サイエンティフィック・システム研究会 2009 年 9 月 3 日
SS 研 HPC フォーラム 2009

JAXA Supercomputer System (JSS) の紹介と性能概要

高木 亮治、藤田直行、松尾裕一
宇宙航空研究開発機構

[アブストラクト]
宇宙航空研究開発機構(JAXA)は航空宇宙分野における基礎研究から研究・利用までを一貫して行っており、前身の宇宙三機関の時代から高性能計算機を用いた数値シミュレーション技術の重要性を認識し、高性能・高機能な大規模計算機システムの整備・運用を積極的に推進してきた。2009 年 4 月に JAXA Supercomputer System(略して JSS)と呼ばれる新しいシステムが稼動を開始した。JSS は複数の計算機システムから構成されるが、その中核は富士通製 FX1 で、マルチコアスカラーCPU を用いた大規模超並列計算機であり、120TFlops の理論演算性能と 94TByte の主記憶容量を持っている。
本講演では JSS の概要を紹介すると同時に、JAXA で実際に使われている航空宇宙分野における CFD プログラムを用いた性能評価結果について報告する。

[キーワード]
航空宇宙、CFD(計算流体力学)、大規模並列計算機、マルチコア

1. はじめに

宇宙航空研究開発機構(JAXA)は、航空宇宙分野の基礎研究から開発・利用までを一貫して行っているが、前身の航空宇宙技術研究所(NAL)および宇宙科学研究所(ISAS)の時代から高性能計算機を用いた数値シミュレーション技術の重要性を認識し、高性能・高機能な大規模計算機システムの整備・運用を積極的に行ってきた。航空宇宙分野における数値シミュレーションは大規模な解析が多く、必然的に大規模並列計算機システムが必要とされてきた。単純な 2 次元翼型まわりの流れの解析から始まり、計算機の発達とともにより複雑な形状、例えば航空機全機まわりの粘性流れ解析や、より複雑な現象、エンジン内での化学反応を伴う燃焼流れなどの解析へと発展していった。これらの数値シミュレーションは学術研究のツールとしてだけではなく、実際の航空機、ロケット、衛星・探査機などの設計や開発に応用されている。特に実際の応用においては開発において発生したトラブルに対するトラブルシューティング的な課題解決型のアプローチから、徐々にではあるが設計探査や最適化といった設計プロセスを革新する様な使われ方にシフトしている。

JAXA 統合前後において、旧 3 機関時代からの経緯で統合後も調布、角田、相模原の 3 箇所で大規模計算機システムが運用されていたが、調布、角田のシステムがほぼ同時期にリースアウトするのを契機に 2009 年 4 月に新しい大規模並列計算機システムを導入した。新しく導入した計算機システムは JSS(JAXA Supercomputer System)と呼ばれ、JAXA 統合後初めての導入となることから、これまで以上に宇宙開発等の JAXA 事業への本格的な活用および宇宙三機関統合のシンボル的な位置づけ(One-JAXA)を意図して導入された。

本報告では、まず始めに JSS の設計思想およびシステム構成、特徴等について紹介する。次に JAXA で利用されている代表的な CFD プログラムを用いた JSS の性能評価結果について報告する。最後に JSS 導入時に実施された大規模解析について紹介する。

2. 設計思想

JSSを設計するにあたり様々な角度からの検討を行った。まず旧システムの課題として大規模SMPの使い難さがあった。JAXAではSMPノード内に複数のジョブが混在する運用を行っていたためジョブの計算時間のぶれが発生した。計算時間のぶれはプログラムチューニングの大きな障害となり最後まで抜本的な解決は行えなかった。またメモリバンド幅不足、自動スレッド並列コンパイラの能力不足、スカラーCPU の経験不足などから期待した性能が出せなかった。

