、地理的に離れた場所からの *JSS* の利便性向上を目指している。

３．システムの特徴

３．１　高並列高効率計算

JSS の *M* システムでは、高並列の計算を高効率に実行するための技術が導入されている

３．１．１　DMP 方式

スカラプロセッサの並列計算機では、プロセッサ単体の実効性能がベクトルプロセッサと比べて高くないこと等から、多数のプロセッサで一つの主記憶メモリ空間を共有する SMP 方式をとることが多い。*JSS* の前システムのひとつである *CeNSS* もこの方式であった。しかし、SMP の場合、数値流体力学計算等メモリアクセス頻度が高い計算の場合、メモリの共有部分がボトルネックになり、CPU に十分なデータ供給がされず、CPU の演算性能を十分に引き出せないという傾向がある。また、SMP の場合、メモリの共有と同様に、I/O 機構も多数のプロセッサで共有することが多く、I/O 機構においてもメモリアクセス競合と同様のボトルネックが発生しがちである。

そこで、*JSS* では、一つの主記憶メモリ空間を一つの CPU が占有する DMP 方式を採用した。更に、CPU と主記憶メモリ間の接続部を通常の 2 倍の能力に強化することにより、CPU から見たメモリアクセス性能を向上させた。これにより、スカラプロセッサ並列計算機で B/F 比=1 を実現している。ここで、B/F 比とは、Byte/FLOPS の略で、演算性能(FLOPS)に対するメモリ転送性能(Byte/s)の比であり、この数字が大きいほど、メモリアクセス性能が高い。図 3 に SMP と DMP の比較を、図 4 に CPU と主記憶メモリ間の接続部強化の様子を示す。

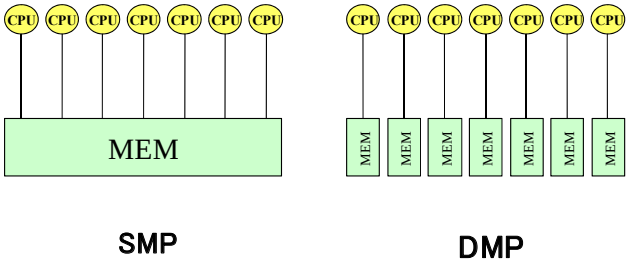

Fig.3　SMP vs. DMP

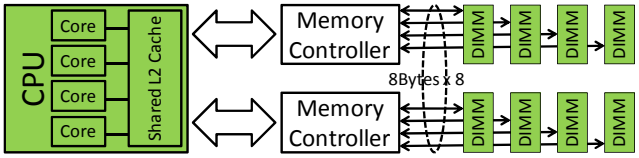

Fig.4　CPU Main memory path

なお、DMP 方式の場合、CPU から見た主記憶メモリ空間が SMP 方式に比べ小さくなる傾向にあり、メモリを大量に使用する大規模数値シミュレーションができなくなる場合がある。この問題点を補うために、*JSS* では、*A* システムを用意している。

３．１．２　マルチコアプログラミングモデル

M システムに限らず、今後の並列計算機の CPU はマルチコア化していくことが予想される。ユーザは、CPU 内の複数のコアの並列性と複数の CPU 間の並列性の両方を意識してアプリケーションコードを開発することが求められることになるが、これはユーザにとって煩雑な作業であり、計算機技術で回避することが望ましい。

FORTRAN の DO ループにおいて、従来のスカラ並列計算機では、一般的に最外 DO ループを手動で並列化する必要があった。*M* システムでは、多重 DO ループをコンパイラが自動的にスレッド並列化(図 5)し CPU 内のコアに割り当て、ユーザは複数 CPU 間の並列性に注力してアプリケーションコードの開発を行うというプログラミングモデルを実現した。このプログラミングモデルを、VisIMPACT：Virtual Single Processor by Integrated Multi-core Parallel Architecture[3] モデルと呼ぶ。ここで、VisIMPACT モデルに対し、従来のプロセス並列モデルのことを FLAT モデルと呼ぶ(図 6)。また、マルチコアの並列実行割り当て技術は、スーパーコンピューティングに大きく貢献したベクトル計算機のコンパイラ技術を適用できるため、ベクトル計算機を凌ぐ高効率なコードを生成することができ、単体 CPU 実行性能の高効率化にも貢献している。

