2004－ARC－158　(6)
2004／5／14

# RHiNETプロジェクトの最終報告

大塚智宏[†]　渡邊幸之介[†]　北村聡[†]
鯉渕道紘[†]　山本淳二[††]　西宏章[†]
工藤知宏[†††,☆]　天野英晴[†]

RHiNET プロジェクトは，オフィス等で利用されている PC の余剰計算能力を利用した低価格高性能な並列処理環境を構築することを目的として実施された．このために，LAN と SAN の双方の利点を併せ持つ新しいネットワーククラス LASN を提唱し，その実装形態としていくつかの RHiNET システムの試作を行った．

プロジェクトは 2001 年度で終了し，LASN 技術を商用レベルまで引き上げることこそできなかったが，CMOS で世界最高レベルのスイッチや，LASN 用のネットワークインタフェースチップの開発に成功し，評価を取ることのできる実験機を実装した．また，これら直接の成果以外にも，多くの副次的な成果を得ることができた．

# Final Report on RHiNET Project

Tomohiro Otsuka,[†] Konosuke Watanabe,[†] Akira Kitamura,[†]
Michihiro Koibuchi,[†] Junji Yamamoto,[††] Hiroaki Nishi,[†] Tomohiro Kudoh[†††,☆]
and Hideharu Amano[†]

A goal of RHiNET project is to build a low-cost and high-performance parallel processing environment using PCs distributed in an office or a building. Through the project, we proposed a new network class LASN which has both advantages of LAN and SAN, and developed RHiNET system prototypes as implementation of the LASN.

Although the project has not been inspired commercial products, we developed network switches which had the largest bandwidth using CMOS technology, a network interface chip for LASN, and a testbed system which can be evaluated. Moreover, various kind of secondary products have been obtained.

## 1. はじめに

近年のパーソナルコンピュータ (PC) の飛躍的な性能向上により，オフィス等で利用されている PC の計算能力の総和は大規模な科学技術計算にも充分対応可能なレベルに達している．また，これらの PC を接続する結合網の転送性能の向上も著しく，オフィス内の PC を高速結合網で接続して並列処理を行うことにより，その余剰計算能力を有効活用するための技術的状況は整っていると言える．

しかし，Gigabit Ethernet などの高性能 LAN (Local Area Network) でこれらの PC を単純に接続しただけでは，効率の良い並列分散処理のための環境を構築することは難しい．これは，LAN で用いられる TCP/IP では，パケットの転送レイテンシが大きく，また，パケットの破棄・重複を認めているため，上位層によるパケットの順序制御・再送制御が必要となり，ソフトウェア各層で大きな損失が生じるためである．このため，PC を用いて実現する高性能計算サーバは，Myrinet[1] に代表される SAN (System Area Network または Server Area Network) を用いて結合した専用の PC クラスタが用いられ，机上で用いている PC の余剰計算能力の有効利用は行われていないのが現状である．ここで，SAN は，パケットを途中で破棄せず，送信順に到着することを保証するネットワークであり，低遅延・高バンド幅の通信を提供する．その一方で最大リンク長やネットワークトポロジに対する制限が厳しいため，オフィスやビル内に分散して配置された PC 同士を接続することは難しい．

このような PC の余剰計算能力を利用して低価格高性能な並列分散環境を構築する目的で，技術研究組合新情報処理開発機構 (RWCP) を中心に慶應義塾大学，日立 IT，東京農工大学などが共同で “RHiNET プロジェクト” を 1997 年度から 2001 年度にかけて実施した．RHiNET プロジェクトでは，目的の達成のため，LAN と SAN の双方の利点を併せ持つ新しいネットワーククラスである Local Area System Network (LASN) という概念を提案した[2]．LASN は，LAN と同様にネットワークのトポロジを比較的自由に設定でき，接続距離も 1km 程度まで延ばすことができると同時に，SAN と同じくパケットの破棄は行わず，送信順に到着することを保証する．

RHiNET は LASN の実装形態であり，試作用の RHiNET-1，実用システムである RHiNET-2，発展システムのプロトタイプ RHiNET-3 (スイッチのみ) の三世代に分けることができる．2001 年に 64 ノードからなる RHiNET-2 クラスタが稼働し，2003 年にはほぼ評価を終了し，これにより，プロジェクトは完全に終結した．本報告では，プロジェクトの終わりに当たって，RHiNET プロジェクト

† 慶應義塾大学理工学部
Faculty of Science and Technology, Keio University
†† 日立製作所 中央研究所
Central Reseach Laboratory, Hitachi
††† 新情報処理開発機構
Real World Computing Partnership
☆ 現在，産業技術総合研究所
Presently with National Institute of Advanced Industrial Science and Technology

が生んだ様々な成果を総括する．

## 2. RHiNET の概観

LASN では，図 1 に示すように，フロア内の机上に配置された PC を接続して並列分散処理のためのシステムを構築する．そのため，LASN を構築するネットワークは，以下の 3 つの性質を持つ必要がある．