本来、どのような計算機(演算性能重視、メモリ性能重視、通信性能重視)を導入すべきかは利用するアプリケーションに大きく依存する。最近のJAXAアプリの傾向として工学系アプリのCapacity計算指向および学術系アプリの Capability 計算指向が挙げられる。工学系はパラメトリック計算などスループット重視のものであり、学術系は性能や規模が重視される。これらのアプリはさらに相対的に計算負荷の大きいアプリ(計算系)、通信負荷の大きいアプリ(通信系)、メモリアクセス負荷の大きなアプリ(メモリ系)に分類できる。どのような特性を持ったアプリをターゲットとするかで計算機のバランス(演算性能、メモリ性能、ネットワーク性能)が決められるが、JSS では工学系アプリを中心に考えつつ学術系アプリにも配慮する方針とし、そのため演算性能とメモリ性能を重視することとした。

技術的な観点からは、先進的過ぎて実績がなかったり、維持管理(手間、コスト)が大変な技術・システムの採用はやはり困難であり、将来動向は見据えつつも確実な技術や持続可能な技術を採用する必要がある。例えばノード形態としてはメモリ性能や電力・コストを考えると大規模共有メモリノードよりも有利な小規模ノードを選択した。また結合ネットワーク(インターコネクト)に関しては、伝統的なクロスバーネットワークは物量的に非現実的であり、ファットツリーなど実績のある多段結合網とした。

最後に統合スパコンとしての要求に応えるため、これまでのプログラムの継続性や遠隔地からの利便性に配慮した。そのため各拠点にはフロントエンド機能やファイルサーバ機能を有する遠隔利用システムや分散データ共有システムを配置した。またベクトルジョブへの配慮から小規模ベクトルシステムを導入した。さらに、前後処理や非並列ジョブ、市販アプリの動作プラットフォームを考えた場合、大きなメモリ空間を有する共有メモリシステムは魅力的であり、巨大なメモリを有する共有メモリシステムを別途用意することとした。

3. システム構成と特徴

様々な要求項目や検討の結果、JSSは図1および表1で示す様に大規模並列計算機システム、ストレージシステム、共有メモリシステム、遠隔利用システム、分散データ共有システムなど複数の計算機システムから構成される複合システムとなった。従来システムに比べて演算性能は約15倍、メモリ量では約25倍、ストレージ量では約 20 倍程度の性能を有する。JSS の中で実際の計算の中核となるシステムは大規模並列計算機システムであり、マルチコアCPUをベースにした富士通製FX1と呼ばれるスカラー超並列計算機で構成される。大規模並列計算機システムは120TFlopsの演算性能と94TBytesの主記憶装置を有するM(メイン)システムと 15TFlops、6TBytes の P(プロジェクト)システムから構成される。これら二つのシステムではIntegrated Multicore Parallel ArChiTecture (IMPACT) と呼ばれるマルチコアCPUを効果的に利用する技術が採用されている。IMPACT を構成するコア間ハードウェアバリア、共有キャッシュ、自動スレッド並列コンパイラの連携により従来性能が出せなかった細粒度スレッド並列(内側ループでのスレッド並列化)でも十分な性能が期待できるようになった。そのため、ユーザーにはノード間をプロセス並列を用いて並列化し、ノード内は IMPACT の自動並列コンパイラによる自動スレッド並列にまかせるという並列化モデルを推奨している。

共有メモリシステムは A(アプリケーション)システムと呼ばれる 1TBytes の共有メモリを有する富士通製SPARC Enterprise M9000とV(ベクトル)システムと呼ばれる、4.8TFlopsの演算性能と3TBytesのメモリを有

するNEC製ベクトル計算機システムSX-9からなる。これらのシステムは巨大な共有メモリを持つ利点を活かして、前後処理を含む非並列ジョブや市販アプリ、ベクトルジョブの実行に利用される。

ストレージシステムとしてはディスクが 1PByte、テープが 10PByte ありディスクへの総実行転送性能は28GByte/s(ioperfにより測定)となる。ディスクとテープは階層型ストレージ管理(HSM)を行っている。

JSSの主要部分は調布事業所に設置されるため、角田、筑波、相模原などの遠隔拠点には遠隔利用システム(L システム)を設置した。これは各拠点からの利便性を向上させるためのもので、ファイルサーバやフロントエンドの役割を果たす。また各拠点間でのファイル共有を実現するためJ-SPACEと呼ばれる分散データ共有システムを構築した。調布システムと各拠点のシステムとはスパコンネットと呼ばれる高速ネットワークで接続されている。

