16

数値風洞のハードウェア

三好　甫*　吉岡義朗**
池田正幸**　高村守幸**

Numerical Wind Tunnel Hardware

by

Hajime MIYOSHI
National Aerospace Laboratory
Yoshiro YOSHIOKA, Masayuki IKEDA and Moriyuki TAKAMURA
Fujitsu Limited

ABSTRACT

In a few years, a computer which processes CFD programs over 100 times faster than the Fujitsu VP400 and has a main memory capacity of more than 32G bytes will be required for CFD technology to play an important role in aerospace research and development.

A distributed main memory parallel processor is free from the memory throughput bottleneck which prevents the implementation of shared memory parallel processors with the necessary speed.

In the light of regular characteristics of CFD codes, a distributed memory parallel processor is likely to deliver the above-mentioned processing speed. Its characteristics include a physically distributed main memory which logically provides programmers with global and local memory views, processing elements with high speed RISC scalar units and high speed vector units with large capacity vector registers, and a crossbar network which interconnects a large number of processing elements. Such a processor can be suitably called the "Numerical Wind Tunnel".

This paper describes the basic main memory structure, system configuration, processing element, and interconnection network and communication mechanism of the Numerical Wind Tunnel.

1. はじめに

計算空気力学(CFD)が航空宇宙技術研究開発の中核基盤技術の地位を築くためには，ここ数年以内に，『主記憶容量が32GB以上，CFDプログラムの実効処理速度がVP400の100倍以上の計算機』が必要であるといわれている[1)]。

密結合多重ベクトル計算機のピーク性能は，多重度4ないし16にて最大で20GFLOPS前後の計算機が発売されているか，もしくは開発が伝えられているのが現状である。

この方式の延長線上で多重度を上げることにより，上記要件を満たす計算機を実現することは，処理速度に見合ったデータ供給能力をもつ主記憶の実現が最大の隘路となり，著しく困難である。

一方，メモリ分散配置型並列計算機では，主記

* 航空宇宙技術研究所
** 富士通株式会社

憶データ供給能力隘路から解放され，強力な局所的主記憶データ供給能力を実現することが可能となる。CFDの主力コードの特徴を考慮すれば，メモリ分散配置型構成にて上記要件を満たす計算機，即ち数値風洞の実現確度は極めて高い。

本稿では，数値風洞の基本方式，ハードウェア構成，要素計算機(PE)構成，主記憶構成，結合ネットワーク構成，およびPE間通信機構について論ずる。

2. 数値風洞基本方式の合理性

計算機の記憶構造および命令・データ列の観点より，数値風洞が採用する基本方式の合理性について論じよう。

2.1 メモリ分散配置方式の合理性

1つの主記憶に複数の計算機を接続した，いわゆる密結合多重計算機（図1）は，各計算機の処理装置の実現もさることながら，主記憶ー計算機間の結合装置の実現が全体の成否をきめる。その計算機がベクトル計算機であるとスカラ計算機に比べ，10ないし20倍の信号線を主記憶との間に接続する必要があり，その実現は一層困難になる。また，ベクトル計算機の多重度を8, 16, 32とあげていくに伴い，またマシンサイクルタイムを高速化するに伴い相乗的にその実現は著しい困難にぶつかる。

この臨界点に存在する，現状の密結合多重ベクトル計算機には，マシンサイクルタイム：6ns，8多重，主記憶：256MBでピーク処理速度が2.66GFLOPSのCRAY-YMP8，マシンサイクルタイム：2.9ns，4多重，主記憶：2048MBでピーク処理速度が22GFLOPSの日電SX3/44がある。YMP16は，マシンサイクルタイム：4ns，16多重，ピーク処理速度16GFLOPSで開発中と伝えられている。

