特集

1チップ10ギガイーサネットスイッチLSI

Single-Chip 10-Gigabit Ethernet Switch LSI

あらまし

柔軟で信頼性の高いITシステムを構築するため，高速サーバや大容量ストレージを接続する小型・低コストでレイテンシも小さな10ギガビットイーサネットスイッチが強く求められている。このような状況で，富士通は，世界初となる1チップの10ギガビットイーサネットスイッチLSIを開発した。本LSIは，12ポートの10ギガビットイーサネットインタフェースを備え，レイヤ2のスイッチ機能を持つ。また，新たに開発した銅線ケーブルで10ギガビットイーサネット信号を25 m伝送できるIO回路（eXAUI）をチップ上に搭載しており，システムの小型化，低コスト化，省電力化を可能にした。本LSIを用いた10ギガビットイーサネットスイッチ製品も既に開発されており，データセンタ，高性能コンピュータなどの分野で使われている。

本稿では，本LSIを実現する主要な技術について述べ，本LSIの機能と構成，eXAUIの構成，LSIの実装について解説する。最後にLSIの性能評価とリファレンスボードを紹介する。

Abstract

Compact, low-cost, and low-latency 10-gigabit Ethernet switches have been urgently needed to connect high-speed servers and large storage systems for constructing flexible, highly reliable IT systems. To meet this need, Fujitsu has developed the world's first single-chip, 10-gigabit Ethernet switch LSI, which features 12 ports for the 10-gigabit Ethernet interface and layer-2 switch functions. This LSI has a newly developed I-O circuit (eXAUI) that can transfer 10-gigabit Ethernet signals over a 25-meter copper cable, thus making it possible to reduce the size, cost, and power consumption of IT systems. We have already developed 10-gigabit Ethernet switches incorporating this LSI, and they are now in use in data centers and high-performance computers. This paper describes the key technologies applied to this LSI, its functions and structure, the structure of the eXAUI circuit, and the integration of its circuits. This paper also includes an evaluation of the LSI's performance and a reference board design.

堀江健志（ほりえ　たけし）

米国富士通研究所高性能インタコネクトテクノロジ部門 所属
現在，高性能インタコネクトの研究開発に従事。

清水　剛（しみず　たけし）

米国富士通研究所高性能インタコネクトテクノロジ部門 所属
現在，高性能インタコネクトの研究開発に従事。

服部　彰（はっとり　あきら）

米国富士通研究所高性能インタコネクトテクノロジ部門 所属
現在，高性能インタコネクトの研究開発に従事。

まえがき

高速サーバや大容量ストレージをネットワークで接続し，柔軟で信頼性の高いシステムを構築するために，高スループットで汎用性のあるインタコネクトが求められている。その要求に応える標準インタコネクトとして10ギガビットイーサネット（10GbE）が注目されている。しかし，従来の10GbEスイッチは，広域ネットワーク用の大規模な装置が一般的で，物理的なサイズやスイッチのレイテンシが大きく，かつ高価であった。このため，ブレードサーバやサーバ・ストレージ間をつなぐ，小型・低コストで，レイテンシも小さな，10GbEスイッチが強く求められていた。

今回，銅線ケーブルで10ギガビットのイーサネット信号を25 m伝送できるIO回路を備えた1チップの10GbEスイッチLSI（MB87Q3070）を開発し，世界で初めて12ポートの10GbEスイッチの1チップ化に成功した。本LSIにより，10GbEスイッチをワンボードで実装することも可能となり，また高価な光モジュールを使わずに電気信号による銅線ケーブル接続やプリント板によるバックプレーン接続が実現できるため，システムの小型化，低コスト化，省電力化が可能となる。

本稿では，本LSIを実現する主要な技術について述べ，本LSIの機能と構成，IO回路の構成，LSIの実装について解説する。最後にLSIの性能評価とリファレンスボードを紹介する。

技術的な特徴

10GbEスイッチMB87Q3070は，以下のような主に四つの技術的な特徴を持つ。

（1） 12ポートの10GbEスイッチを1チップに集積

従来の10GbEスイッチは，汎用の通信装置であるため，多種のインタフェース，レイヤ3以上の機能，長距離伝送のための大容量のバッファメモリなどをサポートしており，1チップ化が困難であった。そこで，機能をITシステムのインタコネクトに必要なものに絞り込み，かつ，インタフェースを10GbEに限定することにより1チップ化を実現した。すなわち，レイヤ2でのスイッチングを基本機能とし，共有メモリとクロスバスイッチの構成方式・制御方式を開発し，スイッチング処理に必要な高速バッファメモリや高速IOマクロを含め，12ポートの10GbEスイッチを1チップに集積した[1]

