DEOS-FY2009-PP-08

高性能・省電力でディペンダブルな通信リンク PEARL

佐藤三久[1]

朴　泰祐[1] 、有本和民[2]

[1]筑波大学、[2]ルネサステクノロジ

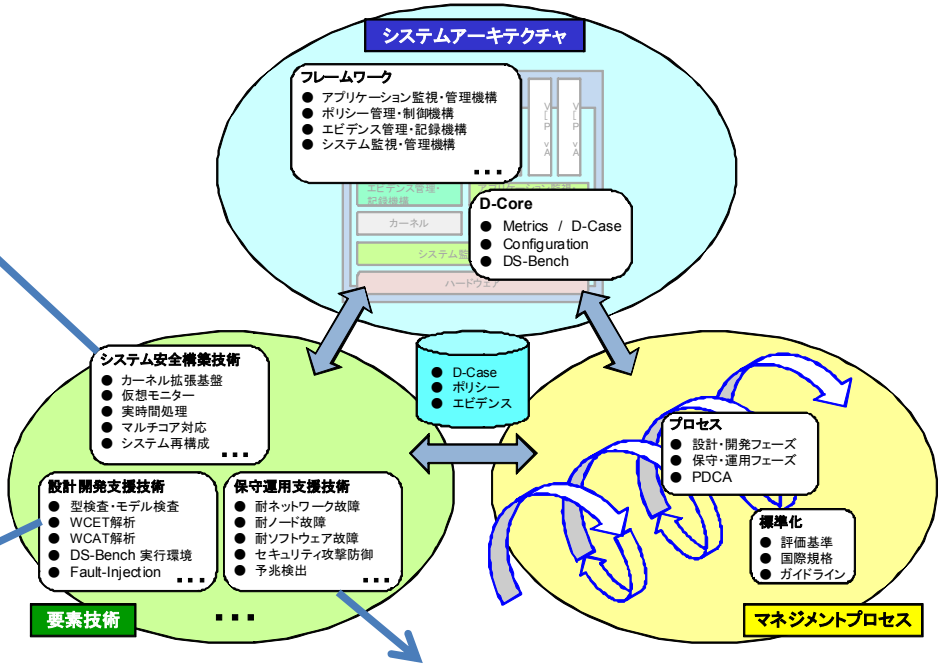
発表の概要

- DEOSの全体構想での位置づけ
  - 「ディペンダブル並列システム」
  - 「ディペンダブル並列システム」に要請されるネットワーク技術
- PEARL
  (PCI Express Adaptive & Reliable Link)
  - PEACH (PCI Express Adaptive Communication Hub) チップ
  - 概要と特徴
  - 詳細、比較、性能評価、…
- 重要成果リスト
- 現状と今後の研究開発
- デモについて

DEOSの全体構想と研究の位置づけ

システム安全構築技術

開発モジュール	チーム名
仮想モニタ モニタリング(VMO) マルコア制御(VMC)	中島
P-Bus Core 論理分割(LPAR)	石川



設計開発支援技術

開発ツール	チーム名
型検査・モデル検査(TCHK/MCHK)	前田
最悪実行時間予測(RETAS)	石川
電力使用量予測(GREEN)	徳田
Fault Injection (D-Cloud)	佐藤
ディペンダブルシステムベンチマーク実行環境(DS-Bench)	石川

保守運用支援技術

開発モジュール	チーム名
動作時間予約機構(TR) 耐故障ネットワーク機構(SCTP+FHO)	徳田
耐故障ネットワーク機構 (RI2N/PEACH)	佐藤
アカウンティング機構(ACT)	中島/センター
シングルIPアドレス機構(SIAC)	石川

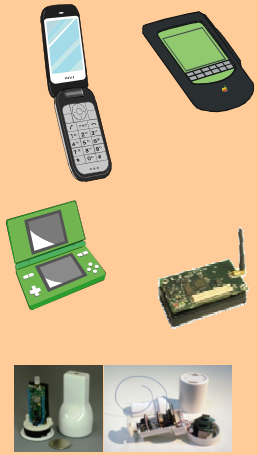
3

組み込み向けプラットフォームのトレンドとディペンダブルシステムの新しい課題

- 「第３世代」の組込み高性能並列システム
  - マルチコア、マルチプロセッサになりつつある
- 高信頼性・高性能化への要請
  - 高度な認識（画像、音声）を用いた認識・監視装置
  - 高性能高信頼・情報家電アプライアンスサーバー
    - PCよりもコンパクト、省電力、高性能、高信頼、高機能

組込み機器

- 消費電力・発熱と性能
- マルチコア化
  - 高周波数による性能向上から並列処理による性能向上
    例：SH-4Aマルチコア、MPCore、モバイルPenryn、Griffin
- ネットワーク・ユビキタス化
  - 無線LAN、アドホックネットワーク
  - セキュリティ