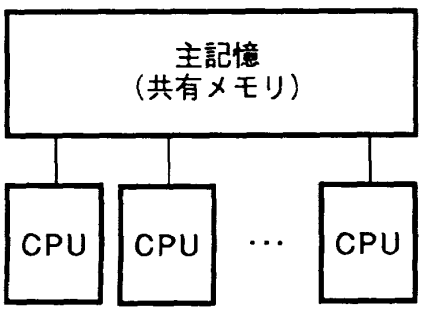

図1 密結合多重計算機

ここ数年の範囲では，マシンサイクルタイムは高々2倍の高速化，および多重度も2倍程度の拡張と筆者は予測している。この予測に基づくと，密結合多重ベクトル計算機においては，ピーク処理速度100GFLOPSを越えることは，ここ3, 4年内はほぼ不可能と言えよう。

主記憶アクセス時間の増大を甘受して仮にピーク80GFLOPS程度が実現できたとしても，以下にのべるスカラユニットのキャッシュ制御方式に関する重大課題に直面する。

即ち，ランダムまたは長いストライドアクセスをした場合のキャッシュミス頻発による極端な性能低下は多重化とは関係ないのでここではさておき，各計算機におけるスカラユニットやベクトルユニット，およびその他の主記憶を参照更新するユニットが主記憶更新を発信するたびに，多重度の数だけ存在しているキャッシュの旧データの無効化および新データの反映登録をハードウェアだけで効率よく制御することは極めて困難を伴う。

この猥雑から解放されたくキャッシュレスにすると，なお一層高速アクセス高データ供給能力をもつ主記憶の必要性から困難は深刻化するというディレンマに陥る。

共用更新可能領域を避けてキャッシュにデータを登録する等の制御をソフトウェアとの連携のもとで行うことがこの解決策として不可欠である。

話は前後するが，前述の広域アクセスに伴うキャッシュミス頻発を軽減するためにもキャッシュ制御ブロック単位を小さくし，かつソフトウェアの制御により必要なワードをプリロードするなどの改善が必要になる。

ソフトウェア制御のもと共用/私用領域を意識したこのような密結合多重ベクトル計算機は，その『共用』主記憶は『ローカル』記憶の様相を呈することになる。

主記憶への集中と，もっぱらハードウェアによるメモリデータ整合管理という構造欠陥を解決す

るには，

i）物理的に主記憶を分散配置し，かつ

ii）論理的に主記憶をローカル化

することが必要である。このような記憶構造をとる並列計算機をメモリ分散配置型並列計算機という（図2）。

PEの主記憶は，i)により，PE処理速度を支えるに充分なデータ供給および高速なアクセス時間を実現することが可能となり，また，ii)によりメモリデータ整合管理や主記憶アクセス競合から解放される。

この結果，メモリ分散配置方式に拠れば，主記憶を含めたPEの台数を大幅に増加することが可能になり高性能化が実現できる基盤が確立されるのである。

高性能実現の成否は，PEを相互結合するネットワークの実現に帰着することになった。

CFDの主力コードは，並列動作可能なプロセスが数100のオーダで存在しており，そのメモリアクセスは局所化可能というレギュラーな特徴をもっている。この特徴を考慮に入れ，メモリ分散配置型構成と強力なネットワークを基本方式とすることにより，上記性能要件の数値風洞の実現が開けるのである。

しかし，ソフトウェアの観点からは大きな課題を抱えることになる。分散配置を意識するプログラミングおよび言語プロセサ（コンパイラ），さらに分散OSなどいずれも未開拓で高度な技術を要するものばかりである。本論文集別稿に詳述されているように並列ソフトウェア技術も急進展しているのでその実現は明るい。よい土壌があれば，その地には豊穣が待っている。かくして数値風洞の基本方式は，メモリ分散配置方式と決めた。

