

# 大規模広域並列分散システムの実現を目指す超高速インターネットの構想

陣崎 明†、林 弘†、中村 修‡、村井 純‡

† 新情報処理開発機構 並列分散システム富士通研究室

〒211-88 川崎市中原区上小田中 4-1-1

*E-mail: {zinzin,hiromu}@flab.fujitsu.co.jp*

‡ 慶應義塾大学 環境情報学部

〒252 神奈川県藤沢市遠藤 5322

*E-mail: {osamu,jun}@keio.ac.jp*

Gigabit ネットワークに対応する新しいプロトコル処理技術を開発することにより「超高速インターネット」を実現する「並列ネットワークサーバ」の構想を述べる。研究は(1)新しいプロトコル処理エンジンである「ストリームプロセッサ」の開発、(2)並列ネットワークサーバの開発ならびにインターネットルータへの応用、の二段階からなる。ストリームプロセッサはプロセッサ内部の有限状態機械を再構成可能とすることにより、パケットを直接解析、処理できるプロトコル処理専用プロセッサである。本研究は国家プロジェクト「次世代情報処理基盤技術開発事業」の一環であり、現在第一段階を進めている。

インターネット、ネットワークサーバ、プロトコルエンジン、ルータ

## “Ultra High-speed Internet” for widely distributed large scale parallel systems: the Plan

Akira Jinzaki †, Hiromu Hayashi †, Osamu Nakamura ‡, Jun Murai ‡

† Parallel and Distributed Systems Fujitsu Laboratory, RWCP

1-1, Kamikodanaka 4-Chome, Nakahara-ku, Kawasaki, 211-88 Japan

*E-mail: {zinzin,hiromu}@flab.fujitsu.co.jp*

‡ Faculty of Environmental Information, Keio University

5322 Endo, Fujisawa Kanagawa, 252 Japan

*E-mail: {osamu,jun}@keio.ac.jp*

This paper describes the plan and current status of “Parallel Network Server” project, that aims a realization of “Ultra High-speed Internet” by developing a new protocol processing technology for Gigabit-Networks. The project sets two phases: (1) development of “Stream Processor”, a new protocol engine, and (2) development of “Parallel Network Server” and its application to the Internet router. The Stream Processor is a protocol-processing oriented processor that has a capability of reconfiguring its internal state machines. This project is now proceeding the phase 1 as the National Project “RWC-RWI/PDC”.

internet, network server, protocol engine, router

## 1. はじめに

100Mbps 台の WAN ( Wide Area Network ) である ATM の実用化にともない「超高速インターネット」ということがいわれるようになった<sup>1)</sup>。LAN ( Local Area Network ) の分野でも Fibre Channel、Gigabit Ethernet など 1Gbps 台のネットワーク技術が実用化されており、数年後には 4Gbps 程度の LAN が実用化されるものとみられる。このような「超高速インターネット」は現在の低速なインターネットと SAN ( System Area Network ) など並列分散システム向け超高速専用ネットワークとの中間に位置付けられ、標準的な技術 ( 標準ネットワーク、標準プロトコル、標準的なコンピュータ ) を利用することによって、従来では専用システムでしか実現できなかったような高度の並列分散処理を実現していく可能性をもつといえよう。

そこで、並列分散システムとして利用可能な「超高速インターネット」を標準技術のプラットフォームの上に構築し、そのようなシステムの限界を探究することを目的とするプロジェクトを開始した。本論文はこのプロジェクトの狙い、アプローチ、予定を明らかにするものである。

本プロジェクトは国家プロジェクト「次世代情報処理基盤技術開発事業」の一環であり、研究期間は 5 年を予定している。大きな開発項目として以下の二つを設定しており、成果は積極的に公開し、標準化していく方針である。

- ・ 標準技術ベースの超高速ネットワーク技術
- ・ 並列分散システムによるインターネットルータ

次世代情報処理基盤技術開発事業では並列分散システムの研究開発に関して WAN、LAN、SAN にまたがる三つのプロジェクトが走る。本プロジェクトでは標準プロトコルを前提として WAN、LAN を扱う。他の二プロジェクトと本プロジェクトの違いは、ネットワーク、プロトコル、コンピュータについて他プロジェクトが性能、機能本位の研究開発を行うのに対して本プロジェクトではあくまで標準技術の制限のもとでの性能、機能を追求する点にある。これら三プロジェクトによって次世代並列分散システムのほぼ全域をカバー

