特集

基幹IAサーバ“PRIMEQUEST”の高性能・高信頼を実現するチップセット

Fujitsu's Chipset Development for High-Performance and High-Reliability Mission-Critical IA-Servers

あらまし

富士通はインテル社と密に協業し，インテル社の高性能・高信頼性を誇る64ビットマイクロプロセッサであるItanium2シリーズの最新プロセッサを搭載し富士通のメインフレームで培った高信頼性技術を導入した基幹IAサーバ“PRIMEQUEST”を開発した。このPRIMEQUESTには1CPUから32CPU（第2世代は64CPU）へのスケーラビリティとメインフレームと同等の高信頼性を実現している。このために富士通の最新ASICテクノロジであるCS101を採用し，六つのチップセットを新規に開発した。

このチップセット開発では，従来の高信頼性技術のほかに，チップセット間の高速リンクやアドレス，システム全体の2重化（チップセット内部も含む）や新しい汎用高速IOインタフェースであるPCI-Expressといった新規の技術にもチャレンジし実装に成功している。

本稿では，今回開発したチップセットの概要について述べる。

Abstract

Fujitsu has developed a new mission-critical IA-server named PRIMEQUEST in close collaboration with Intel using Intel's latest high-performance and highly reliable CPU (Itanium2, the 64-bit microprocessor). PRIMEQUEST offers linear scalability from a single CPU to 32 CPUs (with 64 CPUs in the second generation) and represents a highly reliable technology equivalent to that of a mainframe. In conjunction with this development, we also developed six new chipsets using cutting-edge CS101 ASIC technology and examined and successfully employed new technologies such as high-speed interconnection between chipsets, address and system mirroring (including inside the chipsets), and a new standard high-speed IO interface (PCI-Express). This paper describes an overview of the newly developed chipsets.

柴田泰秀（しばた　やすひで）

IAサーバ開発統括部 所属
現在，IAサーバ用チップセットの開発に従事。

まえがき

富士通はインテル社と密に協業し，インテル社の高性能・高信頼性を誇るItanium2シリーズの最新プロセッサを搭載し，富士通のメインフレームで培った高信頼性技術を融合した新しい基幹IAサーバ“PRIMEQUEST”を開発した。最大32CPU（第2世代は64CPU）への大規模スケーラビリティと，基幹業務でのお客様の運用に対応できる高信頼性を実現するために富士通の最新ASICテクノロジであるCS101を採用して，新規に6種のASIC（チップセット）を開発した。

本稿では，チップセットの概要を高信頼技術を中心に紹介する。

PRIMEQUESTとチップセットの構成

● PRIMEQUESTの構成

PRIMEQUESTは大規模基幹業務に対するお客様の期待に応える機能を提供するために，最新Itanium2プロセッサを採用した。高性能で少数CPUから最大32（第2世代は64CPU）までのスケーラビリティと，どのCPUからもシステムの資源に均一にアクセスできるSMP（Symmetric Multiple Processor：対称型マルチプロセッサ）を基礎アーキテクチャとしている。PRIMEQUESTのシステムブロック図を**図**-1に示す。図中上部に最大8枚のSB（System Board）が配置され，各SBにはCPUを最大4個とDIMM（Dual Inline Memory Module）を最大で32枚装着できる。図中下部にLANやハードディスクなどの周辺装置を接続する最大8枚のIOユニットが搭載され，各SBとIOユニットは図中中央に配置されたクロスバで相互結合される。

実際の装置と構成ユニットを**図**-2に示す。前面から見た外観とその構成ユニットが同図（a）である。装置上部には電源ユニットと冷却ファンユニットが装着され，中央部にSBが縦に最大8枚実装される。下部にはIOユニットが前面から四つ，同図（b）の背面から四つ装着される。背面から見て中央部には横方向にアドレスクロスバ（XAI）とデータクロスバ（XDI）が装着される。

● チップセットの構成

CPUやDIMM，市販のIOコントローラを結合しサーバとしての制御を行うものがチップセットであり，SB上にCPUを制御するASIC（略称FLN）とメモリを制御するASIC（略称LDX）の2種が搭載されている｛図-2（a)｝,（**図**-3)。またIOボード（以下，IOB）上にはPCI-Expressと呼ばれる高速IO制

