計算機アーキテクチャ　107－16
（１９９４．　７．　２２）

並列コンピュータ Cenju-3 のアーキテクチャ

広瀬 哲也　　加納 健　　丸山 勉　　中田 登志之
浅野 由裕†　　稲村 雄†

NEC C&C システム研究所
†NEC 技術情報システム開発

本稿では，新たに開発した並列コンピュータ Cenju-3 のアーキテクチャを概説する．Cenju-3 は，VR4400 を要素プロセッサとする分散メモリ方式の並列コンピュータで，256 プロセッサまでを接続可能である．

要素プロセッサは，実装効率良く性能の向上を図るため VR4400 と 1MByte の 2 次キャッシュを一体化した MCM(Multi Chip Module) を用いた．プロセッサ間ネットワークを構築するに当たって，最大 256 プロセッサをターゲットに最適化を図り，４×４のクロスバスイッチによる多段接続網を用いた．多段接続網は，プロセッサ間が等距離であるという優れた特徴を持ち，256 接続までに置いてはメッシュ等に対してハードウェア量も少い．プロセッサ - ネットワークインタフェースとして専用のハードウェアを開発し，並列化のボトルネックとなる通信速度向上を図った．

Architecture of a Parallel Machine : Cenju-3

Tetsuya Hirose　Yasushi Kanoh　Tsutomu Maruyama　Toshiyuki Nakata
Yoshihiro Asano　Yu Inamura

C&C Research Laboratories NEC Corporation,
†NEC Scientific Information System Development
1-1, Miyazaki 4-Chome Miyamae-ku Kawasaki Kanagawa 216 Japan

In this paper, we present design consideration and machine architecture of a distributed memory parallel machine: Cenju-3.

The processing element(PE) consists of a VR4400 with 1MByte secondary cache memory, which are implemented on a multi-chip-module, in order to realize a compact and high performance PE.

In designing the inter-processor network, we tuned the architecture for up to 256 processor system. PEs are connected through a multi-stage interconnection network, that consists of 4x4 crossbar packet switches.

We also developed special processor-network interface hardware that is used to speed-up the inter-processor communication.

1 はじめに

本稿では，新たに開発した並列コンピュータCenju-3のアーキテクチャについて述べる．Cenju-3は，VR4400を要素プロセッサ(PE)とする分散メモリ方式の並列コンピュータで，256プロセッサまでをスケーラブルに接続可能である．

設計に当たって，最大256プロセッサをターゲットにプロセッサ間ネットワークの最適化を図った．プロセッサ間ネットワークは，4×4のクロスバスイッチによる多段接続網である．Cenju-3では，プロセッサ-ネットワークインタフェースとして専用のハードウェアを開発し，並列化のボトルネックとなる通信速度の向上を図った．

以下，第2章でアーキテクチャ概要，第3章で設計方針，第4章，第5章，第6章で要素プロセッサ，ネットワーク，ネットワークインタフェースについて述べる．次いで，第7章でホストプロセッサとのインタフェース，第8章で故障診断システムについて，第9章で性能評価について述べる．

2 Cenju-3概要

Cenju-3(64PEシステム)のハードウェア構成を図1に示す．256PEシステムの最大性能は38.4 GIPS，12.8 GFLOPS，総メモリサイズは16GBである．

Cenju-3システムは，2つの要素プロセッサを搭載したプロセッサカード(PEC)，2種類のネットワークカード(NWC0，NWC1)，Cenju-3本体とホストプロセッサ(EWS)とのインタフェースカードであるIOカード(IOC)，ホストEWSに付加するAPカード(APC)で構成される．16PEシステムは，8枚のPECと，NWC0，NWC1，IOC各1枚で構成され1つのケージ内にコンパクトに実装される．64PEシステムは1筐体，最大の256PEシステムは4筐体で構成される．