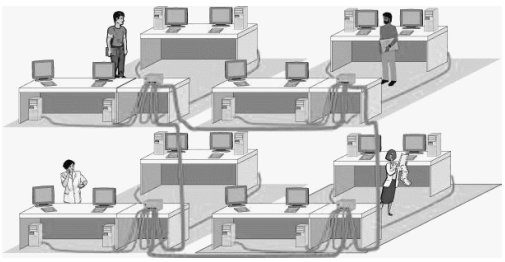

**図 1** LASN の概念図
Fig. 1 An image of LASN

(1) パケットの転送においてハードウェアレベルで信頼性と FIFO 性を保証し，デッドロックフリーで asynchronous wormhole routing を行うことで低レイテンシでの配送を実現する．(2) 長距離ケーブルによる接続と，自由な接続トポロジに対応する．(3) 低レイテンシなパケットの送受信のために，ノードにおいてユーザレベル・ゼロコピー通信を実現し，さらに机上で利用中の PC をノードとして利用することから，アドレス変換および保護機構をハードウェアレベルで実現する．

RHiNET プロジェクトでは，上記の性質を持つネットワーク技術の確立のため，スイッチおよびネットワークインタフェースについて三世代に渡って開発を行った．

### 2.1 ネットワークスイッチ

#### 2.1.1 RHiNET-1/SW

RHiNET-1/SW は，1.4Gbps の転送速度の双方向リンクを 8 本接続可能なスイッチであり，0.35$\mu$m の CMOS プロセスで実現された[3]．リンクには独自の光インタコネクトを用いている．チップ外部メモリに対して仮想チャネルをキャッシュする方式を採用することで大容量のスラックバッファを確保し，また，縮約構造化チャネル法を用いることで，デッドロックフリーと仮想チャネル数の節減を実現している．

FPGA を用いたネットワークインタフェースカードと組み合わせ，小規模なクラスタを構築し，スラックバッファによる Go/Stop 制御，光インタコネクトとの接続技術などスイッチの基本技術のテストに用いられた．

#### 2.1.2 RHiNET-2/SW

RHiNET-2/SW は，実用レベルのクラスタの構築を目的として開発されたスイッチである．リンク速度を最大 8Gbps とし，日立デバイス開発センターの 0.18$\mu$m プロセスの利用により CMOS スイッチとしては開発当時世界最速レベルの転送容量を実現した[4][5]．

基本的な構成およびパケット転送法は RHiNET-1/SW と同じであるが，内部のバッファを充実させることで，外部メモリの利用なしに十分な数の仮想チャネルを確保している．また，様々な転送速度に対応するため，3 種類の転送レートの混在を許した点にも特徴がある．

64 ノードの RHiNET-2 クラスタに利用され，実アプリケーションで動作したが，大規模システムの構築後，後に述べる致命的な不具合が発見された．

#### 2.1.3 RHiNET-3/SW

これまでの二世代のスイッチは，エラーレートの低い光インタコネクトを利用し，さらにフリット毎にエラー訂正符号を付加することで，ハードウェアレベルでの信頼性を確保した．そのため，これらスイッチでは，エラーレートが高い安価なインタコネクトを利用することができないという問題があった．

RHiNET-3/SW は，スラックバッファを用いた Go/Stop 方式のフロー制御に代わって，マイクロウィンドウを用いたスイッチ間の再送機構を用いることでこの問題に対処した[6]．リンク速度も最大 10Gbps に改善されており，商用化に向けて実用的な技術の確立を目指したが，光インタコネクトとの接続用チップの不具合と，プロジェクトの終了により，大規模なシステムは構築されなかった．

### 2.2 ネットワークインタフェースカード

#### 2.2.1 FPGA 版ネットワークインタフェース

RHiNET は，ノード PC のユーザメモリ間でゼロコピー転送を行うことによりレイテンシの低減を図る．しかし，机上の PC では，並列プロセスとは無関係なユーザプロセスも同時に実行されているため，並列処理に用いるメモリ空間を他のユーザプロセスとは独立に確保し，他のプロセスの空間に対する保護を厳重に行う必要がある．このため，RHiNET ではアドレス変換とプロテクションのチェックを行う必要があるが，この操作をネットワークインタフェース等でソフトウェアで行うとレイテンシが増大してしまう．そこで，RHiNET では，一連の操作を全てハードウェアで行うことにより，基本的なゼロコピー転送命令である PUSH (リモートメモリライト) および PULL (リモートメモリリード) をネットワークインタフェースの DMA 機能を用いて完全にハードウェアのみで実現する．

FPGA 版ネットワークインタフェース RHiNET-1/NI および RHiNET-2/NI0 は，Altera 社の FPGA Flex10K を搭載した PCI ボード上に実装されたネットワークインタフェースであり，Flex10K 上のハードウェア論理と内部メモリを用いてこれらの処理を実現した．

RHiNET-1/NI は RHiNET-1/SW に，RHiNET-2/NI0 は RHiNET-2/SW にそれぞれ接続され，実験用の PC クラスタを構築して，論理検証に用いられた．

#### 2.2.2 ASIC 版ネットワークインタフェース

第二世代以降の RHiNET では，RHiNET-1/NI および RHiNET-2/NI0 を大幅に機能拡張したネットワークインタフェースが用いられており，コントローラとして Martini[7] と呼ばれる独自開発の ASIC を搭載している．