—する研究開発成果が期待される ( 図 - 1 )。

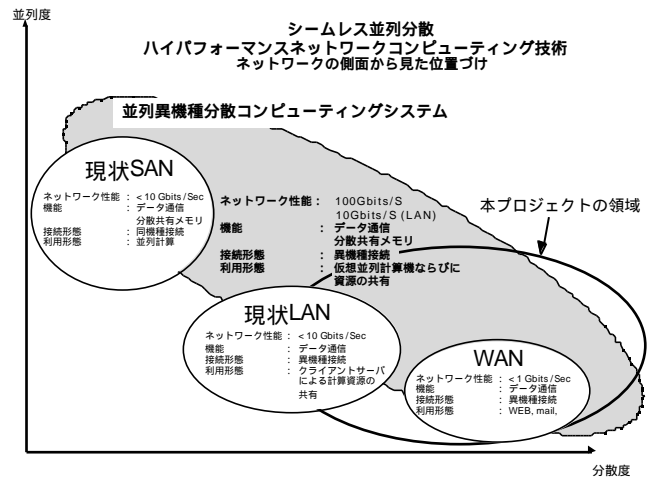


図 - 1 本プロジェクトの位置付け<sup>2)</sup><sup>1)</sup>

## 2. 研究開発の課題

### 2.1. ネットワーク性能問題

コンピュータネットワークではネットワーク伝送路の性能が高速化してもアプリケーションから実効的に利用できる通信性能に反映されないという問題がある。具体例を示そう。Internet Protocol 実装の基準である BSD Unix の UDP/IP 性能をある条件 ( BSD/OS 2.1、Pentium 200MHz、loopback デバイス ) で測定すると単一方向で約 30MB/sec が限界である ( 図 - 2 )。また、これはソフトウェア処理のみの場合で、実際のネットワークに転送する際は同等のコンピュータを使用しても、デバイスドライバの処理、ハードウェアの動作、割り込み処理のために性能は 10MB/sec 程度に低下する。これでは FastEther ( 100Mbps、12.5MB/sec ) をなんとか使いこなせる程度で、Gigabit Ethernet ( 1Gbps、100MB/sec ) クラスには全く及ばない。

すなわち BSD/OS の UDP/IP で 100MB/sec ( 1Gbps ) のネットワーク ( 一方方向のみ ) を使い切るには 600MHz 以上の Pentium が 100% の能力を出さなければならないことになる。双方向の場合は 1.2GHz の Pentium が必要である。ルータなどのようにプロトコル処理そのものが目的である場合はともかく、一般に通信はアプリケーション処理の一部であるから、仮にプロセッサ性能の 50% をプロトコル処理に割くとしても

<sup>1)</sup> 本プロジェクトの領域を追加した。

1.2GHz (双方向では 2.4GHz) 以上の Pentium を使ってやっと 1Gbps ネットワークを使い切ることができることになる。この問題は並列分散システムを構築する上の本質的な障害と見なされ、従来から数多くの研究開発対象となってきた。検討対象もネットワーク、アダプタ、オペレーティングシステム (OS)、プロトコルの幅広い分野にまたがっている[3,4]。

## 2.2. プロトコル処理の解析

プロトコル処理が性能にどのような影響を与えているかを調べるために、BSD/OS 2.1、Pentium 200MHz における UDP/IP 処理を解析した。プロトコル処理要素として、IP 処理<sup>2</sup>、UDP checksum 処理、Socket におけるユーザ空間とカーネル空間の間のメモリコピー処理を選び、それらの処理を除去 (処理時間 = 「零」) したときの性能を実測したところ以下の結果を得た (図 - 2)。

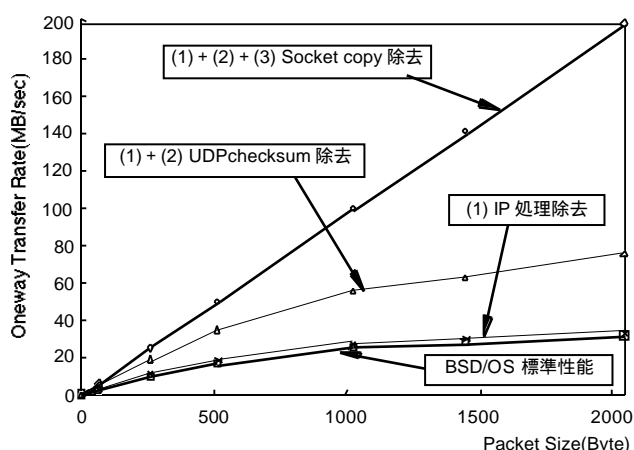


図 - 2 プロトコル処理除去の効果

この結果によれば、評価した処理の全てをなんらかの方法で「零」にできれば、スループットの現在の 4~5 倍の性能を実現可能であることがわかる。プロセッサの高速化を考慮すれば現在よりも一桁以上高速な「超高速インターネット」の実現が期待できる。