Cenju-3では，EWS4800をホストプロセッサ(以下，単にホストと呼ぶ)として用いる．ホストは，IOプロセッサとしての機能(ファイルシステムの提供，ネットワーク(Ethernet等)との接続)と，プログラムの作成・コンパイル，実行の指示・制御，デバッグ支援等の機能を持つ．

3 設計方針

Cenju-3は，Cenju2[1]のアーキテクチャを基本としてアプリケーションレベルの互換性を保っている．

今回Cenju-3を開発するにあたって，次のような方針で設計を行なった．

- 要素プロセッサ単体性能重視
  並列計算機は高速処理を目的とする．より高い性能を実現するには，必然的に要素プロセッサの単体性能をより高いものとしなければならない．

- 最大256プロセッサとしネットワークを最適化
  既に，数千プロセッサのシステムもあり，最大性能(プロセッサ単体性能×プロセッサ数)が注目されがちであるが，並列化による処理速度向上は一般にある程度の台数で飽和する．我々は，電気系CADやプラズマシミュレーションなどのアプリケーションの評価から，現段階での問題の並列度はたかだか数百程度と考える．Cenju-3では，最大256プロセッサをターゲットとし，ネットワークの最適化を図った．

- プロセッサ性能と通信性能のバランス化
  並列計算機では，プロセッサ性能と通信性能のバランスをよく考慮して設計を行なうことが重要である．通信性能としてネットワークのハードウェア転送速度が強調されがちであるが，実際にはソフトウェアでのオーバヘッドがかなりの部分を占めている．Cenju-3では，ネットワークに過大なコストをかける事を避けて，コストパフォーマンスよく出来る限りの高性能化を狙い，ネットワークの性能を決定した．

- ネットワークインタフェース強化
  プロセッサ間通信のオーバヘッドを出来る限り少くすることが高い性能を実現するカギとなる．Cenju-3では，専用の通信用ハードウェアを開発し，転送速度の高速化，ソフトウェアオーバヘッドの軽減を実現した．

