平成 21 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。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、組み込み向け省電力・高信頼・高性能通信リンクであるPEARL及びこれを物理的に実現するコミュニケーションハブであるPEACHチップのプロトタイプ・ハードウェア実装に関する最終段階の設計・開発を行い、最先端45nm CMOS技術による実シリコンの作製を実施した(2010年5月にサンプル完成予定)。同時に、PEACHチップ内のファームウェア、PCI-E仕様に準拠したペリフェラルカードであるPEACHボードの開発を行った。また、GbEthernetのマルチリンク制御に基づく高信頼・高性能インターコネクションであるRI2Nについては、リンク数の異なるマルチリンク環境でも常に性能を最適に保つ動的トラフィック制御機能を組み込んだRI2N+, RI2N++システムを開発した。並列システム電力制御機構CPMD(Cooperative Power Management Daemon)、高信頼ソフトウェア分散共有メモリ機構SCASH-FTについて、引き続き評価を行い、成果発表を行った。

さらに、DEOS の基本コンセプトに基づいた設計・実装・試験・保守時支援ツールとして、仮想マシンを利用して、ハードウェア故障をユーザ透過にOSの検証まで用いることができるフォルトイ

ンジェクション実行環境 FaultVM と、仮想マシンを柔軟に管理するクラウド技術を利用して、大規模な並列分散システムの検証・開発環境を提供するシステム D-Cloud を設計・試作した。この2つを組み合わせることによって、ハードウェア故障を含む並列システムの信頼性機能が動作しているかを確かめるための実験が可能であることを示した。

§2. 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

§3. 研究実施内容

(文中に番号がある場合は(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、チップ、ボード、周辺ソフトウェアについて、最終段階の設計・開発を行い、最先端 45nm CMOS 技術による実シリコンの作製を実施した。

DEOS の基本コンセプトに基づいた設計・実装・試験・保守時支援ツールとして、当グループはフォルトインジェクションを担当しているが、当該年度においては新規のコンセプトとして、仮想マシンを柔軟に管理するクラウド技術を利用して、大規模な並列分散システムの検証・開発環境を提供するシステム D-Cloud を提案し、設計・試作した。これについて、中間評価・報告会にてデモを行った。

また，各研究チームメンバ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）

当該年度においては、特に高信頼ソフトウェアを開発するためのフォルトインジェクション環境の研究に注力した。これまで、フォルトインジェクタについては、仮想マシン上で並列システムを構成しフォルトインジェクションを行い、観測および動作検証を行うためのプロトタイプ FaultVM/Xen を設計・試作を行ってきた。これを元に、仮想マシンである QEMU に汎用デバイスの故障模擬、および故障シナリオを実現する機能を持つ FaultVM/QEMU を実装し、実験・検討を行った。

これらの検討を基に、動的にテスト環境の構築，並列プログラムテスト、デバイスに対する故障エミュレーションが可能な並列分散システムテスト環境を提供する D-Cloud を提案した。高い信頼性確保のためには異なる入力による網羅的なテスト実行やハードウェア故障に対する耐故障性のテストなど様々なテストが存在し、それらを実行するには非常に時間と手間がかかる。D-Cloud ではフォルトインジェクションが可能な仮想マシンを用いて、仮想デバイスレベルでの故障についてのテストが可能であるだけでなく、仮想マシンをクラウドとして管理することにより、多くの計算資源を柔軟に利用することができ、多くのケースについてのテスト作業を自動化することができる環境を提供できる。図 1 にその概要を示す。クラウドシステムとして、Eucalyptus を用いて、フォルトインジェクション機能を持つ仮想マシンとして、上記で検討した FaultVM/QEMU を用いた。FaultVM/QEMU では、メモリ、ハードディスク、ネットワーク等のインジェクション機能を実装した。D-Cloud フロントエンドはこれらの環境を制御し、ユーザインタフェースとテスト実行の自動化を行うサーバである。システムの構成やフォルトインジェクションを含むテストのシナリオの記述について検討し、それを記述する XML 記述を設計した。