Martini は，日立デバイス開発センターの 0.14$\mu$m プロセスを利用し，FPGA 版ではゲート数と搭載メモリの制約により実現できなかったアドレス変換を完全な形で実装している．RHiNET-2/SW, RHiNET-3/SW の両方に対応するスイッチインタフェースを持ち，PCI バスホストインタフェースの他に，DIMMnet[8] 用のコントローラとして利用するためのメモリバス用のインタフェースを備える．また，Martini では，リモート DMA によるデータ転送機構 (PUSH/PULL) 以外に，DIMMnet の機能の一部として実装された PIO によるデータ転送機構である BOTF および AOTF が利用できる．さらに Martini は MIPS R3000 と互換性を持つ CPU コアを内蔵しており，テーブル参照ミス等の例外処理にソフトウェアで対応する．

現在稼働している RHiNET-2 クラスタでは，RHiNET-2/SW に対応した ASIC 版ネットワークインタフェースである RHiNET-2/NI が用いられている．

### 2.3 関連した成果

RHiNET プロジェクトは，上に述べた直接の成果以外にも様々な成果を生んでいる．

メモリバスに接続する DIMMnet は，総務省のプロジェクトでその後継機が開発されている[9]．

FPGA を搭載した RHiNET-2/NI0 は，複数のプロトコルを入れ替えることにより性能向上を実現するリコンフィギャラブルネットワークインタフェースとして用いられ，その効果が確認された[10][11]．

Martini は，組み込み用の CPU がハードウェアの機能をモジュール単位で部分的に代行できる乗っ取り機構[12]を搭載しており，新しいソフトウェア／ハードウェア協調処理の方法として注目されている．また，Martini 内のデータ転送機構の一部は J-joint[13] として IP 化され，公開されている．

RHiNET-2 クラスタは，スイッチが多数の仮想チャネルを持ち，単純なテーブルに基づくルーティング機構を持つため，様々な固定ルーティングプロトコルを組み込むことができ，光インタコネクトを差し換えることでトポロジを変えることもできる．この機能を利用し，今までシミュレーションのみで行われていたルーティングアルゴリズムの評価を実環境で行うことができた．この結果，シミュレーションでは予想できなかった様々な特性が明らかになり，国際学会で注目された[14][15]．

## 3. 性 能 評 価

以下では，現在稼働中の RHiNET-2 の性能評価に関して述べる．RHiNET-2 は，実アプリケーションによる評価[16][17][18][19][20] の他に，実際にオフィス内の PC を接続して使用する際に必要となると考えられる動的負荷分散に関する評価[21][22] や，クラスタ向けネットワークのトポロジ，ルーティングといった特性に関する評価[14][15][23][24] にも用いられている．ここでは，並列ベンチマークを用いた評価結果を示す．

### 3.1 RHiNET のソフトウェア

RHiNET は，基本通信処理へのソフトウェアの介入を極力減らすことで高性能な通信を実現するネットワークであるが，例外処理やホスト上でのデバイス管理等ではソフトウェアが必要となる．また，ユーザが効率的にプログラムを記述するためには，ユーザライブラリによって NI のハードウェアへの直接アクセスや，ハードウェアでは提供されない通信処理をサポートする必要がある．さらに，既存の並列プログラムを動作させるために，メッセージパッシングのような標準的な並列プログラミングモデルに基づく上位ライブラリを提供することが望ましい．

各ノード上に構築される RHiNET-2 のソフトウェアレイヤを図 2 に示す．レイヤの最下層には，ネットワークインタフェース RHiNET-2/NI のコアプロセッサ上で実行されるファームウェアが位置する．それより上位はホスト上で実行されるソフトウェアであり，カーネルレベルで実行されるデバイスドライバとユーザレベルで実行されるユーザライブラリが存在する．

ユーザは，ユーザライブラリを直接用いることで並列プログラムを記述することができる．また，ユーザライブラリ上に実現された SCore[25][26] システムが提供する MPI やソフトウェア分散共有メモリ等の標準的な並列プログラミングライブラリを利用することもできる．

RHiNET-2 のソフトウェアレイアについての詳細は，[19] に述べられている．

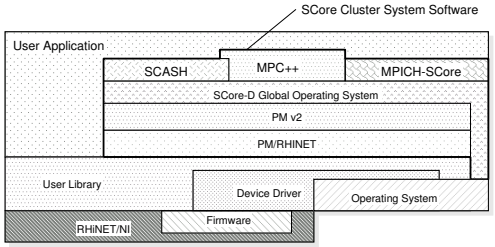


**図 2** RHiNET-2 のソフトウェアレイヤ
Fig. 2 Software Layer of RHiNET-2

### 3.2 評 価 環 境

本来，RHiNET は机上で用いられている PC 同士を接続するためのネットワークであるが，限られた実装スペースで多数のノードを接続して実験するため，図 3 に示すようなラック内に 64 ノードを格納したクラスタを実装した．

**図 3** 64 ノードの RHiNET-2 クラスタ
Fig. 3 RHiNET-2 cluster with 64 nodes

ノードの主な仕様を表 1 に示す．各ノードはラックマウント型の PC であり，PCI バスに RHiNET-2/NI を挿入し，外部より光ケーブルを接続する．ただし，本評価では 1 ノードにつき 1CPU しか使用していない．