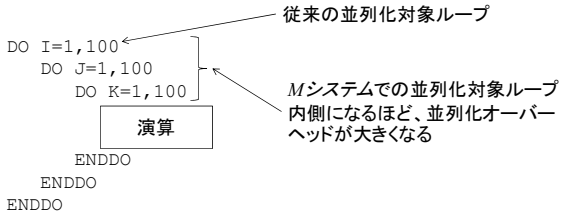

Fig.5 Parallelization DO loop

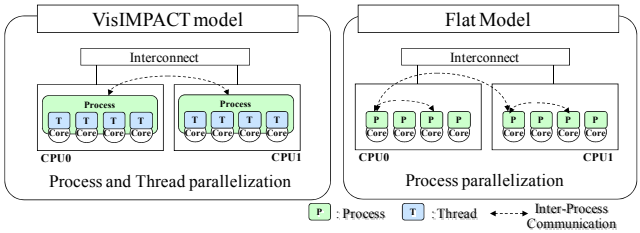

Fig.6 *JSS M-System* Programming Model

３．１．３　チップ内外の高速バリア同期機構

高効率計算を実現するための次の課題は、数百～数千台という多数のコアで、ひとつのアプリケーションコードを動作させる場合の同期処理オーバーヘッドの低減である。*M* システムは、12,032 コアを持つシステムである上に、IMPACT プログラミングモデルの採用により、最内 DO ループでの自動並列化が行われる機会が増えるため、同期処理の機会が FLAT プログラミングモデルに比べ増大している[4]。従って、*M* システムにおける高並列の計算を高効率で実行にするには、同期処理の高速化が必須となる。

M システムでは、これに対応するために、CPU 内の 4 個のコア間にハードバリア機構を備えている。また、CPU 間は、高機能インターコネクトスイッチにより、ハードバリア同期機構や、OS が動作する時間とユーザプログラムが動作する時間を同期させる機構を有している。図 7 に *M* システムのインターコネクトの様子を示す。上半分が計算データを送受信するための DDR InfiniBand インターコネクト網であり、下半分が高機能インターコネクトスイッチによるハードバリア機構のためのインターコネクト網である。また、図 8 に CPU 間におけるハードバリアとソフトバリアの速度比較を示す。これより、並列度が高くなるに従ってハードバリア同期機構の効果が高くなることがわかる。

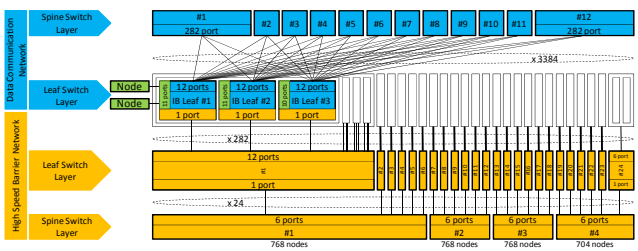

Fig.7 *M-System* interconnect

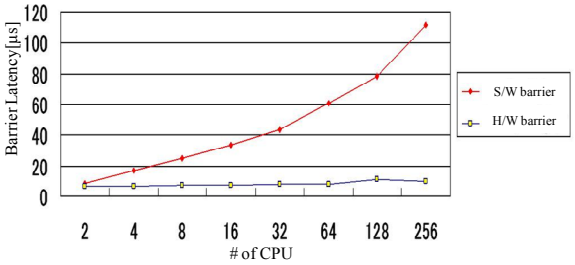

Fig.8 S/W barrier vs. H/W barrier

３．１．４　実行性能検証

ここでは、ベンチマークコードにより実行効率を見ることにする。表 2 は *M* システムで実行した 6 つのベンチマークコードの一覧である。また、表 3 はこれらベンチマークコードの実行効率である。ここで効率は、CPU 数の増加に伴って、問題サイズを大きくした場合のものを測定している。また、効率 e は、式(1)で計算した。

$$e = \frac{\left(a1 * \frac{b2}{a2}\right) / \frac{b3}{a3}}{b1} = \frac{a1 * b2 * a3}{b1 * a2 * b3} \quad \cdots (1)$$

ここで、$\frac{b2}{a2}$ は問題サイズの増加率、$\frac{b3}{a3}$ は計算機資源の増加率、つまり、コア数の増加率である。表 2 から、航空宇宙分野のコードにおいて 2,000 コア以上を使用した場合の効率は 62%から 91%であることが分かる。P6 は、スパコンの性能評価に広く用いられている LINPACK というプログラムであるが、このプログラムにおいて、97%以上の効率を示している。これは、*M* システムが高並列計算においても、高効率を維持していることを示している。

Tbl.2 JAXA's Parallel applications

Code	Application field	Numerical Method
P1	Combustion	FDM+Chemistry
P2	Aeronautics	FVM(Structured)
P3	Turbulence	FDM+FFT
P4	Space Plasma	PIC
P5	Aeronautics	FVM(unstructured)
P6	Linpack	(High Performance Linpack)

