PC クラスタワークショップ in 九州
2007年9月13－14日

PSI(ペタスケール・システムインターコネクト)プロジェクト*の紹介

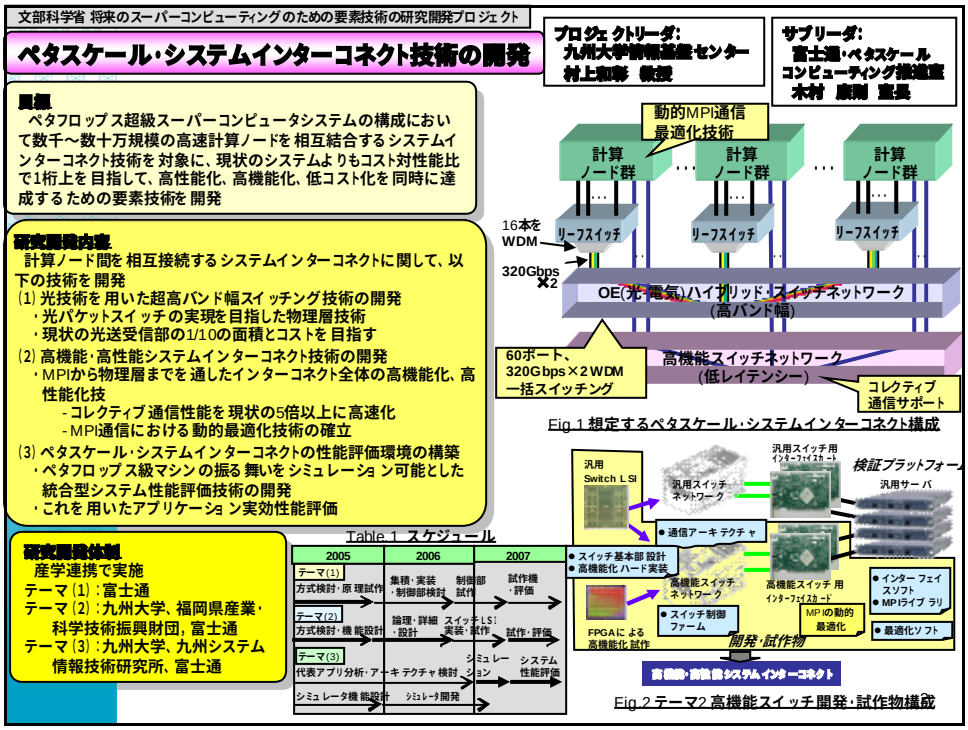
*PSI プロジェクト

リーダー：村上和彰 (九州大学)

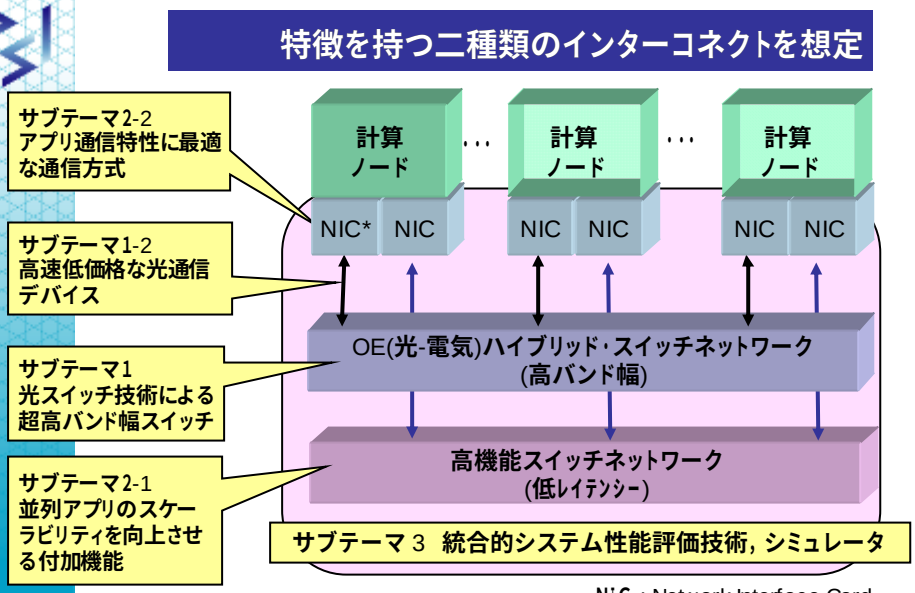
サブリーダー：木村康則 (富士通)

九州大学 システム情報科学府
情報基盤研究開発センター
青柳 睦

PETASCALE SYSTEM INTERCONNECT PROJECT

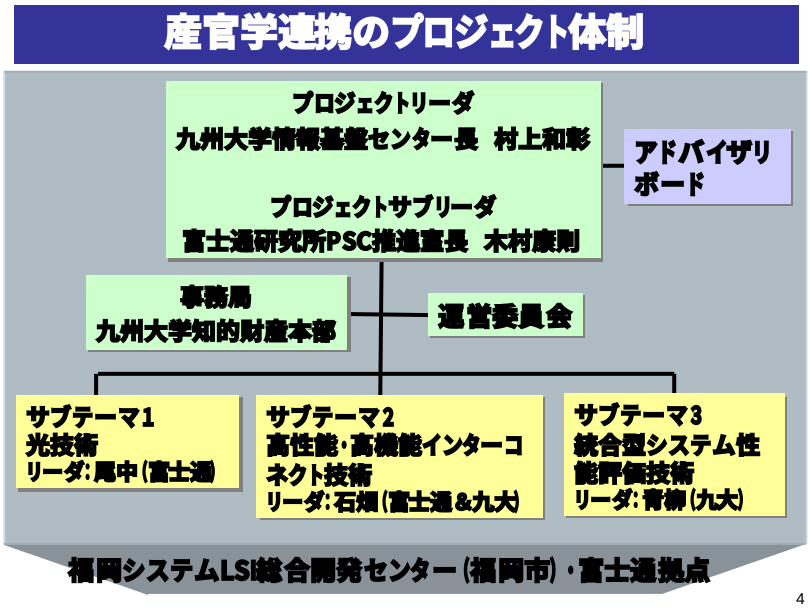


各サブテーマの位置付け



3

PSIプロジェクトの研究体制



4

プロジェクトのゴール

コスト対性能比で現状のシステムインターコネクト技術（たとえば、地球シミュレータ、BlueGene/L）よりも1桁上を目指す

- **サブテーマ1**
  - 32～64ポート程度の光パケットスイッチの実現を視野に、10ポート程度のスイッチ試作
  - 現状のXFPの1/10以下の占有面積とコスト
- **サブテーマ2**
  - コレクティブ通信性能を現状の5倍以上に高速化
  - MPI通信における動的最適化技術の確立
- **サブテーマ3**
  - 「メモリおよび通信性能」対「計算性能」比に優れたペタスケール・アーキテクチャの確立
  - テラスケールシステムでペタスケールシステムの性能を予測可能に