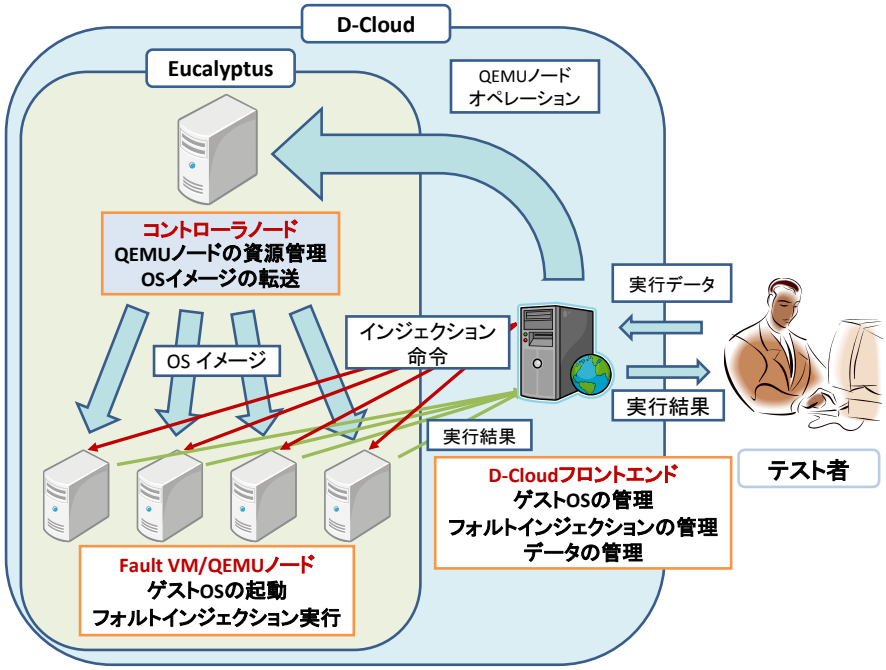


図 1 D-Cloud システムの概要

D-Cloud は、仮想マシンレベルでフォルトインジェクション機能を実装することにより、システムを実際のマシンに近い環境でハードウェア障害に対する耐故障機能をテストすることができる。さらに、仮想マシンを複数用いることにより、高信頼システムとして重要な分散システムのテストが容易にできる。それだけでなく、クラウドによる大量の計算資源の柔軟な管理・利用が可能であり、高信頼化のための網羅的なテストも可能である。クラウドの計算資源を利用した、ログの解析などの展開も検討しており、OSを含むこれからの新しいソフトウエアテスト環境として期待できると考えている。

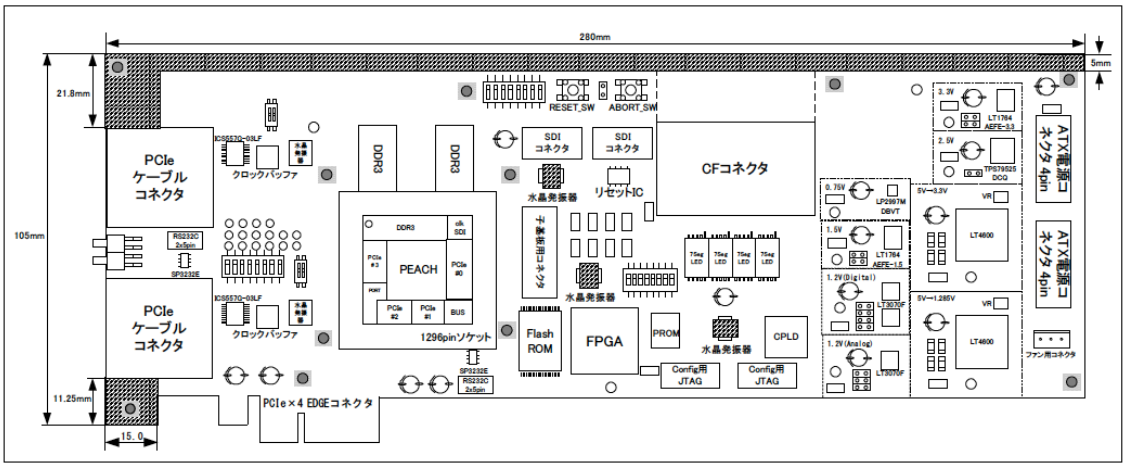
なお、並列システム電力制御機構 CPMD(Cooperative Power Management Daemon)、高信頼ソフトウェア分散共有メモリ機構 SCASH-FT について、引き続き評価を行い、成果発表を行った。