もちろん、プロトコル処理の一部を「零」とすることは、ソフトウェアチューニングなどの改善的手法で実現できるものではない。処理そのものをなくすための

2 この場合単一の宛先と通信しているためフォワーディングキャッシュが常にきき、ルーティングテーブル検索は行われていない。

抜本的なアーキテクチャ変更が必要である。そこでプロトコル処理をコンピュータ /OS から Network Interface Card (NIC) にオフロードするとともにハードウェア化する方法が考えられる (図 - 3)。

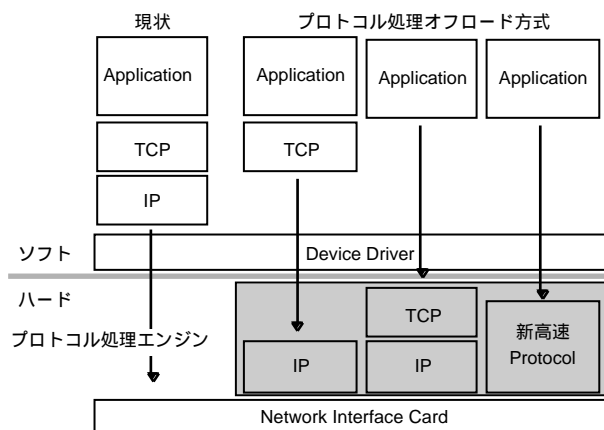


図 - 3 プロトコル処理オフロード

## 3. プロトコル処理ハードウェア化の検討

### 3.1. ハードウェア化の問題

プロトコル処理のハードウェア化は決して新しいアイデアではない[4]。現在でも NIC にマイクロコントローラを仕込んでプロトコル処理をさせるものはよくみかけるし、マルチプロセッサによって高速化させる方法もある。専用のインターネットルータではプロトコル処理機能の一部をハードウェア化するのがむしろ一般的である。

しかしながらプロトコル処理、特に Internet Protocol のような高機能プロトコルのハードウェア化には一つの大きな壁がある。それは Internet Protocol そのものの流動性である。現在最も広く使われているネットワーク層のプロトコルである IPv4 でさえプロトコルオプションが今後変更されない保証はないし、現実には IPsec などの新しい機能が追加されている。ましてや次世代 IP として仕様が固められつつある IPv6 はまだかなりの仕様変更や拡張があると考えなければならない。さらにトランスポート層以上のプロトコルについては確定することなど考えられない。

このように仕様変更が想定される処理をハードウェア化する場合、その内容が制限されるのはやむをえないことである。実例をみると checksum 機能のみをハ

ドウェア化した例、ルーティングテーブル検索をハードウェア化した例、FPGA ( Field Programmable Gate Array ) を利用してある程度プログラマブルなハードウェアとして実現した例、RISC 系のマイクロコントローラを利用した例などがあるが、いずれもある種の制限を含んだ限定的な実現であって、完全に Gigabit ネットワークに対応可能なものは今のところない。

プロトコルの変更 / 拡張に対応するという意味では、NIC にプロトコル処理専用のマイクロコントローラを使用したり、専用ルータのように機能別にマルチプロセッサ化してしまう方法のほうが安全である。またプロトコル処理が基本的にパケット毎に独立であることを利用してパケット単位に並列処理する考え方もある。

しかしこれらのアプローチはスループット、遅延の面で最善の方法とはいえない。NIC に安価で低速のプロセッサを載せて処理させるより、高速なメインプロセッサに処理させたほうが速い場合が多い。マイクロプロセッサの高速化、低価格化が非常な速度で進んでいる今日にあっては、マルチプロセッサアーキテクチャも一年で性能が陳腐化してしまう危険がある。現実に専用ルータの制御プロセッサは最新のプロセッサに比べるとかなり遅い。