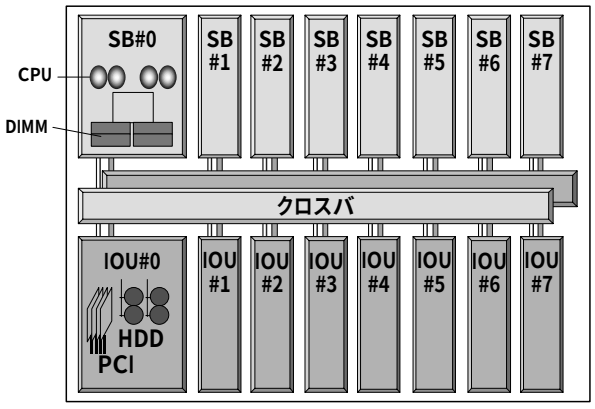
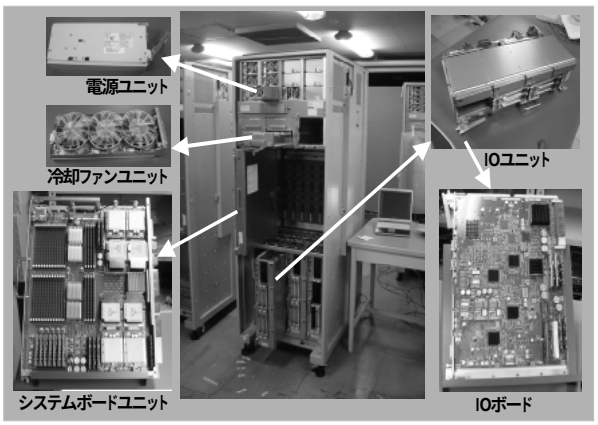
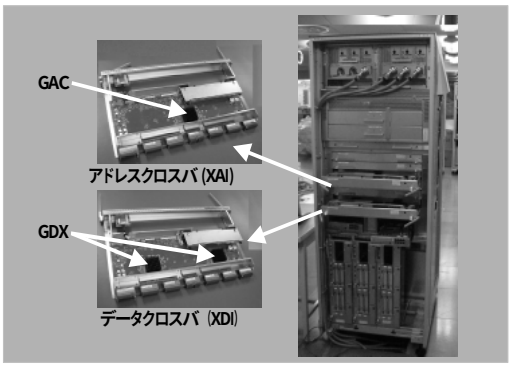


図-1　システムブロック図
Fig.1-System block diagram.

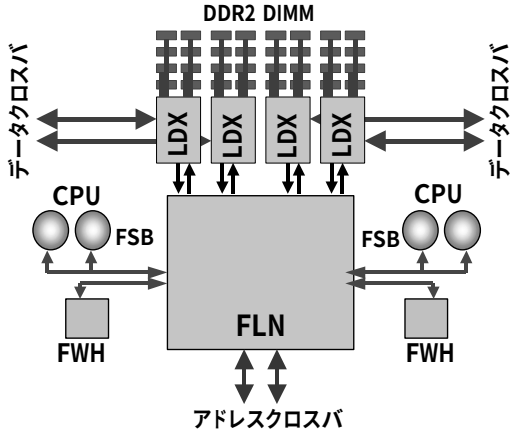


（a）前面

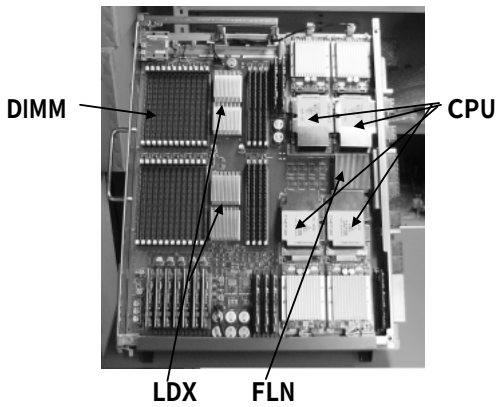


（b）背面

図-2　装置と構成ユニット
Fig.2-System and components.