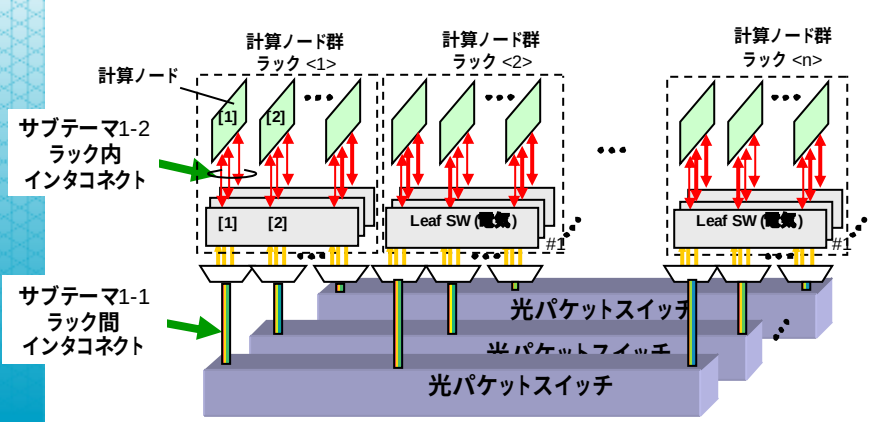
5

サブテーマ1の研究課題

光技術を用いた超高バンド幅スイッチング技術の開発

1-1 :光パケットスイッチング技術の開発

1-2 :光電気変換部集積化技術の開発

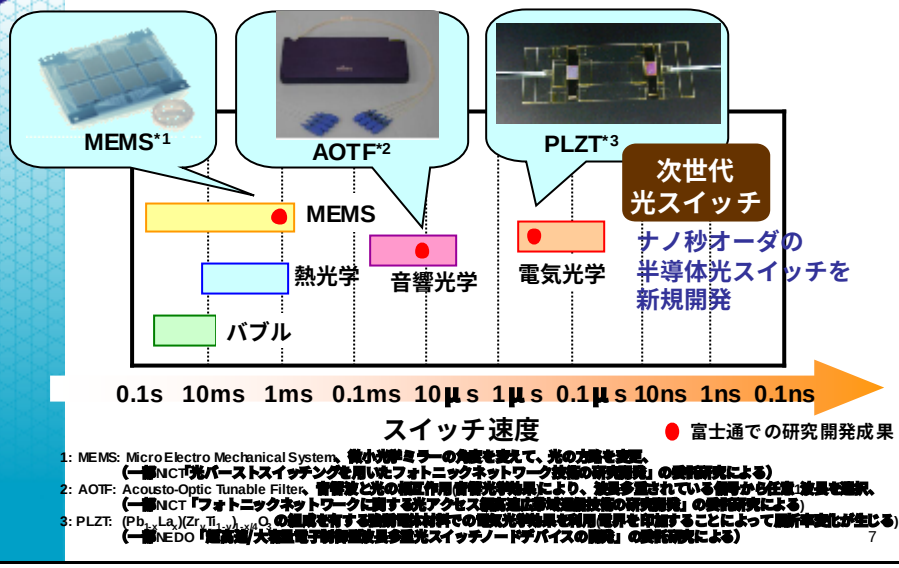


光電気ハイブリッドスイッチ構成

6

サブテーマ1-1 光パケットスイッチ技術の開発(1)

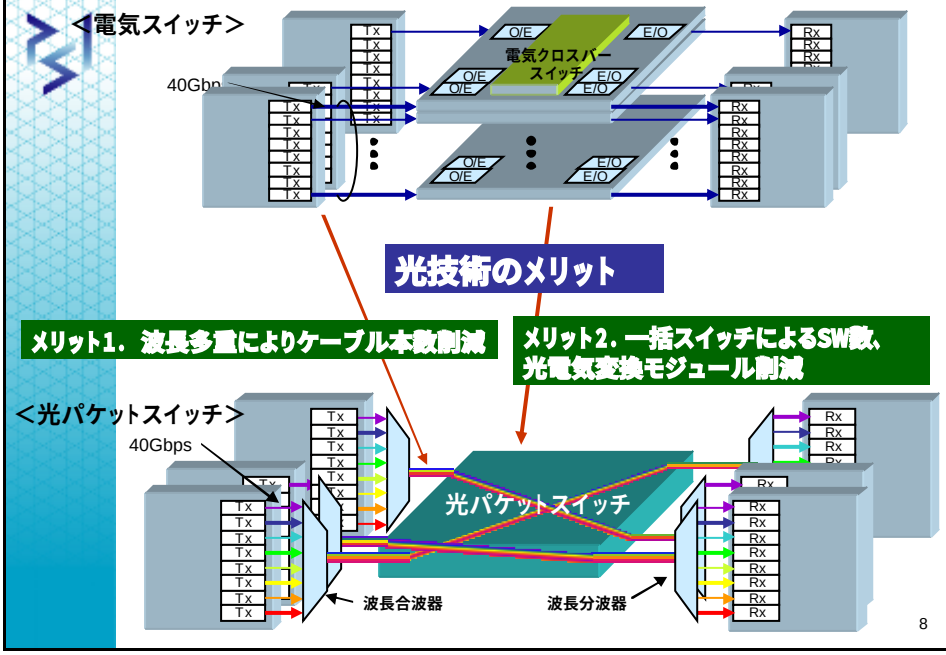
次世代の超高速光スイッチングデバイスを活用



1: MEMS: Micro Electro Mechanical System、微小光学ミラーの角度を変えて、光の方路を変更。
(一部NICT「光バーストスイッチングを用いたフォトニックネットワーク技術の研究開発」の委託研究による)
2: AOTF: Acousto-Optic Tunable Filter、音響波と光の相互作用(音響光学効果)により、波長多重されている信号から任意の波長を選択。
(一部NICT「フォトニックネットワークに関する光アクセス網高速広帯域通信技術の研究開発」の委託研究による)
3: PLZT: $(Pb_{1-x}La_x)(Zr_yTi_{1-y})_{1-x/4}O_3$ の組成を有する透明電気光学材料での電気光学効果を利用(屈折率変化)することによって屈折率変化が生じる
(一部NEDO「超高速/大容量電子制御型波長多重光スイッチノードデバイスの開発」の委託研究による)