**表 1** RHiNET-2 クラスタのノード PC の主な仕様
Table 1 Specifications of the node PCs

| | |
|---|---|
| # of Nodes | 64 |
| CPU | Intel Pentium III 933MHz × 2 |
| Chipset | Serverworks ServerSet III HE-SL |
| Memory | PC133 SDRAM 1Gbyte |
| PCI | 64bit/66MHz |
| Link | 8Gbps optical interconnect 2.0m, 5.0m |
| OS | RedHat Linux 7.2 (kernel 2.4.18) |
| SCore | Version 5.0.1 |

### 3.3 評 価 結 果

今回の評価では，SCore システム上で並列アプリケーションを動作させ，性能測定を行った．現在，RHiNET-2 は 4 節で述べる諸問題により安定稼働に至っていないため，本評価では 16 ノードまでの使用にとどまっている．

評価アプリケーションとしては，NAS 並列ベンチマーク[27][28]2.3 から MG，CG，IS，SP，BT の 5 つと，並列ベンチマーク集 SPLASH-2[29] から LU を用いた．

問題サイズは，CG が Class A，MG，IS，SP，BT については最も小さい Class S とした．また，LU は 2048×2048 の行列とした．これは，CG 以外のアプリケーションでは，ノード数が多くなった場合にシステムの問題によりサイズの大きい問題が動作しないケースがあったためである．

図 4 に各アプリケーションの性能向上率を示す．グラフは，各アプリケーションにおいて 1 ノードでの実行時

間を 1 とした場合の台数効果を表している．

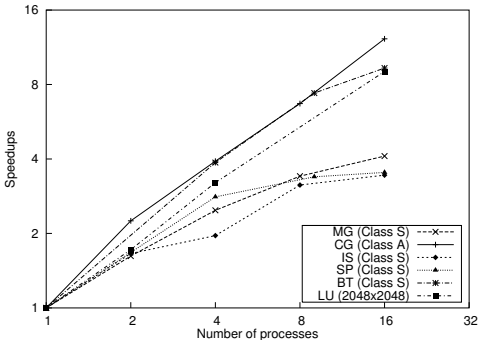


**図 4** SCore 上の並列アプリケーションの性能向上
Fig. 4 Speedups of parallel applications on SCore

CG，BT，LU では，16 ノードでそれぞれ 12.2 倍，9.3 倍，9.0 倍と高い性能向上率を示している．一方，MG，IS，BT では 16 ノードで 3.4 倍から 4.1 倍程度の性能向上に留まっている．これは，これら 3 つのアプリケーションで用いている Class S の問題が非常に小さく，ノード数が多くなった場合に計算量に対する通信量の割合が急激に高くなるためであると考えられる．

## 4. RHiNET-2 実装上の不具合

これまで述べた通り，RHiNET プロジェクトでは最終的に 64 ノード構成の RHiNET-2 クラスタを構築し，様々な評価を行った．しかし，その過程で多数の不具合が発見され，十分に安定したシステムの実現には至らなかった．以下では，RHiNET-2/SW および RHiNET-2/NI において，実装後に発覚した主な不具合について述べる．

### 4.1 設計上の不具合

RHiNET-2/SW や RHiNET-2/NI では，実機完成後にアプリケーションを用いて評価を行う段階になってから，特定の通信パターンやパケットサイズでのみ発生する設計上の不具合が発見された．

#### 4.1.1 RHiNET-2/SW のフロー制御問題

RHiNET-2/SW では Go/Stop 方式でフロー制御を行っているが，アービトレーション部分の設計に問題があり，特定のタイミングで Stop 信号と Go 信号がノードや周辺スイッチから到着した場合にフロー制御が正しく動作しないことが判明している．

これは実機において MPI プログラムを実行した際に，Alltoall のような特定の通信パターンでパケット消失が発生することから発覚し，最終的に RTL シミュレーションでもノード間で特定のパターンでパケットを交換することで容易に発生することが確認された．

ホスト側で MTU を制限してフロー制御自体を不要にすることで問題は回避できるものの，パフォーマンスの低下が著しく本質的な解決策とは言えない．致命的と言えるこの問題の発覚により，安定した RHiNET-2 システムの実現が不可能であることが確定した．

#### 4.1.2 PCI インタフェースの DMA コントローラ問題

Martini の PCI インタフェースに搭載される DMA コントローラには，特定のデータサイズで特定のアドレスに対する DMA を発行した際にデータ抜けが発生する場合があるという不具合が確認されている．

この問題は，特定の条件で DMA コントローラが誤ったバイトイネーブル信号を出力してしまう論理の不具合によるものであり，ユーザプロセスから見ると，256byte 程度のホストへの DMA において (DMA のサイズ + DMA 先のアドレスのページ内オフセット) を 32 で割った余りが 17 から 23 の間の場合にデータの欠落が発生する．

この問題は，このような条件の DMA を伴う通信を行わない限り発生することはなく，コーディングの工夫次第でパフォーマンスの損失なく回避することができる．しかし，ユーザレベルで通信の起動を許可する以上，このようなパターンでの通信を未然に防ぐ手だてはないため，一般ユーザがシステムを利用する際には問題となる．

### 4.2 仕様上の不具合

以下に示すのは，実装されたハードウェアはそれぞれのコンポーネントの設計者の意図した通りに動作するにもかかわらず，仕様上の問題点により，結果的に不具合を生むことになった問題である．

