2005－HPC－103（19）
2005／8／4

10GbE 対応ネットワークテストベッド GtrcNET-10 の構成と評価

児玉祐悦[†]　工藤知宏[†]　清水敏行[†]

我々はネットワークの観測、エミュレーション、新規プロトコルの試験等を目的としたネットワークテストベッドの開発を行っている。今回 10 ギガビットイーサネットに対応した新たなテストベッド GtrcNET-10 を開発した。GtrcNET-10 の構成について述べるとともに、その簡単な評価結果について述べる。

GtrcNET-10 :Network Test-bed Supporting 10GbE – its configuration and preliminary evaluations

Yuetsu Kodama ,[†] Tomohiro Kudoh [†] and Toshiyuki Shimizu[†]

We have been developing a network test bed so as to observe network traffic, emulate networks and test communication protocols. We developed GtrcNET-10 supporting 10 Giga-bit Ethernet. In this report, we introduce the organization of GtrcNET-10 and show its preliminary evaluations.

1. はじめに

近年、広域網ネットワークのバックボーンとして 10 ギガビットイーサネット (10GbE) が利用されるようになってきている。また、ノード PC のネットワークインタフェースとして、ギガビットイーサネット (GbE) が標準で利用可能の場合が多く、GbE により多数の PC を接続する PC クラスタ向けなどに、10GbE をアップリンクとして持つイーサネットスイッチも多く市販されるようになってきた。さらに、各ノードから直接 10GbE でアクセス可能とするネットワークインタフェースカード (NIC) も各種市販されている。10GbE NIC は現在プロセッサとの I/O バスがボトルネックとなりワイヤレートを実現できないが、PCI Express x4 などによりこの問題が解消されれば、現在の GbE のように広く普及し、コストも劇的に安くなることが期待されている。

しかし、このような高速かつ大きな通信遅延をもつ広域網では、単一のアプリケーションでこれらの高バンド幅を有効利用するには様々な問題がある。我々はギガビットイーサネットを用いた広域網における問題点やその解決法などを、ネットワークテストベッド GtrcNET-1[1),2)] によるエミュレーション環境を用いて詳細に評価してきたが、普及が始まりつつ 10GbE を用いたネットワーク環境においても、同様に詳細な評価を可能とするようなネットワークテストベッドが求められていた。

今回、10GbE をサポートしたネットワークテストベッド GtrcNET-10 を開発した。本報告では、その構成と目的、およびこれを用いた評価について報告する。

† 産業技術総合研究所グリッド研究センター
Grid Technology Research Center, AIST

2. GtrcNET-10 の構成

我々は図 1 のような 2 種類の GtrcNET-10 を開発した。1 つは 10GbE を 2 ポート持つ GtrcNET-10p2、もう 1 つは 10GbE を 3 ポート持つ GtrcNET-10p3 である。以下ではそれぞれの GtrcNET-10 についてその構成を示す。

2.1 GtrcNET-10p2

GtrcNET-10p2 の構成を図 2 に示す。GtrcNET-10p2 は、2003 年度より開発を開始した。10GbE の光トランシーバとしては、XSBII ($16bit \times 644.53Mbps$) を持つ MSA300 ピンモジュール (Finisar FTRX-1432-E) を用い、同じく XSBII の入出力を持つ外付 10GbE MAC (Intel IXF18104) を用いている。したがって、MAC と FPGA との接続も XSBII である。

FPGA(Xilinx XC2VP75) を 2 個使用し、1 個が 10GbE の一方向のストリーム（1 つの受信ポートと 1 つの送信ポート、1 組の DDR-SDRAM ポート）を処理する。その他、各 FPGA は 1 ポートの GbE ポートを有している。外部メモリとしては DDR266 SDRAM を 2 ポート接続し、$64bit \times 2 \times 266MHz = 34Gbps$ のピークアクセス性能を有する。 その他、2 つの FPGA 間は FPGA 内のハード IP である RocketI/O により接続されている。 制御 PC とは USB2.0 で接続さ