7

サブテーマ1-1 光パケットスイッチ技術の開発(2)

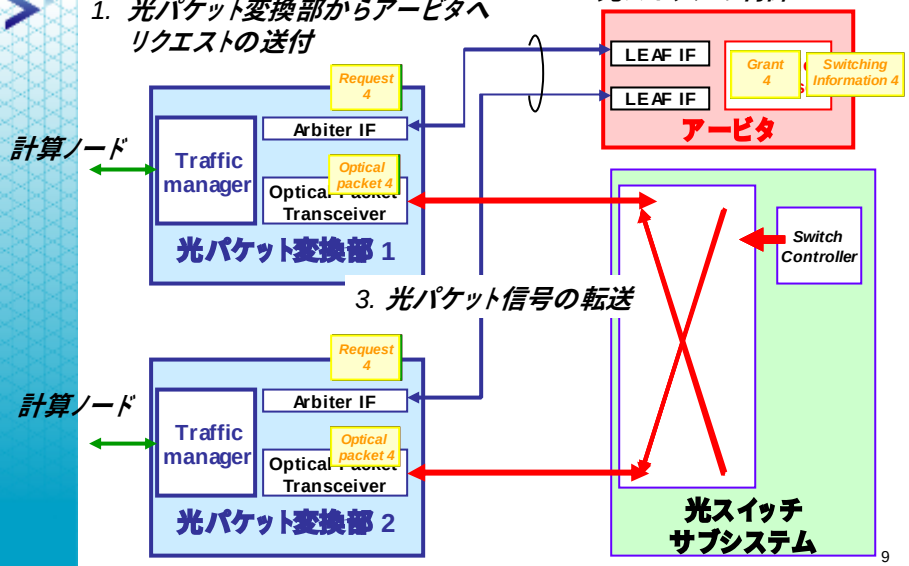


8

光パケットスイッチにおけるデータ転送

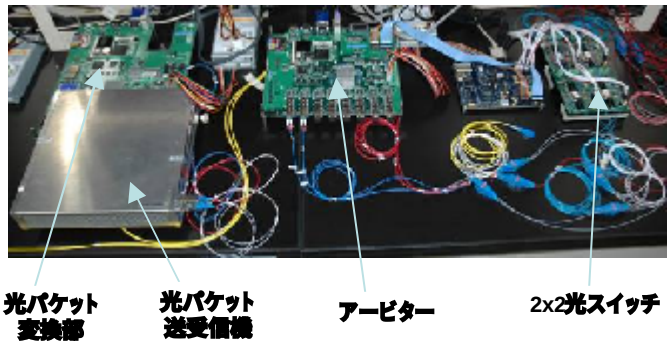
1～3の処理がパイプライン処理

1. *光パケット変換部からアービタへリクエストの送付*
2. *接続調停，送信許可の返信 光スイッチの制御*
3. *光パケット信号の転送*



9

試作システムの構成



1次試作の主要諸元

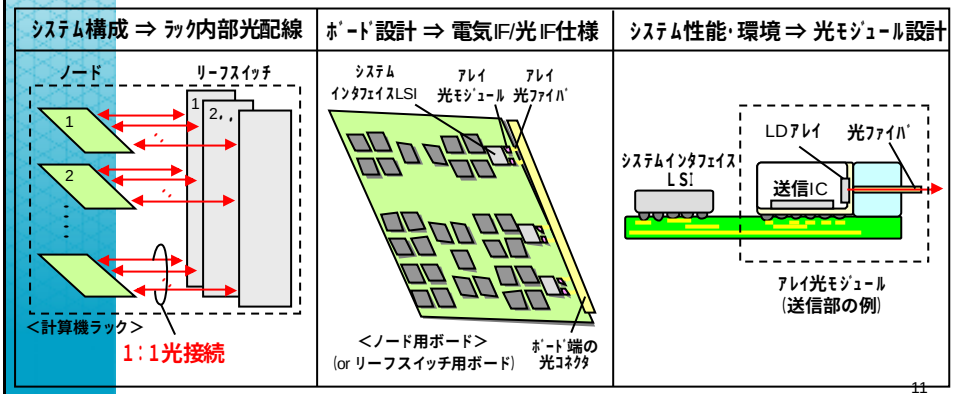
項目	2006年度　1次試作
ポート数	2 ポート
帯域(バンド幅)	10 Gbps x 1波長
スイッチ切替時間	45 ns
光フレーム長	1.2 μs
計算ノードIF	未実装(FPGA内部でデータ生成)

10

ｻﾌﾞﾃｰﾏ1-2 光電気変換部集積化技術の開発

ラック内部のノード～リーフスイッチ間光インタコネクトに適用

- ラック内部の光ファイバ配線：ノード及びリーフスイッチのボード間接続
- ボード上の光モジュール実装： 光コネクタを含むボード内部の実装設計
- 光モジュールの小型･高速化： システムインタフェース回路、動作環境