さらにパケット単位に並列処理する方法もある。この方法では全体のスループットを向上させることができるが、一つのパケットを処理するに要する時間はプロセッサの処理時間で制限されるので、プロトコル処理遅延という点では改善できない点が問題である。

結局のところ、プロトコル処理の高速化について従来実現されてきた方式はある条件に限定して適用されるものであって、本質的かつ永続的な解決策とはなっていないように思われる。そこで究極のアプローチである「プロトコルの変更 / 拡張に柔軟に対応でき、かつネットワーク速度に比例する速度で処理可能」なプロトコル処理機構が求められる。

### 3.2. プロトコル処理エンジン

「ハードウェア化」ということをもう少し具体化してみよう。例えば Gigabit Ethernet で処理しなければならないデータレートは送受信それぞれ 100MB/sec である。

通常 16bit 幅、50MHz で送受信を行うので、片方向で 2 バイトデータを 20ns 以内に処理すればよい。4 バイトデータならば 40ns である。この速度は安価な CMOS テクノロジーで容易に実現できる。

次にプロトコル処理とはネットワーククロックに同期してデータを入出力しつつ、受信の場合はデータのシーケンスを解析、処理し、送信の場合は次のデータを準備することである。この処理はネットワーククロックに同期して動作する同期式順序回路として実現できる。すなわち Gigabit Ethernet 用のプロトコル処理とはたかだか 50MHz のクロックで動作する同期式順序回路に他ならない。そこでプロトコル処理エンジンを再構成可能な有限状態機械として定義し、「ストリームプロセッサ」と名付ける。

従来のノイマン型プロセッサは命令ストリーム (ソフトウェア) を高速に実行する固定的な有限状態機械をハードウェアで実現し、命令ストリームを変更することによって処理を変更する。この結果、命令の処理性能よりも小さいデータ処理性能しか得ることができない (図 - 4)。特にプロトコル処理のように条件分岐が多い処理ではワード毎に演算し、比較し、ジャンプし、必要な操作をするのが基本となるため、データストリームの処理性能は命令ストリームの 10% 以下となることも珍しくない。

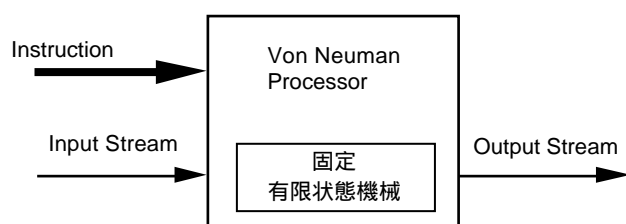


図 - 4 ノイマン型プロセッサ

これに対してストリームプロセッサは再構成可能な有限状態機械であって、プロトコル処理向けの演算器としてチェックサム機能、比較器、テーブル検索などを持つ。ストリームプロセッサではワード毎の演算、比較、操作を一回の状態遷移で行うことができる。すなわちストリームクロックに同期してプロトコル処理を行うことができる。このためノイマン型プロセッサにおける命令ストリーム処理性能レベルのデータ処理性能を実現可能となる (図 - 5)。



図 - 5 ストリームプロセッサ

本プロジェクトではストリームプロセッサのアーキテクチャを決定し、実現することが研究開発の大きな柱である。

### 3.3. プロトコル処理エンジンの実現

本プロジェクトではプロトコル処理エンジンを、ソフトウェアエミュレーションによる機能と性能の評価、ハードウェアの実現、の二フェーズで開発する。現在は第一のフェーズを進めている。

図 - 6 に試作したネットワークアダプタ (Comet) のブロックダイアグラムを示す。このネットワークアダプタは PCI (Peripheral Component Interconnect) 規格準拠の標準サイズボードで、PMC (PCI Mezzanine Card) 規格のドータボードを二枚搭載できる。プロトコル処理エンジンをエミュレートするためのプロセッサとして 166MHz の 64bit RISC を用い、エミュレーションソフトウェアをオンキャッシュで動作させている。また、一部機能をハードウェア化するため 5 万ゲート相当の FPGA を用意している。内部 PCI は現在 32bit、33MHz であるが、次の試作では 64bit 化の予定である。