（a）ブロック図



（b）搭載写真

図-3　システムボード
Fig.3-System board.

御インタフェースを制御するASIC（略称FLIと，略称FLP）の2種が搭載されている。SBとIOユニットを相互結合するクロスバには，アドレス用のXAI上にASIC（略称GAC）が，XDI上にASIC（略称GDX）が搭載されており，計6種のチップセットで装置が構成されている。

SBとFLN/LDXチップセット

SBの主要構成ユニットのブロック図を図-3（a）に，搭載写真を同図（b）に示す。SBはCPU・DIMM・FLN（CPUとほかのSB，IOユニットとの相互結合処理を行う）・LDX（DIMMの制御とデータクロスバとのデータ連携を行う）から構成されている。

CPUはFLNの左右に2個ずつ計4個搭載できる。このCPUとのインタフェースはFSB（Front Side Bus）と呼ばれるもので，インテル社との密な技術協業によりプロトコルおよび電気的仕様を詳細に議論し，チップセットのFSB専用の高速トランジスタマクロを極めて短期間で開発した。FLNにはCPUを起動するために基本ファームウェアを格納したフラッシュROMを8個直結しており，このうち4個は故障時のバックアップとして使用する。

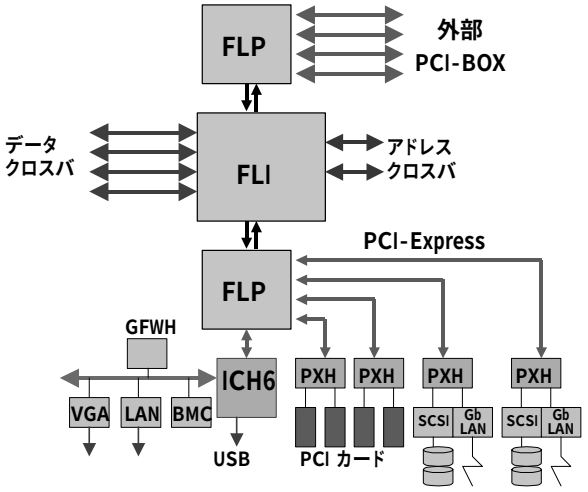
CPUからのメモリまたはIO装置へのRead/Write要求は，いったんFLN内部に蓄えられ，内部で優先順位を取って，アドレスクロスバ上のGACに送信される。アドレスクロスバはこの要求を調停した上でFLN/FLIに送信し，該当するメモリまたはIO装置を持っているFLN/FLIがデータを応答する。

同SB上には4個のLDXチップが配置され，メモリまたはデータクロスバとFLNとのデータの交換を行う。128バイトのデータは四つに分割され，4個のLDXそれぞれで処理される。

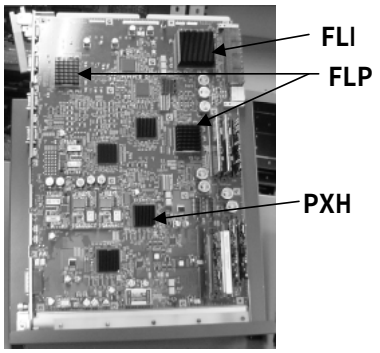
DIMMとのインタフェースはDDR2でこのための高速専用トランジスタマクロも専用に新規設計した。四つのLDXにより一つのSBでは最大32枚のDIMMを実装できる。装置全体で32×8＝256枚搭載でき，2 GバイトDIMM装着時は512 Gバイト，4 GバイトDIMM装着時は1 Tバイトのメモリを最大で搭載できる。