Tbl.3 Result of parallel execution performance

Code	Execution on single node			Execution on multi node			Efficiency
	Exec time [s] a1	# of grids (# of floating point operations) a2	# of cores a3	Exec time [s] b1	# of grids (# of floating point operations) b2	# of cores b3	e
P1	131.0	1,728,000	4	143.3	1,285,632,000	2,976	0.914
P2	71.0	512,000	4	91.5	384,000,000	3,000	0.776
P3	346.8	1,572,864	4	491.7	805,306,368	2,048	0.705
P4	164.0	65,536	4	193.0	49,152,000	3,000	0.850
P5	142.0	4,173	4	181.6	2,492,921	3,000	0.622
P6	3566.4	$(1.3361*10^{14})$	4	218376.38	$(2.4101*10^{19})$	12,032	0.979

３．２　大規模メモリ

M システムは、ピーク性能 40GFLOPS 当たり、32GB のメインメモリを持つシステムであり、システム全体の総メモリ資料は 94TB になる。大きなメモリ空間の利用により、数十億点規模の数値シミュレーションが可能となる。

３．３　遠隔地からの利用

３．３．１　*JSSnet*

分散環境統合部のバックボーン *JSSnet* の帯域は最大 1Gbps であり、SINET3 の４ノード(JAXA 調布、JAXA 相模原、筑波大学、東北大学)を JAXA 各事業所に設置した VPN 装置により IPsec トンネルで相互接続することにより構築している。現在、SINET3 の“L2 サービス”の“L2VPN”及び“QoS”サービスを用いての仮想専用線ネットワーク構築実験を進めている。

３．３．２　*L* システム

L システムは、主要利用拠点に配置した、計算エンジン用のフロントエンド機能を有したシステムであり、ログインやコンパイル機能等のユーザのコード開発環境を、各利用拠点のローカルなネットワーク環境(LAN)上に提供している。また、*L* システムは図 1 のストレージ部の機能であるファイルマウントや高速同期機構を用いて、計算エンジン部とのデータ連携を実現している。

３．３．３　*J-SPACE*（遠隔共有ファイルシステム）

遠隔共有ファイルシステム（*J-SPACE*）は、米国エネルギー省の ASCI PSE(The Accelerated Strategic Computing Initiative, Problem Solving Environment)プロジェクトの成果物のひとつである HPSS[5]を活用している。HPSS は、High Performance Storage System の省略形であり、マスストレージというニッチな分野において、PSE の目的である知識と経験の複数の団体による共有を目指したものである。Lawrence Livermore National Laboratory(LLNL)、Lawrence Berkeley National Laboratory(LBL)、Los Alamos National Laboratory(LANL)、Sandia National Laboratories(SNL)、Oak Ridge National Laboratory(ORNL)、IBM が開発メンバーとして参加している。分散型階層ストレージ管理ソフトウエアで、ASCI を中心とした HPC のユーザを念頭に置いた設計が行われている。全世界 20 組織以上で、40 システム以上が設置され、ASCI、スパコンセンター、気象関係、大学、原子核物理研究所等で使用されている。日本では、理化学研究所、高エネルギー加速器研究機構、と JAXA で利用中である。特徴は高速性・拡張性であり、「下位の H/W 性能の 90%以上を出せる」という設計方針がある。また、H/W を追加すれば、スケーラブルに性能向上ができる分散型の設計である。*JSS* では、この HPSS の分散型の特徴を用い、Mover と呼ばれるユーザが直接 I/O を行うサーバを、主要事業所に分散配置することにより、単一名前空間を分散環境上に構築している。図 9 に *J-SPACE* の利用イメージを示す。

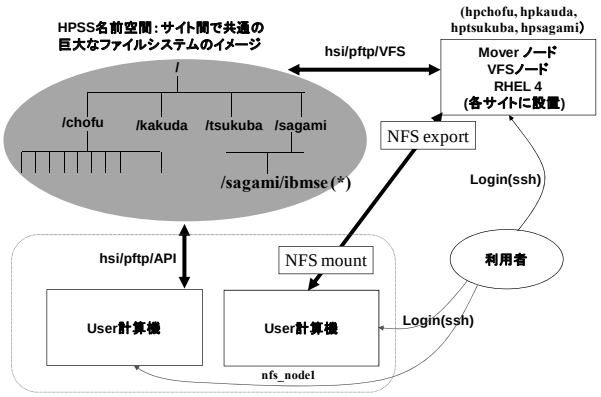

Fig.9 *J-SPACE* Usage Example

３．４　環境への配慮

JSS では、計算機システムと空調等建屋設備の連携、効率的な冷却手法の採用等により、スパコンシステム全体としての停止時間の削減を行い、稼働率の向上を目指している。図 10 に計算機室内の冷却流を示す。冷却効率を高めるため、冷気と暖気が混ざり合わないようにしている。