(a) GtrcNET-10p2　　(b) GtrcNET-10p3

図 1　GtrcNET-10 の写真

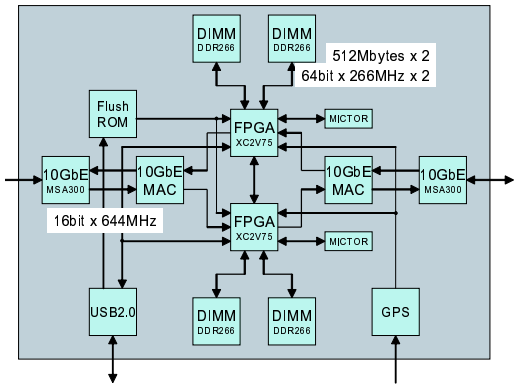


図 2　GtrcNET-10p2 構成図

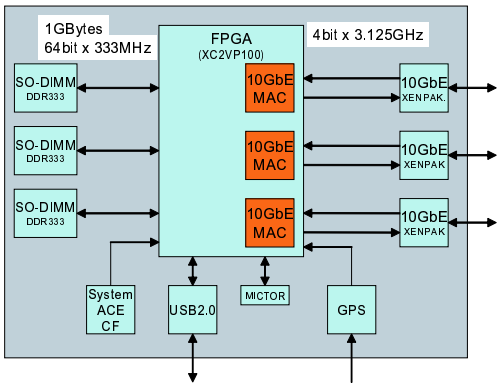


図 3　GtrcNET-10p3 構成図

れ、FPGA のコンフィギュレーション用のフラッシュ ROM の書き換えや、FPGA 内部のレジスタの読み書き、外部メモリへのアクセスなどが USB2.0 の I/F チップ (Cypress CY7C68013-128) を通じて行える。また、GtrcNET-1 と同様に、MICTOR コネクタを各 FPGA に有し、アジレント社の Trace Port Analyzer やロジックアナライザーなどを用いて FPGA 内部の信号を観測することが可能である。GPS モジュールとの接続用シリアルポートを有し、GPS からの時刻情報および PPS（1 秒間に 1 つのパルス信号）を用いて、GtrcNET-10 間の時刻同期をマイクロ秒レベルの高精度で行うことができる。

2.2　GtrcNET-10p3

GtrcNET-10p3 の構成を図 3 に示す。GtrcNET-10p3 は、2004 年度より開発を開始した。10GbE ポートと外部メモリの組を 3 ポート有している。当初は、GtrcNET-1 と同様に、4 ポート分のネットワークインタフェースやメモリインタフェースを持つことを検討したが、外部メモリや 10GbE モジュールのポートの同時動作や、外部メモリの制御回路や 10GbE MAC のゲート規模などから 3 ポートを 1 つの FPGA に接続するのが限界であった。3 ポートは、我々の提案しているマルチパスを用いた高信頼通信 [3] を 10GbE で実装するために必要な最低ポート数である。

10GbE の光トランシーバとしては、XENPAK モジュールを用い、FPGA(Xilinx　XC2VP100) とは FPGA 内のハード IP である RocketI/O を用いて直接 XAUI(4bit x 3.125GHz) により接続している。10GbE MAC としては、Xilinx 社の 10GbE MAC（ソフト IP) を用いている。これが 1 つの FPGA に 3 組搭載されている。また、外部メモリとして DDR333 SDRAM を 3 ポートもつ。これはパリティ用の 8 ビットのタグ領域を持つ SO-DIMM で、タグ部を除いたピークのアクセス性能は $64bit \times 333MHz = 21.2Gbps$ である。FPGA のコンフィギュレーションにはコンパクトフラッシュを用いた SystemACE を用いる他、ホスト PC との接続の USB2.0 や MICTOR コネクタ、GPS モジュールとの接続シリアルポート等を有する。