JSS 導入にあたり、既存建屋には入りきらないため新建屋を建設した。新建屋では冷却効率の向上を目指し、排気拡散防止版、空調ダクトを設置するなどして暖気と冷気をできるだけ分離し、冷却効率が高まるような工夫を行った。また一般的な電力による空調の代わりにガスによる空調機の採用や遮音板の設置、室内設定温度の検討など環境に配慮したものとなっている。

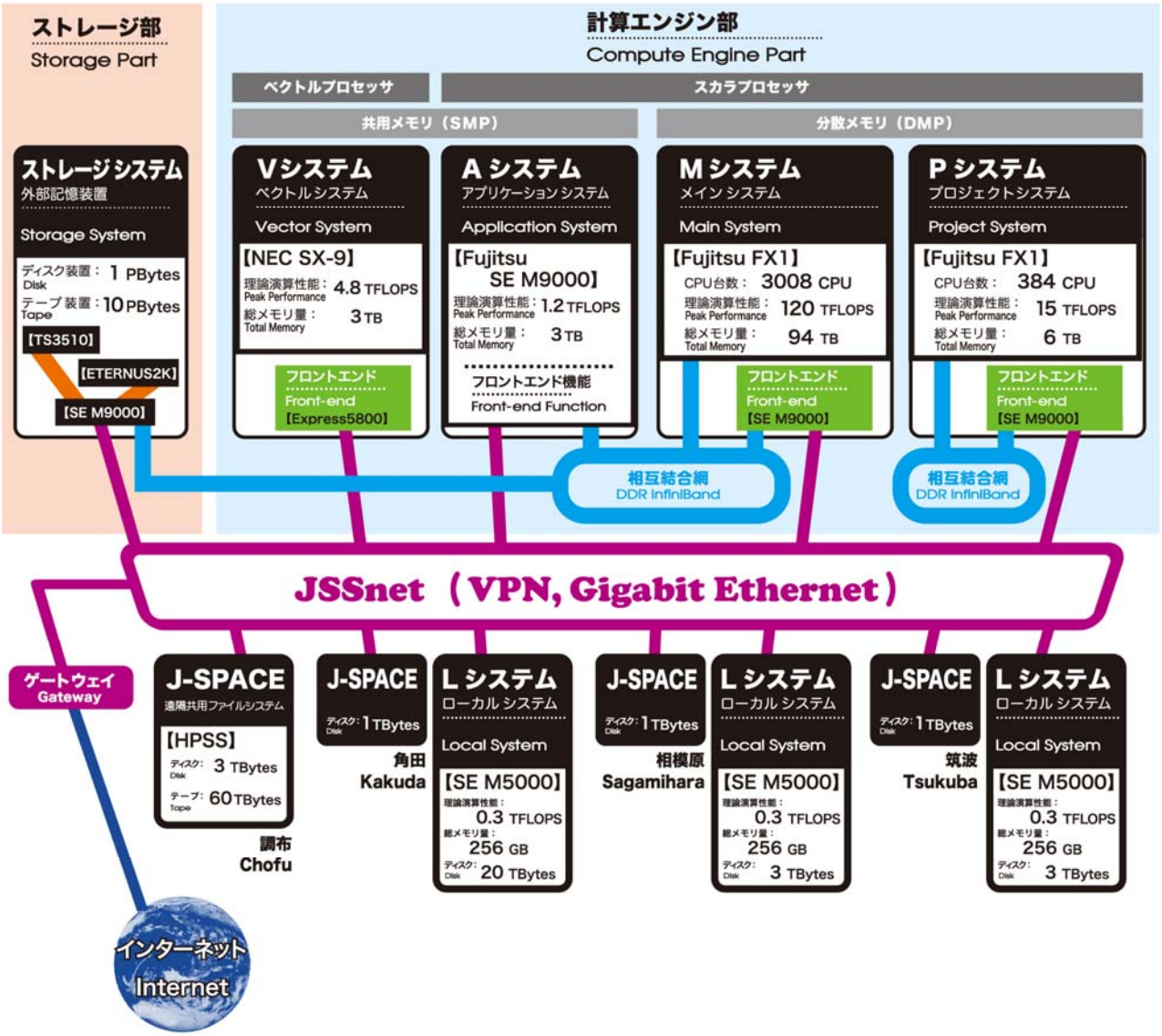


図1 : JAXA Supercomputer System (JSS)のシステム図

表1:JSS の主要計算機システム

名称	JSS-M/P	JSS-A	JSS-V
CPU	Scalar	Scalar	Vector
System	MPP	SMP	SMP
ノード数	3008/384	1	3
CPU/ノード	1	32	16
Core/CPU	4	4	1
論理性能 [TFlops]	120/15	1.2	4.8
ノード性能 [GFlops]	40	40	1,600
総メモリ [TBytes]	94/6	1	3
ノードメモリ [GBytes]	32/16	1	1
メーカー	富士通 FX1	富士通 SEM9000	NEC SX-9