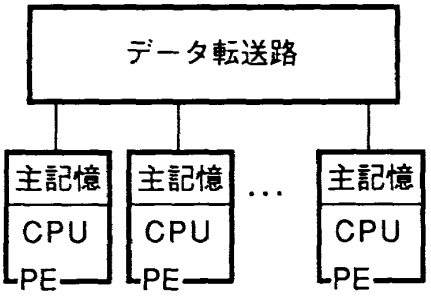

図2　メモリ分散配置型並列計算機

2.2　MIMDの合理性

1つのプログラムで主翼等航空機の多数の空力要素をそれぞれ独立に計算し境界値を調整するマルチドメイン法では，各空力要素の計算手続きが異なることがあるのは自然であるが，シングルドメイン法の場合でも両端の部分の計算手続きはその他の部分の計算手続きと異なるのが普通である。

さらに，数値風洞を複数のPEグループに分割し複数の異なるジョブを独立に走らせる機能が必要である。物理PEへのマッピングは，PE稼働率向上，ジョブ運用性向上および信頼性向上のために動的である必要がある。

このようなプログラミング要件および運用要件を数値風洞が備えることは，航空宇宙技術開発の開発現場において，数値風洞が実用されるために必須である。

SIMD，MIMDという命令列データ列の分類の観点から，この要件を満たす方式は，MIMDしかない。MIMDは，各PEで実行されているプロセス間で同期をとる必要が生じる。同期オーバヘッド削減のため，数値風洞は高速同期機構を備えた。

3.　ハードウェア構成

図3に数値風洞のハードウェア構成を示す。数値風洞は，コントロールプロセサ(CP)，演算処理を行うための100～200台程度のPE，およびそれらを相互接続する結合ネットワークより構成される。ネットワークは，クロスバネットワークである。

CPは，システム記憶(SSU)を介してVP2000シリーズのフロントエンドプロセサ(FEP)と接続され，プログラムおよびデータは，FEP-SSU-CP-

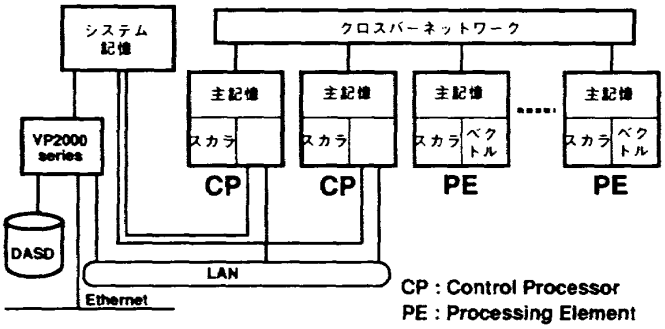

図3　ハードウェア構成

結合ネットワークを経由して，PEとやりとりされる。またCPは，標準LANとも接続される。

数値風洞が実行するジョブの起動/停止およびその依頼は，FEPがおこなう。CPは，依頼受付，数値風洞の管理，PE割当およびPEジョブの起動/停止をおこなう。

SSUは，CPからはDASDのキャッシュとして見え，プログラムおよびデータのプレステージングおよびディステージングの機能をはたす。この結果PE稼働率を98%程度に維持できる可能性がシミュレーションで検証されている。

4. 要素計算機（PE）構成

CFDの主力プログラムにおける大部分の処理は，2次元以上の並列性をもっている。1次元の並列性をPE台数方向の並列性に活用しても依然としてベクトル並列性がのこっている。ベクトル処理技術は，ハードウェアおよびソフトウェアともに成熟し充分にコストパフォーマンスの高い性能を得ることが可能であること等を考慮して，PEにはベクトル処理機構を装備した。

PEは，図4に示すようにスカラユニット，ベクトルユニット，主記憶，PE間通信機構とで構成される。

4.1 スカラユニット

スカラ計算機は，1マシンサイクル当たり複数命令の実行，即ち，1命令の実行サイクル数(CPI)を1以下にする技術の実現に向かって，長足の進歩をとげつつある。1989年2月発表のIntel社860