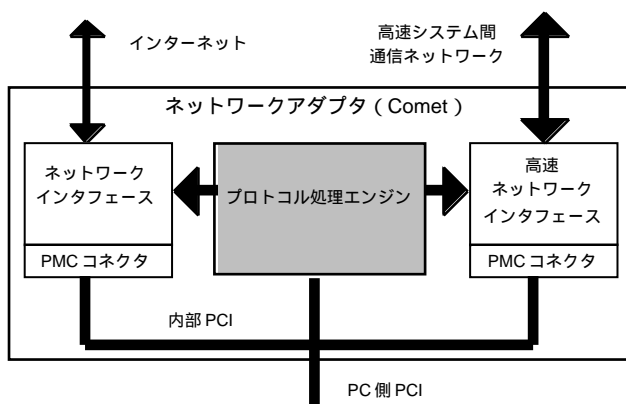


図 - 6 ネットワークアダプタ構成

ネットワークとしては現在 FastEther (100Mbps) を用いており、Comet 用のデバイスドライバを開発したところである。Comet デバイスドライバは OS に対して

ネットワークの種類に依存しない汎用的なドライバインタフェースを提供する。このため単一のデバイスドライバで様々なネットワークに対応可能である。

今後は WAN として ATM Megalink (135Mbps)、LAN として IEEE 1394 (200~400Mbps)、Gigabit Ethernet (1Gbps) を採用する予定である。将来 POLO などのさらに高速な標準ネットワークが利用可能になれば随時プラットフォームとして採用していく。

## 4. ネットワークサーバへの応用

### 4.1. 並列ネットワークサーバ

プロトコル処理エンジンを搭載したネットワークアダプタを一般的なコンピュータに追加するだけで、そのコンピュータは強力なネットワークサーバとなる。そればかりか、実効 100MB/sec 以上の通信性能は一世代前の MPP で用いられていた SAN と同レベルであり、高速システム間通信ネットワークとして使用できる領域に達する。そこでネットワークサーバを高速かつ高信頼なネットワークで結合して並列システムとして動作させることが期待される (図 - 7)。

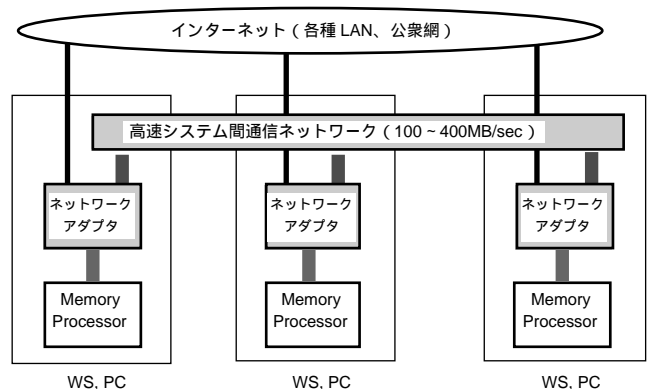


図 - 7 ネットワークサーバの高速相互接続

このようなシステムでは従来のインターネット通信パラダイムに加えて、クラスタ向け軽装プロトコル[5]や分散共有メモリ[6]のような並列分散処理の通信パラダイムを効果的にサポート可能となる。例えばプロトコル処理エンジンは分散共有メモリのコンシステンシ制御を高速かつプロセッサの負荷なしに実現できる。このような技術を用いることでネットワークサーバを高速、高信頼なネットワークで結合した「並列ネットワークサーバ」を実現できる (図 - 8)。

