平成 19 年度
実績報告

「実用化を目指した組込みシステム用ディペンダブル・オペレーティングシステム」
平成 18 年度採択研究代表者

佐藤　三久

筑波大学　システム情報工学研究科　教授
計算科学研究センター　センター長

省電力でディペンダブルな組込み並列システム向け計算プラットフォーム

1. 研究実施の概要

本研究では、ユビキタス情報社会における高度な情報処理の要請に対し、これからの高性能組込みシステムはマルチコア・マルチチップになることを想定し、ディペンダブルOSの一部として、並列システムの高信頼化機構および電力制御機構、省電力高性能高信頼通信機構を研究開発する。高信頼ソフトウェア分散共有メモリ機構を用いた並列システムでの耐故障性機能により高信頼化を図り、電力制御機構により実時間制約下で並列性制御と省電力化を行う。低電力高性能高信頼通信機構では、低電力で並列処理を効率的に行うために複数のネットワークリンクを適宜用いることにより、電力制御・性能制御・耐故障性を包括的に実現する。そのための組み込み向けの通信ハードウェア及び通信機構の開発を行う。これらの技術を統合し、省電力高信頼組込み並列プラットフォームの実証プロトタイプを開発する。

当該年度においては、プロセッサ電力制御機構、高信頼ソフトウェア分散共有メモリ機構を含む高信頼化機構、低電力高性能高信頼通信機構の各項目について、試作実装を開始し、評価検討を行った。低電力高速通信のためのハードウェアとして、PCIexpress Gen2 によるインターコネクトチップ、その通信レイヤについての詳細設計を行った。

2. 研究実施内容

（文中にある参照番号は４．(1)に対応する）

本研究の全体目標である実証プロトタイプならびに核となる低電力高速インターコネクトに関しては、高速通信アーキテクチャグループが中心となって、チーム全体で議論を行い、詳細設計を行った。並列システムの高信頼化機構および電力制御機構、省電力高性能高信頼通信機構については、個別のケーススタディによる検討、試作・実装を行った。プロジェクト全体

のアーキテクチャである P-Bus については、チーム間の打ち合わせを定期的に行っており、次年度以降、この試作に基づき、P-Bus への実装・統合を進める予定である。

以下に、各研究項目の研究実施内容をまとめる。

１、並列組込み向け高信頼共有メモリ機構および省電力実時間並列実行制御機構（電力制御・高信頼並列グループ）

当該年度においては、並列システム電力制御機構と並列組込みシステム向け高信頼化機構について、試作実装を行った。

並列システム電力制御機構については、API について検討した。詳細実装については、次年度において、本プロジェクトの全体アーキテクチャである P-Bus での実装を行う。試作のためのケーススタディとして、並列サーバーアプリケーションでのシステム全体の電力性能最適化システムを設計・評価した。Linux Virtual Server(LVS) による分散 web サーバシステムを取り上げ、負荷状況によりプロセッサの動的電源制御を行う CPMD(Cooperative Power Management Demon)を設計・開発し、DVSF による周波数制御だけでなく、stand-by 状態までノードの状態を制御することにより、QOS を維持しつつ 17%の電力削減が達成できた。この試作で並列システムおいてはノード単体だけでなく、システム全体の電力制御を行うフレームワークが重要であることがわかった。また、電源制御のコードを自動的に埋め込むための手法とそれを用いて動的にDVFS用いて電力制御する方法について試作検討した。

信頼ソフトウェア分散共有メモリ機構を用いた並列システムでの耐故障性機能については、前年度整備した SH4 による並列システム上でソフトウェア分散共有メモリを実装し、評価した。このソフトウェア分散共有メモリシステムでは、組み込みプロセッサ環境に対応するためにページを積極的に開放するなどの工夫を行った。また、シングルプロセスの fail-over 機能については、SH-4 用の check-point 機能を実装した。次年度以降、ページ冗長化による耐故障機能については実装をすすめ、P-Bus 上で実装を行う。

また、開発したシステムの信頼性を定量的に評価するために、仮想的な並列システムにおいて fault injection 行い動作状況を観測するシステム fault VM を設計・試作を行い、その有効性について検証した[1]。

2、並列システム内高信頼高性能通信機構（通信システムアーキテクチャ・グループ）

通信システムアーキテクチャ・グループにおける平成19年度の研究は、(1)組み込みシステム向け低電力・高性能・高信頼通信システムのアーキテクチャ設計と下位ハードウェア制御機構の設計、(2)Gigabit Ethernet を用いた汎用高性能・高信頼通信システムの実装、の2つに分けられる。

