

大規模ブレードサーバ向け10ギガビットイーサネットスイッチブレード

10 Gigabit Ethernet Switch Blade for Large-Scale Blade Servers

あらまし

近年、ブレードサーバの導入が急速に進んでいる。ブレードサーバのネットワーク機能を提供する重要な構成要素であるスイッチブレードは、以下のような要件が求められる。まず、FCoE (Fibre Channel over Ethernet) といったストレージ用ネットワークのイーサネットへの統合技術の普及に対しても対応できる高い通信性能、つぎに実装密度の高いブレードサーバに対応する小型化、省電力化、そして、サーバとネットワークが一体化したブレードサーバの構造に最適化されたスイッチ機能と運用設定を簡単化するスイッチソフトウェアの実現である。著者らは、これらの要件を満たす大規模ブレードサーバ向け10ギガビットイーサネットスイッチブレードを開発した。

本稿では、ブレードサーバに最適化した10ギガビットイーサネットスイッチLSI “MB86C69” の特長、スイッチLSIに内蔵された高密度・省電力の高速送受信回路、およびスイッチブレードの設定簡単化を実現したスイッチソフトウェアを中心に解説する。

Abstract

Blade server application fields have recently been expanding rapidly. The switch blade module—a key network component of the blade server—must provide high performance to accommodate growing data communication demands for network consolidation such as Fibre Channel over Ethernet (FCoE), have a compact low-power-consumption design, and have switch control software optimized for blade server systems. To meet these requirements, we have developed a 10 Gigabit Ethernet (10GbE) switch blade suitable for large-scale blade servers. This paper focuses on the 10GbE switch LSI MB86C69, whose features have been optimized for blade servers, the high-speed transceiver circuit integrated into this LSI, and the switch control software for easy network configuration of switch blades.



小柳洋一 (こやなぎ よういち)
ITシステム研究所サーバテクノロジ
研究部 所属
現在、10ギガビットイーサネット
ブレードのハードウェアの研究、開
発に従事。



新家正総 (にいのみ ただふさ)
ITシステム研究所サーバテクノロジ
研究部 所属
現在、10ギガビットイーサネット
ブレードのソフトウェアの研究、開
発に従事。



梅澤 靖 (うめざわ やすし)
ITシステム研究所サーバテクノロジ
研究部 所属
現在、10ギガビットイーサネット
チップの研究、開発に従事。

ま え が き

ブレードサーバは、高性能かつ高密度な特長からデータセンタや企業サーバの集約など、今後ますます導入の拡大が見込まれる。そこで著者らは、大規模ブレードサーバの高性能化、高密度化、運用簡単化を実現する小型、省電力の10ギガビットイーサネットスイッチブレードを開発した。

ブレードサーバのネットワーク機能を提供する重要な構成要素であるスイッチブレードは、以下のような要件が求められる。

(1) 高性能

サーバ間、あるいはサーバと外部ストレージやネットワーク間の通信ボトルネックを解消する高い性能が必要である。とくに、仮想化によるサーバの運用効率の向上や、FCoE (Fibre Channel over Ethernet) といったストレージ用ネットワークをイーサネットに統合する技術の普及も期待され、従来に比べて飛躍的に高い通信性能を持つスイッチブレードの実現が強く望まれている。

(2) 高密度、省電力

高密度に実装されるブレードサーバに適應する小型化と省電力化を性能と両立して実現する必要がある。

(3) 簡単運用

サーバとネットワークが一体化した構造であるブレードサーバに最適化されたスイッチ機能、運用機能を提供するスイッチソフトウェアが必要になる。

これらの要件を満たすために、著者らは、高性能、低レイテンシの10ギガビットイーサネットスイッチLSI “MB86C69”を開発し、スイッチブレード

に搭載した。本LSIは、ブレードサーバのスイッチに要求される性能と機能を実装し、さらに10 Gbpsの高速信号伝送を実現する高速送受信回路をLSIに内蔵することにより、高密度実装、省電力の両立を実現した。また、スイッチブレードに搭載した制御用プロセッサのソフトウェア (ファームウェア) に、ブレードサーバ用スイッチの設定運用を簡単化する機能を開発し搭載した。