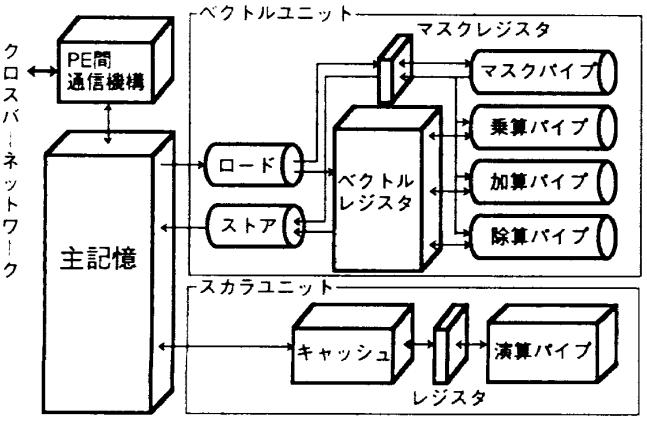

図4 要素計算機（PE）

チップ（デュアルモード）[2]がまず先鞭をつけ，同年10月には4命令同時実行可能のCPUを搭載したIBM社のRS6000ワークステーション[3]が発表され，製品化された。

これ以外のメーカにおいても，複数命令同時実行のプロセサの開発が精力的に推進されているのが現状である。

ベクトル計算機ではスカラユニットの性能が実効性能を支配する重要要素の1つであることは，数値風洞のPEについても同様に正しい。1PEで1ジョブ実行の場合には尚更である。これを考慮に入れ，数値風洞PEのスカラユニットは，

i) 1サイクルごとに複数命令を同時に実行可能なハードウェア機構を具備する，
ii) メモリアクセス命令，浮動小数点演算命令，およびベクトル命令は非同期実行制御を行う，
iii) ベクトルユニットと密に結合できるインタフェースをもつ

構成をとる。

4.2 ベクトルユニット

大容量のベクトルレジスタとロード・ストアパイプラインをもつ演算パイプライン方式を採用した。

演算パイプラインの効率の良さ，ベクトルレジスタ制御方式の効率の良さ，これらハードウェア資源を最大限活かすコンパイラ技術が高度化・洗練化されている，その総合としてCFDプログラムにたいして極めて高い実効性能を発揮しているという事実を考えても，本方式が最適の構成である。

その構成は，図4に図示のごとく，乗算，加算，除算，マスクの各パイプライン1本，およびロード，ストアの各パイプライン1本，およびマスクレジスタ/ベクトルレジスタより構成される。

ハードウェアシミュレータでCFDコードを走行させて，性能決定要因のうち最も支配的であるパイプラインのスループット・多重度，立ち上がり時間，およびベクトルレジスタの構成・アクセス制御方式のパラメトリックスタディをおこない，詳細にわたり構成の最適化を行った。

表1 PE諸元

スカラユニット	キャッシュ容量	64KB
ベクトルユニット	ベクトルレジスタ容量	128KB
	演算パイプライン性能	1.6GFLOPS
	ロード・ストア性能	12.8GB/S
主記憶	容量	256MB

PEの諸元を表1に示す。

5. コントロールプロセサ（CP）構成

CPは，PEよりベクトルユニットを除外し，SSUおよび標準LANのインターフェースを追加した構成をとる。従って，スカラユニット，主記憶，対クロスバインタフェースは，PEのそれらと論理的および物理的に同一である。CPハードウェアの開発およびこれに載るOS開発の観点よりみても同一にする意義は大きい。

数値風洞の信頼度およびデータ転送能力の向上のため複数台のCPを搭載することができる。

6. 主記憶構成

前述のとおり，主記憶は各PEおよびCPに配置されている。PEにおいてはスカラユニット，ベクトルユニットおよび通信機構が，またCPにおいてはスカラユニット，通信機構，SSUおよびLANインタフェースが主記憶のアクセス要求源となる。

バンクサイクル時間の大幅な短縮により，主記憶の小型化，その結果として高速アクセス高速データ供給能力を実現した。

プログラマーがローカル配列を用いてプログラムを記述すれば，スカラユニットおよびベクトルユニットの各種レジスタと直接に高速アクセス高速データ供給が可能という分散主記憶の優位性が有効に働き，高い実効処理速度が達成できる。