2. 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当該年度の研究は、(1)省電力・高性能・高信頼通信機構 PEARL の詳細設計及びこれを実現する PEACH チップのテープアウト向け最終設計仕様の作成、(2)PEACH チップ完成後、一般の PC 環境において PEACH 及び PEARL のテストと評価を容易に行うための PCI-E カードとして PEACH ボードの作成、(3)Gigabit Ethernet を用いた汎用高性能・高信頼通信システム RI2N の改良、の3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、PEACH チップを実現する 45nm ルール及びシャトル発注で許されるチップ・ダイ・サイズを勘案し、最終的な機能及び内部バス性能、割り込み制御回路、そして最も重要な中央制御用プロセッサである 4 core M32R に付随する IP を確定し、テープアウト向け仕様作成を完了した。PEACH チップ自体の実装は高速ネットワークグループが行うが、限られたハードウェア制約の中で PEARL の想定仕様を満たすためのノード間通信機能を実現するため、内部の詳細仕様の全てについて高速ネットワークグループとの打ち合わせを毎月行い、共同でアーキテクチャ設計を行った。また、それらの内部仕様に基づき、PEACH チップ内の制御を行う 4 core M32R の制御ファームウェアの基本設計と実装テストを行った。この作業は昨年度開発した FPGA テストシステム上で行った。

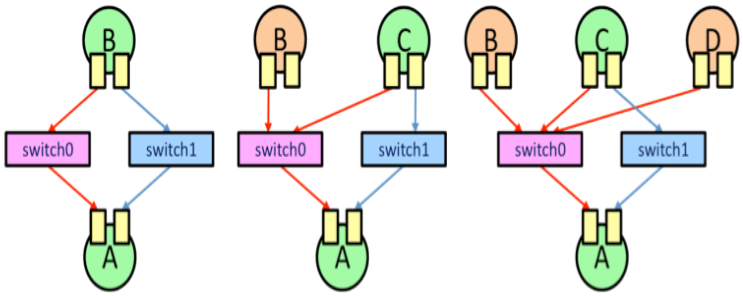
(2)については、PEACH プロトタイプチップ完成後直ちに一般の PC サーバ上で PEACH のテストと PEARL としてのノード間接続機能・性能評価を行えるよう、PEACH を搭載した PCI-E ボードである PEACH ボードを開発した。PEACH が提供す4つの PCI-E ポートのうち1つを PC サーバの PCI-E スロットに直接接続可能とし、残り3ポートを外部ノードとの接続用に PCI-E extension cable を接続するためのコネクタとして実装した。この他、消費電力測定や JTAG プローブ等、PEACH の全ての機能テストを行う周辺回路を実装している。図 2 に PEACH ボードのレイアウトを示す。