11

サブテーマ2 概要

サブテーマ2：高機能･高性能システムインターコネクト技術の開発

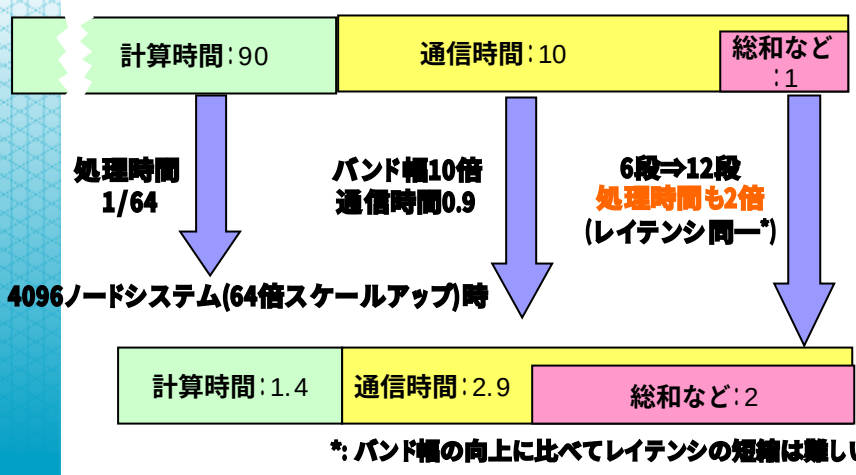
- 2－1：コレクティブ通信をサポートする高機能スイッチの開発
  - コレクティブ通信時間を大幅に削減することを目的に、ノード間の通信経路上で通信データに対して種々の演算を施すことが可能な高機能スイッチ装置の開発を行う。
- 2－2：動的最適化を用いたMPI高速化技術の開発
  - 並列プログラム中のコレクティブ通信の性能は、通信経路や通信パケットサイズ等の通信パラメータに影響される。コンパイル時や実行時に得られるこれらの情報を利用して、適応的に通信パラメータを調整する動的通信最適化技術の開発を行う。

12

コレクティブ通信の高速化

並列度が高いほど高性能な通信が必要

64ノードシステムでの処理時間分布



*: バンド幅の向上に比べてレイテンシの短縮は難しい

出典 ： 第1回 PSIプロジェクト研究者会議資料、石畑、August 4, 2005

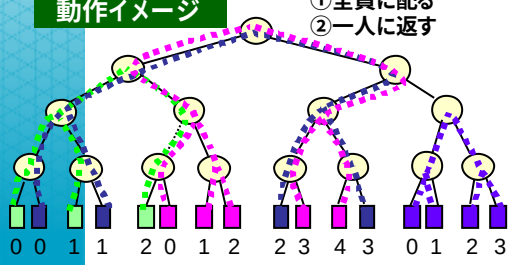
13

サブテーマ 2-1 コレクティブ通信サポート高機能スイッチの開発

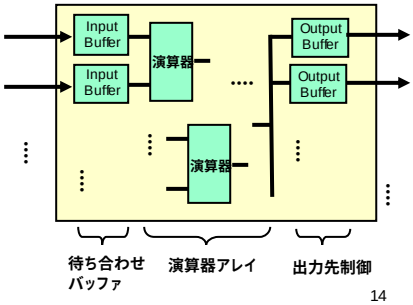
演算が必要な通信を高速化&低レイテンシー化

- 高機能スイッチではコレクティブ通信(AND/OR/MIN/MAX/SUM など)をサポートする。ノードにまたがる演算を高並列で効率よく実行するため、内部に演算機能を持つ
- 通信と演算が必要な操作を一括してネットワークスループットでデータフロー的に実行

高機能スイッチ 動作イメージ



高機能スイッチ構成案

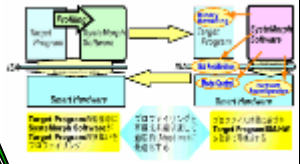


14

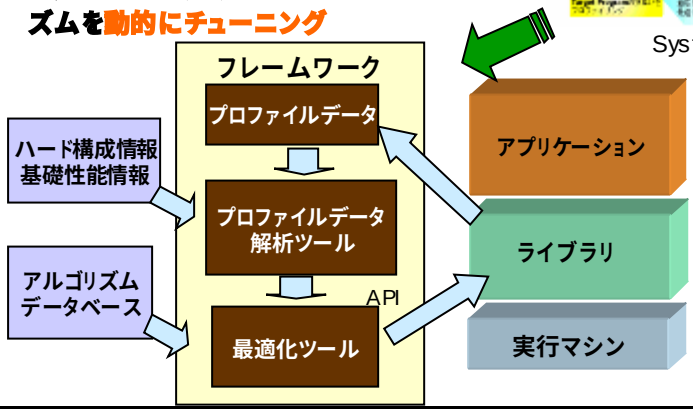
ｻﾌﾞﾃｰﾏ2-2 動的最適化を用いたMPI高速化技術の開発

MPI通信ﾊﾟﾗﾒｰﾀ、通信ｱﾙｺﾞﾘｽﾞﾑの動的最適化による性能向上

- SysteMorph技術の動的最適化フレームワークの考え方をMPI通信の自動チューニングに適用
- 実行中にプロファイルデータ（通信所要時間等）を取得し、解析して、MPI内部のアルゴリズムを動的にチューニング



SysteMorph技術



15

サブテーマ3 概要

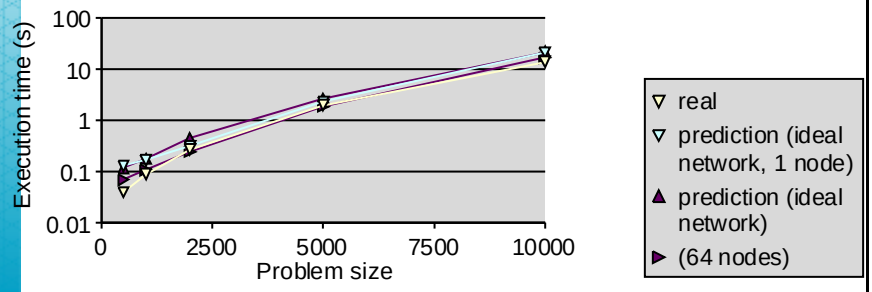
サブテーマ3：ﾍﾟﾀｽｹｰﾙ・ｼｽﾃﾑｲﾝﾀｰｺﾈｸﾄの性能評価環境の構築

- 3ー1：ペタスケール・アーキテクチャの開発
  - 代表的な大規模科学技術計算アプリの計算や通信パターンを分析
  - 主要計算処理部の超並列分散化を検討
  - 主要計算部を効率よく実行できるペタスケール・システム・ノードアーキを検討
  - 計算ノードのシミュレータを開発し、主要計算部をシミュレーションにより評価
  - 性能ボトルネックを解消し、大規模科学技術計算アプリケーションの実行に適したペタスケール・アーキを検討。
- 3ー2：ペタスケール・システムの性能予測技術の開発(PSI-SIM)
  - 小規模システムでのプログラム実行によって得た各種統計情報に基づき、ペタスケール・システムの性能を現実時間内で精度良く見積もる技術を開発
  - 特にシステムインターコネクトに焦点を当て、システム設計時に利用できる性能評価環境を構築する。
  - 実際のアプリを用い、開発した性能予測方式の評価を行う。
  - 最終的には、実在する高性能コンピュータ・システムの性能予測を想定した評価実験を実施する。