しかし，プログラム記述容易性の観点より，複数またはすべてのPEからアクセスする配列をも許容する必要があると考え，図5に示すとおり，分散主記憶の上に論理的に共有（グローバル）主記憶をローカル記憶と共存させることも可能な機構とした。

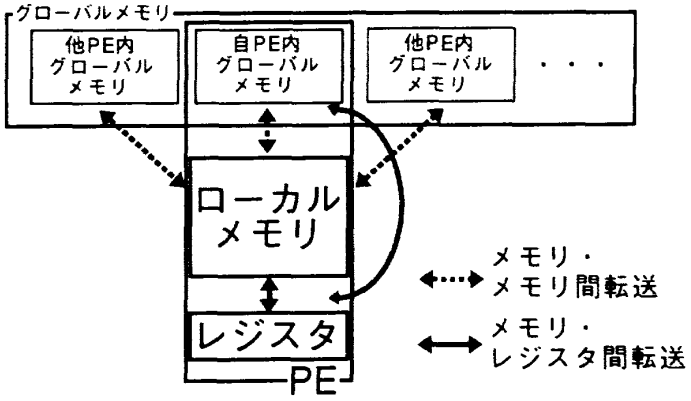

図5 メモリ構成

グローバル配列の分割区分は，それを使用する手続のPEの主記憶に一致させるよう配置するのは，コンパイラの責務ではあるが，自主記憶は他主記憶に比べて一層高速なアクセスおよびデータ供給能力をもつので，手続とデータの一致は高性能化に不可欠である。

グローバル配列とローカル配列とが同一の記憶位置を指し示す機能が有効な場合は，グローバル配列はローカルと同一の高速アクセスが可能となる。

このように主記憶分散配置方式の基盤の上に，高性能特化と記述性の追求が柔軟にできる機構が確保されている。

7. 結合ネットワーク

結合ネットワークには，解法，転送スケジューリングおよび運用より厳しい要件が課せられる。

Rq 1（解法からの要件）：自然現象はすべて隣接作用の原理のもとで推移するのだから，計算解法としては隣接原理に基づいた陽解法のみを効率的にとりあつかえればよいという主張はあまりにも短絡的である。現在航空宇宙技術開発におけるCFDプログラムに採用されている最有力のIAF解法においては，I軸方向→J軸方向→K軸方向というようにスイープ方向を交代して計算をすすめる。スイープ方向に対しては逐次処理，その他の2軸に対しては並列処理が可能である。スイープ方向の計算領域が各PEに分担されている（即

ちスイープ方向とPE分割方向が同一の）場合は，各PE主記憶がもっている大部分の配列データの総配置換え（転置）が計算に先立って必要になる*。

転置をおこなうと，計算に参加している全PEの間で大量の配列データが結合ネットワークを『一斉に』移動する。転置によるオーバヘッドを削減できる強力な結合ネットワークが，主力解法適合性から要求される。

Rq 2 （データ転送スケジューリング最適化からの要件）：転置をはじめ，PE間データ転送に起因する並列化効率低下を防止する必要がある。そのため転送の計算の陰に隠す等スケジューリングを最適化するためには，PE間等距離・対等・競合なし結合が要求される。

Rq 3 （PE割当の同型性からの要件）：

i) 計算需要に対応し，PEを切り離しまたは組み込む（動的）
ii) 故障PEを切り離す/新PEを組み込む（動的）
iii) 緊急ジョブ実行のPE確保（動的）
iv) 計算規模に対応した複数のPEグループ（ジョブクラス）の設定

など運用性および可用性の観点より，論理PEから物理PEへのマッピングは動的である必要がある。

動的割当てにおいて，最適化された転送スケジューリングを維持するためには，いかなるマッピングを行っても同一の結合トポロジを持つ必要がある。即ち，いかなるPEの部分集合をとってみても全体集合と同一の結合トポロジを持つこと（同型性）が，結合ネットワークに要求される。