IOBとFLI/FLPチップセット

IOユニットは内蔵IOであるハードディスクスロットとPCIカードスロット，および制御ボードであるIOBにより構成される。IOBのブロック図を**図**-4（a）に，その搭載写真を同図（b）に示す。FLIチップセットシステムはアドレスクロスバとデータクロスバと結合され，CPUからIO装置へのアクセスとIO装置からメモリへの転送（DMA）を制御する。またIO装置へのバス変換チップであるインテル社製のICH6とPXHチップに対しPCI-Expressで接続する。PCI-Expressは最新の高速I/OバスでPCIバスに替わる2.5 GHz転送能力を持つ高速バスである。FLIはシステムバスからPCI-Expressバスにバス変換を行うが，極めて高速のPCI-Expressの電気的インタフェースとしてFLPと呼ばれるチップセットを使用している。ICH6はインテル社から供給される汎用IO制御チップで管理LAN，タイマなどの機能を提供している。近い将来，IO制御チップは直接PCI-Expressで接続されるが，現在はまだPCIバスのIO制御チップがほとん



（a）ブロック図



（b）搭載写真

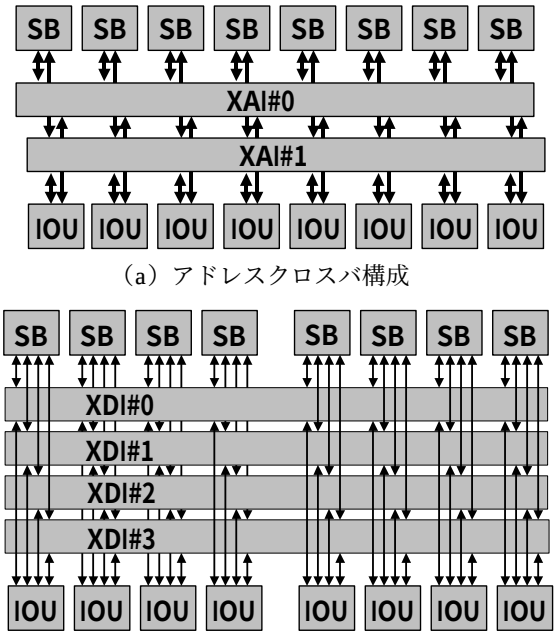
図-4　IOボード
Fig.4-IO board.

どであり，オプションカードもPCI規格のものが大半を占めるため，PCI-ExpressからPCIバスへ変換するPXHチップを搭載している。このため自社およびISVベンダから豊富なIO制御コントローラやPCIカードをそのまま装置に組み込むことが可能となっている。

また，FLPからはケーブルにより装置外部へPCI-Expressを介して出力しているので，装置外部のPCI拡張ボックスにより更に多くのPCIカードをサポートすることが可能である。さらに，近い将来PCI-Expressに直接接続できるIO制御チップやPCI-Expressカードにもスムーズに移行・対応できる。

XAI，XDIとGAC，GDXチップセット

図-1に示したように，SBとIOユニットはクロスバで相互接続される。高性能と高スループットを達成するためにアドレスクロスバとデータクロスバは



（a）アドレスクロスバ構成

（b）データクロスバ構成

図-5　アドレスクロスバとデータクロスバの接続図
Fig.5-Address crossbar and data crossbar network diagram.

分離して構成されている。アドレスクロスバ構成を**図**-5（a）に，データクロスバ構成を同図（b）に示す。

アドレスクロスバは性能を向上させるために二つのGACで構成され，図-2（b）に示すように，XAI上に各1個のGACが搭載され，システムでは2枚のXAIが実装される。データクロスバは図-5に示すように合計8個のデータクロスバチップ（GDX）で構成されている。アドレスに比べ一度に転送される情報（データ）が多いためこのような構造になっている。クロスバチップは電話交換機のクロスバスイッチと概念的に同じで，送信元と受信先のペアを接続するため，システムでは同時複数の情報伝達が実現できる。

チップセットと新技術

PRIMEQUESTにおけるチップセット開発は，インテル社製Itanium2プロセッサを使用した大規模・高性能サーバとしての取組みと同時に，基幹業務向けにメインフレームで培ってきた高信頼性技術も融合した。またPCI-Expressなどの最新技術も積極的に取り入れている。本章ではこれらの最新技術について述べる。