16

結果（抜粋）
24GFlop/sマシンで 384GFlop/sマシンの性能を予測

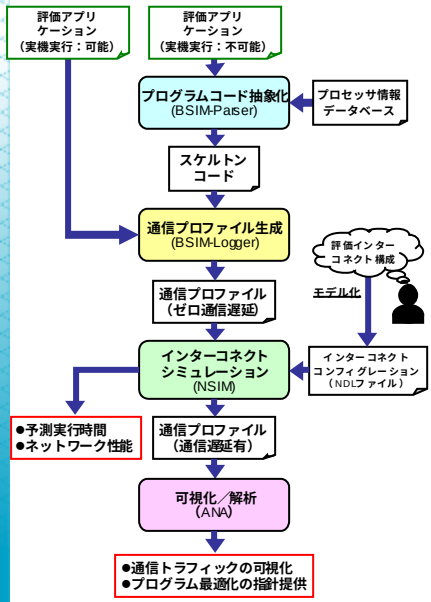
- HPLオリジナルコード（実機実行）
- 「Xeon 4ノード」で「Xeon 64ノード」での実行時間を予測



（注）上図ではネットワーク･シミュレーションの結果は含まない

17

PSI-SIMのワークフロー



1. **BSIM-Parser**
   - 評価アプリケーションのプログラムコード抽象化（通信プロファイルの高速生成を目的）
2. **BSIM-Logger**
   - 通信プロファイルの生成（中規模システムによる大規模システムの通信プロファイル生成を目的）
3. **NSIM**
   - ネットワークシミュレーション（ゼロ通信遅延プロファイルへの実遅延時間付加が目的）
4. **ANA**
   - アプリケーションの可視化／解析（アプリケーションの評価や開発支援が目的）
5. **Open-FMO**
   - 次世代ペタスケールソフトウェアの開発と評価

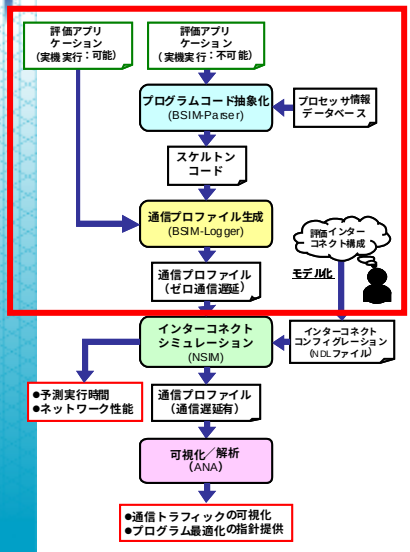
18

BSIM-Parser/Logger

プログラムコード抽象化技術に基づく

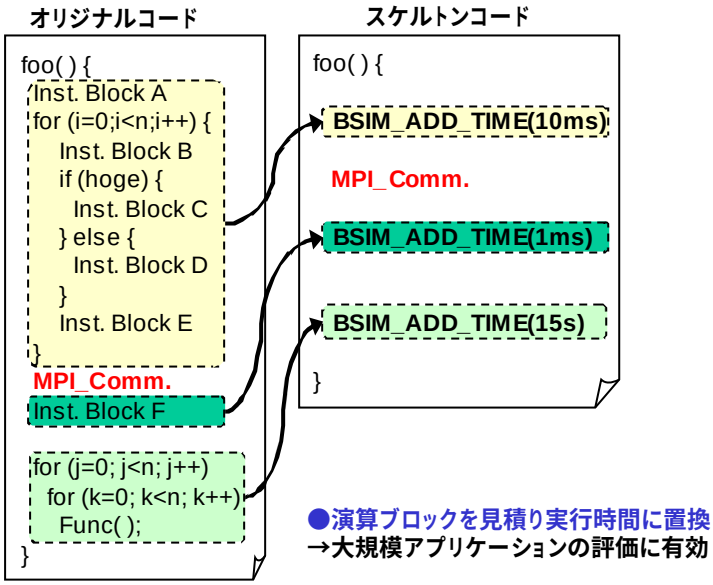
通信プロファイルの高速生成環境

BSIM-Parser/Logger



- **評価用アプリケーションの制御フローを維持したスケルトンコードの生成** (BSIM-Parser)
  - **演算と制御・通信の分離**
  - **命令ブロックから演算ブロックを抽出**
  - **出力コードへの見積り実行時間の埋め込み**
- **通信・計算処理の履歴を含む通信プロファイルの生成** (BSIM-Logger)
  - **実機 (クラスタ計算機) によるスケルトンコード化されたアプリケーションの擬似実行**
  - **通信遅延時間0 (理想的なネットワーク環境) の通信プロファイルを出力**
  - **通信イベントの依存関係を保持**

通信プロファイルの高速生成に向けたプログラムコード抽象化



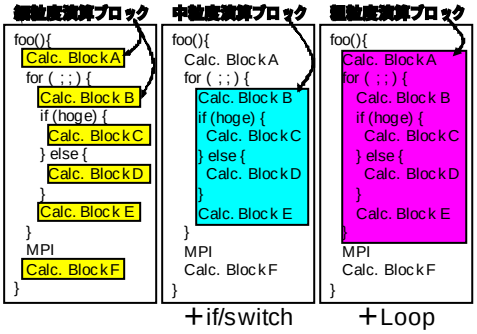
●演算ブロックを見積り実行時間に置換
→大規模アプリケーションの評価に有効

命令ブロックと演算ブロックの定義

- 命令ブロック
  - プログラムコード上に連続して出現する命令列
- 演算ブロック
  - MPI通信を含まない命令ブロック
  - 既存のプロセッサ情報や実測値を活用した実行時間の見積が可能