3. GtrcNET-10 の目的

GtrcNET-10 は、FPGA コンフィグレーションの変更により様々な用途に用いることができる。現在想定している用途は以下のようなものである。基本的には GtrcNET-1 と同様であり、それを 10GbE に拡張したものとなっている。

(1) 新規プロトコルのプロトタイプ
GtrcNET-10 では FPGA の回路を変更することにより、様々な機能をハードウェアで実現できる。例えば、マルチパスを用いた高信頼通信 [3] を実現できる。これは受信したフレームの複製を作り、元のフレームとともに両者を異なるパスを通るようにはカプセル化して 2 つのポートから出力する。2 つのパスから到着したフレームをマージすることにより、もし片方のパスでフレームが破棄されても、もう一方のパスを通って到着したフレームを用いて再送を行わずにパケットロスを回復できる通信方式である。GtrcNET-10 を用いると、このような新たなプロトコルを実ネットワーク上で実フレームレートで評価することが可能である。

(2) 10Gigabit Ethernet 上のパケットの観測
10GbE のリンクに GtrcNET-10 を挿入することにより、リンクを流れるイーサネットフレームを詳細に観測する。たとえば、各フレームのヘッダ部や受信時のタイムスタンプなどをキャプチャして、外部メモリに保存することにより、数マイクロ秒の遅延の増加だけで、他にリンクに影響を与えることなくフレーム単位の詳細な情報を入手できる。また、送信・受信転送バイト数と共に高精度な時刻情報を内部に保持することにより、ミリ秒レベルのデータ転送量を測定することもできる。この時刻情報を GPS からの信号に合わせることにより、WAN をまたがる通信の送信端と受信端に GtrcNET-10 を用意し、片方向の通信遅延をマイクロ秒程度の高い精度で知ることができる。

(3) 広域網のエミュレーション
広域網上でのソフトウェアやミドルウェアの性能評価には、遅延やネットワーク品質を自由に制御できるネットワークエミュレーションが有用である。GtrcNET-10 は、外部メモリにフレームを蓄え、一定の時間だけフレームの送信を待たせることにより、広域網の遅延を模擬することができる。外部メモリは 1GB の容量を持ち、10Gbps のデータ転送時に 800ms 強の遅延を与えることができる。また、通信誤りの模擬や、中間経路でのルータにおける輻輳や RED (Randum Early Detection) などのレートコントロール機能の模擬も可能である。

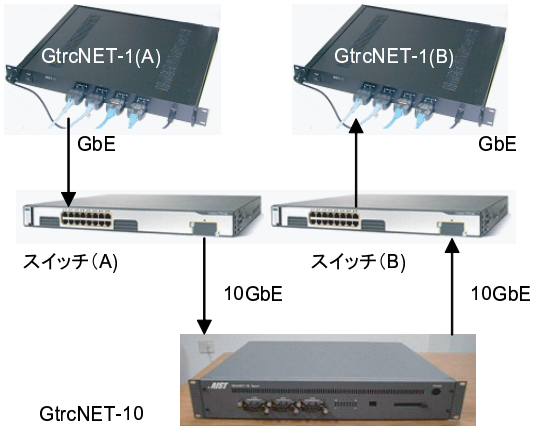


図 4 遅延測定環境

これらは、例えば広域網をエミュレーションしながらパケットを観測するなど同時に組み合わせて行うことができる。

4. GtrcNET-10 の評価

4.1 GtrcNET-10 自身の遅延

まず、図 4 に示す環境を用いて GtrcNET-10 自身の遅延オーバヘッドを調べた。GtrcNET-1 とスイッチ (Cisco 3750G-16TD) の間は GbE、スイッチと GtrcNET-10 の間は 10GbE である。ここでは、GtrcNET-1 の片道遅延測定機能を用いた。また、比較のために、スイッチ間を 10GbE で接続した場合と、GbE で接続した場合についても測定した。