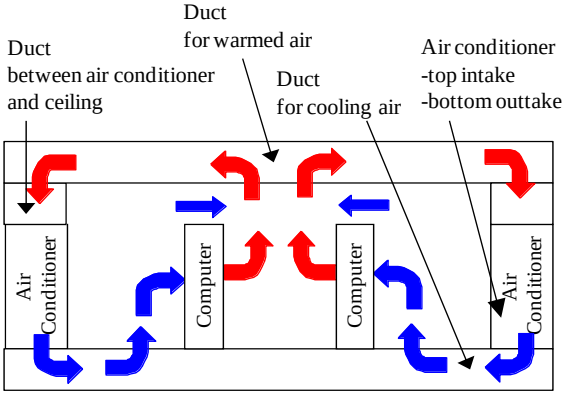

Fig.10 *JSS* Cooling System

また、空調機吹き出し温度と計算機消費電力の関係を試算し、システム全体として省エネルギー運転を行うための検討も行っている(図 11)。表 4 に試算の結果を示す。空調機吹き出し口温度を 20℃と 25℃に設定した場合の、計算機消費電力の変化を推定し、空調機と計算機を合わせたシステム全体の消費電力量を比較した。なお、電力量は、空調機吹き出し温度が 20℃の場合の空調機消費電力を 1 とした時の比で表している。この結果から、空調機吹き出し温度は 20℃の方がシステム全体の電力消費は抑えられるということがわかる。

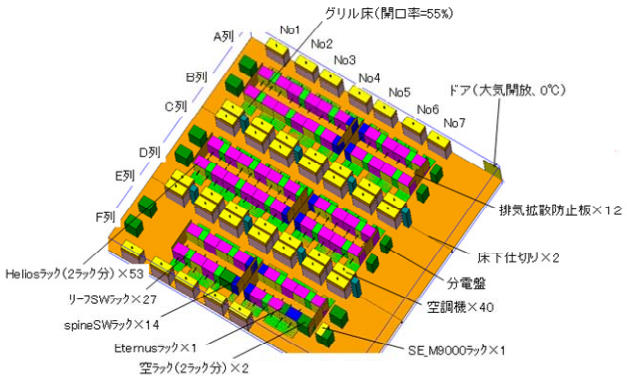

Fig.11 System wide Cooling Simulation Example

Tbl.4 System wide Power Consumption

Air con. outtake temperature [℃]	Air con. power ratio A	Computer Power ratio B	System wide Power ratio C=A+B
20	1(base)	2.00	3.00
25	0.95	2.08	3.03

4. おわりに

2009 年 4 月から全システムの稼動を開始した JAXA の新スパコン *JSS* のシステム構成を紹介すると共に、その特徴である並列化効率 62～97%を実測した高並列高効率計算、総メインメモリ容量 94TB の大規模メモリ、4 種類の計算機群による多様な計算需要への対応、主要機能の遠隔配置と SINET3 による遠隔地利用環境、計算機室の省エネへの配慮について述べた。

今後は、計算性能の分析、ストレージ部の性能評価、遠隔環境の性能チューニング等スパコンシステム本来の性能向上へ向けた研究を行うと共に、大規模システムにおける故障発生の分析や保守の効率的実施方法等運用面の分析も進めていく予定である。

最後に、本稿作成にあたり，シミュレーション結果や図表の提供をいただいた，富士通株式会社に感謝の意を表する.

参考文献

1) “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 , 2008
2) 国立情報学研究所, “学術情報ネットワークとは”, http://www.sinet.ad.jp/about_sinet3
3) Fujitsu Limited, ”ホワイトペーパー:富士通 SPARC64TMVII プロセッサ”, http://img.jp.fujitsu.com/downloads/jp/jhpc/sparc64vii-wpj.pdf, 2008
4) 藤田直行、高木亮治、松尾裕一、“JAXA 次期スーパーコンピュータシステム″JSS″の設計思想と構成概要”、第 41 回流体力学講演会/航空宇宙数値シミュレーション技術シンポジウム 2009、2D1、2009
5) HPSS Collaboration, ”High Performance Storage System”, http://www.hpss-collaboration.org/hpss/index.jsp