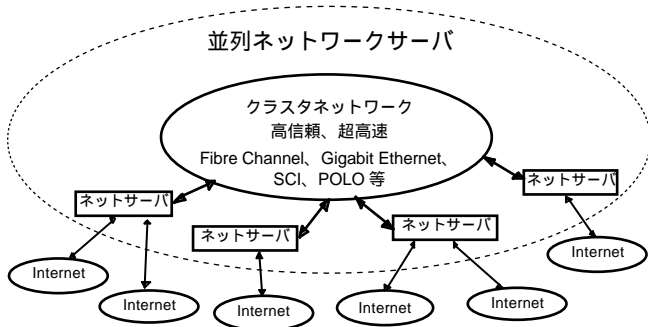


図 - 8 並列ネットワークサーバのシステムイメージ

今年度構築予定の実験ネットワークの構成を図 - 9 に示す。ATM Megalink (135Mbps) を用いて広域分散型の並列ネットワークサーバを構築する。

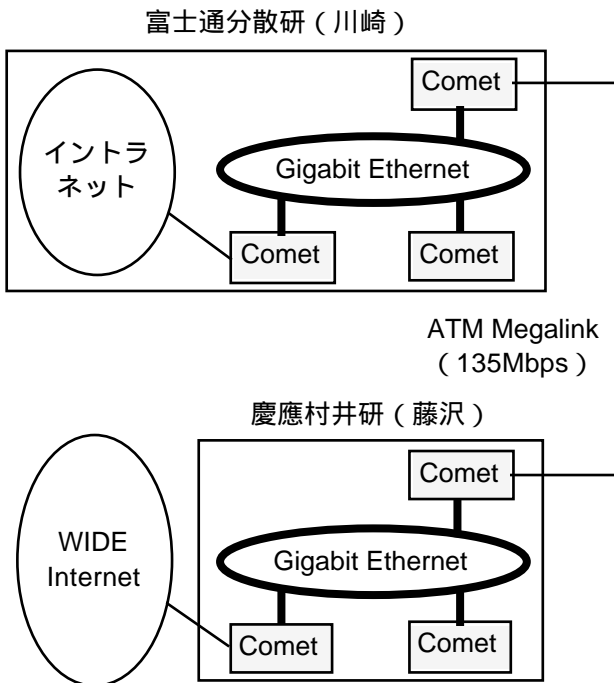


図 - 9 実験ネットワーク

#### 4.2. 超高速インターネットルータ

インターネットでは「拡張性」ということが重要な課題となっている。ここで「拡張性」とはより多くのネットワークやコンピュータをつなぎ、より多くのトラフィックを扱うことを意味する。インターネットにおいて拡張性を実現するのはルータであるが、このルータに求められる機能、処理能力は厳しいものとなっており、将来にわたって機能、性能ともに拡張可能なアーキテクチャが強く求められている。

ルータの処理は経路の管理とネットワークから別のネットワークの packets の中継である。この処理を並列

ネットワークサーバで実現する時、経路の管理は分散したネットワークサーバ間で協調して行うが、パケット中継は個々のサーバが共通の経路に基いて独立に実行できる。パケットの中継はサーバ同士を結合する高速システム間ネットワークを用いて行う。

経路管理のためには経路情報をコンシステンシを保ちつつ共有することが必要である。このためには本質的に並列システムの仕組み、例えば分散共有メモリなどの技術を利用できる。

パケット中継処理についてはストリームプロセッサによるハード処理が大きなポイントになる。ストリームプロセッサはパケットの形式判定、経路情報の検索、転送の一連の処理をネットワークからの到着に従って逐次実時間 (on the fly) 処理できる。これにより中継の遅延を削減することが期待できる。現在試作している Comet では IP パケットを受信してからルーティングテーブルを検索し、中継送信を開始するまでの遅延を 5 μ秒以下で実現することを目標としている。

並列ネットワークサーバによって多くの高速なネットワークを結合し、それらのネットワークの間を 5 μ秒以下という非常に小さな遅延で結合する時「超高速インターネット」が現実のものとなる。

### 5. プロジェクトまとめ

#### 5.1. 研究開発技術

標準技術として以下のプラットフォームを使用する。

表 - 1 プラットフォーム

コンピュータ	PC/AT 互換機、Workstation
オペレーティングシステム	UNIX系、Windows系
ネットワーク	ATM (155Mbps) 100BaseTX (100Mbps) IEEE1394 (200Mbps ~ ) Fibre Channel ( ~1Gbps) Gigabit Ethernet (1Gbps) SCI (1.6Gbps) POLO (4Gbps ~ )
通信プロトコル	インターネットプロトコル IPv4、IPv6
アプリケーション	インターネットルータ

上記プラットフォームにおいて「超高速インターネット」を実現するための具体的な研究開発項目としてはアプリケーションレベルの通信性能向上があり、これを実現するためのブレークスルーとして、システム階層に対応した以下の技術を研究開発する。これらの技術はあらゆる計算機システムに適用可能である。

表 - 2 開発技術

システム階層	ブレークスルー技術
ネットワークアダプタ	ハードによる高速通信処理 ソフトインタフェースの効率化
デバイスドライバ	標準プロトコルの高速処理技術 新高速通信プロトコル
通信ライブラリ	高速化実装技術

次に応用システム（インターネットルータ）の構築を通して「超高速インターネット」の完成を目指すとともに開発成果の実用化促進ならびに標準化を行う。

## 5.2. 研究開発体制

本プロジェクトでは開発技術を標準へとフィードバックすることが重要である。そこで研究成果を積極的に公開し、進捗に応じて他の研究機関との情報交換、討論を可能な限りオープンに行っていく。さらに試作など成果物を積極的に他機関に提供し、評価を依頼することで開発技術の普遍性を確保していきたい。