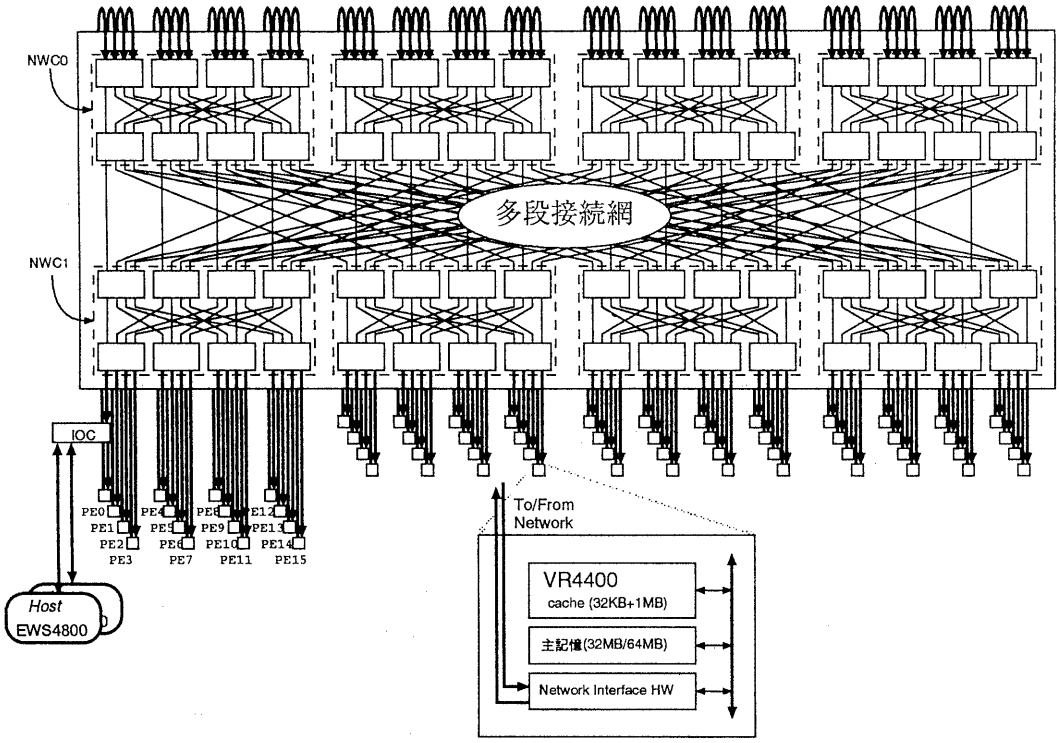

図 1: Cenju-3(64PE) のハードウェア構成

- 故障診断機能のサポート
  Cenju-3 では，専用の診断パスによる診断機構を用意した．これによって，システムの初期診断，故障診断，ソフトウェアのデバッグ支援等を行なう．

4 要素プロセッサ

並列処理の効果を十分に引き出すためには，ある程度の大きさの問題を扱うことで各プロセッサでの処理粒度を大きくする必要がある．より高い性能を実現するには，必然的に要素プロセッサの単体性能をより高いものとしなければならない．近年汎用マイクロプロセッサの処理速度向上はめざましく，1～2年で倍に延びている．一方，専用のプロセッサを開発するためには，多くの時間，コストがかかる上，コンパイラ等の開発も必要となる．現在のところこれらのコストを考えると，使用可能な最も速い汎用プロセッサを要素プロセッサとして並列計算機を実現することが，もっとも速い並列計算機を実現することになる．

Cenju-3 では，VR4400 をプロセッサに採用した．VR4400(75MHz) は，150MIPS，50MFLOPS の処理能力を持つ．

実効処理能力の向上のため，MCM(Multi Chip Module) の技術を用いて，1MB の 2 次キャッシュをプロセッサと一体化して実装した．MCM は，基板特性も良く配線長も短くできることから，最小遅延時間の 2 次キャッシュシステムを実現することが出来る．また，MCM は実装効率も非常に高く，VR4400 単体とほぼ同じスペースにコンパ

クトに実装される．MCM化によって実装効率が上がったため，プロセッサ当たり64MBの主記憶を実装でき，かつ2つの要素プロセッサをVMEトリプルハイトの標準カード1枚に実装する事が出来た．

要素プロセッサのブロック図を，図2に示す．

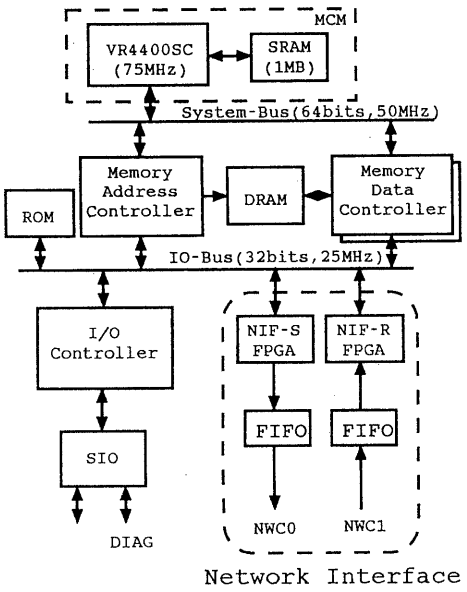

図 2: 要素プロセッサのブロック図

プロセッサ - メモリ間は64bit，50MHzのSystem-Busで接続される．専用ネットワークインタフェース(第6章参照)は，IO-Bus(32bit, 25MHz)に接続される．

5 プロセッサ間ネットワーク

5.1 トポロジー

我々は，256PEまでのプロセッサを接続するのに最も適したネットワークは多段接続網であると考える．多段接続網は，PE間が等距離でありアプリケーションのプログラミングの容易さや柔軟性，デッドロックの原因となるPE間の閉じたパスが無いという優れた特徴を持つ．さらに，表1，表2，表3に示した通り，2D,3Dメッシュ，3Dハイパークロスバと比較してハードウェア量も少い．