	細粒度演算ブロック	中粒度演算ブロック	粗粒度演算ブロック
MPI通信	×	×	×
分岐命令（後方向）	×	×	○
分岐命令（前方向）	×	○	○

○：含む ×：含まない



スケルトン・コード化可能部分の抽出

1. 細粒度演算ブロックの抽出
2. 分割統治法 (divide & conquer) による粗粒度化

例）FFTのスケルトン化

```
subroutine fft(dir, x1, x2)
          :
else if (layout_type .eq. layout_2d) then
 call cffts1(-1, dims(1,3), x1, x1, scratch)
 call transpose_x_z(3, 2, x1, x2)
 call cffts1(-1, dims(1,2), x2, x2, scratch)
 call transpose_x_y(2, 1, x2, x1)
 call cffts1(-1, dims(1,1), x1, x2, scratch)
endif
          :
```

```
subroutine transpose_x_y_local(d, xin, xout)
do k = 1, d(3)
      do i = 1, d(1)
        do j = 1, d(2)
!!!CALC      xout(j,k,i)=xin(i,j,k)
        end do
      end do
    end do
return
```

細粒度演算ブロック
（粗粒度化可能）

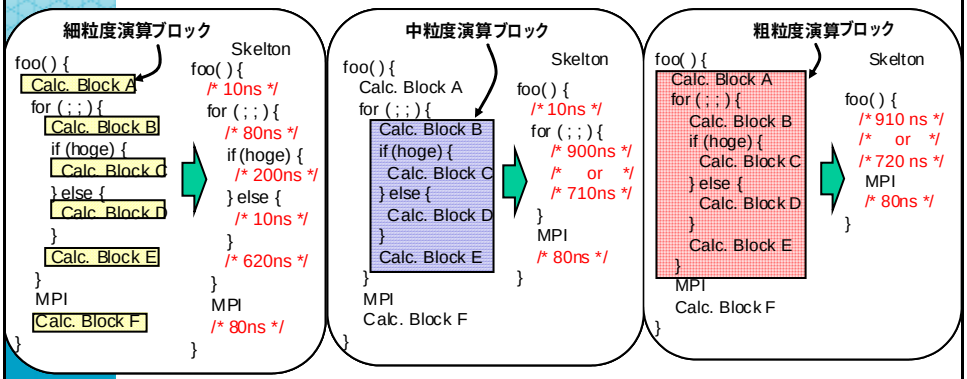
```
subroutine transpose_x_y_global(d, ..xout)
          :
call mpi_alltoall(xin, … commslice2, ierr)
```

粗粒度化不可能

```
subroutine transpose_x_y(l1 ... xin, xout)
          :
call transpose_x_y_local(dims(1,l1),xin, xout)
call transpose_x_y_global(dims(1,l1), xout, xin)
call transpose_x_y_finish(dims(1,l1), xin, xout)
```


スケルトン・コードの生成

- 各演算ブロックの実行時間に相当する見積実行時間を取得・計算
  - 実行命令数に基づく見積り（命令数×CPI×クロックサイクル時間）
    - プロセッサ情報データベースの利用
  - 実機による実時間測定（ハードウェアカウンタやRTCの利用）
  - サイクルレベル・シミュレーション
- 対応する各演算ブロックのコードを見積実行時間に置換



性能評価方法とイベントログの採取

- 性能評価方法
  - 設計空間探索（パラメータ・サーベイ） ⇒ 相対性能
  - 一点詳細評価 ⇒ 絶対性能
- イベントログの採取
  - 採取範囲：全実行 vs. 部分実行 vs. 未（非）実行
  - 実行方式：実実行 vs. 擬似実行（スケルトン・コード実行）

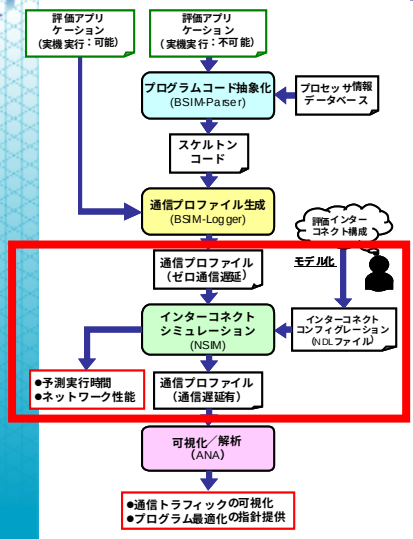
ログ採取法 / 実行方式 / 評価目的	プログラム全実行		プログラム部分実行		未実行
	実実行	擬似実行	実実行	擬似実行	人工的なイベントログ生成
設計空間探索	×	○	○	−	○
一点詳細評価	×	○	−	−	−