Rq 4 （スケーラビリティからの要件）：

PE台数はリニアに増加する結合ネットワークが，所要トータル性能，計算規模，価格，環境付帯設備等への適合性から望ましい。PE粒度が大になると益々リニアスケーラビリティがよい。

数値風洞には最低限，以上の要件Rq1，Rq2，およびRq3を満たすことが不可欠である。この条件をみたす結合ネットワークは，数多くある結合ネットワークのなかで完全結合とクロスバしか候補にならない。クロスバを2進木状に結合したり[4]，クロスバを2次元結合[5]した組み合せネットワークも提案されているが，いずれも上記条件を満たさない。

完全結合とクロスバについて物量一定のもとで，接続できるPE台数を検討した結果，

i) 完全結合は，PEの送り口および受け口の物量制限により接続できるPE台数が，クロスバ結合にくらべ1/10のオーダに減少する，
ii) 1ビットデータ幅完全結合にしても，完全結合の3，4倍程度しか台数が増加しない，
iii) PEの送り口および受け口の物量削減のため，セレクタ付き完全結合にすると，これはクロスバになる。

以上の結果より，数値風洞の結合ネットワークはクロスバで実現した。数値風洞のクロスバの諸元を表2に示す。

クロスバにおいて転置を行う場合，

i) PE_i は PE_{i+1} からはじめて昇順にデータを転送するスキュー方式，
ii) PEペアをたすき掛け状にデータを中継しつつ転送するバタフライ方式

が有力である。図6に示すとおり，転送のスタートアップ時間にくらべてデータ転送の正味の時間が大きいほど方式i)が，またデータ転送の正味の時間が小さいほど方式ii)が高速である。データ転送量に対応して転送方式を選択する必要がある。

方式i)にて転置を行ったときのクロスバの稼

表2 結合ネットワーク諸元

トポロジ	クロスバ
転送速度	PE当り800MB/S
PE台数	100から200台程度

* 転置をしないで，計算結果を隣のPEに『逐次』伝播していく方法もある。各PEが一斉に計算開始できないため性能が極端に落ちて採用できない。

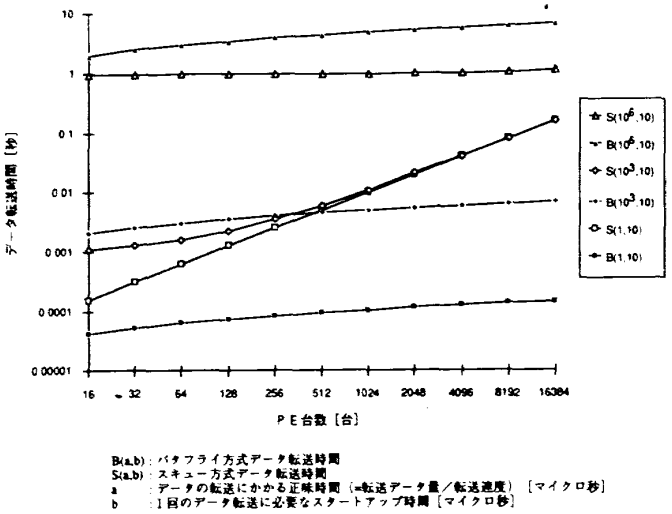

図6 転置のデータ転送時間

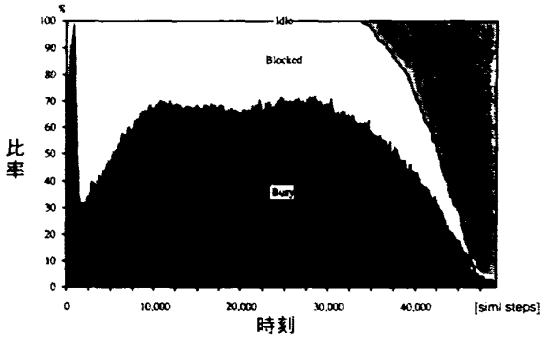

図7 クロスバ稼働状況

動状況のシミュレーション結果を図7に示す。

また，SUM（総和），MAX（最大値）/MIN（最小値），ブロードキャスト（指定複数PEへの放送）などのPEにまたがるグローバル演算は，バタフライまたは2進木演算として，クロスバを使用して実行される。