Cenju-3では，4×4のクロスバスイッチ(以下，SU (Switching Unit)と呼ぶ)による多段網を用いた．また，次のような実装方法の工夫から多段網の小型化，スケーラブル化を実現した．

- 8個のスイッチングユニットを1カードに実装(16 × 16スイッチ / カード)
- PEとの接続はBP(Back Panel)を使用

この結果，16PEにつきネットワークカード2枚，ケーブル配線量はプロセッサ数と同じ(オーダーN)と非常に少なく抑えられている．

表 1: Switch数

プロセッサ数	64	256	512	N
2Dメッシュ	64	256	-	N
3Dメッシュ	64	-	512	N
3Dハイパークロスバ	$48_{(4x4)}$ +64	-	$192_{(8x8)}$ +512	$3\sqrt[3]{N}$ +N
多段(4x4)	48	256	-	$N/4 \times \log_4 N$
多段(8x8)	16	-	192	$N/8 \times \log_8 N$
Cenju-3多段	64	256	-	$N/16 \times 16$

表 2: Link数

プロセッサ数	64	256	512	N
2Dメッシュ	256	1024	-	4N
3Dメッシュ	384	-	3072	6N
3Dハイパークロスバ	384	-	3072	$3N \times 2$
多段(4x4)	256	1280	-	$(\log_4 N + 1)N$
多段(8x8)	192	-	2048	$(\log_8 N + 1)N$
Cenju-3多段	320	1280	-	5N

一方，段数はプロセッサ数に寄らず4段である．しかし，表4から解るようにSU1段の遅延時間はわずか250nsであり，ソフトウェアでの処理を考えると，現状では無視できる．また，16PEから128PE構成の場合はカード間の接続が通常の多段

表 3: PE 間 最大距離

プロセッサ数	64	256	512	N
2D メッシュ	8	16	-	$\sqrt{N}$
3D メッシュ	4	-	8	$\sqrt[3]{N}$
3D ハイパー クロスバメッシュ	7	-	7	7
多段 (4x4)	3	4	-	$\log_4 N$
多段 (8x8)	2	-	3	$\log_8 N$
Cenju-3 多段	4	4	-	4

網より多くなり (64PE で 4 本，図 1参照)，パスの競合が避けられより強力な接続となる．

5.2 転送性能

Cenju-3 では，プロセッサ性能とのバランスを考慮してネットワークの設計を行なった．この際のプロセッサ - メモリ間のスループット，バスのスピード，ソフトウェアでの処理を考慮してネットワークの性能を決定した．

Cenju-3 の PE における実行メモリスループットは，128MB ～ 160MB/s である．

ネットワークから書かれたものは必ず CPU がアクセスするデータであるので，プロセッサからの書き込み / 読み出し >= ネットワークからの書き込み / 読みと考えられる．従って，ネットワークアクセスが最大の時で，ネットワークへの書き込み 32MB ～ 40MB/s ネットワークからの書き込み 32MB ～ 40MB/s CPU からのアクセス 64MB ～ 80MB/s となり，ネットワークのスループットはメモリスループットの 1/4 あれば十分である．従って，ネットワーク速度は 40MB/s でバランスがとれる．

また，専用ネットワークインタフェース (第 6章参照) は，最大バンド幅 100MB/s の IObus に接続されるので，それ以上のネットワークスピードがあってもバスネックで性能を活かす事が出来ない．

そこで，Cenju-3 ではネットワークの転送速度を 40MB/s とした．表 4にネットワークの諸元を示す．この SU では，パケット交換方式でウォームホールルーティング方式を用いており，プロセッサ間通信の遅延時間が非常に短く抑えられている．

表 4: ネットワークの諸元

トポロジ	多段 (ベースライン)
交換方式	パケット交換
転送ワード幅	16bit
動作周波数	20MHz
スループット	40MB/s
遅延	5clk(250ns)/stage
パケット長	可変:8 ～ 518 byte
SU	4 × 4 クロスバ