(1)に関しては、高速ネットワークグループで開発中の PCI-Express Gen2 のリンク(PHY)を制御する組み込みプロセッサ(M32R)での制御プログラムの設計と基本部分の開発が中心テーマである。次年度以降の研究で、最終的に4本のPCI-Express Gen2リンクとこの制御プロセッサを全てIP

化し、1つのチップ（Communicator Chip と呼ぶ）に実装する計画であるが、今年度は PHY 部分が FPGA 実装されることを踏まえ、制御プロセッサの基本機能と制御プロトコル、ルーティング機構等に関する基本設計を行った。メインプロセッサからの通信要求レートに応じた省電力通信を行うための PCI-Express のステート制御方法とリンク故障時の対応を従来の PCI-Express 規格以上に高めるためのレーン入れ替え制御に関する検討を行い、高速ネットワークグループへの要求仕様を確定させた。また、現在利用可能な M32R プラットフォームにおいて制御プログラムのプロトタイプ実装を行い、通信レートや処理時間の予備評価を行った。この結果として、制御プロセッサである M32R コアは、最終的に実装可能な動作周波数を考慮しても、single core では性能不足で、最低限2コアを持つ multi-core 構成とすることが望ましいという結論を得た。また、PCI-Express Gen2 の PHY の FPGA 実装（1ボード当たり1リンク分）の2リンク分と既存の M32R プラットフォームを接続するブリッジボードを開発し、ルーティング機能と並列通信機能を確認するための最小構成である3プロセッサ分の開発プラットフォームを作成した。

(2)に関しては、x86 ベースの標準 PC サーバにおけるユーザレベルでの複数 Gigabit Ethernet による高性能・高信頼通信システム RI2N/UDP を完成させ、マルチリンク Gigabit Ethernet を、無故障時にはバンド幅増強に、故障時は fail-over 機能として活用するシステムを構築した（文献[1],[2]）。さらに、RI2N/UDP を改良し、Linux カーネルにおける仮想ネットワークドライバとして実装した。これにより、マルチリンク Gigabit Ethernet を高バンド幅化と高信頼化の両者に活用する、システムレベル実装 RI2N/DRV を開発した。RI2N/UDP ではユーザプログラムの一部ライブラリの入れ替えが必要であり、また無故障時のバンド幅は高いがレイテンシが通常の TCP/IP 通信に比べ増大するという問題があったが、RI2N/DRV では完全な TCP/IP ライブラリ互換として利用可能で、ユーザプログラムに一切変更を加えず利用可能となった。さらに、システムレベル実装としたことと、TCP におけるマルチリンク利用時のパケット入れ替え問題が解消され、1リンク時の TCP 通信と同等なレイテンシでの通信が可能となった。この結果、概念的に本システムと同じ機能を標準 Linux で提供している Linux Channel Bonding 機構より高いバンド幅と低いレイテンシを実現し、標準 Linux 実装を上回る性能を達成した（以上、文献[3]）。また、DEOS において標準的に利用される P-Bus のコンポーネントとして RI2N/DRV を実装する際の P-Bus デバイスドライバ機能に関する検討を行った。

3、低電力高速インターコネクトと省電力高密度並列ハードウェアプラットフォームの開発（高速ネットワークグループ）

前年度に検討したハードウェアプラットフォームの全体構成（プロセッサ、DRAM、フラッシュメモリ、及び PCI-express Communicator）の中核となる高速ネットワークインターコネクト構成する PCI-Express Communicator 機能部につき、平成20年度に実施するデモをターゲットとするデモ実装及び評価システムボードの設計を実施した。本評価システムボードは、FPGA に実装されたリンク層と物理層(PHY)テストチップから構成される PCI Express 評価ボード、ルーティング機能（上記レイヤ）を実装する M32R T-Engine ボード、PCI-Express バックプレーンボード等から構成される