4. 性能評価

JSS における計算の中核となる JSS-M/V に関して実際に JAXA で使われている実アプリケーションによる性能評価を行った。まず JSS-M に対する評価について示す。表 2 にアプリケーションの概要を示す。どのアプリケーションも日常的に使われている実用アプリケーションであり、各アプリケーションの特性としては P1 と P4 は演算負荷が大きいアプリケーション、P2 とP5はメモリアクセス負荷が大きいアプリケーション、P3 はネットワーク負荷が大きいアプリケーションである。

表 2:JAXA アプリケーション一覧

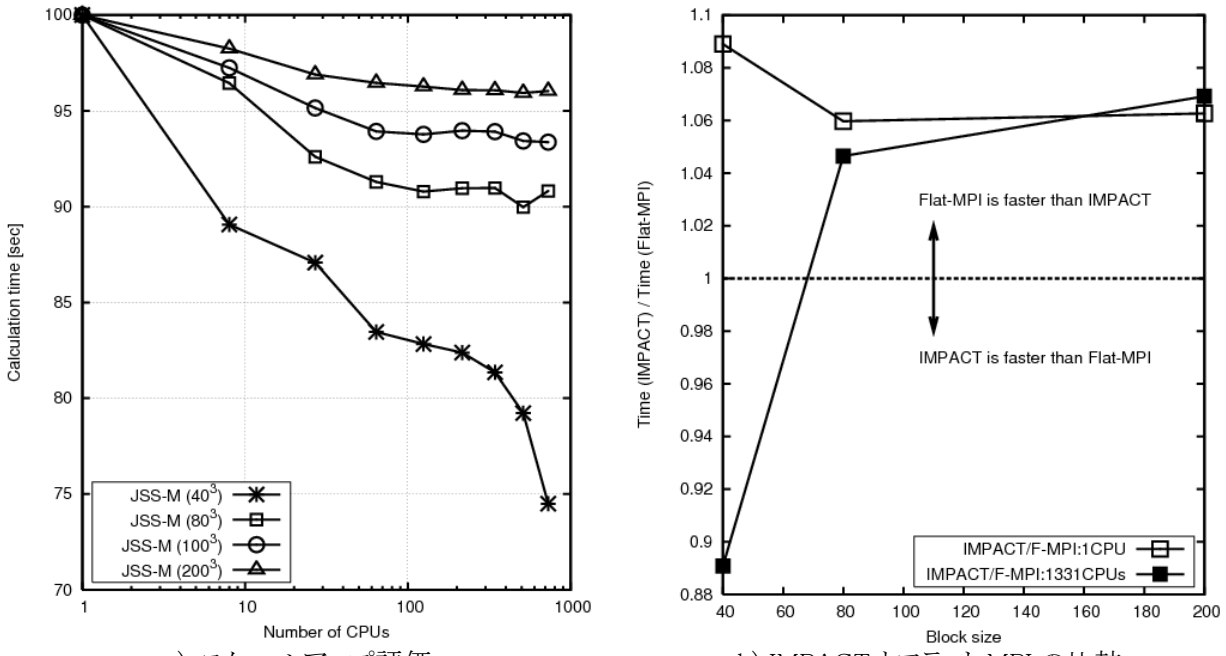
名称	適用先	計算手法	並列化	特性
P1	燃焼	FDM+化学反応	MPI+IMPACT	演算負荷が大
P2	航空	FVM (構造)	MPI+IMPACT	メモリアクセス負荷が大
P3	乱流	FDM+FFT	XPF+IMPACT	ネットワーク負荷が大
P4	プラズマ	PIC	MPI+IMPACT	演算負荷が大
P5	航空	FVM (非構造)	MPI+IMPACT	メモリアクセス負荷が大