本稿では、まず富士通の大規模ブレードサーバBX900と本スイッチブレードの概要について述べ、続いて開発したスイッチLSIの特長、LSIに搭載された高速送受信回路、およびソフトウェアについて解説する。

ブレードサーバとスイッチブレード

BX900ブレードサーバの外観を図-1に示す。CPU、メモリ、チップセットを搭載したサーバブレードがシャーシ前面から最大18枚実装される。また、サーバブレード間およびサーバブレードと外部装置との間のデータ通信のためのスイッチブレードがシャーシ背面から最大8枚実装される。さらに、ブレードサーバ装置全体を管理する機能を持つマネジメントブレードがシャーシ背面から2枚実装される。

ブレードサーバの内部構成を図-2に示す。サーバブレードのマザーボードまたは拡張基板 (メザニンカード) に実装された10ギガビットイーサネットアダプタと、スイッチブレードに搭載されたMB86C69が、10 Gbpsの高速信号で相互に接続される。スイッチブレードを複数枚実装することにより、サーバブレード間およびサーバ外部との通信性能を拡張していくことが可能であるとともに、相互

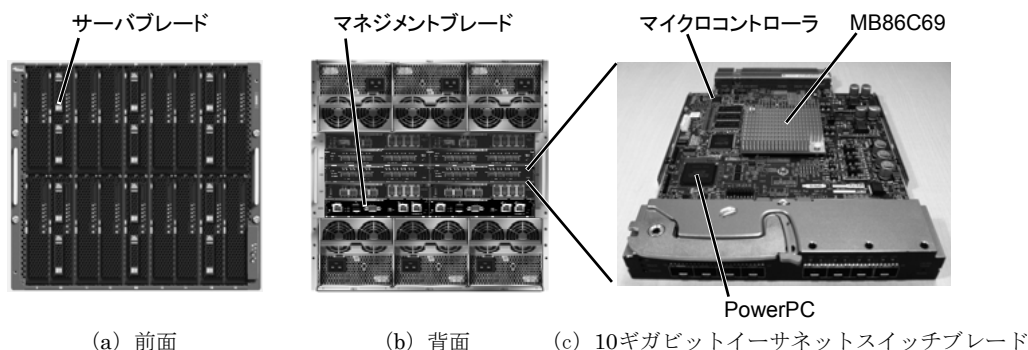


図-1 BX900ブレードサーバと10ギガビットイーサネットスイッチブレード
Fig.1-BX900 blade server and 10 Gigabit Ethernet switch blade.