（図 1）。

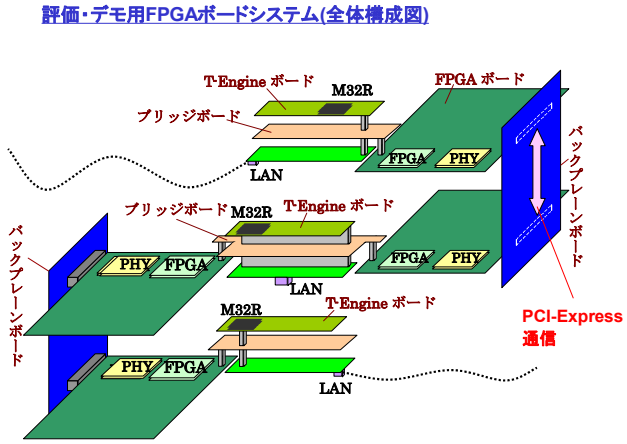

図 1　評価・デモ用 FPGA ボードシステムの全体構成図

PCI Express Genaration2 物理層（PHY）テストチップは、65nmCMOS プロセスを用いてテストデバイスを設計・試作評価し、パワマネージメントを可能とする 6Gbps のデータ転送及び、動的制御による Generation 2(5Gbps)/Generation 1(2.5Gbps)のデータ転送レート切り替え対応等の基本動作を確認した。図 YY は振幅制限機能（De-Emphasis）を含め、十分な Eye パターンが確保されている様子を示している。また Rx（受信側）の性能指標である入力信号ジッタに対して 6Gbps 時動作時でも十分なジッタ耐量を有することが確認され、PCI-Express Generation2 として十分な動作マージンを確保できた。

実用化として通信 IP として重要な市場での接続実績保障のため、本チップは 2007 年 12 月には、PCI-SIG 主催（PCI 関係のインターフェース IP の接続認証機関）のコンプライアンステストに参画し、FYI（試行テスト）ながら波形品質試験をパスした（公式な認証試験は 2008 年 8 月以降に実施予定）。

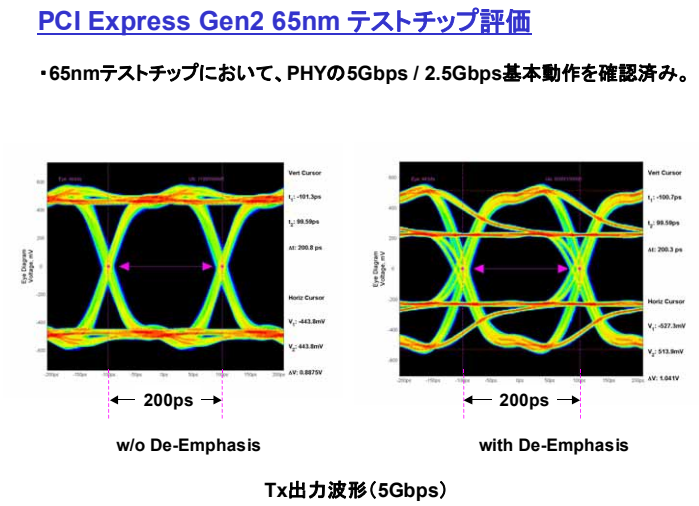

図2　PCIe Gen2 65nm テストチップの評価結果

3. 研究実施体制

(1)電力制御・高信頼並列システムグループ

①研究分担グループ長:佐藤三久(筑波大学、教授)

②研究項目

・並列組込み向け高信頼共有メモリ機構および省電力実時間並列実行制御機構

(2)通信システムアーキテクチャグループ

①研究分担グループ長:朴　泰祐(筑波大学、教授)

②研究項目

・並列システム内高信頼高性能通信機構

(3)高速ネットワークグループ

①研究分担グループ長:有本　和民(ルネサステクノロジ、副統括部長)

②研究項目

・低電力高速インターコネクトと省電力高密度並列ハードウェアプラットフォームの開発

4. 研究成果の発表等

(1) 論文発表(原著論文)

１、並列組込み向け高信頼共有メモリ機構および省電力実時間並列実行制御機構（電力制御・高信頼並列グループ)

[1] Takayuki Imada, Mitsuhisa Sato, Yoshihiko Hotta and Hideaki Kimura、"Power Management of Distributed Web Servers by Controlling Server Power State and Traffic Prediction for QoS", HPPAC 2008 in conjunction with IPDPS2008, to appear, 2008.

2、並列システム内高信頼高性能通信機構（通信システムアーキテクチャ・グループ)

[1] T. Okamoto, S. Miura, T. Boku, M. Sato, D. Takahashi, ″RI2N/UDP: High bandwidth and fault-tolerant network for a PC-cluster based on multi-link Ethernet″, Proc. of CAC2007 (included in Proc. of IPDPS2007), CD-ROM, Long Beach, 2007.

[2] 岡本 高幸, 三浦 信一, 朴 泰祐, 佐藤 三久, 高橋 大介, ″Ethernet マルチリンクによるPC クラスタ向け高バンド幅・耐故障ネットワーク RI2N/UDP″, 情報処理学会論文誌コンピューティングシステム, Vol. 48, No. SIG 8(ACS 18), pp.153-164, 2007.

[3] 岡本 高幸, 三浦 信一, 朴 泰祐, 塙 敏博, 佐藤 三久, ″ユーザ透過に利用可能な高性能・耐故障マルチリンク Ethernet 結合システム″, 2008 年年ハイパフォーマンスコンピューティン

グと計算科学シンポジウム HPCS2008 論文集, 2008.