表 3 に測定結果を示す。この結果より JAXA の実アプリケーションに対して JSS-M は CeNSS(従来システムである NS-III の中核計算機)より平均で 10 倍以上高速であることがわかる。P2 と P4 の性能比が他よりも高いが、P2 は自動スレッド並列コンパイラの性能向上によりスレッド並列の範囲が広がったためと考えられる。また P4 は MPI の集合通信の改善が主な理由と考えられる。逆に P3 が悪いのは他アプリケーションと異なり、演算負荷に対して相対的にデータ転送負荷が大きく経過時間ではほぼ同程度となる。JSS-M では通信性能比(対 CeNSS 比で 2 倍にしかならない)は演算性能比(周波数:2 倍、コア数:4 倍、その他:?倍)に比べて悪いため、全体の性能比が悪くなったと考えられる。

表 3:JAXA アプリによる性能評価

名称	CPU 数	CeNSS [sec]	JSS-M [sec]	性能比
P1	744	1380.4	143.3	9.63
P2	750	1468.6	91.5	16.05
P3	512	3517.0	491.7	7.15
P4	750	3061.7	193.0	15.86
P5	750	1447.2	181.6	8.13
平均 (相乗平均)	-			11.36 (10.73)

JAXA の代表的な実アプリの一つである UPACS を用いて JSS-M の性能評価を行った。UPACS は構造体、動的配列、ポインター、モジュールといった Fortran90 の機能を活用して作成された汎用的な圧縮性流体解析プログラムであり、3 次元マルチブロック構造格子および重合格子を扱うことができる。UPACS を用いたスケールアップ評価を図 2a)に示す。使用する CPU を増やす度に計算規模を増大させるスケールアップ評価では 729CPU まで 74%(ブロックサイズは 40^3)～96%(ブロックサイズは 200^3)といった良い並列効率を示した。ブロックサイズが大きくなると相対的に並列性能は良くなる。何故なら、ブロックの1辺を N とすると演算量は N^3 に対して通信量は N^2 に比例するため、N が大きくなる、つまりブロックサイズが大きくなると相対的に通信によるオーバーヘッドが減少するからである。



a) スケールアップ評価　　b) IMPACT とフラット MPI の比較

図2:UPACS による JSS-M の評価

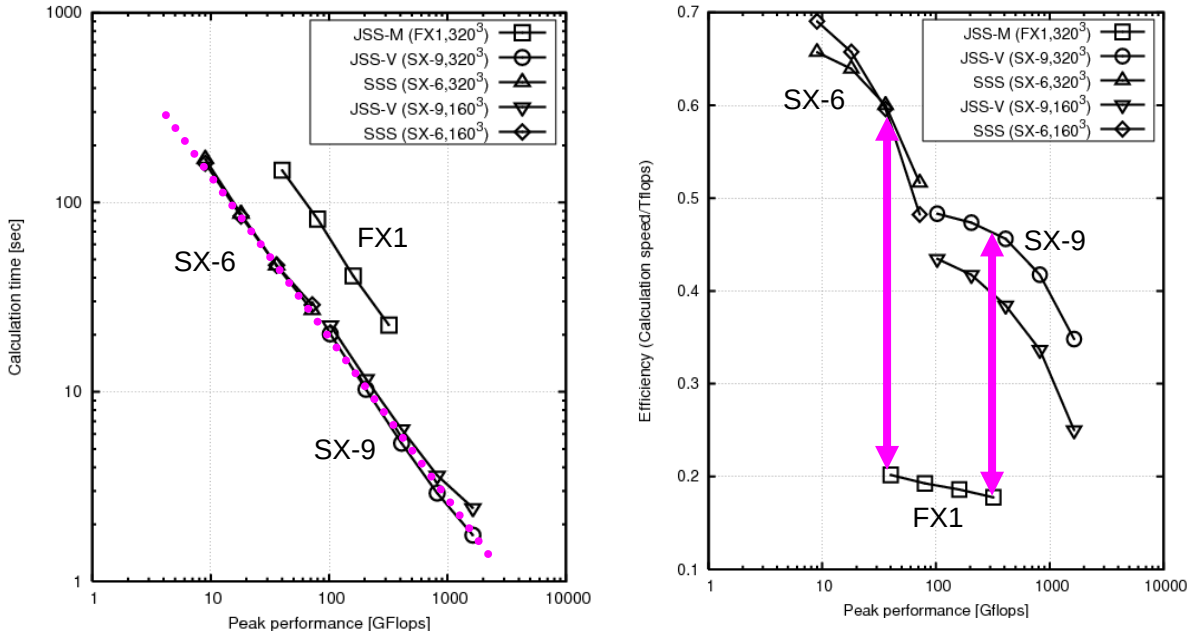
IMPACT とフラット MPI の比較を図 2b)に示す。IMPACT とフラット MPI で、どちらが良いかは色々な条件の影響を受ける。例えば、CPU 内並列を考えた場合、IMPACT の並列性能は自動並列コンパイラの能力に依存するが、フラット MPI ではユーザーが明示的に並列化を行うことで高い並列性能が期待できる。一方、プロセス間通信を考えた場合、IMPACT はプロセス数を減少させることができるので、高プロセス並列において