片道遅延の測定方法は次の通りである。GtrcNET-1(A) から測定用の UDP パケットを生成し、スイッチ (A)、GtrcNET-10、スイッチ (B) を経由して GtrcNET-1(B) でパケットを受信できるように IP アドレス等を適宜設定する。この UDP パケットには、GtrcNET-1(A) から送信したときのタイムスタンプ、および GtrcNET-1(B) で受信したときのタイムスタンプが、それぞれの GtrcNET-1 のハードウェアにより挿入される。ここで挿入されるタイムスタンプは、RFC1305 に準拠した 64bit 形式 (1900 年 1 月 1 日からの秒数部 (32bit) と固定小数点形式の秒端数部 (32bit)) からなる。ただし、GtrcNET-1 の実装では秒端数部は上位 24bit のみ有効で下位 8bit は 0 固定である。時間分解能は $2^{-24} = 59.6$ns であるが、31.25MHz のクロックを、GPS からの PPS 信号 (pulse per second) により計測した周波数で、2^{24}Hz カウンタに変換しており、また、PPS 自体の精度がマイクロ秒程度なので、タイムスタンプの精度も同程度と考えられる。

この受信パケットを解析することにより片道遅延を測定した結果が、表 1 である。スイッチでは基本的

表 1 GtrcNET-10 の遅延

	latency(マイクロ秒)						
	スイッチ直結			GtrcNET-10p2		GtrcNET-10p3	
IP 長	10GbE	GbE	diff	w sw	w/o sw	w sw	w/o sw
64	15.1	13.5	1.6	16.6	1.5	16.7	1.6
1500	54.8	41.9	12.9	58.8	4.0	57.9	3.1
9000	257.1	184.0	73.0	273.0	16.0	267.0	10.0

表 2 GtrcNET-10p3 内部回路の遅延

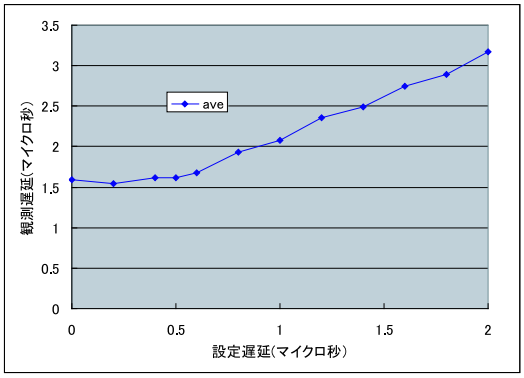
	遅延 (マイクロ秒)		
IP 長	全体	ユーザ回路	光モジュール等
64	1.621	0.509	1.112
1500	3.139	2.089	1.050
9000	9.965	8.947	1.018

にフレームをいったんメモリに格納してから、転送を行うストア-アンド-フォワードを行っているため、フレーム長によりその遅延が変化する。表では、IP 長 64、1500、9000 バイトのそれぞれについて 1000 回の遅延を測定し、その平均を示している。スイッチ直結の項は、10GbE および GbE でスイッチを直結した場合の片道遅延、およびその差を示している。これによると、IP 長が長くなった時のスイッチの遅延はかなり大きいことが分かる。一方、各 GtrcNET-10 の項は、スイッチの遅延を含む測定値 (w sw) と、スイッチの遅延を除いた GtrcNET-10 自体の遅延 (w/o sw) を示している。IP 長が 9000 バイトの時には、10Gbps でのフレームの転送だけで 7.2 マイクロ秒かかることを考えると、GtrcNET-10 の遅延は十分小さいといえる。