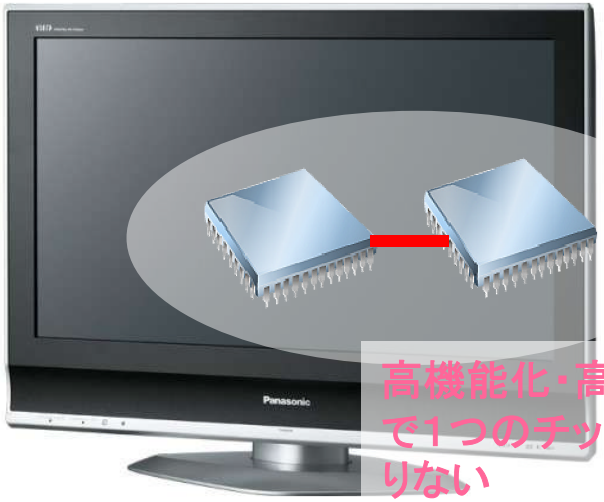


「ディペンダブル並列システム」

- 並列システムでのディペンダビリティのサポート（高性能化と信頼性）
- 並列システムを利用したディペンダビリティの提供（冗長性など）

本チームでは、ディペンダブルOSの実現例として、「ディペンダブル並列システム」の構築に取り組む

近い将来、期待されているコンパクトな高性能・高信頼リンク

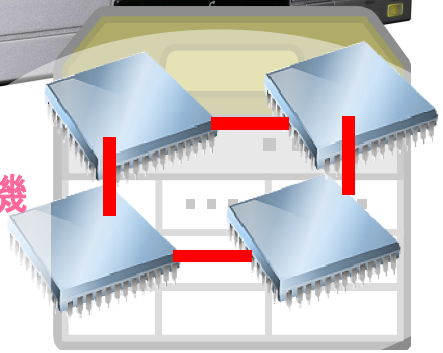


HDTV
などの
メディア機器

高機能化・高性能化で1つのチップではたりない

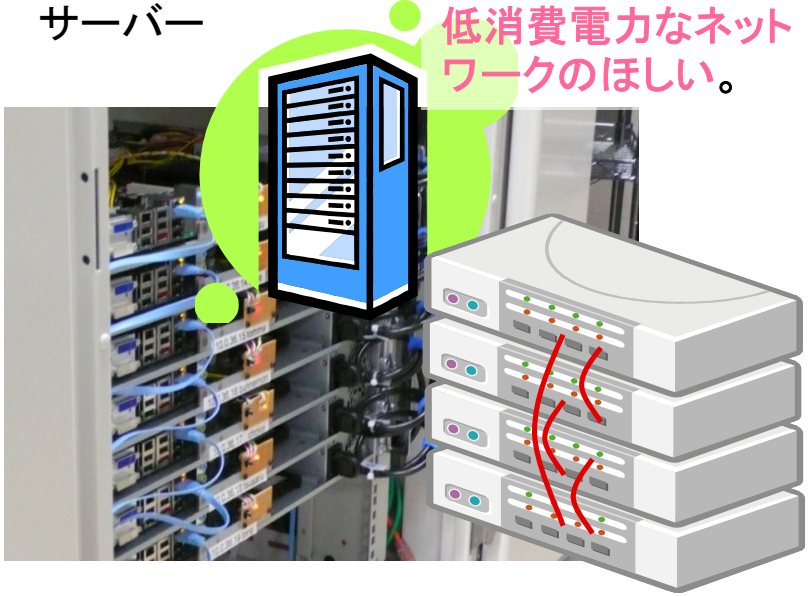
セットトップボックス
メディアサーバ

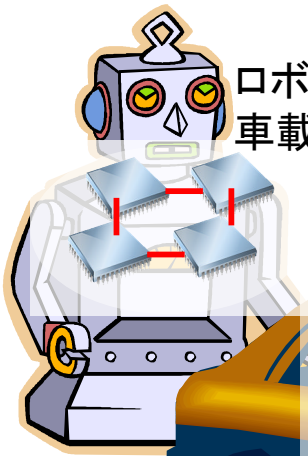
家庭内のメディア高度化・情報化で、高機能化・高性能化が期待される



データセンター
サーバー

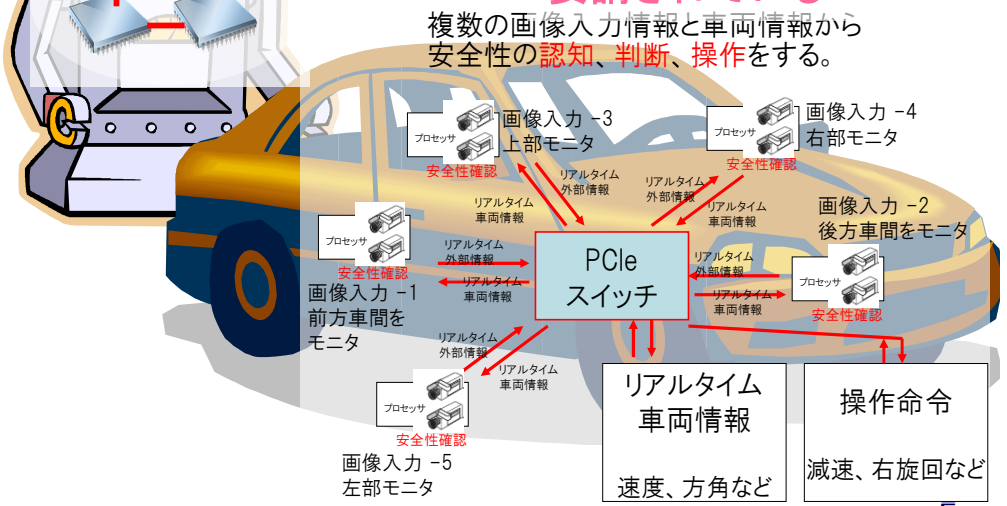
高性能でコンパクト、低消費電力なネットワークのほしい。





ロボット
車載情報機器

IOもできる、
高性能化かつコンパクト・高信頼なネットワークが要請されている



（オープンシステム）ディペンダビリティと高信頼通信リンク

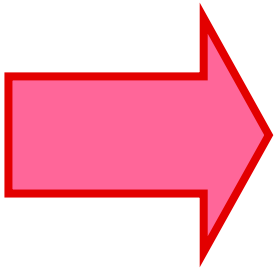
- （オープンシステム）ディペンダビリティ向上のために
  - 開放系障害を起こす要因の最小化（設計時）
    - …要求、仕様、設計、実装、テストなどの設計時の改善…
  - 開放系障害による影響の最小化（稼動時）
    - 実環境・実時間での仮稼働
    - 稼働中の予知
    - 障害の最小化、迅速な復旧支援

- ソフトウエアだけでは対処できない障害・対策
  - ハードウエア故障対策としての冗長化
  - 物理的な制約（電力や実装スペース）
  - 処理性能の調整・向上

  ⇒　通信リンクのハードウエアはこれらを解決する有効な手段

「ディペンダブル並列システム」に要請されるネットワーク技術

- 並列システムによる高性能化が可能であること
  - 並列処理による性能向上を達成するための十分な性能
- 並列システムの信頼性の向上をサポートすること
  - 冗長性(プロセッサ・機器)による信頼性確保
  - 一部が壊れていても、動作できる柔軟性
  - プロセッサだけでなく、入出力機器についても信頼性を確保できること
- 省電力性

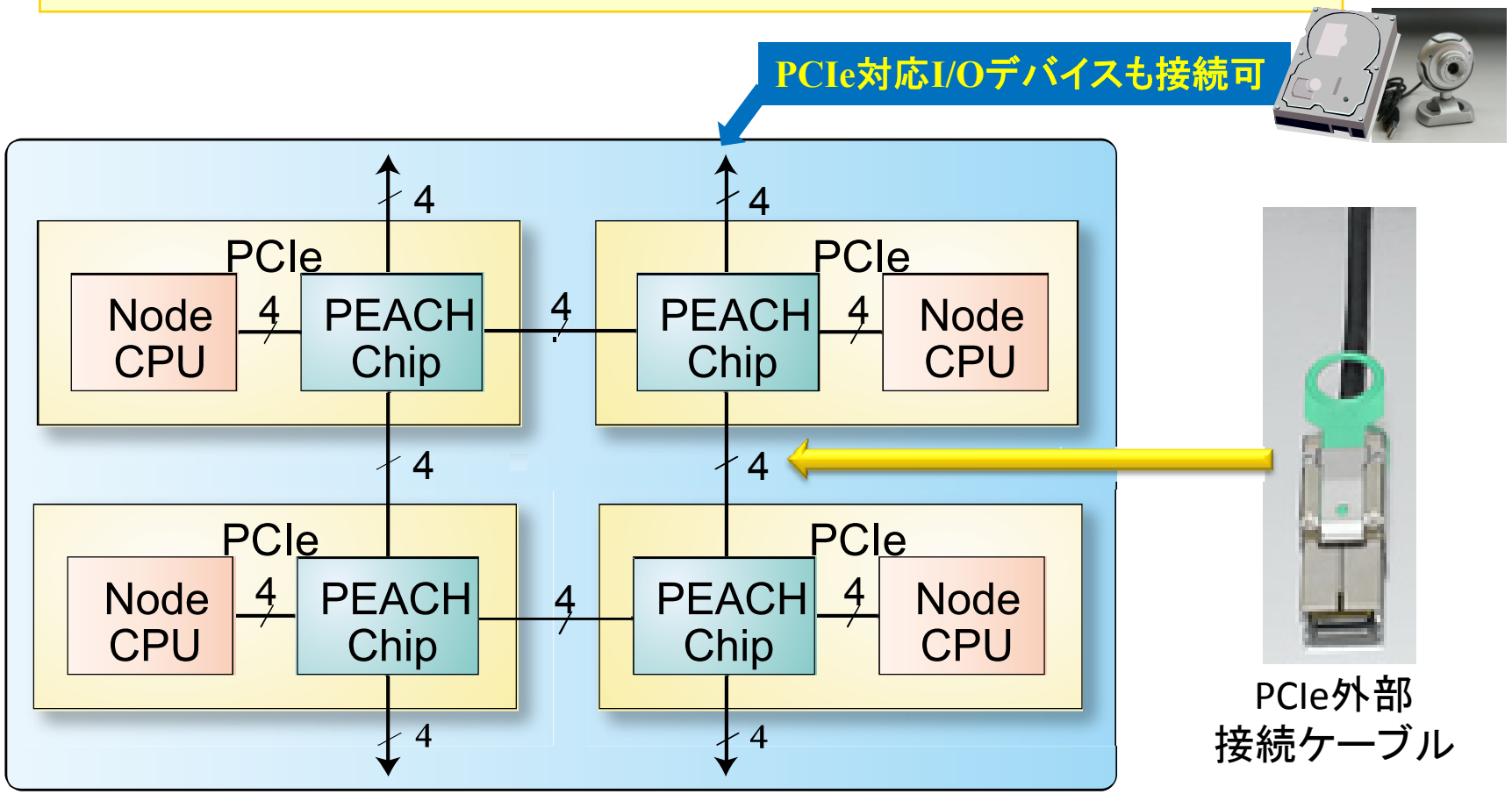


- RI2N (Redundant Interconnection with Inexpensive Network)
  - Ethernetによるマルチリンク技術(統合デモ)
- PEARL (PCI-Express Adaptive and Reliable Link)
  - PCI-Express Gen2を近距離ノード接続に利用、PEACHチップを開発中