8. PE間通信機構

PE_iの主記憶データとPE_jの主記憶データの送受を行うことによりPE間の通信を実現する。

送信（ライト）動作は：

自PE (PE_i)はPE_iグローバルまたはローカル記憶から要求およびデータを読みだし，他PE (PE_j)へ発信し，PE_jはPE_jローカル記憶に書きこむ。

受信（リード）動作は：

自PE (PE_i)はPE_iローカル記憶より要求を読みだし要求を他PE (PE_j)に発信し，PE_jはPE_jグローバルまたはローカル記憶データをよみだしPE_iへと発信し，PE_iはPE_iローカル記憶に書き込む。

このようなPE間通信の起動，制御および終結動作をPEスカラユニットが行っていると，PEのスカラおよびベクトル演算との同時実行ができない。通信と演算の同時実行は並列効率向上のため最優先要件であり，そのためにはPE間通信制御専用プロセサを設け，通信からスカラユニットを解放することが必須である。PE間通信プロセサを設けることによりPE間データ転送が非同期化でき，計算に必要な他PEデータの要求を十分早めに起動しておくこと，または他PEが十分早めにデータを送りつけておくことによりデータ到着待ちでPEがアイドルになることを回避できる。転送のスケジューリング最適化をユーザまたはコンパイラが存分にできる構成がとられている。

PE間通信プロセサの主記憶データ転送は，連続，一定ストライド，リスト，サブ配列の各モードを実現している。また論理アドレスから物理アドレスへの変換も通信プロセサのハードウェアで実現している。

9. おわりに

本稿では，数値風洞の基本方式，ハードウェア構成，要素計算機(PE)構成，主記憶構成，結合ネットワーク構成，およびPE間通信機構について設計思想，その実現および諸元を述べた。

PEの論理的および物理的な切り離し（縮退），組み込み（復元）を数値風洞の稼動中に動的に実行する機構などの高信頼化機構，またCFDプログラム性能評価，半導体技術，実装冷却技術については紙面の制限で割愛する。

マイクロプロセサを要素計算機とする並列計算機の市販がすすんでいる。CM2, iPSC/860, nCUBEなど世にある並列計算機を数値風洞に採用しない主たる理由は，以下の5項目である。

D1) 実効性能が上記性能要件を満たさない。
D2) ネットワークがIAF解法に適合していない。
D3) 転送スケジューリングの最適化が極めて困難。
D4) 運用性および可用性が満足できない。

D5) 要素計算機がスカラ計算機（キャッシュ付き）であるため広域データアクセスが発生すると効率が1%以下にまで激減するといった脆弱な計算機である。

数値風洞が稼働のあかつきには，これを利用する研究者と技術者は，ベクトル化率とともに並列化率の高いプログラムを使用することにより，数値風洞の高速性を享受されることを祈念して止まない。

なお，本稿は平成3年2月より開始された航技研・富士通の共同研究『数値風洞の開発研究』に基づいている。

参 考 文 献

1) 三好 甫："CFDの推進に必要な計算機性能"，第8回航空機計算空気力学シンポジウム論文集，SP-13, pp.1-26, 1990年9月.

2) i860 64-BIT MICROPROCESSOR, Intel Corp. Feb. 1989.

3) Groves, R et al : "An IBM Second Generation RISC Processor Architecture", Proceedings IEEE ICCD' 89, pp.134-137, Oct. 1989.

4) Butner, S : "A Fault-Tolerant GaAs/CMOS Interconnection Network for Scalable Multiprocessors", IEEE J. of Solid-State Circuits, Vol.26, No.5, May 1991.

5) 田中輝雄，面田耕一郎："高並列計算機による空気力学シミュレーションの構想"，第8回航空機計算空気力学シンポジウム論文集，SP-13, pp.99-108, 1990年9月.