また、GtrcNET-10p3 にはフレームを FPGA 内部の MAC から受信開始した時刻と、FPGA 内部の MAC へ送信開始した時刻を保持する機能がある。この機能は、送受信する度に内部レジスタの値を変更するため、単発のフレームについてのみ、その内部遅延を測定できる。測定分解能は 25.6ns である。表 2 に測定結果を示す。GtrcNET-10p3 全体の遅延が、上記の GtrcNET-1 を用いて測定した結果であり、ユーザ回路が外部メモリへの読み書きを含むユーザ回路部分の遅延の測定結果 (10 回の平均) である。全体の遅延からユーザ回路の遅延を差し引いた遅延が、光モジュールや FPGA 内のシリアルリンクハードコア IP である RocketI/O、XAUI インタフェース、10GbE MAC インタフェースの遅延である。ユーザ回路を除いた部分は、IP 長によらずにほぼ 1 マイクロ秒であることが分かる。

4.2 遅延挿入機能の確認

GtrcNET-10 には、ネットワーク遅延を模擬する機能を実装している。ネットワーク遅延の実装方式は次の通り。受信したフレームを外部メモリに格納する前に、指定した遅延時刻を受信時刻に加えた値を、フレームの先頭部にタイムスタンプとして追加し、送信時に



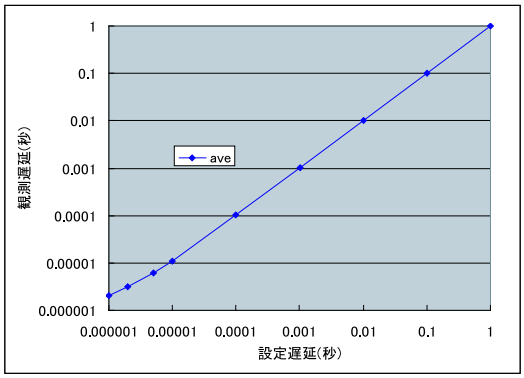


図 5 評価ネットワーク環境

現在の時刻がそのタイムスタンプより大きくなるまで送信を待たせる。現在の実装では、タイムスタンプは 25.6 ナノ秒ごとに 1 カウントアップするカウンタを用いている。将来的には GtrcNET-1 と同様に GPS 信号で同期する固定少数点形式に変更する予定である。

遅延機能が正常に働いていることを確認するために、図 4 に示した環境で、GtrcNET-10 で遅延を設定し、その片道遅延の変化を測定した。IP 長は 64 バイトとし、GtrcNET-10p3 を用いて測定した。図 5 はその結果である。x 軸が指定した遅延時間、y 軸がスイッチ分の遅延を除いた測定結果である。下のグラフは、指定範囲を 1 マイクロ秒から 1 秒としたグラフであり、両方の軸とも対数で表わしている。ほぼ指定した値が観測されていることが分かる。一方、上のグラフは、指定範囲を 0 秒から 2 マイクロ秒とした時のグラフである。ユーザ回路の遅延である 0.5 マイクロ秒以下を

(a) 評価ネットワーク環境1

(b) 評価ネットワーク環境2

図 6 評価ネットワーク環境

指定しても、GtrcNET-10 自体の最小遅延である 1.5 マイクロ秒が観測されるが、それ以上の値を指定すると、指定値＋ 1 マイクロ秒の遅延が観測されていることが分かる。IP 長を変化させた場合も、その IP 長のユーザ回路遅延以下では遅延が変化せず、それ以上では設定値＋ 1 マイクロ秒の遅延が観測される。

4.3 高精細バンド幅測定機能の確認

GtrcNET-10 にはバンド幅を高分解能で計測する機能を実装している。バンド幅計測機能は、受信バイト数および送信バイト数を計測するレジスタと共に、内部時刻を保持するレジスタをホスト PC から一定間隔毎に読み出すことにより、測定可能である。