も性能劣化が低いが、フラット MPI の場合はプロセス数がコア数倍増えるため、高プロセス並列においては性能劣化が大きいと考えられる。またアプリ側の問題として計算格子のブロックサイズが大きくなるとフラット MPI が有利となる。これは前述したようにブロックサイズが大きくなると通信のオーバーヘッドが小さくなるからである。ブロックサイズの影響は図 2b)からも読み取れる。現状の JSS の規模ではアプリケーションにもよるが IMPACT よりもフラット MPI の方が若干有利かもしれないが、今後更なるスケールアップ（コア数、ノード数）を考えた場合、フラット MPI の限界が見えてきたと考えている。

次にこれも JAXA の代表的なアプリの一つである LANS3D を用いて JSS-M および JSS-V の評価を行った。LANS3D は航空宇宙分野で比較的初期に開発された先駆的な CFD プログラムで Fortran77 をベースに構造化プログラミング的な考え方で設計され、主にベクトル計算機向けに実装された典型的な圧縮性流体解析プログラムである。並列化に関しては基本的に自動スレッド並列による並列化を行っている。プロセス並列化としては MPI を用いた領域分割に一部対応しているが、多数プロセスによる並列計算は現実的でないためここでは最大 8 プロセスまでとした。この LANS3D を用いて JSS-M/V のスピードアップ評価を行った。問題規模として約 3300 万点の計算格子を固定して、CPU 数を増やすことによる計算速度の向上を評価した。表 4 に計算の概要を示す。JSS-M では IMPACT を用いて、プロセス数として1～8プロセスまで測定を行った。プロセス数に応じて計算格子をブロック分割し、1CPUに1ブロックを割り当てた。JSS-V および SSS（現在も相模原で運用中の NEC 製 SX-6）ではノード内に限定し自動並列でスレッド数1～16（SSS は8まで）で計測を行った。またベクトル長の影響を見るため、320x320x320 が1ブロックの計算と 160x160x160 が8ブロックの計算を行った。測定結果を図 3 に示す。図 3a）は計算時間を比較したもの、図 3b）は相対的実行効率を比較したものである。ここで相対的実行効率は計算時間の逆数を利用したCPUのピーク性能で割った値であり、単位ピーク性能あたりの計算速度を示す。そのためこれらの値の比較は実行効率の比較と同じ意味を持つ。相対的実行効率は通常の実行効率とは異なり、利用範囲が限定されるが、ユーザーの実感に近い、計測が容易といったメリットがあるため、ここでは相対的実行効率を比較した。図 3b）より SSS(SX-6)と JSS-M(FX1)では実行効率で 3 倍程度の違いが見られるが、JS-V(SX-9)と JSS-M(FX1)では約 2.5 倍程度の違いに縮小していることがわかる。また JSS-V(SX-9)は SSS(SX-6)に比べてベクトル長が短くなると性能が悪化することもわかる。