#### 4.2.1 キックアドレスへの書き込み

Martini における通信要求は，Window[7] と呼ばれるメモリ領域に対して通信要求のパラメータを書き込み，Window のアドレス空間上にマップされているキックアドレスと呼ばれる特定ライン 8byte に対して，通信コマンドと転送サイズを書き込むことで発行される．Martini は，この Window をユーザ領域にマップすることで，ユーザレベル通信を実現している．

その際，データの書き込み方によっては，キックアドレスに対する書き込みが 4byte ずつ 2 回に分かれてしまうことがある．一方，Martini 側で通信要求の起動を受け付ける部分はこのような状況を想定しておらず，コマンドとサイズが同時に PCI インタフェースより受け渡されるものとして設計されている．そのため，コマンドと転送サイズが別々に書き込まれた場合，サイズ情報が壊れ，正しいサイズでのデータ転送が行われない．

Martini は，64 もの Window を備え，これを各プロセスに個別に割り当てることで複数のプロセスが同時に通信要求を発行できるようになっている．しかし，この不具合により，SMP ノードで複数のプロセスが同時に Window へアクセスすると正常に通信要求ができず，最悪の場合ノードがフリーズしてしまう．結果として，Martini を同時に使えるのは高々1 プロセスとなり，また，キックアドレスへは必ず 8byte をまとめて書き込むコーディングをする必要が生じてしまった．

#### 4.2.2 受信チャネルの問題

RHiNET-2/SW は 16 本の仮想チャネル (以下ネットワークチャネル) を持ち，構造化チャネル法を用いてデッドロックフリーのルーティングを提供する．一方，Martini が備える受信チャネルは 4 本であるため，パケットはネットワークチャネルを 4 で割った商の受信チャネルに受信される．また，Martini の受信チャネルは 0 が PUSH 等のデータパケット受信用，2 が ACK 等の応答パケット受信用となっており，1 と 3 はハードウェア利用しないという設計となっているため，チャネル 0 でパケットを受け取った場合にのみ ACK を正常に返信できる構造となっている．

しかし，スイッチを 4 個以上経由する場合，受信パケットのネットワークチャネルが 4 以上となるため，Martini の受信チャネル 1 で受信されてしまい，ACK パケットの返信ができない．

ネットワークのトポロジによってはノード–スイッチ間での接続で仮想チャネルを増やさないようにすることでこの問題を回避できるが，4×4 の 2 次元メッシュ等の場合は，通信を行うノードの組合せによっては経由スイッ

チ数が 4 個以上となるため，データパケットが受信チャネル 1 で受信されてしまう．

この不具合は，設計段階でこのようなネットワーク構成を十分に考慮に入れていなかったことに起因する．結果として，仮想チャネルが 16 本もあり，トポロジフリーを謳っているにもかかわらず，安定した通信を行うためのトポロジが大幅に制限されてしまうこととなった．

### 4.3 電気的問題

RHiNET-2/SW では，パケットが通過する際に，通信相手との相性や個体差により，データが 1 ビット化ける，あるいは 2 ビットずれるという現象が確認されている．この現象は特定のポート (ポート 4) において特に発生率が高く，内部で発生しているためか ECC による誤り訂正も全く行われない．

この問題は電気的な要因によるもので，RHiNET-2/SW のチップの電圧を上げることである程度改善される．現在は，チップの動作周波数を 100MHz から 75MHz に落とすことでこの問題を回避しているが，これによりネットワークの最大バンド幅が 3/4 に低下してしまい，スイッチのレイテンシも増加してしまう結果となった．

### 4.4 不具合発生の原因分析

設計上の不具合は，RTL 検証時のタイミング検証の甘さに起因する．また，仕様上の不具合は，設計グループの担当者間での意志疎通が十分でなかった点に起因する．

しかし，大規模なネットワークにおいては，それぞれのコンポーネントに対するイベントの発生タイミングの組合せをシミュレーションレベルで完全に検証するのは不可能に近い．また，仕様の整合性も，上位レベルのソフトウェアで動作させてみてはじめて堅牢なものとすることができる．

RHiNET-2 を安定稼働まで持っていくことができなかった最大の原因は，最初のプロトタイプ作成時に PCI 周辺に生じた大きな不具合により，他の不具合を洗い出すことのできるレベルまでシステムを稼働させることができなかった点にある．逆に言えば，RHiNET-2/SW および Martini をもう一度リメイクすることができたならば，RHiNET-2 を安定稼働させることができたであろう．

## 5. 総　　括

RHiNET プロジェクトは，以下の点で一通りの成功を収めたと言って良い．(1) CMOS で世界最高レベルのスイッチを実装して LASN 転送の基本技術を確立し，Hot Interconnect を何度か賑わせた．(2) LASN 用のネットワークインタフェースチップを開発し，評価を取ることのできる実験機の実装に成功した．実験機はルーティング，トポロジ，マルチキャストなどのネットワーク特性の評価に貢献し，これらの結果は，ICPP，Cluster などで発表された．(3) 多くの副産物を得た．特に関連プロジェクトである DIMMnet は後継プロジェクトが現在も進行している．