○：サポート　×：サポート困難　−：未サポート

NSIM

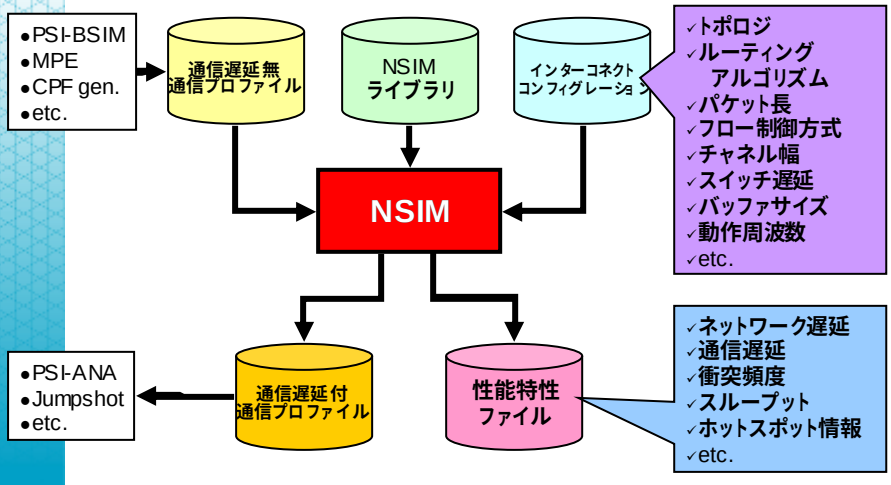
通信プロファイルを入力とする
大規模インターコネクトシミュレータ

NSIM



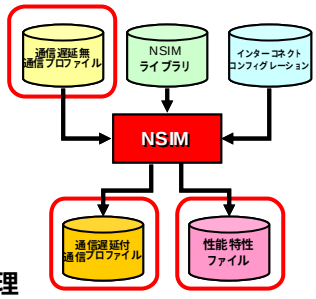
- 通信プロファイルに基づいたインターコネクトシミュレータ
  - 超大規模インターコネクトへの対応
  - 実用時間内におけるシミュレーションの完了を目的
  - 設計開発現場での実用性 (シミュレーション解像度：1ナノ秒～)
  - 評価インターコネクトをコンフィグレーションファイルによってモデル化
  - ゼロ通信遅延時間の通信プロファイルを入力
  - 通信遅延時間を付加した通信プロファイルを出力 → PSI-ANAへ
  - 並列離散事象シミュレーション

NSIMのシミュレーションフロー
(今後の拡張予定を含む)



PSI-NSIMの入出力ファイル

- 通信プロファイル
  - 通信遅延無通信プロファイル（入力）
    - ネットワーク遅延ゼロとした、通信メッセージの送受信内容を含むイベントログファイル（PSI-BSIMから出力ログ）
  - 通信遅延有通信プロファイル（出力）
    - ネットワーク遅延を反映した通信ログ
    - 可視化･解析ツールに入力可能な形式
- 性能特性ファイル
  - 各通信メッセージの送受信時刻、送受信ノード、通信遅延、ホップ数、衝突回数などを出力
  - ネットワーク全体における遅延、通信処理能力、ホットスポットなどの解析に利用

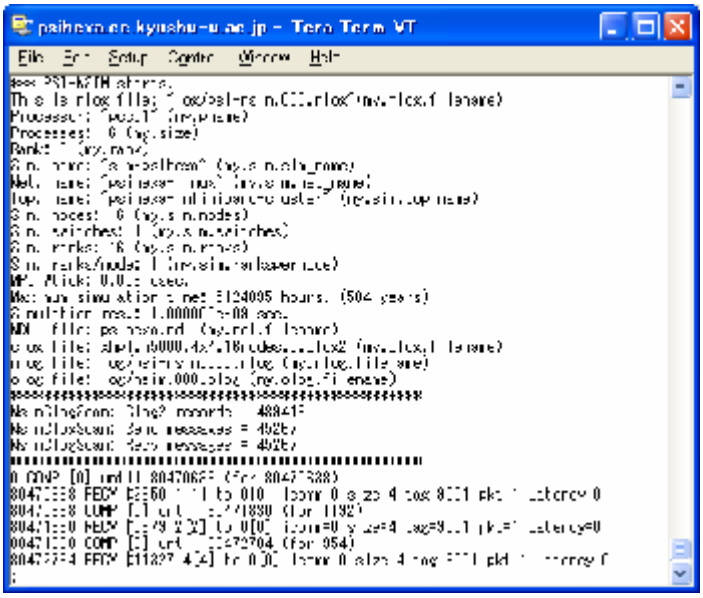


インターコネクト コンフィグレーション（例）

```
; Reference Network Description for PSI-NSIM
;   by shibamura@isit.or.jp

(simulation "sim-psihexa" ; Simulation name
;  (clog_filename "xhpl.n2000.4x4.16nodes.0.clog2")
  (clog_filename "xhpl.n5000.4x4.16nodes.0.clog2")
  (nlog_filename "log/psi-nsim")
  (olog_filename "log/nsim")
  (stdoutput true)
  (debug false)
) ; end simulation

; Node configuration (Intel Xeon 3.0GHz, Single Core)
(network "psihexa-linux"
  (node
    (name "pcc")
    (number node 16)
    (number port 1)
;   (powerconsumption 0.001mW)
  ) ; end node

; InfiniBand switch configuration
  (switch
    (name "ibsw0")
    (number switch 1)
    (number port 16)
    (bandwidth 4Gbps)
    (packet 2048B:size)
    (packet 1024B:payload)
    (latency 17usec:startup)
    (latency 200nsec:pre)
    (latency 10nsec:post)
;   (powerconsumption 0.002mW)
  ) ; end switch

  (topology
    (name "psihexa-infiniband-cluster")
;
; Node-Switch interconnection part
;
    (connect (pcc:0:0 ibsw0:0:0) (pcc:1:0 ibsw0:0:1)
             (pcc:2:0 ibsw0:0:2) (pcc:3:0 ibsw0:0:3))
    (connect (pcc:4:0 ibsw0:0:4) (pcc:5:0 ibsw0:0:5)
             (pcc:6:0 ibsw0:0:6) (pcc:7:0 ibsw0:0:7))
    (connect (pcc:8:0 ibsw0:0:8) (pcc:9:0 ibsw0:0:9)
             (pcc:10:0 ibsw0:0:10) (pcc:11:0 ibsw0:0:11))
    (connect (pcc:12:0 ibsw0:0:12) (pcc:13:0 ibsw0:0:13)
             (pcc:14:0 ibsw0:0:14) (pcc:15:0 ibsw0:0:15))
  ) ; end topology
) ; end network
```