6 プロセッサ - ネットワーク・インタフェース

汎用の DMA コントローラではネットワークのハードウェアに見合う十分な性能を実現することができなかったため，Cenju-3 では専用のインタフェースハードウェアを開発した．

この専用インタフェースは，送信インタフェースと受信インタフェースで構成され，メッセージ通信と DMA 通信をサポートする．メッセージ通信では，受信側 PE で予め指定されたバッファ領域にデータが格納される．このメッセージ受信バッファは 2 組用意されていている．DMA 通信では，送信側で指定したアドレスに直接データが書き込まれる．

プロセッサと独立に，送信インタフェースはキューイングされた送信データを自動的にパケットに分割してネットワークに送り出す．一方，受信インタフェースは届いたパケットヘッダを自動的解析してデータをメモリに格納し，必要に応じ各種割り込みを発生する．これらの機能により，ソフトウェアによる処理が大幅に軽減される．

ここで，パケットに分割される前のソフトウェアが扱う通信データの単位を Parcel，実際に Cenju-3 ネットワークを通る通信データ単位を，単に Packet と呼ぶ．

送／受信インタフェースは，それぞれFPGA (Field Programmable Gate Array)で実現される．これらは，メモリへのバーストアクセス機能を持ち，ネットワークの最大性能40MB/sでパケットを送受信可能である．

6.1 送信インタフェースの機能

送信インタフェースでは，キューに継れたParcelを自動的にネットワークに送り出す処理を行なう(図3)．

- 送信データの自動分割
  Cenju-3のネットワークでは，パケットのサイズは512Byteに制限されている．送信インタフェースは，最大128KBのParcelを自動的にPacketに分割して送信する機能をもつ．
- 送信データのキューイング
  最大255個までのParcelをキューにつなぐことが可能．プロセッサの介在無しに自動的にハードウェアがキュー中で指定されたデータ領域を送信する．従って，128KB × 255個 = 約32MBまでのデータを自動的に送信することができる．
- 優先度付きの送信キュー
  送信キューには，優先度の高いキューと低いキューの2種類が用意されている．送信インタフェースは，優先度の高いキューのParcelの処理を必ず先に行ない，優先度の高いキューが空の場合にのみ優先度の低いキューのParcelの処理を行なう．これにより，急行メッセージ等を実現することが可能である．
- メッセージ，DMA送信終了割り込み
  この割り込みが指定されたメッセージ，DMAを送信し終ったら割り込みを発生する．

6.2 受信インタフェースの機能

受信インタフェースでは，届いたパケットのヘッダを解析し自動的に指定されたメモリに書き込む処理を行なう(図4)．

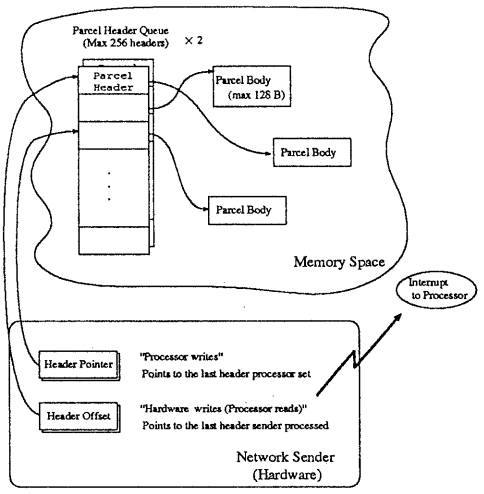

図 3: 送信インタフェースの機能

また，必要に応じて，以下のような割り込みをサポートする．これらの割り込みをかけるか否かは送信側で指定する．

- メッセージ，DMA到着割り込み
  この割り込みを指定されたメッセージ，DMAを受信したらデータを格納する前に，割り込みを発生する．
- メッセージ，DMA受信終了割り込み
  この割り込みを指定されたメッセージ，DMAを受信したらデータを格納し終った時点で割り込みを発生する．メッセージを複数のパケットに分割して送信する場合には，一番最後のパケットが到着した時点で割り込みが発生する．
- バッファオーバーフロー割り込み
  各バッファの約半分もしくはほぼ全てが使用されるとプロセッサに割り込みを発生する．