※ 図中の斜線部は部品配置、配線禁止領域。

図 2　PEACH ボードレイアウト

(3)については、これまで開発してきたGbEthernetマルチリンクによる汎用高信頼・高性能インターコネクションであるRI2Nの適用範囲を拡張し、より広範囲なシステム構成で最高性能を出すような抜本的なシステム改良を行った。従来のRI2Nでは、PCクラスタのような均一なマルチリンク構成、すなわち全てのノードのNIC数が等しく、均等にネットワークが構成されている場合のみを想定していた。しかし、実際の研究室や計算センターの環境では、例えば重要なサーバ系ネットワークは二重化して高信頼・高性能を実現しつつ、クライアント系マシンでは通常のシングルリンクのみで結合するような、コスト性能バランスを考慮した構成を取る事が予想される。従来のRI2Nは、このような非対称ネットワークトポロジにも原理的に対応するが、TCP/IPとの連携を考えると、非対称リンク構成上のトラフィックを最適化できず、十分な性能が引き出せないことがわかった。そこで、動的なリンク構成変更に対応するRI2N+、さらにトラフィックパターン自体の動的変動にも追従するRI2N++という2つの改良システムを実装した。各種非対称トポロジで評価した結果、RI2N⇒RI2N+⇒RI2N++の順で、非対称性が強い場合の性能向上が顕著に見られた。これらの改良により、RI2Nを一般のあらゆるマルチリンク構成に柔軟に適応可能となり、幅広い局面で利用することが可能になった。図3に典型的な非対称リンク構成とそれぞれの場合のRI2N, RI2N+, RI2N++の性能比較を示す。



RI2N	223.1 [MB/s]	142.9 [MB/s]	120.4 [MB/s]
RI2N+	223.1 [MB/s]	161.7 [MB/s]	153.8 [MB/s]
RI2N++	218.3 [MB/s]	191.3 [MB/s]	192.5 [MB/s]

図 3　RI2N,RI2N+,RI2N++の性能比較

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）

本研究の実証システムとして、組込み向け低電力プロセッサを用いたハードウェアプラットフォームを開発において、前年度に設計を行った、PCI Express Gen.2 バックプレーンボードから構成された FPGA デモボードシステムのハードウェアの評価と中間報告会デモの支援を実施した。また、上記デモ開発結果のフィードバックを実施し、最終形ハードウエアプラットフォーム（PEARL）の仕様策定を実施し、上記 FPGA デモボードを1チップ化したシステムの中核となる PCI-Express Communicator チップ（コード名：PEACH(PCI Express Adaptive Communication Hub)とパッケージ、チップ評価ボードの設計を行い、最先端 45nm CMOS 技術による実シリコンの作製を実施した（現在、ウエハプロセス中であり、2010 年 5 月にサンプル完成予定）。また、搭載された PCIexpressGen.2 の物理層（PHY）は、FPGA デモボード用に開発された 65nm 版から 45nm 版にポーティングを実施した。

PEACH チップは、コントローラとしての M32R CPU と DDR3 コントロールインターフェース、および最大 20 ギガビット/秒の転送レートを有する4レーンの PCI Express Gen2 を 4 ポート搭載したインターフェースを装備し、これに対応した高速内部データバスを中心に構成される。チップ内の各モジュールは高速内部システムバスで結合されている。また 1 メガバイト規模の高速キャッシュメモリ等のオンチップ SRAM が搭載されている。PEACH チップに搭載されているインターコネクト技術は消費電力と伝送レート・距離に対応した可変プログラムリンクを可能とし、本研究の主旨である省電力・ディペンダブル機能に対応したハードウェアを可能とする構成である。図 4 は PEACH のブロック図である。