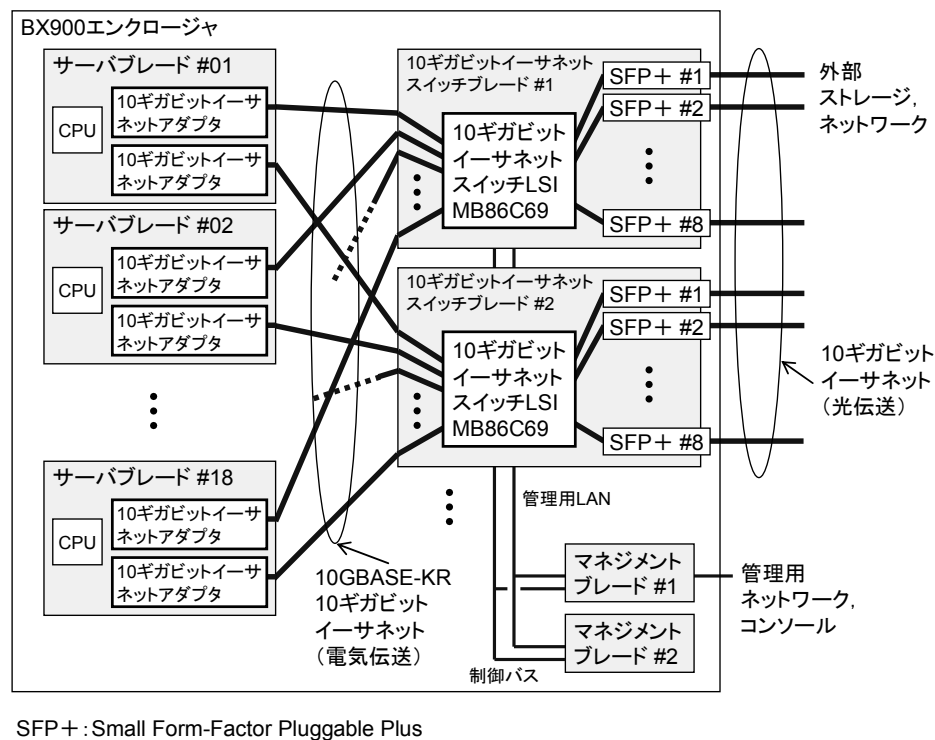


図-2 BX900ブレードサーバの内部構成
Fig.2-Blockdiagram of BX900 blade server.

に独立した通信路となるので冗長系として高信頼なシステムを構築することも可能となる。また、スイッチブレードはマネジメントブレードと管理用LANおよび制御バスで相互に接続されており、マネジメントブレードが集中管理するエラー監視、温度監視といったRAS機能に対応している。

本スイッチブレードには、スイッチの機能を提供するMB86C69に加え、スイッチソフトウェアが動作するPowerPCコアの組込みプロセッサ、および制御用マイクロコントローラが搭載されている。マイクロコントローラは、マネジメントブレードと連携したRAS機能とともに、消費電力の通知や10ギガビットイーサネットアダプタカードが接続されていないサーバブレードに対応したポートの高速送受信回路と制御論理回路をOFFにして消費電力を削減する機能なども提供する。

開発したスイッチブレードの諸元を表-1に示す。本スイッチブレードは動作モードとして通常のL2スイッチモードに加え、設定運用の簡単化を実現するIBP (Intelligent Blade Panel) のモードを備えている。IBPについてはソフトウェアの章で詳しく述べる。

表-1 10ギガビットイーサネットスイッチブレードの諸元

項 目	仕 様
ダウンリンク	18ポート (10GBASE-KR)
アップリンク	8ポート (SFP+)
スイッチスループット	520 Gbps
スイッチ遅延	300 ns
消費電力	30 W (最大40 W)
物理サイズ	幅 : 193 mm, 奥行 : 268 mm, 高さ : 28 mm
モード	レイヤ2 および IBP
VLAN機能	ポート, タグ, プロトコル
冗長機能	STP (STP, MSTP, RSTP), リンクアグリゲーション, バックアップポート, リンクダウンリレー
QoS	IEEE 802.1p (COS) / TOS (IP Precedence) / DSCP / ACL (IPv4, IPv6), Strict/DRR
ネットワーク認証	MAC, IEEE 802.1X, Web, RADIUS, TACACS+
アクセス制御	レイヤ2~4
マルチキャスト	IGMPスヌープ機能
モニタリング	ポートミラーリング
ネットワーク管理	CLI, Web UI, SNMP/RMON, LLDP, ロギング

10ギガビットイーサネットスイッチLSI “MB86C69”

MB86C69はMB86C68⁽¹⁾をブレードサーバに最適化したチップで、90 nmのCMOSテクノロジーを用い、約2200万ゲート相当の論理回路と、2.9 Mバイトの内蔵データバッファを262 mm²の大きさのチップ上に集積し、35 mm×35 mmのFCBGA (Flip Chip Ball Grid Array) パッケージに封入した。MB86C69の諸元を表-2に示す。また、以下に主要な特長について説明する。

