

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。省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、本プロジェクトの基本コンセプト ***-proof** を実現するためのOS機能抽象化層 **P-Bus** について、最初の **P-Component** としてマルチリンク **Gigabit Ethernet** による高性能・耐故障通信機構 **RI2N/DRV** の **P-Component** 版実装を行った。また、並列システム電力制御機構、高信頼ソフトウェア分散共有メモリ機構、プロセッサ電力制御機構の各項目について、試作実装を行い、評価検討を行った。省電力高性能高信頼通信機構 **PEARL** については、低電力高速通信のためのハードウェアとして **PCI Express Gen2** によるインターコネクトチップ **PEACH** 及びその通信レイヤについての詳細設計を引き続き行い、同時に **FPGA** を用いた **PEACH** 評価環境を試作実装した。また、各研究チームと共同でディペンダブル OS のフレームワークおよびディペンダブルシステムベンチマークについて検討を行った。

2. 研究実施内容(文中にある参照番号は 4.(1)に対応する)

本研究の核となる低電力高速インターコネクト **PEACH (PCI Express Adaptive Communication Hub)** に関しては、通信システム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 PEARL(PCI Express Adaptive and Reliable Link)については、個別のケーススタディによる検討、試作・実装を引き続き行った。本プロジェクトで導入された OS 機能拡張のための抽象化層 P-Bus API に基づき、P-Component として RI2N の機能を実装した。また、各研究チームメンバーからなるコアチームにおいて、ディペンダブル OS のフレームワークとディペンダビリティ支援、さらにディペンダブルシステム全体を評価するためのベンチマークについての検討を行った。

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)

並列システム電力制御機構については、昨年度に引き続き CPMD(Cooperative Power Management Daemon)に関して実際のリクエストレートの時間変化を模倣したベンチマークソフトを作成し、複数の実環境想定下における CPMD の QoS 性能および省電力性能について評価を行った。これにより、サーバノードの Stand-by 状態を積極的に利用することでリクエストレート変化の度合いによらず高い QoS 性能の維持と効果的な消費エネルギーの削減を達成できることが分かった[1][3]。さらに Intel の組込み向けプロセッサである Atom を搭載した小規模な並列システムを構築し、省電力化を実現する手法についても検討を行い、実行するプログラムの並列度に応じて計算ノード数を制御することでエネルギー削減が効果的に行えることを確認した。また、プログラム中の計算や通信などの特性に応じてコードを挿入し、プログラム実行時に動的電力制御を行う手法を提案、実現した。プログラム中の各領域において DVFS(Dynamic Voltage and Frequency Scaling)機構を適切に制御することにより、消費エネルギーを削減できることを確認した[2]。CPMD、ギャングスケジューリングの P-Component 化については引き続き石川チームおよびディペンダブル組込み OS 研究開発センターと連携および議論を行っている。

高信頼ソフトウェア分散共有メモリを用いた並列システムでの耐故障性機能として、耐故障ソフトウェア分散共有メモリシステム SCASH-FT の設計と実装を行った。SCASH-FT は PC クラスタ向けに開発されたソフトウェア分散共有メモリ SCASH をベースにチェックポインティング・リスタート機構を実現したものである。ユーザの指示によりシステム全体のスナップショットを複数ノードに冗長化して保存し、故障が発生すると予備ノードを用いて回復を行う。一連の処理は全てソフトウェア分散共有メモリシステムが行うため、アプリケーションでは少ないプログラミングコストでチェックポインティングし、故障時にはリスタートさせることができる。性能評価の結果、チェックポインティングや回復のコストは保存するデータの大きさに比例しており、チェックポインティングの頻度をユーザが明示的に指定することで、十分に現実的なパフォーマンスが得られることを確認した。

フォールトインジェクタについては、利用を検討している企業と議論を行い、汎用デバイスや組込み機器固有のデバイスの故障模擬やシナリオの実現可能性について検討した。仮想マシンを用いた並列テスト環境についても検討を行った。また、組込みマルチコアプロセッサでの OpenMP の性能についても調査を行った[4]。