しかし一方で，LASN 技術を商用レベルにまで引き上げることはできなかった．これは以下の点が原因であると考えられる．(1) RHiNET では高速インタコネクトとして日立製作所製の光インタコネクトを用いたが，独自規格の製品であり価格が安くならなかった．これに対し，InfiniBand[30] などでは当初は光インタコネクトを用いていたが，2.5Gbps/ch の電気配線による高速インタコネクトを用いるようになって普及が早まった．RHiNET も製品化のためには，このような電気配線による高速通信技術を採用する必要があった．(2) InfiniBand は RHiNET と同時期に開発された LASN の性質を持つネットワークである．InfiniBand は，米国の大企業が中心となって標準化が進められたのに対して，RHiNET プロジェクトでは標準化に対する取り組みが不足していた．(3) 新情報処理開発機構が解散する時点で，製品レベルまで達するにはさらに数回の試作を重ねて実用試験を行う必要があった．電気配線技術の採用や標準化作業も含めて，企業のサポートが必要な状況であったが，当時の日本では InfiniBand に対しても懐疑的な企業が多く，RHiNET プロジェクトへのサポートは得られなかった．

現状では，机上の PC の余剰計算能力を利用する場合，10Gbit Ethernet や InfiniBand の普及に従い，これらに準拠した標準部品を利用して行うのが適切であると考えられる．この場合，RHiNET で確立された技術の一部は利用可能であろう．また，ネットワークインタフェースの一部の設計は，DIMMnet の後継プロジェクトでも利用される予定である．

**謝辞**　RHiNET プロジェクトに関わったすべての方々に感謝いたします．

## 参　考　文　献


1) N. J. Boden, D. Cohen, R. E. Felderman, A. E. Kulawik, C. L. Seitz, J. N. Seizovic and W. Su: Myrinet: A Gigabit-per-Second Local Area Network, *IEEE Micro*, Vol. 15, No. 1, pp. 29–36 (1995).
2) Tomohiro Kudoh, Shinji Nishimura, Junji Yamamoto, Hiroaki Nishi, Osamu Tatebe and Hideharu Amano: RHiNET: A network for high performance parallel processing using locally distributed computers, *Proceedings of IWIA 99*, pp. 69–73 (1999).
3) Hiroaki Nishi, Koji Tasho, Tomohiro Kudoh, Junji Yamamoto and Hideharu Amano: RHiNET-1/SW: an LSI switch for a local area system network, *An International Symposium on Low-Power and High-Speed Chips (COOL Chips III)*, pp. 175–187 (2000).
4) Shinji Nishimura, Tomohiro Kudoh, Hiroaki Nishi, Junji Yamamoto, Katsuyoshi Harasawa, Nobuhiro Matsudaira, Shigeto Akutsu, Koji Tasho and Hideharu Amano: High-speed network switch RHiNET-2/SW and its implementation with optical interconnections, *Technical Digest of Hot Interconnects 8*, pp.31–38 (2000).
5) Shinji Nishimura, Tomohiro Kudoh, Hiroaki Nishi, Junji Yamamoto, Katsuyoshi Harasawa, Nobuhiro Matsudaira, Shigeto Akutsu and Hideharu Amano: 64-Gbit/s highly reliable network switch (RHiNET-2/SW) using parallel optical interconnection, *IEEE journal of Lightwave Tehnology (Special issue on Optical Networks)*, Vol. 18, No. 12, pp. 1620–1627 (2000).
6) Shinji Nishimura, Tomohiro Kudoh, Hiroaki Nishi, Junji Yamamoto, Katsuyoshi Harasawa, Nobuhiro Matsudaira, Shigeto Akutsu, Koji Tasho and Hideharu Amano: RHiNET-3/SW: an 80-Gbit/s high-speed network switch for distributed parallel computing, *Hot Interconnects 9*, pp. 119–123 (2001).
7) 山本 淳二, 渡邊 幸之介, 土屋 潤一郎, 原田 浩, 今城 英樹, 寺川 博昭, 西 宏章, 田邊 昇, 上嶋 利明, 工藤 知宏, 天野 英晴: 高性能計算をサポートするネットワークインタフェース用コントローラチップ Martini, 情報処理学会論文誌ハイパフォーマンスコンピューティングシステム, Vol. 5, No. 018, pp. 122–133 (2002).
8) 田邊 昇, 山本 淳二, 濱田 芳博, 中條 拓伯, 工藤 知宏, 天野 英晴: DIMM スロット搭載型ネットワークインタフェース DIMMnet-1 とその高バンド幅通信機構 BOTF, 情報処理学会論文誌, Vol. 43, No. 4, pp. 866–877 (2002).
9) 田邊 昇, 濱田 芳博, 三橋 彰浩, 中條 拓伯, 天野 英晴: メモリスロット装着型ネットワークインタフェース DIMMnet-2 の構想, 情報処理学会研究報告 HOKKE2003 (2003).
10) Naoyuki Izu, Tomonori Yokoyama, Junichiro Tsuchiya, Konosuke Watanabe and Hideharu Amano: RHiNET/NI: A reconfigurable network interface for cluster computing, *12th International Conference on Field Programmable Logic and Application* (2002).
11) Tomonori Yokoyama, Naoyuki Izu, Jun-ichiro Tsuchiya, Konosuke Watanabe, Hideharu Amano and Tomohiro Kudoh:

Design and implementation of RHiNET-2/NI0: a reconfigurable network interface for cluster computing, *IEICE Transaction*, No. 5, pp. 789–795 (2003).
12) Konosuke Watanabe, Hideharu Amano, Junji Yamamoto, Junichiro Tsuchiya, Tomohiro Otsuka and Tomohiro Kudoh: Taking over mechanism: a Cooperation Methodology of Hardware and Software in Network Controller, *Proceedings of the Workshop on Synthesis And System Integration of Mixed Information Technologies*, pp. 386–393 (2003).
13) 渡邊 幸之介, 土屋 潤一郎, 天野 英晴: ミスアラインメントを補正するモジュール間転送用インタフェース, 電子情報通信学会技術研究報告 VLD2002-117 (2002).
14) Michihiro Koibuchi, Akiya Jouraku, Konosuke Watanabe and Hideharu Amano: Descending Layers Routing: A Deadlock-Free Deterministic Routing using Virtual Channels in System Area Networks with Irregular Topologies, *the International Conference on Parallel Processing (ICPP'03)*, pp. 527–536 (2003).
15) Michihiro Koibuchi, Konosuke Watanabe, Kenichi Kono, Akiya Jouraku and Hideharu Amano: Performance Evaluation of Routing Algorithms in RHiNET-2 Cluster, *Proceedings of IEEE International Conference on Cluster Computing*, pp.395–402 (2003).
16) 原田 浩, 山本 淳二, 土屋 潤一郎, 渡邊 幸之介, 天野 英晴, 工藤 知宏, 石川 裕: RHiNET の高速通信ライブラリ PMv2 による評価, 情報処理学会研究報告 2002-ARC-147/2002-HPS-89, pp. 145–150 (2002).
17) 大塚 智宏, 渡邊 幸之介, 土屋 潤一郎, 原田 浩, 山本 淳二, 西 宏章, 工藤 知宏, 天野 英晴: RHiNET ネットワークインタフェースの性能評価, 電子情報通信学会技術研究報告 CPSY2002-44, pp. 23–28 (2002).
18) Tomohiro Otsuka, Konosuke Watanabe, Jun-ichiro Tsuchiya, Hiroshi Harada, Junji Yamamoto, Hiroaki Nishi, Tomohiro Kudoh and Hideharu Amano: Performance Evaluation of a Prototype of RHiNET-2: A Network-based Distributed Parallel Computing System, *Proceedings of the IASTED International Multi-Conference on Applied Imformatics (AI2003)*, pp. 738–743 (2003).
19) 大塚 智宏, 渡邊 幸之介, 北村 聡, 原田 浩, 山本 淳二, 西 宏章, 工藤知宏, 天野 英晴: 分散並列処理用ネットワーク RHiNET-2 の性能評価, 先進的計算基盤システムシンポジウム SACSIS 論文集, pp. 45–52 (2003).
20) 大門 優, 松尾 亜紀子, 大塚 智宏, 渡邊 幸之介, 天野 英晴: 反応を伴った圧縮性流体計算による RHiNET-2 の評価, 電子情報通信学会技術研究報告 CPSY2003, Vol. 103, No. 249 (2003).
21) 北村 聡, 天野 英晴, 渡邊 幸之介, 大塚 智宏: PC クラスタ用ネットワーク RHiNET-2 上における動的負荷分散アルゴリズム の評価, 情報処理学会研究報告 ARC-152, pp. 73–78 (2003).
22) Akira Kitamura, Konosuke Watanabe, Tomohiro Otsuka and Hideharu Amano: The evaluation of dynamic load balancing algorithm on RHiNET-2, *Parallel and Distributed Computing and Systems (PDCS'03)*, pp. 262–267 (2003).
23) 鯉渕 道紘, 渡邊 幸之介, 河野 賢一, 上樂 明也, 天野 英晴: RHiNET-2 クラスタを用いたルーティングアルゴリズムの実機評価, 電子情報通信学会技術研究報告 CPSY2003-13, pp. 43–48 (2003).
24) 鯉渕 道紘, 大塚 智宏, 渡邊 幸之介, 天野 英晴: RHiNET-2 クラスタにおけるユニキャストを基にしたマルチキャストアルゴリズムの評価, 情報処理学会研究報告 2004-EVA-8, pp. 25–30 (2004).
25) Yutaka Ishikawa, Hiroshi Tezuka, Atsushi Hori, Shinji Sumimoto, Toshiyuki Takahashi, Francis O'Carroll and Hiroshi Harada: RWC PC Cluster II and SCore Cluster System Software – High Performance Linux Cluster, *5th Annual Linux Expo*, pp. 55–62 (1999).
26) Toshiyuki Takahashi, Shinji Sumimoto, Atsushi Hori, Hiroshi Harada and Yutaka Ishikawa: PM2: High Performance Communication Middleware for Heterogeneous Network Environment, *SC2000*, pp. 52–53 (2000).
27) D. Bailey, T. Harris, W. Saphir, R. Wijngaart, A. Woo and M. Yarrow: The NAS Parallel Benchmarks 2.0, *NAS Technical Report NAS-95-020* (1995).
28) W. Saphir, R. Wijngaart, A. Woo and M. Yarrow: New Implementations and Results for the NAS Parallel Benchmarks 2, *8th SIAM Conference on Parallel Processing for Scientific Computing* (1997).
29) S. C. Woo, M. Ohara, E. Torrie, J. P. Singh and A. Gupta: The SPLASH-2 Programs: Characterization and Methodological Considerations, *ISCA95*, pp. 24–36 (1995).
30) Association, I. T.: InfiniBand architecture. Specification Volume 1, Release 1.0.a, *available from the InfiniBand Trade Association, http://www.infinibandta.com* (2001).