（2） 240 Gbpsの高い帯域幅を実現

チップ上の複数のメモリブロックを効率良く利用して高速・大容量で多ポートの共有メモリをチップ上に実現する新たな方式（マルチポートストリームメモリ）を開発した（**図**-1）。このような構成にすることで，複数のメモリブロックを同時に読み書きできるため，10GbEスイッチの12ポートから同時に読出しと書込みの動作を実行でき，240 Gbpsの高いバンド幅を実現した。

（3） レイテンシを大幅に低減させるメモリ制御方式の開発

到着したパケットを短いレイテンシで出力側に送るため，共有メモリの高速スケジューリング制御方式を開発した（図-1）。これにより，従来，数μs以上かかったスイッチのレイテンシを450 nsと大幅な短縮を実現した。

（4） 銅線ケーブルで25 m伝送可能な高性能IO回路の開発

送信側と受信側両方に高周波ロスを補償する高性能なイコライザ回路を搭載したIO回路（eXAUI：enhanced 10 gigabit Attachment Unit Interface）を開発した[2] このIO回路により，10GBASE-CX4（IEEE標準規格）の15 mを上回る25 mの銅線ケーブルによる10GbEの信号伝送を実現した。これにより，高価な光モジュールを使わずに，筐体間または筐体内を接続でき，システムの大幅な低コスト化が可能となった。

LSIの機能

MB87Q3070の主な機能を**表**-1に示す。本LSIは，10GbEのインタフェースとレイヤ2のスイッチ機能

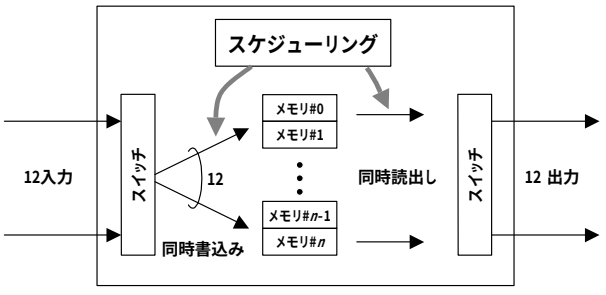

図-1　マルチポートストリームメモリ
Fig.1-Multiport stream memory.

表-1　機能概要

項目	仕様
ポート数	12個
準拠インタフェース	XAUIおよび10GBASE-CX4
スイッチスループット	240 Gbps
スイッチレイテンシ	450 ns（無負荷時）
スイッチング方式	カットスルー/ストアアンドフォワード
MACアドレス	8 Kエントリ
	自動学習とエージング
VLAN	IEEE802.1Q対応
	多重VLAN対応
QoS	IEEE802.1Q/p（4優先度）
マルチキャスト	MACマルチキャスト対応
ジャンボフレーム	最大15 Kバイト
スパニングツリー	IEEE802.1D対応
フロー制御	IEEE802.3ae全2重
ポートミラーリング	対応
ネットワーク管理	RMON，SMON統計情報
CPUインタフェース	モトローラMPC860バス

というITシステムのインタコネクトに必要なものに絞り込むことにより，1チップ化を実現している。

レイヤ2のスイッチ機能として，8 Kエントリの MAC（Media Access Control）アドレステーブルを持ち，アドレスの学習と削除をハードウェアでサポートしている。さらには，4 Kアドレスまでの VLAN（Virtual LAN）をサポート可能であり，ネットワークを論理的に異なるサブネットに分割することができる。また，スパニングツリー（ループ状に形成されたネットワークでデータが無限に循環するのを防止する機能）をサポートすることにより，ネットワークの冗長構成を可能としている。CPUインタフェースとして，MPC860バスをサポートしているが，EEPROMからのLSIの初期化も可能であり，スイッチ管理が不要な非管理型スイッチシステムも構成可能である。

LSIの内部構成

MB87Q3070の概略ブロック図を**図**-2に示す。

12個あるポートブロックは，IO回路（eXAUI），10GbEのMAC，フレームのフィルタ部，入力バッファから構成される。この入力バッファは，ストアアンドフォワードおよびスピードマッチングのために用いられる。12ポートの共有メモリを実現するマルチポートストリームメモリは，メモリバッファ，クロスバスイッチ，共有メモリ制御部から構成される。それ以外のブロックとして，ルーティングテーブル，統計カウンタ，CPUインタフェースがある。