オープンシステムディペンダビリティをサポートする場合にも、システム内のハードウエア障害に対処する運用時の信頼性を確保するのは重要

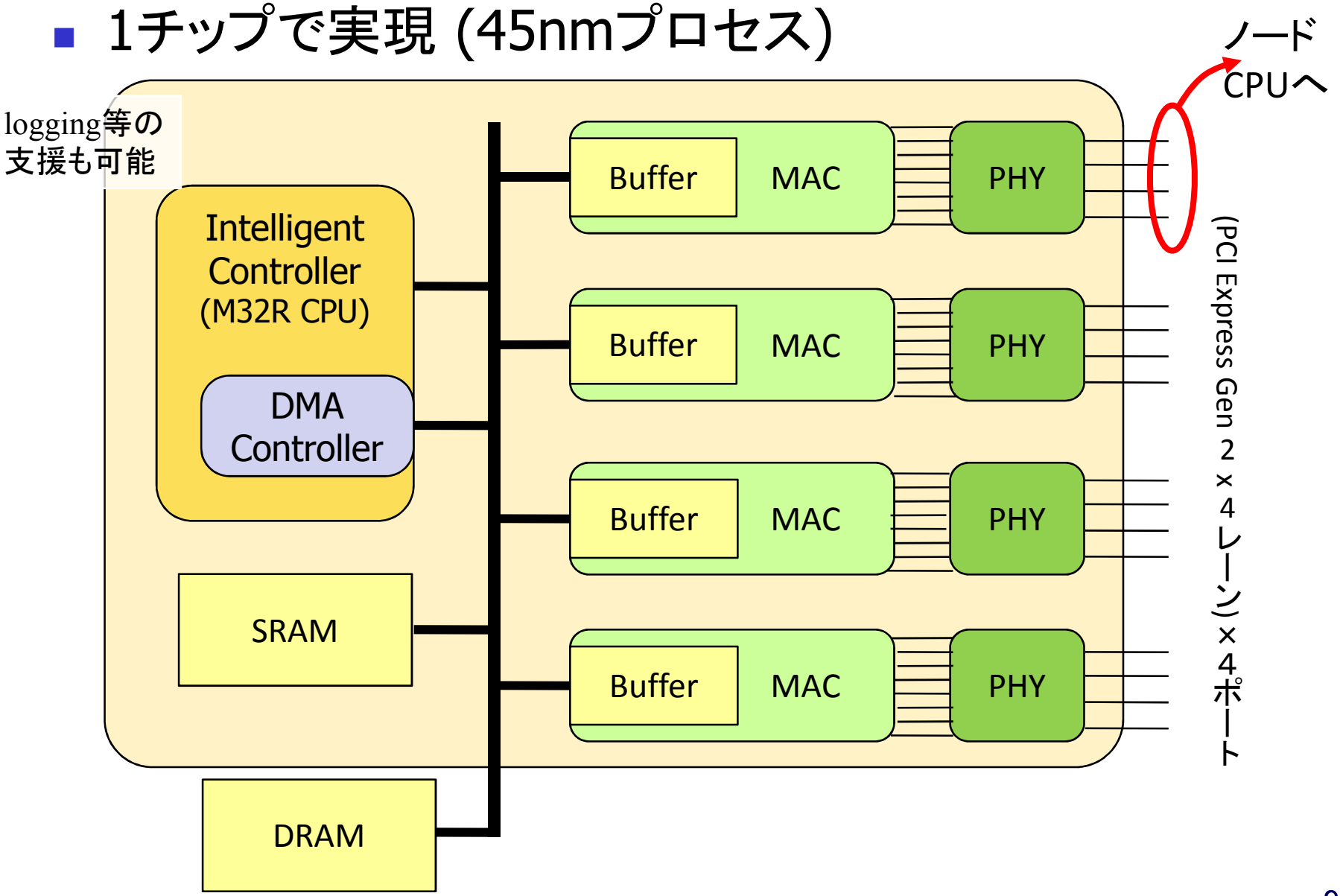
PEARL
(PCI Express Adaptive & Reliable Link)

- ノード間をPCI Expressで直接接続
  - PEACH (PCI Express Adaptive Communication Hub) チップ
  - PCI外部接続ケーブルを用いてノード間を接続（数mまで）
  - 入出力デバイスも接続可能



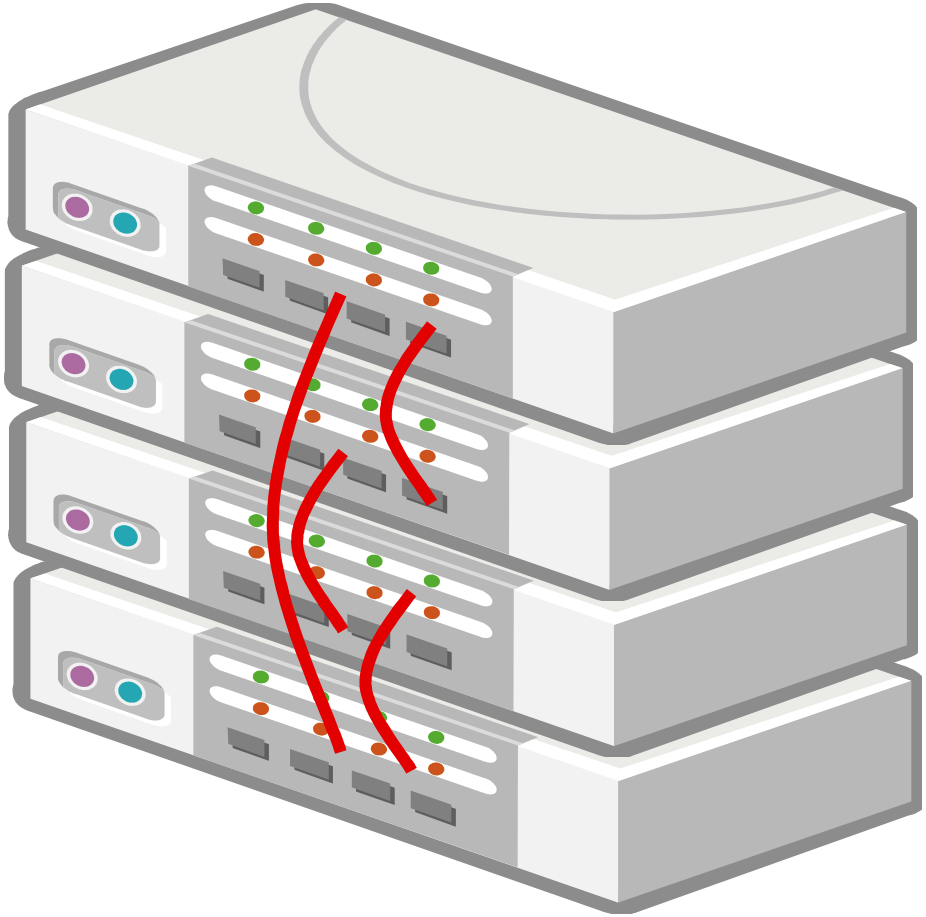
PEACHチップの概略ブロック図

- 1チップで実現 (45nmプロセス)