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 20 年度の研究は、(1)省電力高性能高信頼通信機構 PEARL 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と共同で PEARL のアーキテクチャについて検討した。最終的に、PCI Express カードエッジコネクタを持ち PEARL の核となる PEACH チップを搭載したネットワークインタフェースカード(NIC)を実装する計画である。この NIC をノードの PCI Express スロットに挿入し、さらに 3 本の PCI Express 外部ケーブルを用いて隣接ノードの NIC との間を接続する。また、リンクの先にはノードの代わりに入出力デバイスを接続することもでき、PEACH をスイッチとしてノード毎に入出力デバイスを切り替えて使用することが可能になる。

また、高速ネットワークグループによって作成された PCI-Express Gen2 の PHY テストチップと、FPGA による Link 層コントローラを搭載した評価ボードを用いて、M32R プロセッサ搭載評価ボードと組み合わせて PEACH の評価環境を作成し、PEACH での通信制御プログラムの試作実装を行った。

(2)に関しては、Linux カーネルにおける仮想ネットワークドライバとして、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 RI2N/DRV を実現し、評価を行った[1][2]。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。同時に、いずれかのリンクが故障した場合にも残りのリンクを用いて通信を継続し、故障が取り除かれると正常時のバンド幅に回復することを確認した。さらに、RI2N/DRV を既存の UNIX ネットワークサービスに応用し、ネットワークファイルシステム(NFS)に適用した場合に高い性能を示すことを確認した[3]。また、本プロジェクトにおいて共通に利用する OS 機能拡張のための抽象化層 P-Bus API に基づく P-Component として、RI2N/DRV を移植し、RI2N/P-Bus を実装した。これは P-Component として初めての実用例であり、2008 年 11 月に行われた組込み総合技術展 ET2008 においてデモンストレーションを行った。

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(高速ネットワークグループ)

高速ネットワークインターコネクトを実現するための要素技術として開発した PCI Express Generation2 Link 層 IP を実装した FPGA と、物理層(PHY) テストチップとを搭載した評価ボードを製作した。本評価ボードにより、Link 層、PHY 層を結合した形で、プロトコル動作、電気的特性について PCI Express 規格に対するチェック、評価を行い、規格準拠していることの最終確認をするとともに、市販製品での接続性検証も実施した。また、通信 IP として実用化後に重要な要素となる、市場での接続実績を保障するため、本評価ボードで2008年9月に 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、世界で初めて、全テスト項目にパスし、同年11月に公式な認証を得た。

一方、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を構成する PCI-Express Communicator チップ (コード名: PEACH(PCI Express Adaptive Communication Hub))機能部につき、デモ及び評価用のシステムボードの設計を完了し、ボード製作を完成させた。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)。

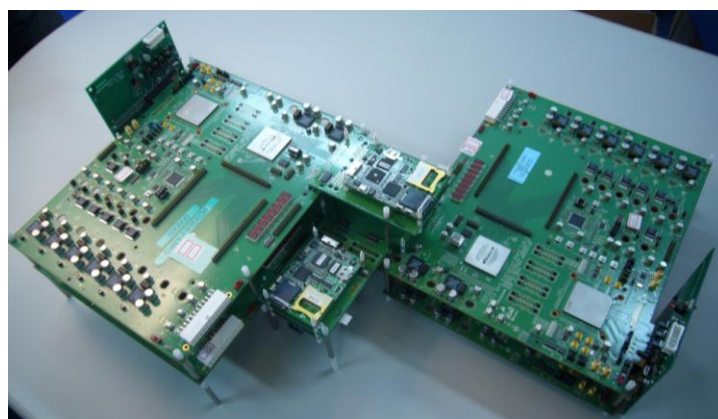


図 1 試作 FPGA ボードと M32R ボードによる開発システム

また、PEACH のバス構造を含む内部アーキテクチャ検討を実施し基本仕様を策定した(図 2)。さらに PCI Express の詳細性能評価を実施すると共に、その結果を基礎に、ルーティング性能を向上させるための PCI Express IP と内部バスとの間のブリッジ部分の詳細仕様検討に着手した。