7 ホストプロセッサ・インタフェース

Cenju-3とホストプロセッサ間は，IOカード(IOC)を介して接続され(図1参照)，ホストはソ

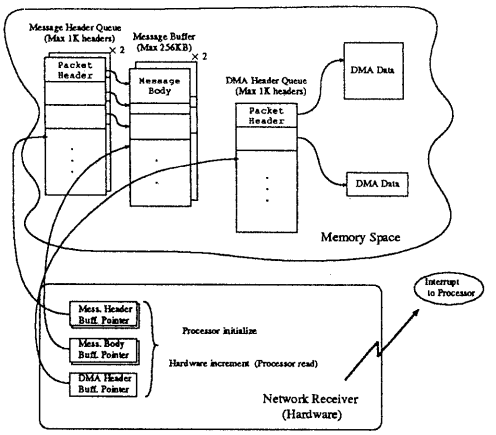

図 4: 受信インタフェースの機能

フトウェア的にネットワーク上の1つのプロセッサとして見える．ホスト側には，PE に於けるネットワークインタフェースと同様の機能を果たす AP カードが付加される．

Cenju-3 では，ネットワークの入出力ポートにそれぞれ要素プロセッサを接続し，かつホストプロセッサとの通信路を提供するために，16PE に1枚 IO カードが実装されている．IO カードは，ネットワークの一部として要素プロセッサと2台のホストプロセッサ間のネットワークスイッチの役割を果たす．転送速度は 40MB/s で，ホストは各 PE と同等に通信を行なう事が出来る．すなわち，16PE につき2台のホストプロセッサ (IO プロセッサ) を結合することが可能となる．(但し，現在はソフトウェアが未対応である．)

8 診断機構

単一プロセッサ計算機に較べ並列計算機の場合，計算機の内部状態が解り難くプログラミング (デバッグ) が難しい．例えば，ネットワークの1部が閉塞状態になった場合，それが次々に波及し，ネットワークによる通常の通信は使えなくなる可能性が高い．また，ハードウェア量の多さから故障確率も高い．

Cenju-3 では，ネットワーク上のハードウェアの障害を検知するため，またソフトデバッグ支援のために，専用の診断パスによる診断機構を用意している．これは，アドレス付きのシリアルライン (RS232C) で実現されている．この診断パスは，全ての PEC，NWC，IOC とホストを結合しており，ホストから診断を行なうことで計算機の内部情報を得る事が可能である．

この診断機構によって，NWC，IOC の各データパスに於ける ネットワークの閉塞状態，データ化けが起きたことを示すパリティエラー，各 PE の状態 (初期化時の診断結果，実行時エラーなど) などの情報が得られる．また，このパスを通して PE 上の特定プログラムの実行も可能である．

9 評価

現在，システムソフト / ライブラリ関数のチューンアップを行なっている段階であるが，現状での性能評価を行なった．

9.1 通信レイテンシー

まず，OS カーネルレベルでの通信のレイテンシーを測定した．測定結果は，12 μ s である．

これは，送信側プロセッサでの Parcel Header の作成 (キャッシュ上)，キャッシュの write-back，メモリコントローラの buffer-flush，送信ネットワークインターフェースへの Parcel Header 登録 (外部アクセス) の処理．送信ネットワークインタフェースでの，Parcel Header の読み込み (メモリアクセス)，Parcel Body の読み込み (メモリアクセス) の処理．ネットワークのハードウェア遅延．受信側プロセッサのネットワークインタフェースによる Parcel Header の書き込み (メモリ) の処理時間の合計である．