表 4：LANS3D による評価

システム	プロセス数	スレッド数	ピーク性能 [GFlops]	(ブロックサイズ) ×ブロック数	並列手法
JSS-M	1	4	40	(320x320x320) x1	IMPACT
	2	8	80	(320x320x160) x2	MPI+IMPACT
	4	16	160	(320x160x160) x4	
	8	32	320	(160x160x160) x8	
JSS-V(SX-9)	1	1,2,4,8,16	102.4～1638.4	(320x320x320) x1 (160x160x160) x8	自動並列
SSS(SX-6)	1	1,2,4,8	9.0～720	(320x320x320) x1 (160x160x160) x8	



a) 計算時間の比較　　b) 相対的実行効率の比較

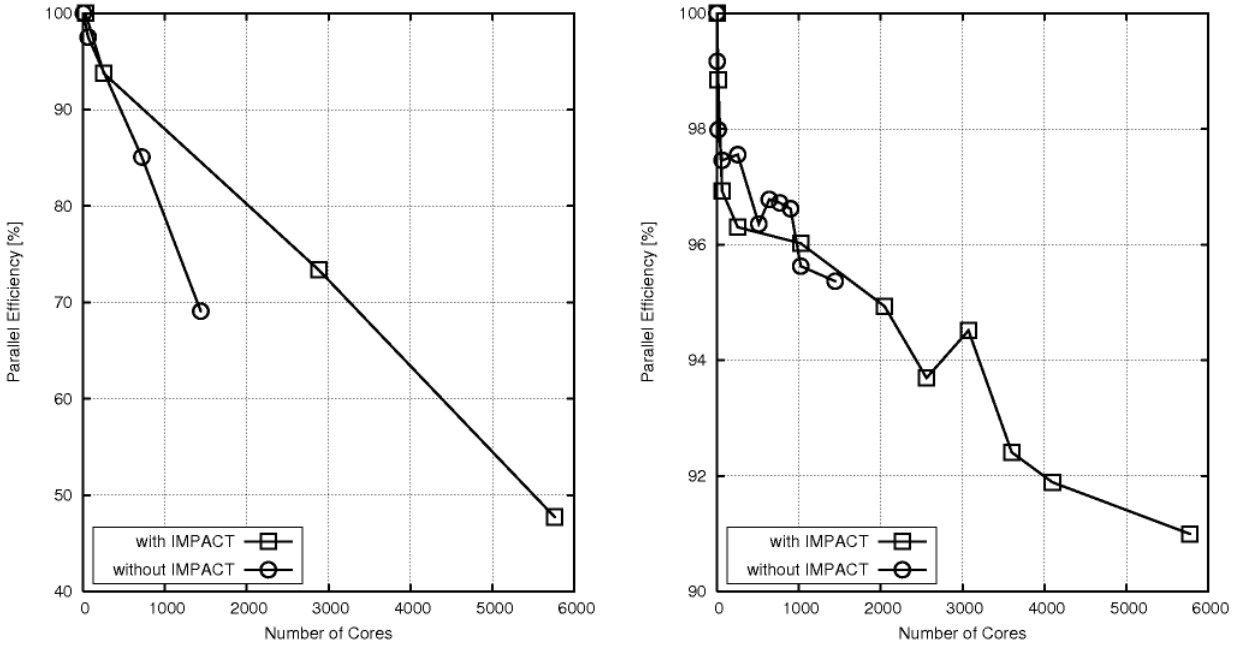
図 3:LANS3D による評価結果(JSS-M(FX1), JSS-V(SX-9), SSS(SX-6))