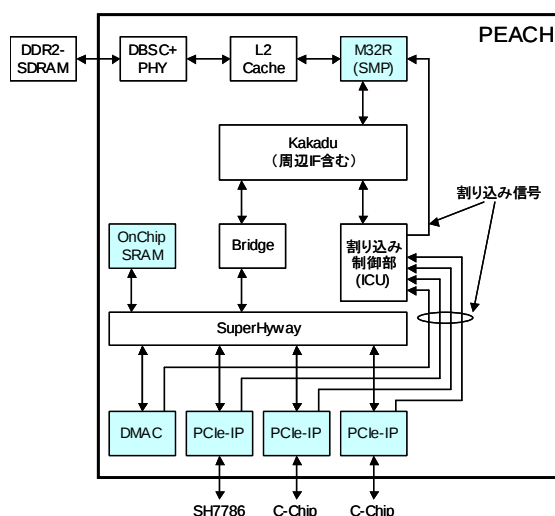


図 2 PEACH チップの基本構成図

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、副統括本部長)
- 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, Mitsuhsa Sato, Yoshihiko Hotta, and Hideaki Kimura, "Power Management of Distributed Web Servers by Controlling Server Power State and Traffic Prediction for QoS," Proc. 22th IEEE International Parallel and Distributed Processing Symposium (IPDPS) Workshop on High-Performance, Power-Aware Computing (HP-PAC), pp. 1-8, Apr. 2008.
- [2] Hideaki Kimura, Mitsuhsa Sato, Takayuki Imada, and Yoshihiko Hotta, "Runtime DVFS Control with instrumented code in Power-scalable Cluster System," Proc. 10th IEEE International Conference on Cluster Computing (CLUSTER 2008), pp.354-359, Sep. 2008.
- [3] 今田 貴之, 佐藤 三久, 木村 英明, 堀田 義彦, 「分散型 web サーバでの負荷変動を考慮した省電力化のためのノード状態制御」, 情報処理学会論文誌コンピュータインテリジェントシステム, Vol.2 No.2 (ACS 26), 2009 年. (to appear)
- [4] Toshihiro Hanawa, Mitsuhsa Sato, Jinpil Lee, Takayuki Imada, Hideaki Kimura, Taisuke Boku, "Evaluation of Multicore Processor for Embedded Systems by Parallel Benchmark Program using OpenMP," Proc. Of International Workshop on OpenMP (IWOMP 2009), 2009. (to appear)

- 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] 岡本 高幸, 三浦 信一, 朴 泰祐, 埴 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」, 情報処理学会論文誌 (ACS),

Vol.1, No.1, pp.12-27, 2008 年 6 月.

- [2] Shin'ichi Miura, Takayuki Okamoto, Taisuke Boku, Toshihiro Hanawa, Mitsuhisa Sato, "RI2N: High-Bandwidth and Fault-Tolerant Network with Multi-link Ethernet for PC Clusters," Proceedings of 2008 IEEE International Conference on Cluster Computing (Cluster 2008), pp.1-6, Sep. 2008.
- [3] Shin'ichi Miura, Toshihiro Hanawa, Taiga Yonemoto, Taisuke Boku, Mitsuhisa Sato, "RI2N/DRV: Multi-link Ethernet for High-Bandwidth and Fault-Tolerant Network on PC Clusters," Proceedings of Workshop on Communication Architecture for Clusters (CAC2009) in IPDPS 2009, 2009. (to appear)
- 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(高速ネットワークグループ)

(2) 特許出願

平成 20 年度 国内特許出願件数 : 0 件 (CREST 研究期間累積件数 : 0 件)