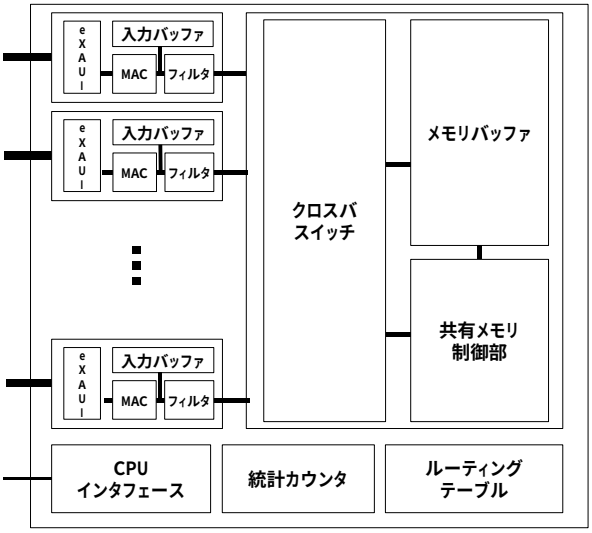

図-2　MB87Q3070のブロック図
Fig.2-MB87Q3070 block diagram.

マルチポートストリームメモリは，複数のメモリバンクを多段ネットワークで接続した構成をとっている。その制御方式は，可変長データの連続アクセスを実現し，さらに，カットスルー（送信先アドレスだけ見てフォワード）を可能とするクレジットベース（空き状況を確認して転送）のフロー制御を採用している。マルチポートストリームメモリは，共有メモリを実現することにより，高いスループットとともに，マルチキャストを効率良く実現できる。

なお，マルチポートストリームメモリそのものは，プロトコルに依存しない設計になっており，また，特別なメモリを必要としないため，イーサネット以外のプロトコルをサポートするLSIへの適用を容易にしている。

eXAUIの構成

eXAUIは，10GbEの仕様（IEEE 802.3ae）で規定されているインタフェースXAUI（光モジュールなどのデバイス間接続に使用される）と10GBASE-CX4（IEEE802.3ak）の両方に準拠するIOである。したがって，eXAUIの接続先には，XAUIのインタフェースを持つ光モジュールや10GbEのLSIを接続できるだけでなく，10GBASE-

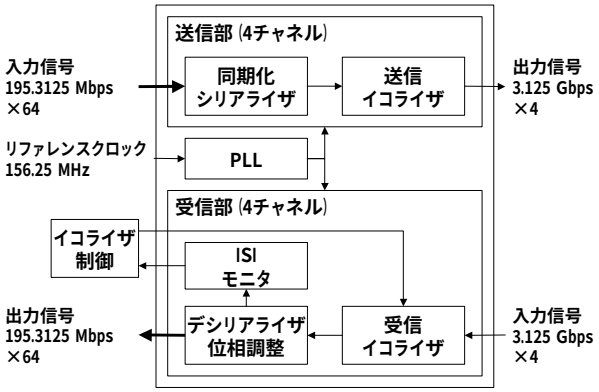

図-3　eXAUIの構成
Fig.3-eXAUI configuration.

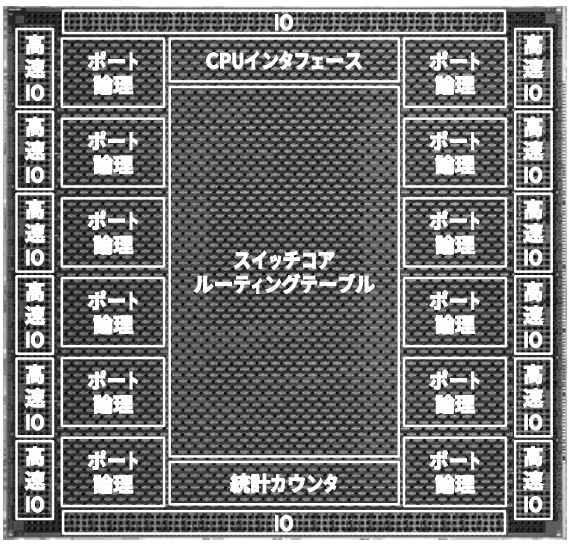

図-4　MB87Q3070のチップ
Fig.4-MB87Q3070 chip.