(1) 10GbEポート数

BX900ブレードサーバのスイッチブレード要件に合わせ、26の10GbEポート（アップリンク18ポート、ダウンリンク8ポート）をシングルチップに実装した。

(2) インタフェース

XAUI、10GBASE-CX4に加え、10GBASE-KR準拠としてBX900のバックプレーン伝送を可能とした。また、今後普及が見込まれるSFP+ (Small Form-Factor Pluggable Plus) を直接接続することを可能とし、コスト、遅延時間と消費電力の削減を可能にした。さらに、1000BASE-KXにも準拠し、顧客要望の高かったGbE/10GbEの両スピードによる伝送を全26ポートで実現した。

表-2 MB86C69の諸元

項 目	MB86C69
10GbEポート数	26
インタフェース	XAUI 10GBASE-CX4 10GBASE-KR 1000BASE-KX
スイッチング帯域	520 Gbps以上
遅延時間	300 ns
MACアドレス数	16 Kエントリ
内蔵データバッファ容量	2.9 Mバイト
内部優先度	8レベル
最大フレームサイズ	16 Kバイト
フロー制御	IEEE 802.3a Priority PAUSE BCN
制御インタフェース	GMII/MII×2
セキュリティ	ACL (L2～L4) DoSアタック検出
パッケージ	FCBGA1156 (35 mm×35 mm)
消費電力	23.1 W (標準)
使用テクノロジー	90 nm CMOS

(3) 性能

スイッチブレードは、近年のサーバの仮想化やマルチコア化の進展に伴って、多数のユーザ業務からのI/O要求を処理する必要がある。MB86C69のスイッチング帯域は、シングルチップとしては業界最大である520 Gbps以上、また、入出力間の遅延時間は、業界最小の300 nsを実現した。これらの性能によって、通信性能のボトルネックを解消した。

(4) データセンタ向けイーサネット

スイッチブレードへの機能要件として、LAN/SANの統合によるコスト削減、および運用簡単化を実現できるFCoE対応への要求が高まっている。MB86C69は、FCoEに必須の要素技術であるCEE (Converged Enhanced Ethernet) を実現するために、データセンタ向けの輻輳制御機能として、出力ポート側の輻輳情報を送信側端末に伝えるためのメッセージ機構 (BCN : Backward Congestion Notification) と優先度ごとのフロー制御機構 (Priority PAUSE) を実装した。

(5) セキュリティ

スイッチブレードに要求される高度なセキュリティを実現するために、入力側ACL (Access Control List) を実装した。ACLは、入力フレームからレイヤ2～4の標準フィールド、およびユーザ定義フィールドを抽出してフィルタを行う。マッチした入力フレームは、フィルタタイプによって、許可、廃棄、CPUへのコピー、CPUへのフォワードのいずれかのアクションが適用される。

10 Gbpsバックプレーン対応高速送受信回路

サーバブレードとスイッチブレードは、図-3に示すように、バックプレーン（またはミッドプレーン）と呼ばれるプリント配線板でコネクタを通して電氣的に接続されている。従来、この配線を通る信号は1～5 Gbpsの通信速度であり、サーバブレードの性能向上につれて増大する通信容量に対応するには、配線本数を増やして対応する必要があった。その結果、バックプレーンの配線密度の限界により、サーバ全体の高密度化、高性能化が困難となっていた。そのため、信号線あたりの伝送速度を10 Gbpsに向上させる技術が必須となる。