● チップセット間高速リンク

これまで紹介してきたチップセット間は新規に開発したMTL（Morimuta-Transceiver-Logic）と呼ばれる高速リンクで接続され，設計値で1.3 GHz（実験室環境では1.6 GHz）をシングルエンド（不平衡）転送で実現した。ハードウェアでトレーニングフェーズを実使用の前に実施することで，同一信号グループの信号間ディレイ差異やプリント板やASICの製造ばらつき，環境ばらつきを自動的に調整し，コストと時間がかかる出荷・設置時の人手による調整を一切廃止した。またこの高速リンクはクロックボードから各チップセットに配信される基準クロックをもとに，チップセット同士が互いにどれくらいの距離に配置されているかを解析しフィードバックすることで，第1世代で32CPU，第2世代で64CPUという非常に大規模な構成においてもチップセット同士の同時刻性を保証し，完全対称型のSMPを実現している。

● チップセット内高信頼性

PRIMEQUESTは基幹業務向けに設計されており，チップセットもメインフレーム技術を導入している。チップセット内に実装されているメモリやバッファ，キューなど，データや制御情報を蓄えておく機構にはすべてECC（Error Correcting Code）が付加されており，1ビットエラーはハードウェアで自動訂正し継続運転が可能である。2ビットエラーについては確実に問題を検出し，安全にシステムを停止することができる。制御系の重要度に応じて，3重化，2重化，パリティの保護も併せて実装している。

万一障害が発生したときも，障害レベルに応じて三つのレベルでのソフトウェア・ファームウェア処理を行えるようにエラー通知機構が実装されており，さらに深刻な問題が発生した場合は，業務で動作しているOSやアプリケーションと全く独立したシステムマネジメントシステムから専用のインタフェースで，障害レベルに応じて三つの独立したエラー報告を行い，故障箇所を分析することができ，システムの停止時間を最小限に抑えることが可能となっている。

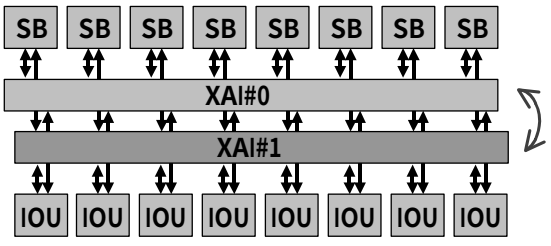


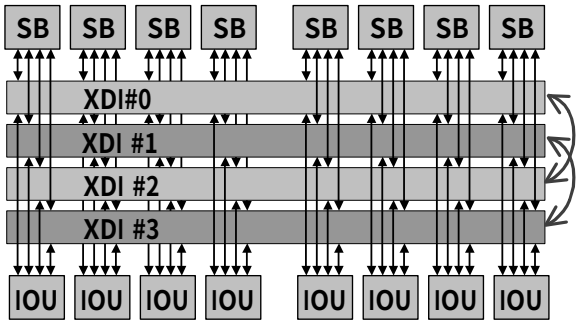
図-6　アドレスクロスバ2重化モード
Fig.6-Address mirroring mode.

● システム2重化技術

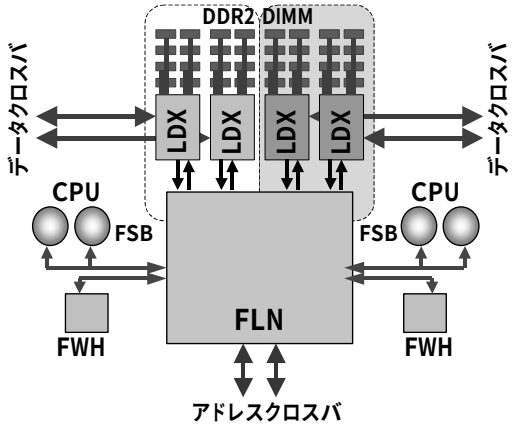
図-6は通常の運用形態に対してアドレスを2重化した高信頼性動作モードである。

通常モードでは，二つのアドレスクロスバ（GAC#0と#1）は独立に動作しているが，本モードでは同時に同じアドレス制御を行っている。FLNとFLIは二つのGACに対し同時に同じアドレスリクエストを発行し，二つのGACから応答されるアドレス応答をチェックし，比較する。万一どちらかのアドレス応答に2ビット以上のECCエラーがあった場合は，エラーのない方を採用し，動作を継続する。

図-7はさらに進んで，アドレスだけでなくデータも2重化するもので，一般に呼ばれるメモリ2重化



（a）データクロスバの2重化



（b）チップセットの2重化

図-7　システム2重化モード
Fig.7-System mirroring mode.