31) Hiroaki Nishi, Koji Tasho, Junji Yamamoto, Tomohiro Kudoh and Hideharu Amano: A local area system network RHiNET-1: a network for high performance parallel computing, *High Performance Distributed Computing (HPDC-9)*, pp. 296–297 (2000).
32) Konosuke Watanabe, Junji Yamamoto, Jun-ichiro Tsuchiya, Noboru Tanabe, Hiroaki Nishi, Tomohiro Kudoh and Hideharu Amano: Preliminary Evaluation of Martini: a Novel Network Interface Controller Chip for Cluster-based Parallel Processing, *Proceedings of the IASTED International Multi-Conference on Applied Imformatics (AI2002)*, pp. 390–395 (2002).
33) Konosuke Watanabe, Tomohiro Otsuka, Jun-ichiro Tsuchiya, Hiroshi Harada, Junji Yamamoto, Hiroaki Nishi, Tomohiro Kudoh and Hideharu Amano: Performance Evaluation of RHiNET-2/NI: A Network Interface for Distributed Parallel Computing Systems, *Proceedings of International Symposium on Cluster Computing and the Grid*, pp. 318–325 (2003).
34) Shinji Nishimura, Tomohiro Kudoh, Hiroaki Nishi, Katsuyoshi Harasawa, Nobuhiro Matsudaira, Shigeto Akutsu, Koji Tasho and Hideharu Amano: A network switch using optical interconnection for high performance parallel computing using PCs, *Proceedings of The 6th International Conference on Parallel Interconnects (PI'99)*, pp. 5–12 (1999).
35) 西 宏章, 多昌 廣治, 工藤 知宏, 天野 英晴: 効率良い並列処理をサポートするローカルエリア向けネットワーク スイッチ, 電子情報通信学会論文誌, No. 2, pp. 245–254 (2000).
36) Shinji Nishimura, Tomohiro Kudoh, Hiroaki Nishi, Katsuyoshi Harasawa, Nobuhiro Matsudaira, Shigeto Akutsu, Koji Tasho and Hideharu Amano: High-throughput network switch for the RHiNET-2 optically interconnected parallel computing system, *Proceedings of Optics in Computing 2000 (OC2000)*, pp. 562–569 (2000).
37) Shinji Nishimura, Tomohiro Kudoh, Hiroaki Nishi, Katsuyoshi Harasawa, Nobuhiro Matsudaira, Shigeto Akutsu and Hideharu Amano: High-throughput network switch using high-speed highly reliable optical interconnection, *Technical Digest of Fifth Optoelectronics and Communications Conference (OECC 2000)*, pp. 466–467 (2000).
38) Shinji Nishimura, Katsuyoshi Harasawa, Nobuhiro Matsudaira, Shigeto Akutsu, Tomohiro Kudoh, Hiroaki Nishi and Hideharu Amano: RHiNET-2/SW: a large-throughput, compact network-switch using 8.8-Gbit/s optical interconnection, *New Generation Computing*, Vol. 18, No. 2, pp. 188–197 (2000).
39) Ueno Ryuichiro, Inasawa Satoru, Nishi Hiroaki, Kudoh Tomohiro and Amano Hideharu: Flow control method in high speed transfer using optical interconnect, *Proceedings of Applied Informatics (AI2001)*, pp. 271–276 (2001).
40) Shinji Nishimura, Tomohiro Kudoh, Hiroaki Nishi, Koji Tasho, Katsuyoshi Harasawa, Shigeto Akutsu, Shuji Fukuda and Yasutaka Shikichi: A high-speed, highly-reliable network switch for parallel computing system using optical interconnection, *The IEICE Transaction on Electronics (Special issue on Optical Interconnects/Optical Signal Processing)*, No. 9, pp. 756–765 (2001).
41) Takashi Yoshikawa, Ichiro Hatakeyama, Kazunori Miyoshi and Kazuhiko Kurata: Optical Interconnection as an Intellectual Property of a CMOS Library, *Hot Interconnects 9*, pp. 31–35 (2001).
42) Tanabe Noboru, Yamamoto Junji, Nishi Hiroaki, Kudoh Tomohiro, Hamada Yoshihiro, Nakajo Hironori and Amano Hideharu: MEMOnet: Network interface plugged into a memory slot, *Proceedings of IEEE International Conference on Cluster Computing*, pp. 17–26 (2000).
43) 田邊 昇, 濱田 芳博, 山本 淳二, 今城 英樹, 中條 拓伯, 工藤 知宏, 天野 英晴: DIMM スロット搭載型ネットワークインタフェース DIMMnet-1 とその低遅延通信機構 AOTF, 情報処理学会論文誌ハイパフォーマンスコンピューティングシステム, Vol. 44 (2003).
44) Junji Yamamoto and Tomohiro Kudoh: Analysis of communication traffic of shared memory based programs, *Proceedings of the International Conference on Parallel and Distributed Processing Techniques and Applications*, pp. 843–850 (1998).