利用イメージ

- PCI Expressに対応したネットワークボードとして実現
  - 一般的に用いられているIntel x86アーキテクチャの汎用PCに対しても利用可能
- 1Uサーバ
  - 近距離を直接結線



PEARL & PEACHの特長

- 高性能
  - 最大20Gbps(実効2Gbyte/s)のリンク性能
- 高い信頼性
  - PCI Expressによるパケット到着保証と
    フロー制御
  - 複数レーンの冗長性を活かし
    レーン故障から回復
  - 制御プロセッサによる監視，
    リンク故障回避
- 低消費電力・省電力
  - 従来の高速ネットワーク（e.g. Infiniband）
    に比べ低い消費電力
  - 必要な性能に応じてレーンを選択する
    ことで電力削減が可能

レーン速度・本数の選択
(物理層消費電力比)

レーン速度→ ↓本数	Gen1	Gen2
x1	2.5Gbps (21)	5Gbps (28)
x2	5Gbps (38)	10Gbps (50)
x4	10Gbps (75)	20Gbps (100*)

（ ＊現在の試作版では、20Gpbs／ポートで1.7W程度）

既存のネットワークとの比較

	Gigabit Ethernet	Infiniband DDR 4x	PEARL
性能	×	○	○
消費電力	△	×	○
省電力機構	×	×	○
信頼性	×	○	○
スイッチ	必要	必要	不要（但し、近距離）

Infiniband

○高性能　DDR 4x: 20Gbps

○Subnet Managerによる自動故障検出・回復

×　消費電力が大きい

- DDR 4x (20Gbps) x2ポート　約12W
- 24ポートDDRスイッチチップ … 34W

Ethernet

○（比較的）長距離

×　到着保証がない
→上位プロトコルに依存

△Gigabit Ethernet (1Gbps) 1ポート 1～1.5W

PEARL

- 低消費電力
  - 20Gbps x 3ポートで5W程度
- 短距離伝送でスイッチ不要

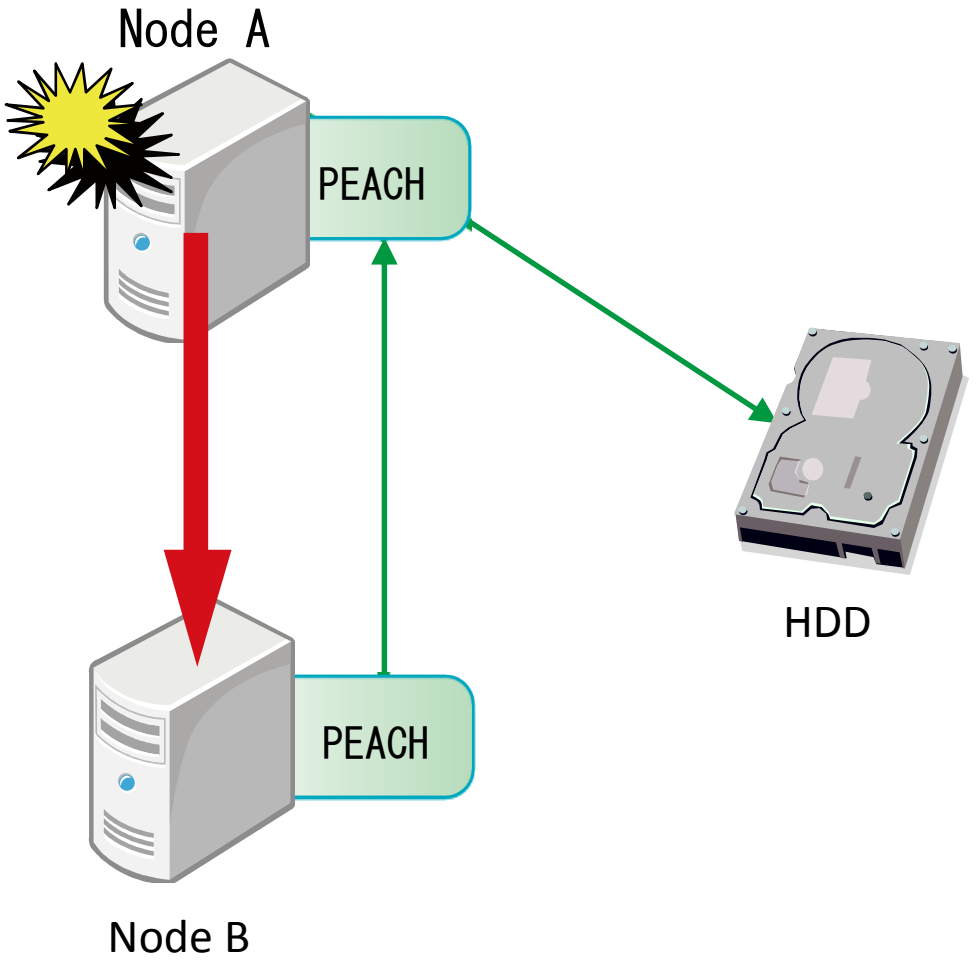
PEARLを使うと...

- 近距離の通信を高速かつ低消費電力で実現
- 性能に余裕がある場合には、動作モードを変更して電力を削減可能
- ノードとPCI Express デバイスとの間を接続し、ネットワーク経由でデバイスをアクセス可能
- リンクに故障が起こっても、別のリンクを使って通信を継続
- ノードが故障しても、接続されたデバイスのフェイルオーバが可能
- 制限
  - ノード数は数十程度、各リンク距離は数m