具体的には並列分散システム向けオペレーティングシステム、分散共有メモリに関して新情報処理開発機構つくば研と、超高速インターネットの実現および運用評価に関して WIDE Project 参加組織との協力を考えている。また標準化を進めるために IETF での積極的行動を行う。

## 5.3. スケジュール

全体としては3年後に並列ネットワークサーバの試作および評価を完了する。ネットワーク性能は1~4Gbpsを想定する。その後並列ネットワークサーバ上にインターネットルータ機能を実装し「超高速インターネット」の運用評価を行う予定である。

ストリームプロセッサは今年度にアーキテクチャを確定する。当初 FPGA で試作し、3年後には Gate Array

化することを目標にしている。

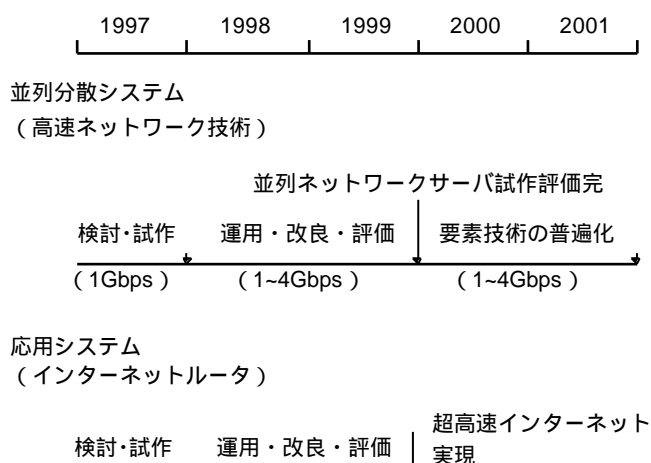


図 - 10 開発予定

## 6. おわりに

通商産業省が行う次世代情報処理基盤技術開発事業の一環として今年度より5年間の予定で開始した並列ネットワークサーバプロジェクトの概要と予定を述べた。本プロジェクトにより、普及しているPCやインターネットによる極めて低コストな並列分散システムを実現することが期待できる。またGigabitネットワークに対応可能なインターネットルータなどのネットワーク装置を低コストの標準技術で実現し、「超高速インターネット」の構築を可能とする。

現在世界中でインターネットに関わる研究開発が盛んに行われている。その中において国家プロジェクトの枠組みで本プロジェクトを行う意義は本当の意味でのオープンな技術開発を実現することにある。

インターネット技術は開かれた場で議論され、実現され、評価されることによって発達してきた。インターネット技術が世界に受け入れられた大きな理由の一つはこの公開性にあると考えられる。ネットワークシステムでは普遍性（だれでも使える、どこにもある）、相互接続性（だれとでもつながる）が重要である。一研究者、一企業、一研究機関の枠内で独占的、閉鎖的に開発された技術は、それがいかに優秀であっても必ずしも有効利用されるとは限らない。我々は実際に社会に貢献できる技術開発を行うことを念願し、開かれた研究開発を試みたい。

## 参考文献

- [1] 後藤, 村上: ギガビットネットワークの壁 情報処理学会誌、Vol. 36, No. 7, Jul. 1995
- [2] 次世代情報処理基盤技術開発 (RWC-RWI/PDC) 事業推進委員会: 次世代情報処理基盤技術開発 (RWC-RWI/PDC) 事業計画添付資料、May 1997
- [3] T. Wicki “A Multiprocessor-based controller architecture for high-speed communication protocol processing”, Swiss Federal Institute of Technology Zurich, Doctorial Thesis, 1990
- [4] P. Druschel “Operating System Support for High-speed communication, CACM, Vol. 39, No. 9, Sept. 1996
- [5] A. Jinzaki, T. Niinomi, S. Kobayashi “Illinois Fast Messages on 1Gbps Fibre Channel”, ACM ASPLOS-VII, NOW/Cluster Workshop, Oct. 1996
- [6] M. Morris, A. Jinzaki, T. Niinomi, Y. Ageo, S. Kobayashi "Fast Communication in Distributed Systems Using the Networked Virtual Memory System", IEEE Computer Society TCCA Newsletter, 63, 6, Aug. 1993