しかし、プリント配線板で10 Gbpsもの高速信号を伝送すると、信号の減衰による符号間干渉

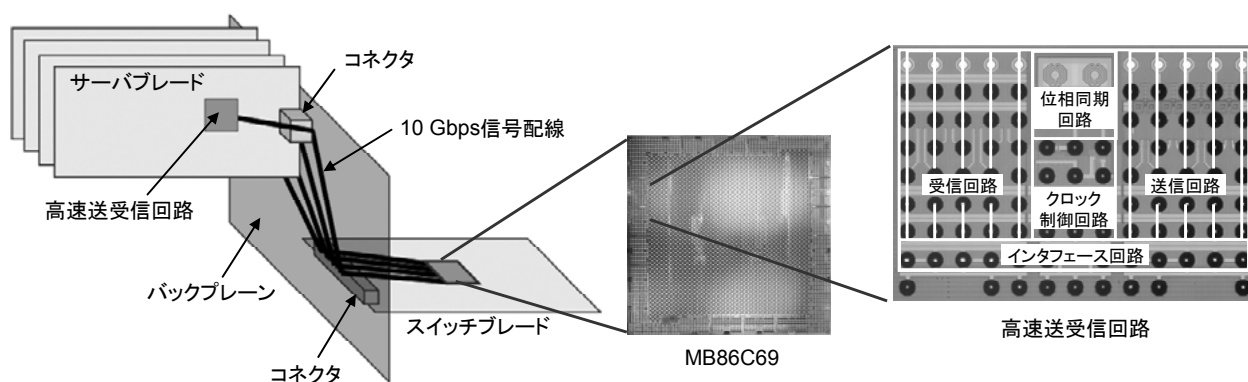


図-3 バックプレーンと高速送受信回路
Fig.3-Backplane and high-speed transceiver circuit.

(ISI) と呼ばれる波形の歪^{ひずみ}の問題が顕著となり、エラーなくデータを伝送することが難しくなる。この問題を解決する手段として、イコライザと呼ばれる信号の減衰を補償する回路技術がある。その機能を持った専用の信号送受信ICが存在するが、多数のサーバブレードとの接続を持つスイッチブレードのそれぞれの配線チャンネルにイコライザICを搭載することは物理的に困難で、かつ消費電力も大きくなり、高性能、高密度でかつ省電力のスイッチブレードを実現するには現実的ではなかった。

そこで著者らは、ブレードサーバのバックプレーンにおいて10 Gbpsの伝送速度を実現するイコライザ機能を持った、小型、省電力の高速送受信回路を開発し²⁾ スイッチLSIに組み込むことに成功した。これにより、配線本数を増やすことなくサーバブレードと10ギガビットイーサネットの通信を実現することができ、かつスイッチLSI 1個で外付けのイコライザICを用いずにスイッチブレードを実現することができるため、小型化、省電力化も達成することが可能となった。

開発した高速送受信回路は、以下のような特長を持つ。

- (1) 信号の減衰により歪を受けた波形を補正するイコライザ回路として、リニアイコライザとデジジョンフィードバックイコライザ (DFE : Decision Feedback Equalizer) と呼ばれる2種類のイコライザ回路を組み合わせることで、従来困難であったバックプレーンでの10 Gbps信号の補正を少ない消費電力で実現した。
- (2) イコライザの補正強度を動的に制御するため

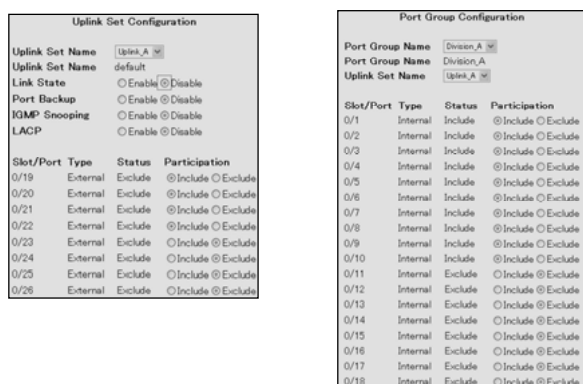
の新しい制御方式を開発した。本方式は計算量の少ないデジタル回路で実現できるため回路面積が小さく、ポート数の多い本スイッチLSIに集積することが可能となった。

スイッチブレードのソフトウェア

ブレードサーバ内のスイッチブレードは、ネットワーク管理者が管理する状況を想定するよりもむしろサーバ管理者が管理する状況に備える必要がある。このためスイッチブレードのソフトウェア機能にはサーバ管理者にとって使いやすいという要件が求められる。