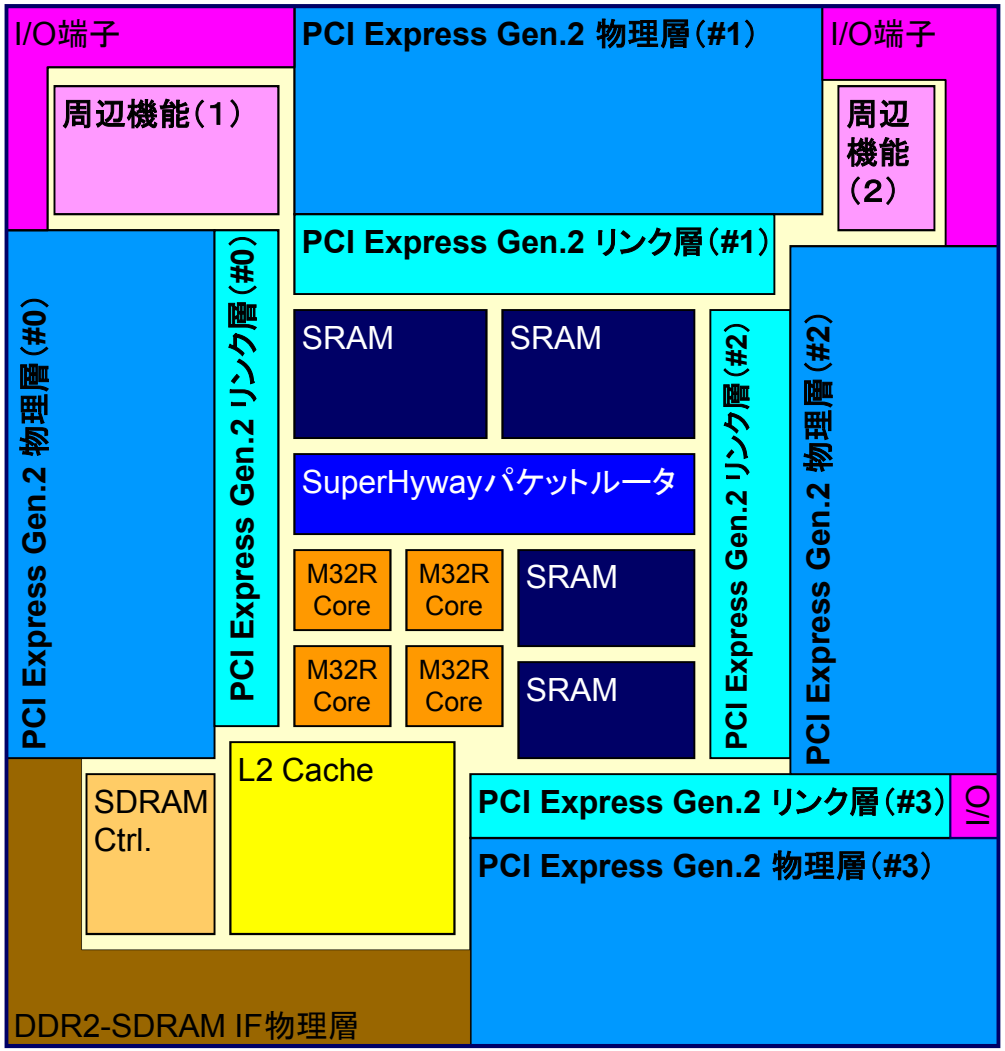
PEARL応用例：ノード故障とフェイルオーバ

- PEARLは、もともとI/Oリンクであるため、PCIeのインタフェースを持つ入出力機器が直接接続可能

- ノードが故障した際にデバイスを含めたフェイルオーバが可能
  - PEACHチップが「生きて」いれば他のノードがコントロールのtake-overできる。



PEACHチップ（開発中）



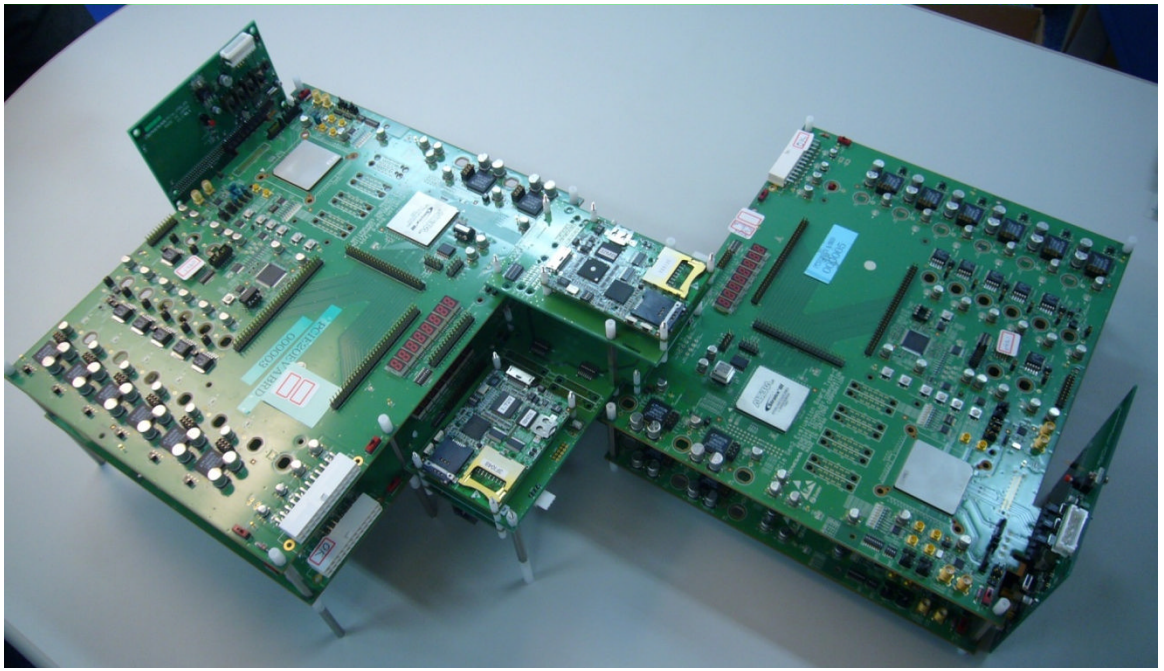
- プロセス：45nm LowPower, triple-Vth, 8-Layer Metal
- パッケージ：FCBGA-1296*pin, 37.5*mm x 37.5*mm
- チップサイズ：12*mm x 12*mm
- CPU：M32R（4CPU, SMP対応）, L1/L2-cache容量（8*+8*/512*Kbyte）
- PCI Express IF：Revision 2.0（5/2.5GHz）, 4レーン/チャネル, 4チャネル
- オンチップSRAM：256*KByte