ホスト PC と GtrcNET-10 の接続は USB2.0 を用いており、現在の実装では 0.3 ミリ秒に 1 回程度のレジスタアクセスが可能である。しかし、現在の実装では、一度に 1 つのレジスタしかアクセスできず、かつ、レジスタアクセスの間隔は OS 等の影響によりばらつきが生じるため、送信バイト数と時刻を別々に読み出したのでは、正しいバンド幅を得ることはできない。そのため、時刻と送信バイト数を 16bit ずつ組み合わせたレジスタを新たに実装した。これにより、読み出し間隔がばらついたときでも、正しいバンド幅を得ることができる。ただし、送信バイト数が 16bit であるため、バイト単位でカウントすると 10Gbps のデータ転送時は 50 マイクロ秒でオーバフローしてしまう。そのため、送信バイト数は 256 バイト単位としている。これでも 10Gbps のデータ転送時は 13 ミリ秒でオーバフローしてしまうため、それより短い間隔でのレジスタ読み出しが必要である (データ転送がそれより低ければそれだけ制限は緩められるが)。このため、今後複数のレジスタ読み出しを一度の USB アクセス要求で行えるように改良する予定である。

また、現在の実装では、転送バイト数はフレーム転送完了時にしか更新されない。したがって転送途中でレジスタを読み出した場合は、そのフレームの転送バイト数はカウンタに反映されておらず、次に読み出したときにはフレームサイズ分が一度にカウンタに加算されることとなる。このため、最大フレーム長だけの誤差が転送バイト数には含まれることになり、レジスタ読み出し間隔を短くすればするだけ、相対的な影響は大きくなってしまう。例えば、フレーム長を 1500 バイトとして、1 秒毎に測定すれば、その誤差は 12Kbps であるが、1 ミリ秒毎に測定すれば、誤差は 12Mbps に広がる。10Gbps のデータ転送が行われていれば相対誤差は 0.12%と小さいが、1Gbps のデータ転送では、1.2%、100Mbps のデータ転送では 12%と大きくなるため、測定データの評価の場合は注意が必要である。現在の実装では、上で述べたようにデータ bit 数が少なく、測定間隔をあまり広げられないため、測定したデータをあとで平均化するなどの処理が必要である。今後、このような処理が不要なように、転送バイト数のカウントをバイト転送毎に行うように改良する予定である。

表 3 評価ネットワーク環境の諸元

環境 1	
Processor	Pentium4 2.8GHz
Mother Board	Intel D865GLC
Main Memory	1GB (DDR400)
NIC	Intel 82547EI (CSA)
OS	RedHat Linux 9 2.4.20
Switch	Cisco 3750G-16TD
環境 2	
Processor	Xeon 2.4GHz dual
Mother Board	SuperMicro X5DAE
Main Memory	2GB (DDR266)
NIC	Netrion Xframe
PCI	PCI-X 133MHz
OS	FedraCore3 2.6.9

バンド幅測定機能を確認するために、図 6 に示す 2 つのネットワーク環境を構築し、各 PC 間で iperf コマンドによりバンド幅の変化を詳細に測定した。環境 1 は 10GbE のアップリンクを持つ 2 つの GbE スイッチ間に GtrcNET-10 を挿入し、スイッチにそれぞれ 8 台の PC を接続した環境、環境 2 は 2 つの PC に 10GbE NIC を搭載してそれらを GtrcNET-10 に接