サーバ管理者にとってやりたいことは、ブレードサーバをネットワークにつなぐという要件が主であって、ブレードサーバの外部の機器間の通信をブレードサーバ内のスイッチで行いたいというような要件は付随的と考えられる。またサーバ管理者はVLANやSTP (Spanning Tree Protocol) などのネットワークの知識が必要な機器の設定には馴染みがない場合があり、そうした管理者にとってネットワークに大きな影響を及ぼし兼ねないL2スイッチの設定は困難と考えられる。このように考えると、スイッチブレードに求められる要件としてL2スイッチよりも簡単に利用できるような、サーバブレードの集線装置的な設定ができる機能を兼ね備えることが挙げられる。

そこで通常のL2スイッチ機能のソフトウェアに加え、集線装置的な簡単設定が可能なIBPのソフトウェアを開発した。お客様はニーズに合わせて、L2スイッチとIBPの動作モードを選択し利用する



(a) アップリンクの設定

(b) ポートグループの設定

図-4 IBPの設定画面

Fig.4-Configuration screen shot of IBP.

ことができる。

IBPの設定画面の一部を図-4に示す。IBPの設定画面では、(a)のアップリンクの設定画面で外部へのネットワークポートに名前（Uplink Set Name）を付け、その名前のアップリンクに属させるポートをIncludeボタンで指定する。同一アップリンク名の複数のポートはリンクアグリゲーションで束ねられ、論理的に一つのポートになる。その上で、(b)ポートグループの設定画面でサーバブレードへのネットワークポート（ダウンリンク）をどのアップリンクに接続するかを指定する。図-4(a)では、アップリンク名としてUplink_Aが指定され、Port19からPort22の四つのポートが論理的に一つのポートに設定されている。同(b)ではポートグループDivision_Aが設定され、ダウンリンクPort1からPort10がUplink_Aに接続される。このようにして、GUI（Graphical User Interface）で簡潔に接続を指定できる。

IBPではポートグループに対して論理的な複数のアップリンクを設定することはできないようになっており、またアップリンク間での通信はできないようにしている。したがってアップリンクを外部のネットワークと接続してループが生ずることはない。管理者はSTPのようなループ防止の機能を設定する必要がないため、IBPではSTP機能自体を排除している。ポートグループは簡単な操作で作成することができるため、VLANやSTPといったL2スイッチ

的設定をせずに業務や部門別にアップリンクへのサーバ割当てが簡単に、安全にできる。

ポートグループの機能はMB86C69のパーティショニング機能（拡張VLAN）を使って実現した。拡張VLANはVLANよりも下位のレベルでスイッチを分割する。したがって必要があればポートグループ内でVLANを使用することもできる。

さらにIBPは、富士通のサーバ管理システムServerViewの仮想化ソフトウェアVIOM（Virtual I/O Manager）と連携して利用でき、VIOMとの連携によってブレードサーバシステムの導入、運用、保守を容易化できる。

IBPを用いるとこのように簡単な設定が可能となる一方、スイッチブレードを通常のL2スイッチとして利用し、STPを含めた機能を利用したい管理者も存在する。その場合には、L2スイッチモードに切り替えて動作させることも可能となっている。本スイッチブレードは表-1で示したような一般的なL2スイッチ機能を備えており、ネットワーク管理に慣れている管理者にとって必要な機能を備えている。

む す び

本稿では、大規模ブレードサーバに適した、高性能、小型、省電力の10ギガビットイーサネットスイッチブレードについて解説した。

今後は、スイッチLSIの機能として既に実装されているFCoEをスイッチブレードでサポートするためのスイッチソフトウェアの強化を進め、データセンタ向けイーサネットへの最適化やストレージ通信との統合の実現に向けて研究開発を推進していく。

参 考 文 献

- (1) 清水 剛ほか：20ポート10 GイーサネットスイッチLSIとその応用．Fujitsu, Vol.58, No.3, p.246-250 (2007)．
- (2) Y. Hidaka et al. : A 4-Channel 10.3 Gb/s Backplane Transceiver Macro with 35 dB Equalizer and Sign-Based Zero-Forcing Adaptive Control. ISSCC Dig. Tech. Papers, paper 10.5, (February 2009)．