（*は仮仕様、試作バージョン）

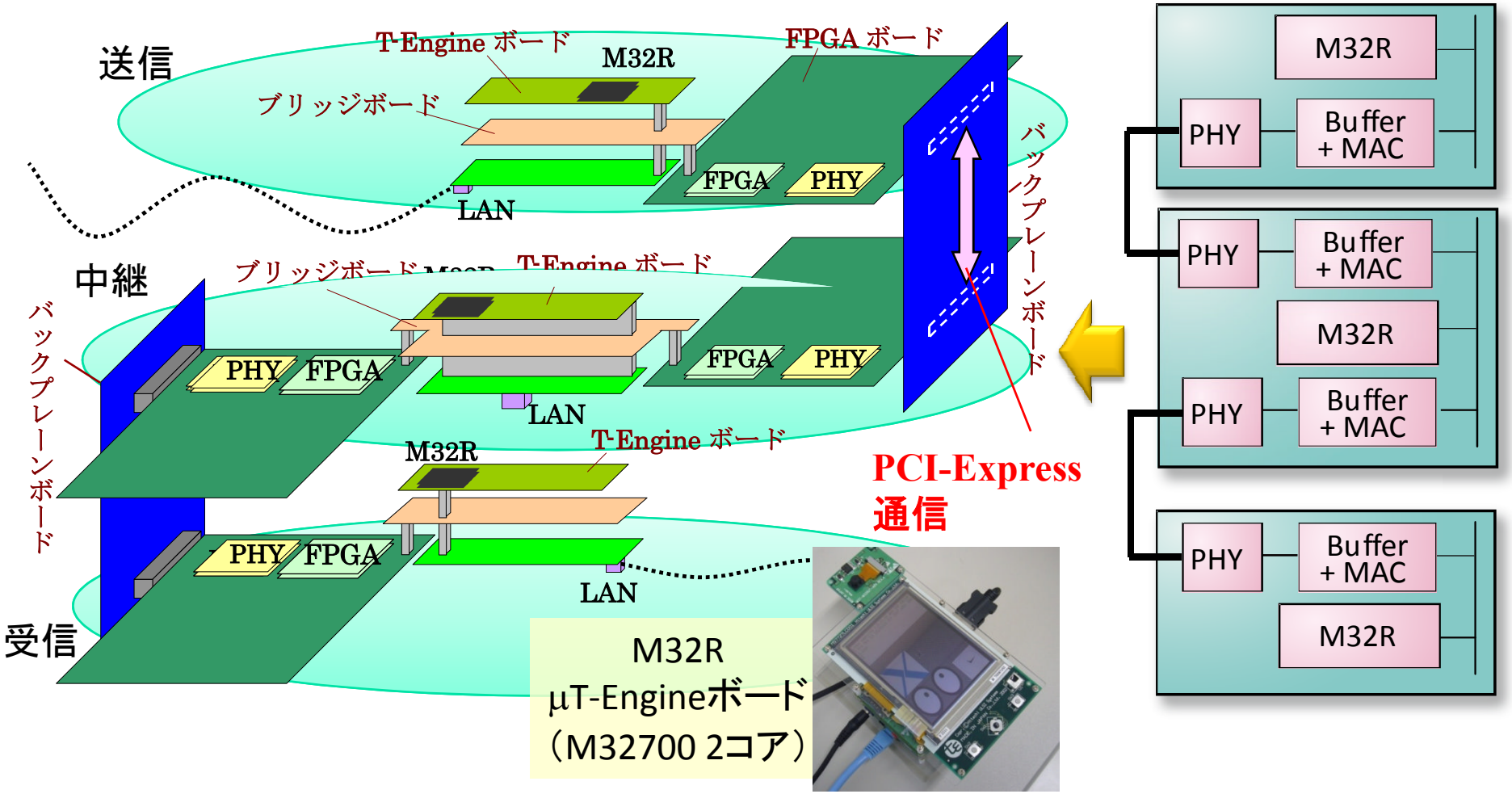
実際に、さらに最適化することで、面積、消費電力ともに改善可能

PEACH展示・デモ

- PEACH チップ FPGA評価環境
  - 3ノード（PEACHチップのみ）に相当
  - W 60cm x D 50cm x H 20cm
- 現在、FPGAベースのPEACH開発プラットフォーム上でファームウェア開発

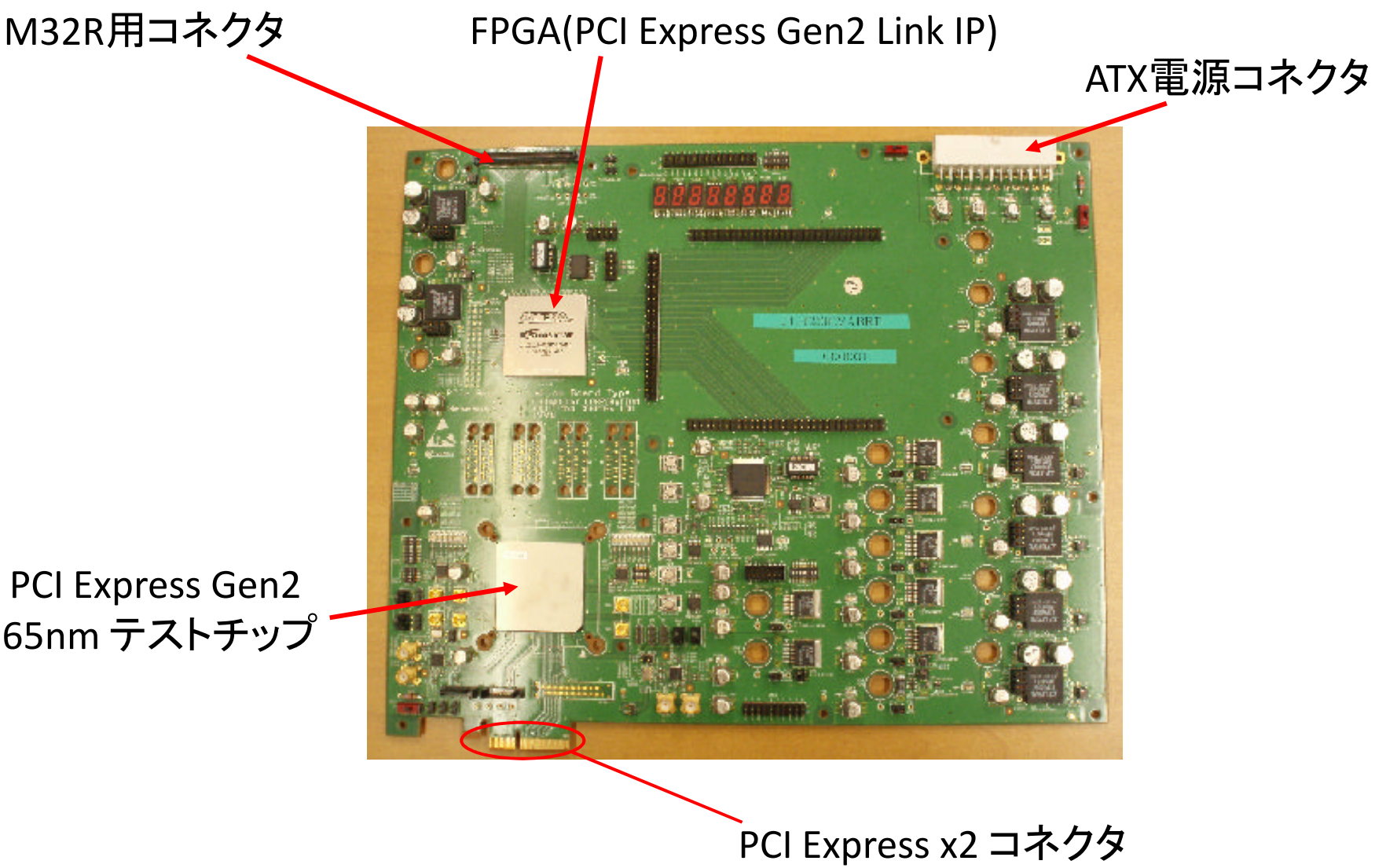


評価・デモ用FPGAボードシステム(全体構成図)