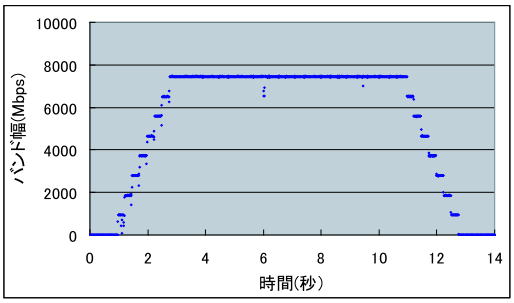


図 7 環境 1 におけるバンド幅測定結果

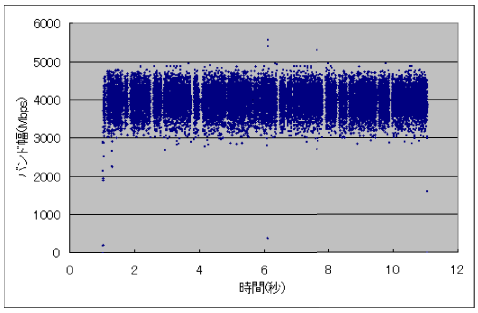


図 8 環境 2 におけるバンド幅測定結果

続した環境である。各環境における諸元を表 3 に示す。

図 7 は、環境 1 を用い、8 個の 900Mbps の UDP ストリームを 10GbE 上に流したときの様子である。各ストリームは 250 ミリ秒ずれて開始している。カウンタのオーバフローを防ぐため、レジスタアクセスを 10 ミリ秒間隔としている。測定結果からは、900Mbps づつバンド幅が増加し、総計 7.2Gbps のデータ通信が安定して行われていることが観測できる。

図 8 は、環境 2 を用い、MTU 9000、ソケットバッファサイズ 9.5MB で iperf を実行した結果である。レジスタアクセスを連続的に行ったところ、平均 500 マイクロ秒間隔でアクセスが行えた。所々グラフの空いている部分はデータ転送が行われていないのではなく、ホスト PC 側の OS 等の影響で、転送バイト数レジスタのアクセス間隔が空いてしまっているためである。最大 100 ミリ秒の空きが観測された。そのため、カウンタがオーバフローしてしまっており、本グラフでは測定後に補正を行っている。GtrcNET-1 では精密バンド幅測定機能として、ホスト PC からレジスタをアクセスするのではなく、GtrcNET-1 で定期的にレジスタをアクセスして FIFO メモリに蓄え、ホスト PC からはそのメモリをまとめて読み出す実装を行っている。GtrcNET-10 にもこの機能を実装する予定である。測定結果によると、約 4Gbps のデータ通信が行われているが、±約 1Gbps のバンド幅の揺れが観測されている。転送バイト数がフレーム単位であることの影響は、144Mbps(約 500 マイクロ秒の測定間隔で 9000 バイトの揺れがあることより、$9000 \times 8/0.0005 = 144Mbps$) 程度であり、観測される揺れはそれよりもかなり大きい。

5. おわりに

GtrcNET-10 は、FPGA と高速大容量メモリを搭載したシンプルな構造のネットワーク実験装置である。FPGA のプログラムにより様々な用途に用いることができる。ギガビットイーサネットの場合は、PC 上のソフトウェアでもある程度可能であったが、10 ギガビットイーサネットではソフトウェアでワイヤレート処理することは困難で、ハードウェアによる実現が不可欠である。

本稿では、GtrcNET-10 の構成と目的について述べ、GtrcNET-10 を用いて広域網に相当する遅延の挿入とミリ秒レベルの高分解能なバンド幅測定機能を実装した初期的な GtrcNET-10 の評価を報告した。

今後 GtrcNET-10 の目的で述べた各種機能を FPGA に実装し種々のネットワーク環境での評価を行っていく予定である。

謝辞

本研究の一部は、新エネルギー・産業技術総合開発機構基盤技術研究促進事業（民間基盤技術研究支援制度）の一環として委託を受け実施している「大規模・高信頼サーバの研究」の成果である。GtrcNET-10 の開発にご協力いただいた三精システム (株) の藤代氏、東京エレクトロンデバイス (株) の森川・長井・菅原氏、および (株) アテックスの三島氏に感謝する。

参 考 文 献

1) Y. Kodama, T. Kudoh, R. Takano, H. Sato, O. Tatebe and S. Sekiguchi, “GNET-1: Gigabit Ethernet Network Testbed,” Proc. of 2004 IEEE International Conference on Cluster Computing (Cluster2004), pp.185-192, 2004.
2) http://www.aist.go.jp/gnet/
3) Y. Kodama, T. Kudoh and S. Sekiguchi, “Dependable Communication Using Multiple Network Paths,” Proc. of 2005 IEEE International conference on Dependable System and Network (DSN05), Supplement pp. 2005.