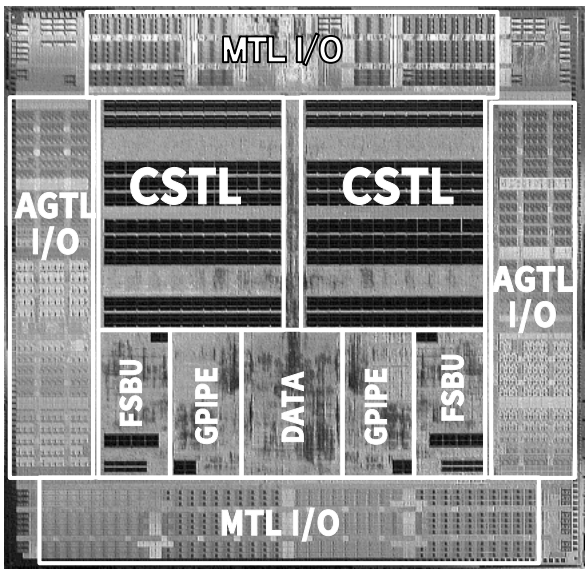


図-8　FLNのトランジスタダイ顕微鏡写真
Fig.8-FLN DIE photo.

だけでなく，データクロスバも2重化され，さらにはチップセット内部も2重化されたものを示している。つまり装置全体が2重化されるのでシステムミラー機構と呼ばれる。

両モードとも単に2重化しているだけでなく，チップセットの内部も2重化しており，チップセット間，およびチップセット内部の2重化比較，エラー検出回路を要所ごとに実装しており，エラー検出時に正常系から情報を伝播させ，1箇所のエラーによりシステム全体が1重運転に縮退しないように制御している。

トランジスタ技術

本チップセットの開発では，前述の新規の高速インタフェースを多く採用した。このため，高速IOマクロをトランジスタレベルで新規設計を行い，短期での開発とテストチップでの動作確認を完了し，実際のチップセットに搭載した後は問題が全く発生していない高信頼設計を実現した。

また六つのチップセットを同時に開発し，高速IO部はGHzオーダで動作させるため，ベースとなるトランジスタテクノロジは富士通の最先端90 nm CMOS：CS101を採用，当チップセット開発に向けては，専用設計ルール（クロック配線，電源グランド配置配線など）を新規設計し，またスタンダードセルと呼ばれる基本ゲートも新規に開発した。

図-8は六つのチップセットの中で最も複雑で規模の大きいFLNのトランジスタダイの顕微鏡写真である。ダイの大きさは16.5×17.0 mmであり，バンプと呼ばれる電源とグランド，および信号を取り出す端子は9,024個に上る。写真上で黒い四角はRAMであり，大小合わせて約200個搭載している。

配線層は銅の9層である。論理トランジスタ数はRAMを含まないで約500万ゲート（2NAND換算）である。周辺の4辺にはIOマクロが配置され，左右にはCPUへのインタフェースであるAGTL＋IOマクロを配置，上下にはチップセット間インタフェース（上部がLDX間，下部がGAC間）用のMTL IOマクロを配置した。

む　す　び

本稿では，PRIMEQUESTで新規に開発した六つのチップセットの概要とこれらに実装されている新技術と高信頼性技術について述べた。

今後は第2世代機（最大64CPU）の開発と，次期インテル社製Itanium2シリーズCPUのサポートに取り組んでいく。また中長期的に更なる高性能・高信頼性を追求したPRIMEQUESTを提供し続けていくために，新たなアイディアと新技術に挑戦しチップセットの開発を継続していく。

この研究に対して「半導体アプリケーションチッププロジェクト」の一環として助成していただいた経済産業省と独立行政法人 新エネルギー・産業技術総合開発機構に感謝します。