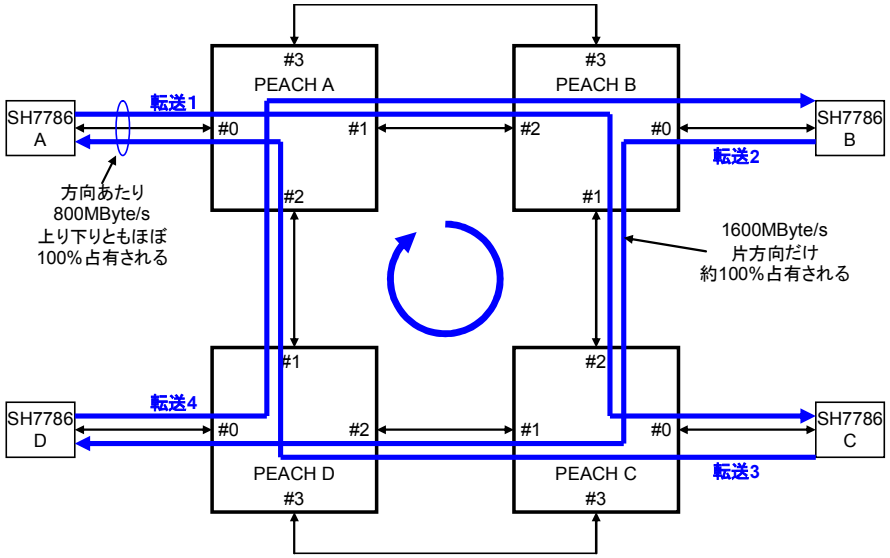
M32R
μT-Engineボード
（M32700 2コア）

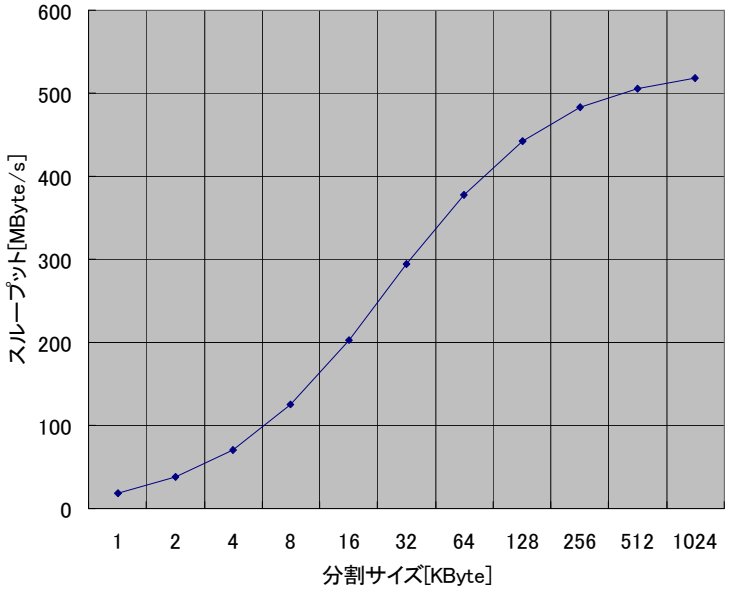
PCI Express 評価用FPGAボード

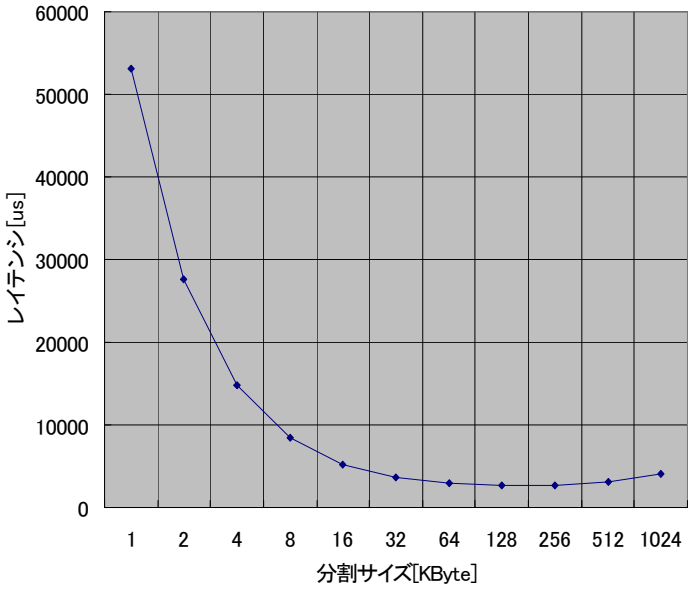


初期性能評価

- VerilogをRTL（Register Transfer Level）シミュレーションを中心に評価を実施
  - 1Mbyteのデータを、左上のSH7786から右下のSH7786まで、3個のPEACHを経由してDMA（Direct Memory Access）転送したときの転送性能評価結果







PEARLのための通信ソフトウエア

- MCAPI (Multicore Communications API) を用いて記述することで、PEARLの性能を最大限に活かしたRemote DMA 通信を提供
  - MCAPIが組み込み並列システムの標準的な通信ライブラリになることを期待

- Socket ライブラリによるTCP/IP 通信も提供
  - TCP/IP over PEARL

現状 と 今後の計画

現状:

- PCI Expressリンクを使った1対1通信が可能（デモ）
  - Linux/M32Rのデバイスドライバを実装，様々な種類の送信，受信のテスト中
  - ノードCPU(SH, x86)用通信API(MCAPI, Socket)サポートの実装中

今後の計画:

- 現在のFPGAベースのPEACH開発プラットフォーム上でファームウェア開発
- PEACHチップは45nmプロセスでの設計を進めており、2009年12月テープアウト、2010年3月 ワーキングサンプル完成の予定
- PEACHチップを搭載したPEARLネットワークボードについても、基板設計を行い、PEACHチップのワーキングサンプル完成後に、2010年4月に完成する予定
- PCI Expressに対応したネットワークボードとして実現
  - 一般的に用いられているIntel x86アーキテクチャの汎用PCに対しても利用可能
- 次年度以降、PEACH実チップとこれを搭載したPCI-Expressネットワークボード上で動作させて、PEARLの持つ様々な機能・性能を検証
  - レーン動作モード切り替え
  - リンク故障に基づく迂回ルーティング
  - （ノード以外の）PCIeデバイスのハンドリング
- 「ディペンダブル並列システム」の構築

重要成果リスト

- 並列システム内高信頼高性能通信機構 (RI2N&PEARL) 関連
  - 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.
  - 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.
  - 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", 情報処理学会論文誌(ACS), Vol.1, No.1, pp.12-27, 2008年6月.
  - 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.
  - 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.
  - Toshihiro Hanawa, Mitsuhisa 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.
- D-cloud関連
  - 神林 亮, 佐藤三久, "仮想マシンを用いた分散システムの耐故障性評価環境の検討", 第70回 情報処理学会、全国大会、3P-1, 2008.(情報処理学会第70回全国大会優秀賞)
  - 神林 亮, 坂西 隆之, 小泉 仁志、佐藤三久、"クラウド環境を用いた大規模テストファームの検討"、DSW99,日本ソフトウエア科学会、2009.
  - 坂西隆之, 小泉仁志, 神林亮, 佐藤三久," プログラムテスト環境を提供するクラウドコンピューティングシステムの検討", SWoPP2009, 2009-OS-112, 仙台、8月4日、2009.
- 電力制御関連
  - 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, 2008.
  - Hideaki Kimura, Mitsuhisa 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), Tsukuba, Japan, 30 Sep. 2008.
  - 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 (ACS 26), 2009年.
- 組み込みOpenMP関連・高信頼ソフトウエア分散共有メモリSCASH-FT
  - Toshihiro Hanawa, Mitsuhisa 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.
  - 李 珍泌, 木村 英明, 佐藤 三久:メモリ効率を考慮した組み込み向け高信頼ソフトウェア分散共有メモリの検討, 情報処理学会研究報告, 2007-HPC-112, pp. 13--18, 3月 (2007)