現在システムソフトのチューンアップ中であり，ユーザプロセスレベルからのネットワークの性能評価は今後の課題である．

9.2 アプリケーションによる評価

次に，流体解析のアプリケーションによる評価を行なった．非圧縮 Navier-Stokes 方程式解法の1つである SMAC(Simplified-Marker-and-Cell) 法を，領域分割法を用いて並列化し，Cenju-3 で実行した [4][5]．ここでは結果のみを図 5に示す．72 × 72，144 × 144 等は領域分割の格子点数で問題の規模を表す．

この結果から，問題規模が十分大きい場合には，プロセッサ数に見合う速度向上が得られる事がわかる．特に，216 × 216 の場合は，16 プロセッサまではプロセッサ数以上の速度向上が確認された．これは問題規模が大きくなるにつれ計算に対する通信コストが相対的に小さくなることに加え，プロセッサを増やすと 1 プロセッサ当たりの処理が小さくなりキャッシュヒット率が向上することによるものと考えられる．36 プロセッサ以上の場合は，キャッシュヒット率が飽和していると考えられ，あまり良い台数効果は得られない．しかし，この問題においては 216 × 216 という規模は特に大きなものではなく，より大きな問題を解く場合には，さらに台数効果が期待される．

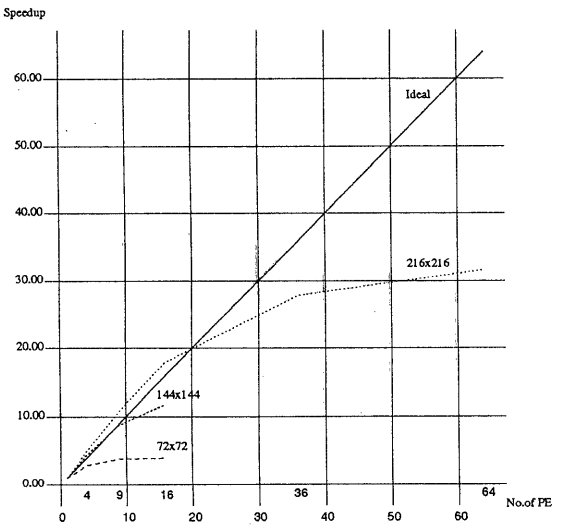

図 5: 流体解析における速度向上

10 おわりに

本稿では，Cenju-3 を紹介した．Cenju-3 では，要素プロセッサに最新のマイクロプロセッサ VR4400，プロセッサ間ネットワークに多段接続網を用い，最大 256 プロセッサまでをスケーラブルに接続可能である．最大構成時で 38.4 GIPS，12.8 GFLOPS という高性能を実現している．

Cenju-3 では，要素プロセッサ単体性能を重視し MCM(Multi Chip Module) による大容量かつ高速の 2 次キャッシュを搭載した．アプリケーションによる評価により，キャッシュの効果が確認された．

256 プロセッサまでをターゲットにネットワークの最適化を考え，非常にコンパクトなネットワークを実現した．

また，プロセッサ間通信におけるソフトウェア処理の軽減のため専用のネットワークインタフェースハードウェアを開発しプロセッサ間通信の高速化を図った．現在システムソフト / ライブラリ等の高速化をおこなっており，今後より詳細な評価を行なう予定である．

参考文献

[1] 松下，山内，中田，小池: 並列マシン Cenju2 のアーキテクチャ，情報処理学会 計算機アーキテクチャ研究会資料 Vol.92，No.64，pp.17-23，Aug. 1990

[2] N. Koike: NEC Cenju-3: A Microprocessor-Based Parallel Computer，Proc. of IPPS，Apr. 1994，pp.396-401

[3] 財団法人 新世代コンピュータ技術開発機構: 並列アーキテクチャに関する技術開発動向調査研究報告書，Apr. 1994

[4] 村松，鷲尾，土肥: 並列マシン Cenju2 による非圧縮性流体解析，情報処理学会 研究報告 Vol.94 No.22 pp9-16 Mar. 1994，

[5] S. Doi，K. Muramatsu，T. Washio，T. Nakata: Implementing a CFD Solver on Cenju-3 Parallel Computer，Proc. of PARALLEL CFD '94, pp31-36