CX4のインタフェースを持つ10GbEのアダプタカードも直接接続可能である。

長いケーブルまたはプリント板のパターン，コネクタを使用して10GbE信号を伝送したとき，高周波の減衰による符号間干渉が問題になる。eXAUIは，符号間干渉による波形の乱れを補正するためにイコライザ回路を送信側と受信側の両方に搭載している（**図**-3）。このイコライザ回路により，システムの柔軟な構成が可能となる。例えば，筐体内の1 m以上のバックプレーン伝送，筐体間の25 mまでの銅線ケーブル伝送を実現可能であり，XAUIまたは10GBASE-CX4の規格を上回る伝送距離をサポートしている。

eXAUIマクロ1ポートは，3.125 Gbps転送可能なチャネルを送受信にそれぞれ4チャネルずつ持っている。eXAUIの回路には，送信側イコライザとして5タップのFIR（Finite Impulse Response）フィルタ回路，受信側イコライザとして2階微分フィルタ回路が実装されており，それぞれのイコライザ回路により約30 dBまでの減衰を補正することが可能である。イコライザのパラメタは柔軟に制御可能となっており，さらに，eXAUIに搭載されている符号間干渉をモニタする回路を使うことにより，伝送線路の高周波ロス特性に応じて制御する適応型イコライザ制御も可能としている。eXAUIのマクロは，ポートごとにパワーダウンモードが制御でき，本LSIの未使用ポートはパワーダウンすることにより，省電力化できる。

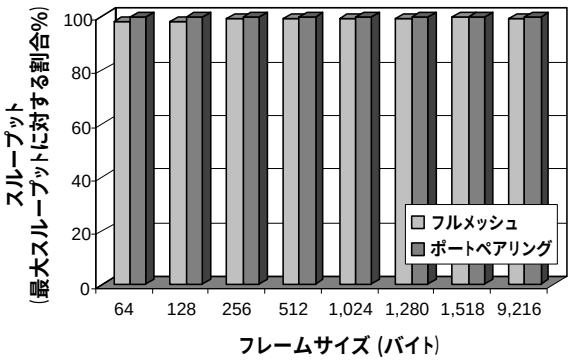

図-5　MB87Q3070の性能評価
Fig.5-MB87Q3070 evaluation.

LSIの実装

MB87Q3070のチップを**図**-4に示す。

LSIは0.11 μm CMOSテクノロジを用いて実装した。内部は，312.5 MHzで動作し，動作電圧は，1.2 Vおよび2.5 Vである。LSIの物理サイズは，16 mm×16 mm，LSIチップのパッケージは，728ピンのFCBGA（Flip Chip Ball Grid Array，35 mm×35 mm）である。パッケージでは，送信信号と受信信号間のクロストークノイズが問題とならないよう高速信号を配線した。

LSI内部は，両側にeXAUIマクロを搭載し対称的な構成をとることにより，規則的な内部ブロックの構成となるようにし，チップ内部のブロック間のタイミング設計が収束しやすいようにした。

MB87Q3070のように多数の高速IO回路を搭載したLSIを1チップで実現するための課題の一つに，

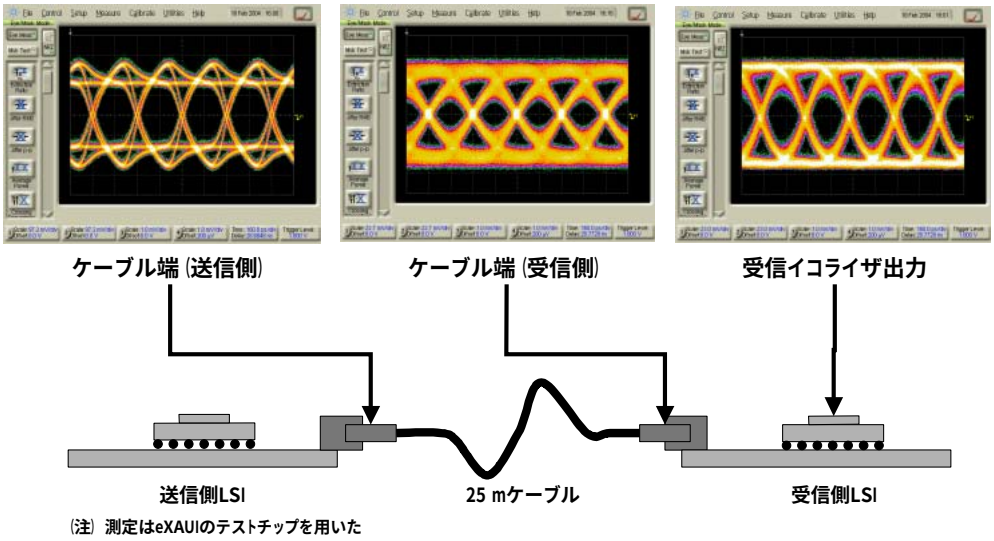

図-6　eXAUIの評価結果
Fig.6-eXAUI evaluation.