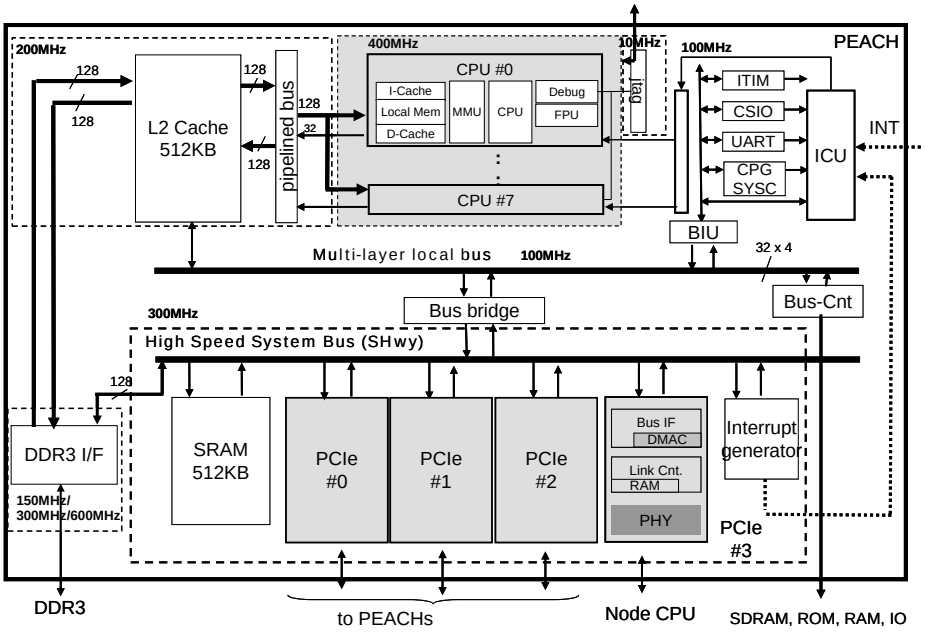


図 4　PEACH ブロック図

PEACH チップは、45nm の8層配線、マルチ Vth トランジスタ対応の Low power CMOS プロセスで試作されている。チップサイズは 11x11mm2 で、1008 ピンの BGA パッケージに収納される。電源電圧は、コア部の 1.2V, DDR3 インターフェース部の 1.5V、周辺 IO 部の 3.3V の3電源を供給する。（図 5 にチップ図、表 1 に、チップ諸元を示す）

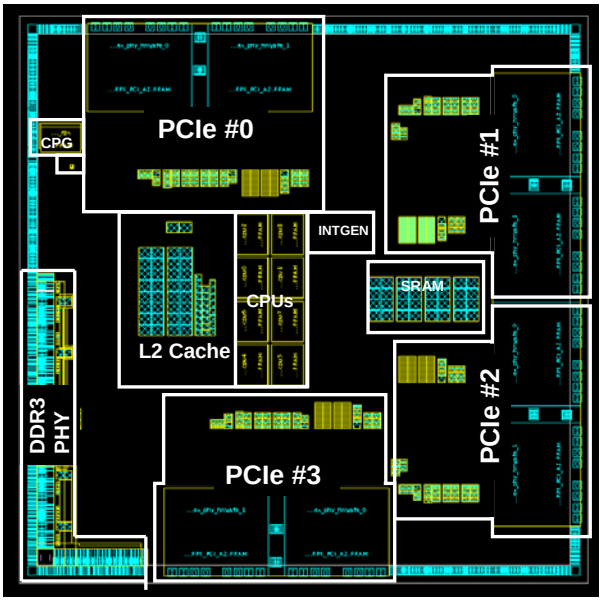


図 5　PEACH　チップ図

CPU	32-bit Processor (400MHz) x 8 SMP L1-cache:8kB(I)+8kB(D), LM:8kB, MMU, FPU
Memory	L2-cache: 512kB Internal SRAM: 32kB, 512kB
DRAM I/F	DDR3 I/F x 1 SDRAM I/F x 1
PCIe I/F	PCI Express standard Rev.2.0 Transfer speed: 5.0GT/s, 2.5GT/s per lane 4 lanes (20Gbps) x 4 ports Maximum payload size:1024bytes Upconfiguration function Automatic retransmission function Root port / Endpoint selectable
Interrupt Generator	Transfer address, size information register x 3 Automatic transfer mode
Bus	Packet router Multi-layer bus (4-layer) Pipelined bus

表 1 チップ諸元

§4. 成果発表等

原著論文発表

- 論文詳細情報

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] 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ティングシステム, Vol.2, No.2, pp.75-88

[2] Jinpil Lee and Mitsuhisa Sato: Reliable Software Distributed Shared Memory using Page Migration, The Fifteenth International Conference on Parallel and Distributed Systems (ICPADS'09), pp.456-463

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. Yonemoto, S. Miura, T. Hanawa, T. Boku, M. Sato, "Flexible Multi-link Ethernet Binding System for PC Clusters with Asymmetrical Topology", Proc. of Int. Conf. on Parallel and Distributed System 2009 (ICPADS2009), pp.49-56, DOI: 10.1109/ICPADS.2009.104, 2009.