ポスター・デモ発表の紹介・お誘い

- RI2N
  (冗長Ethernetによる高性能・高信頼通信機構)
  - 統合デモ：RI2N P-Component for P-BUS

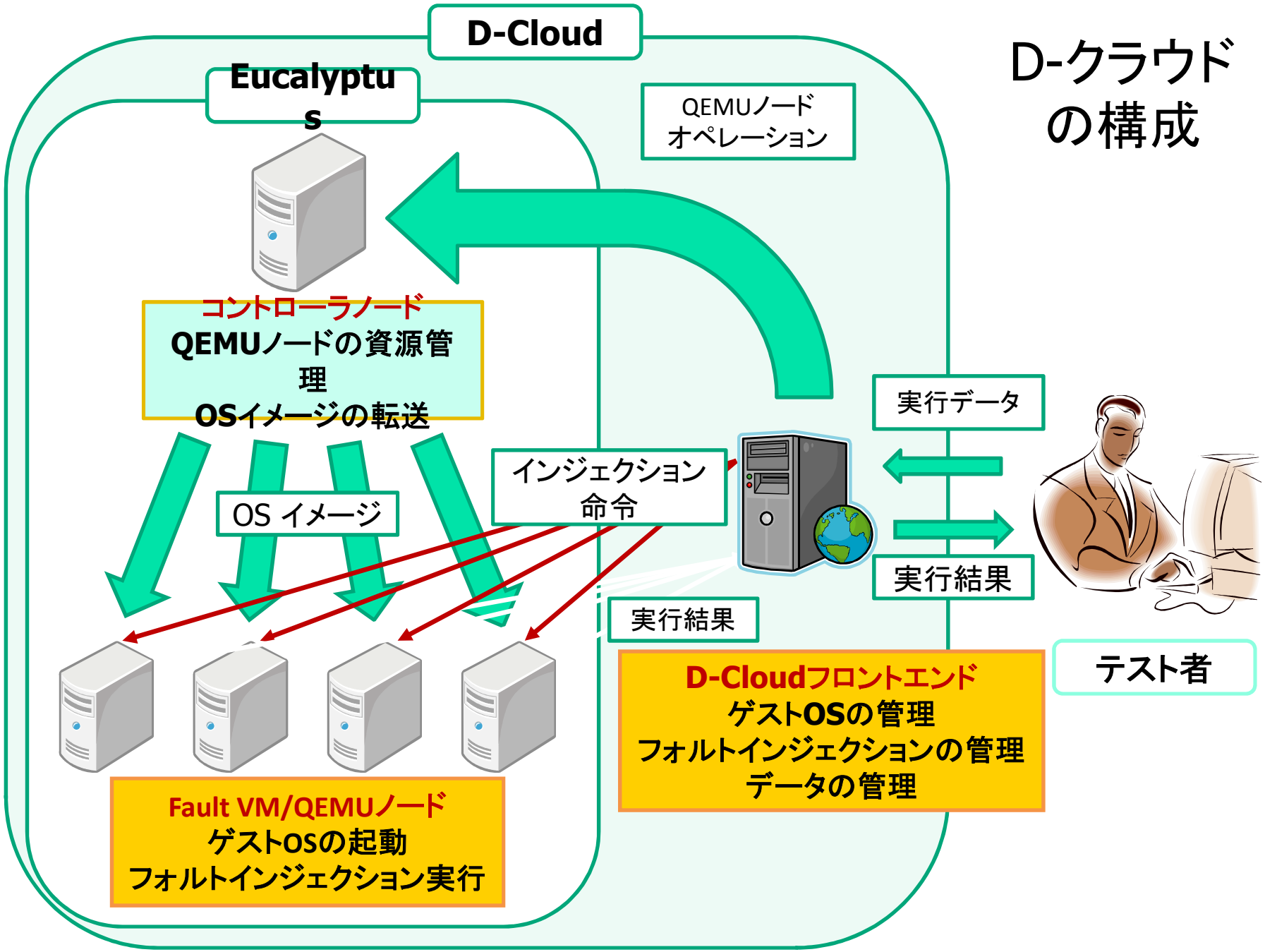
- 高性能・省電力でディペンダブルな通信リンク
  PEARL(PCI Express Adaptive & Reliable Link)
  - PEACH チップ FPGA評価環境

- フォルトインジェクション・テストを加速する環境
  D-クラウド
  - クラウド技術を使った、OS・並列システムのためのテスト環境

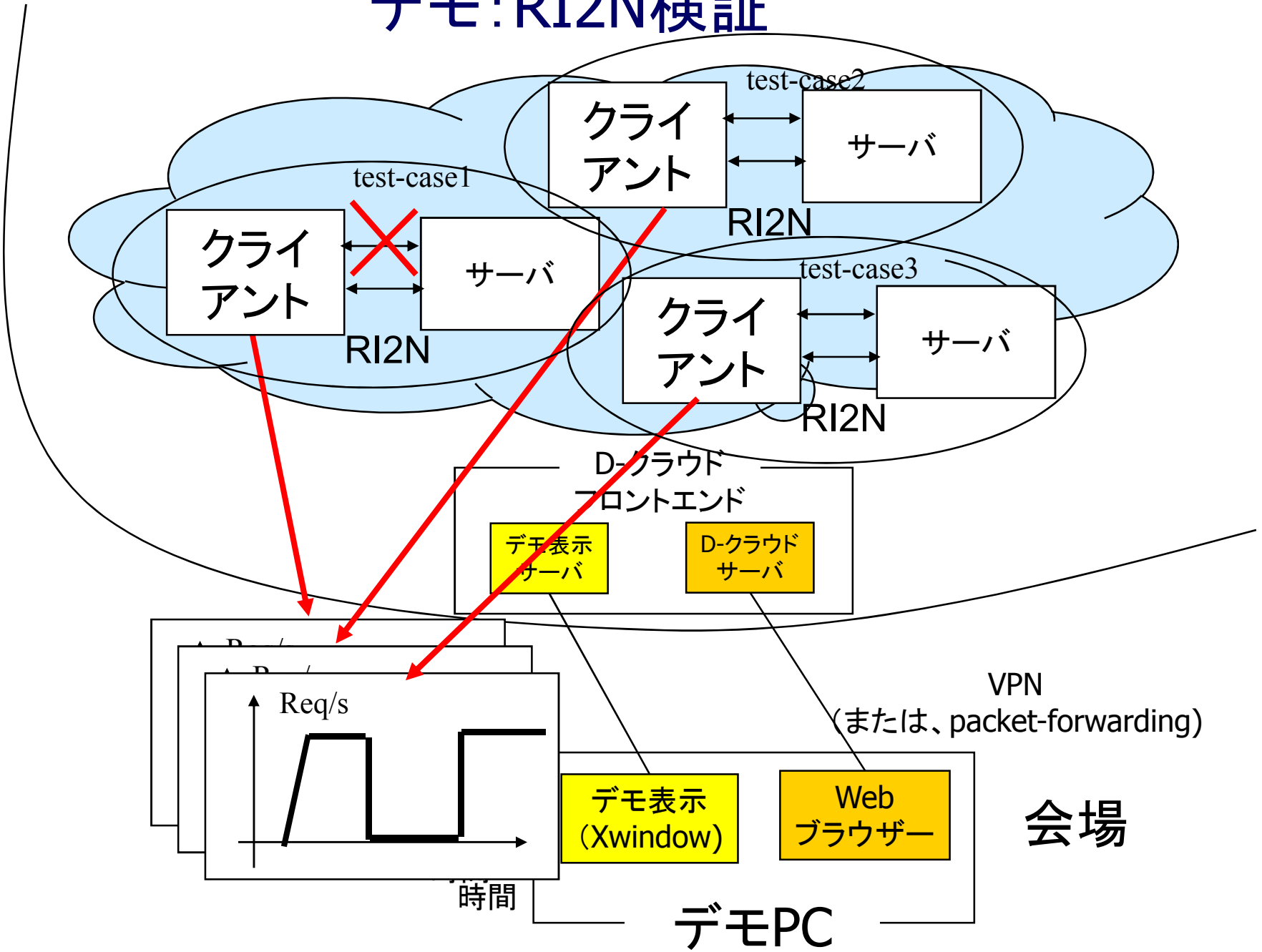
D-クラウドをつかうと…

- クラウドによって提供される計算資源にシ多数のテスト項目を同時に実行し，テストを加速することが可能
- ハードウェア故障を自在に何度でもエミュレーションができる。
- テスト対象の並列・分散システムをクラウド上に構築してタイミングバグの発見および再現を支援
  - システム設定やテストシナリオ記述により、複雑なテストの自動化を支援

D-クラウドの構成


D-Cloud
Eucalyptus
QEMUノード
オペレーション
コントローラノード
QEMUノードの資源管理
OSイメージの転送
OS イメージ
インジェクション
命令
実行データ
実行結果
実行結果
テスト者
D-Cloudフロントエンド
ゲストOSの管理
フォルトインジェクションの管理
データの管理
Fault VM/QEMUノード
ゲストOSの起動
フォルトインジェクション実行

デモ：RI2N検証


test-case2
クライアント
サーバ
RI2N
test-case1
クライアント
サーバ
RI2N
test-case3
クライアント
サーバ
RI2N
D-クラウド
フロントエンド
デモ表示サーバ
D-クラウドサーバ
Req/s
時間
VPN
（または、packet-forwarding）
デモ表示
（Xwindow）
Web
ブラウザー
会場
デモPC