デジタル回路からアナログ回路への干渉の問題がある。eXAUI内のアナログ回路に対してデジタル回路から発生するノイズの流れ込みを防ぐため，アナログ回路とデジタル回路の電源とグラウンドは，完全に分離させ，また，十分なオンチップキャパシタを搭載した。さらには，eXAUIなどの回路側でも，CMOS構造として3層構造のウェル（トリプルウェル）の使用など，ノイズの影響を受けにくい工夫を施している。

LSIの性能評価

MB87Q3070の性能評価として，まず，スイッチのスループット評価の結果を示す。本LSIは，最大負荷をかけた状態で，とくに条件が厳しいイーサネットフレームのサイズが小さい場合でも，各ポートがほぼ100％の10ギガビットの帯域を実現している（**図**-5）。

なお，フルメッシュでは，各ポートから他ポートすべてにイーサネットフレームを順に送り，ポートペアリングでは，ポートのペア間でイーサネットフレームを送受信する。

つぎに，eXAUIの評価を示す。送信側信号は，送信イコライザ回路を用いて10GBASE-CX4に準拠させておき，その信号が，規格を上回る25 mケーブル伝送できるかを測定した。その結果を**図**-6に示す。図では，左から送信側のケーブル端でのEyeパターン，受信側のケーブル端でのEyeパターン，受信イコライザ回路の出力Eyeパターンを示す。この測定の結果から，受信イコライザがEyeを開けていることが確認できた。

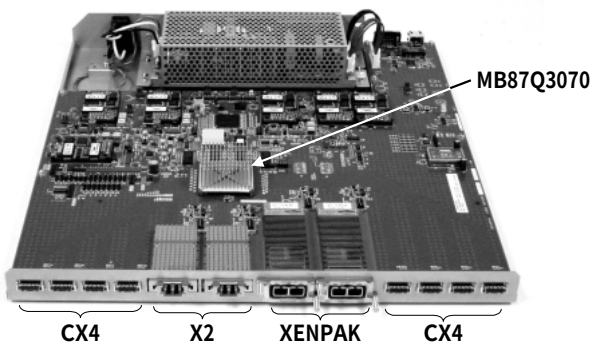

図-7　リファレンスボード
Fig.7-Reference board.

リファレンスボード

お客様によるMB87Q3070の評価，またはMB87Q3070を使ったシステム開発の短TAT化を実現するため，評価用としてリファレンスボードを開発した（**図**-7）。本ボードには，10GbEのインタフェースとして，XENPAKとX2という光モジュール，および10GBASE-CX4を用意している。さらに，本ボードにはスイッチソフトウェアが搭載されており，10GbEインタフェースを持つ機器との相互接続試験を容易に行うことが可能である。

む　す　び

本稿では，ITシステムの信頼性・高性能を実現させるインタコネクトとして，1チップ化に成功した10GbEスイッチLSIについて紹介した。

今回開発したLSIを使用することにより，従来は10GbEの適用が困難であったITシステムを構成するサーバやストレージ装置間のスイッチ，また，ブレードサーバ内のスイッチを小型化かつ高性能化，低コスト化することが可能になった。本LSIを用いた10GbEスイッチ製品も既に開発されており，データセンタ，高性能コンピュータの分野などで使われている。今後，さらなる機能と性能の強化を図り，省電力化したスイッチチップを開発していく予定である。

なお，本LSIの開発は，新エネルギー・産業技術総合開発機構（NEDO）の一部助成（テーマ名：高信頼・低消費電力サーバの研究開発）を受けて行った。

参 考 文 献

(1) T. Shimizu et al.：A Single Chip Shared Memory Switch with Twelve 10Gb Ethernet Ports. Hot Chips 15, August 2003.

(2) W. Gai et al.：A 4-Channel 3.125Gb/s/ch CMOS Transceiver with 30dB Equalization. 2004 Symposium on VLSI Circuits. June 2004, p.138-141.