5. 大規模解析

JSS 導入時に JSS-M の半分を1ヶ月程度利用する大規模解析を実施した。「液体燃料噴流微粒化過程解明の大規模計算(LSC1)」と「大規模粒子計算で探る宇宙空間衝撃波のダイナミクス ～科学衛星観測成果の理解に向けて～(LSC2)」の 2 件でそれぞれの計算概要を表 5 に、スケールアップでの並列化効率を図4に、計算結果の一例を図5に示す。LSC1 は計算時間ネック、LSC2 はメモリネックの解析である。また並列化効率も LSC1 では 50%弱であるが、LSC2 は 90%以上の高い値となった。それぞれの計算結果は当該分野で大きな成果を挙げているが、それらの成果もさることながら、導入初期に大きなトラブルもなくこれらの計算が延べ2ヶ月に渡って実行できたことはシステムの安定性・信頼性が非常に高いことの表れである。これとは別に一般的な Linpack HPL の計測も行ったが、その際も延べ 60 時間以上の高負荷計算においてもトラブルなく計算を完遂することができ、システムの安定性が非常に高いことを実証できた。なお、Linpack に関しては Rpeak=121.282TFlops に対して Rmax=110.600 となり高い実行効率(91.19%)を示した。

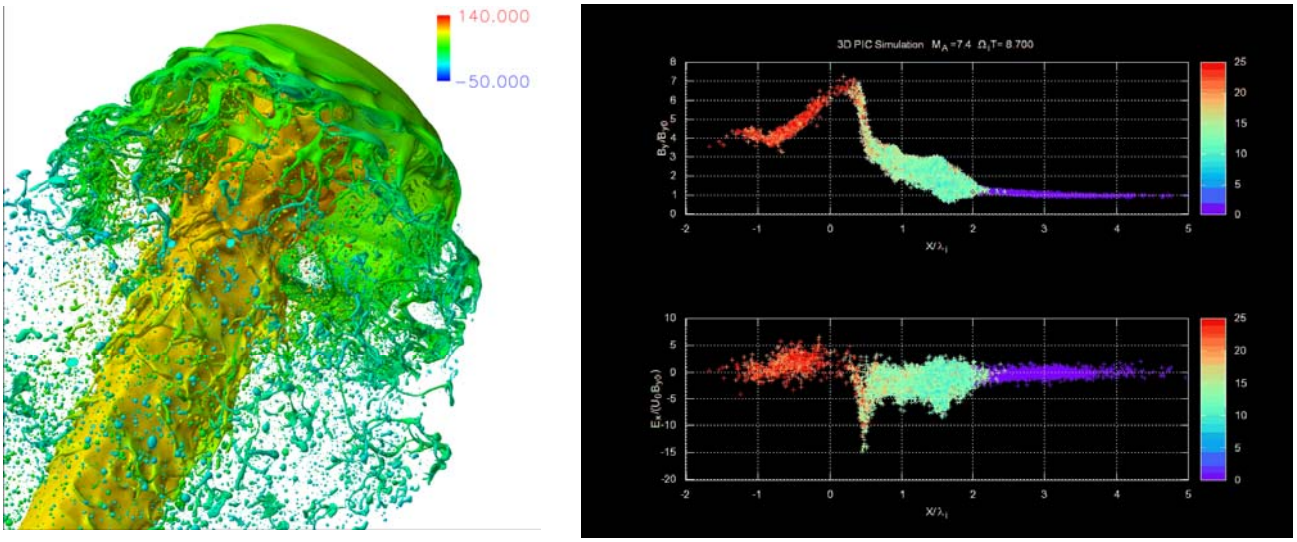
表 5:大規模解析の概要

	LSC1	LSC2
並列規模	1440 プロセス(5760 コア)	1444 プロセス(5776 コア)
計算規模	格子点:58 億点	格子点:4.5 億点、粒子数:500 億個、メモリ:40TB
計算時間	410 時間	740 時間
出力ファイル	153TB(25 時間)	180TB(総量:430TB、43 時間)
実行効率	約 4%程度	約 8%程度



a) LSC1　　b) LSC2

図4：大規模解析の並列化効率（スケールアップ）



a) LSC1 の結果　　b) LSC2 の結果

図5：大規模解析の計算結果例

6. おわりに

JAXA の新スパコンシステム（JSS）について、設計思想、システムの概要、初期性能評価、大規模解析例などについて紹介した。今後は更なる性能評価、ユーザーのチューニング支援、可視化システムの整備などを進めて行くと同時に、JSS を活用した数値シミュレーション技術により、JAXA 事業における設計開発プロセス自体の革新を目指し、「もの造り」への貢献をより強力に推進して行く。

以上