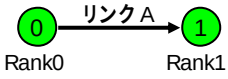
インターコネクトの仕様を容易に変更できるため、スケーラブルかつ柔軟な評価が可能

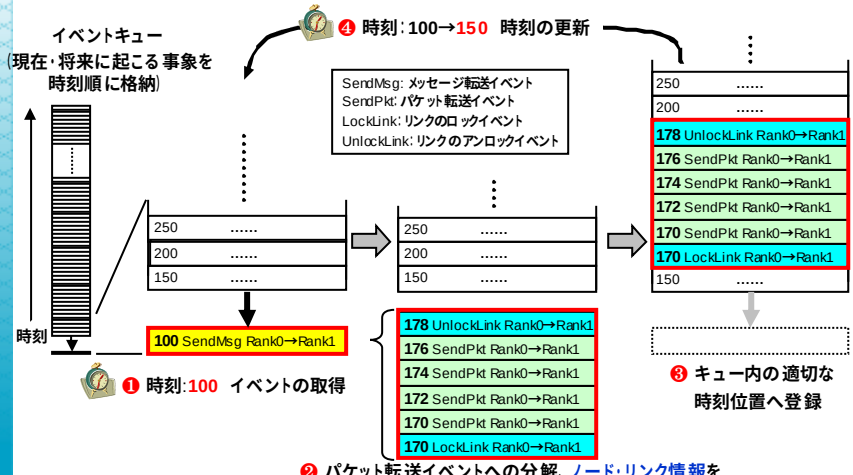
通信プロファイル（出力例）



イベントキューを用いた離散事象シミュレーション

例）メッセージの送信
・ノード・リンク情報
リンクAの開放予定時刻：170、通信遅延時間：2





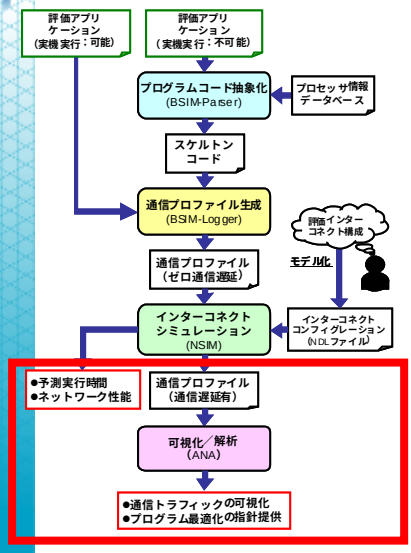
NSIMの拡張可能な機能

- **対故障性の評価**
  - 動的なリンクの切断指示を設定可能
    - ➪リンク故障などによる性能低下の見積り
    - ➪適応ルーティング手法の評価
- **消費電力の概算評価**
  - リンク、スイッチ、ノードに関する電力パラメータを設定可能
    - ➪ネットワークやシステム全体における電力消費量の見積り
- 開発時に基本構造を実装済み➪必要に応じて利用可能

ANA

次世代の超大規模アプリケーションに向けた
可視化／解析ツール

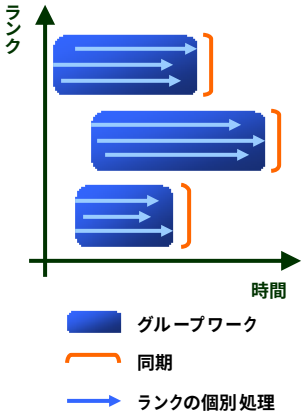
ANA



- グループワークと呼ぶ新しいプログラミング単位に基づいた解析・可視化機能を提供
- ペタスケールシステムでの実行を前提としたアプリケーションのチューニング支援機能を提供
- 高機能エンジン
  - 可視化エンジン (ANA-Viewer)
    - プログラマのためのチューニング支援
  - 検索エンジン (ANA-Search)
    - 可視化ツールと連携した類似性検索

新しいプログラミング単位の提案

超並列時代には、個々のMPIランクのロードバランスと大きな演算単位を基にした全体のフローの把握が必要



グループワーク

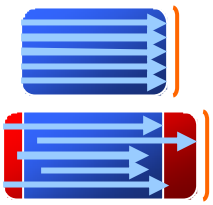
◆複数のMPIランクによって構成される、同様の処理群

◆グループワークに属したすべてのMPIランクにおいて、個別処理（ワーカー）の実行結果をまとめ、一つの実行結果とする

◆アプリケーション開発者は、ソースコード中にグループワークを明示的に記述する

グループワークを用いた解析

グループワーク内の処理効率の把握



効率の良いグループワーク

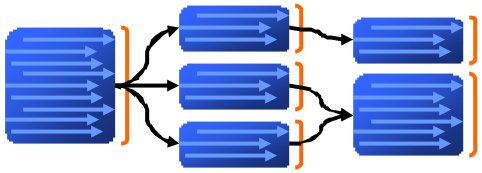
- すべてのワーカーが、同時に開始、同時に終了
- 待ち時間がない

効率の悪いグループワーク

- ワーカーの開始、終了にばらつきがあり
- いくつかのワーカーに待ち時間が発生

グループワーク単位の処理効率の把握

並列アプリケーションをグループワークの連鎖とみなす



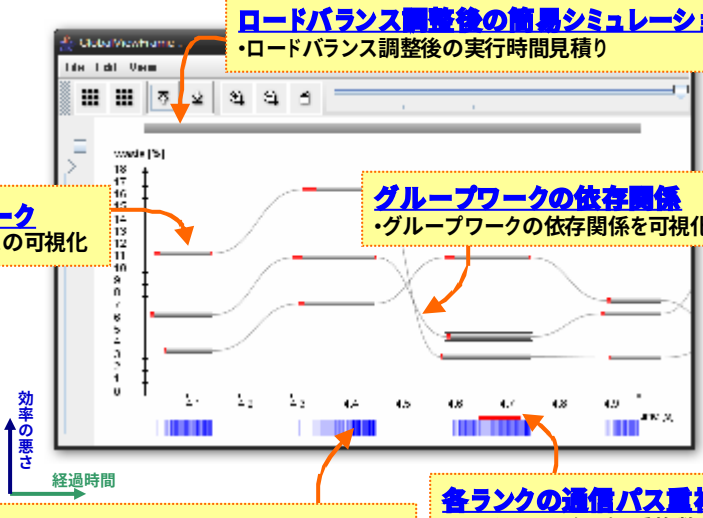
- グループワーク単位のロードバランス解析が容易
- 優先して改良すべきグループワークを把握できる
- クリティカルパスの早期発見に役立つ

ANA GroupWork Viewer

ロードバランス調整後の簡易シミュレーション
・ロードバランス調整後の実行時間見積り

グループワーク
・ロードバランスの可視化

グループワークの依存関係
・グループワークの依存関係を可視化



通信量
・単位時間当たりの通信量を可視化
・通信混雑が発生する時間帯を検索エンジンにて高速サーチ

各ランクの通信パス重複数
・ランク間の通信パス重複数を可視化し、ホットスポット発生の可能性を示唆

PSI-SIMの性能評価

評価実験 (BSIM-Logger + NSIM)

– 実験：

評価アプリケーションの実行時間の予測性能

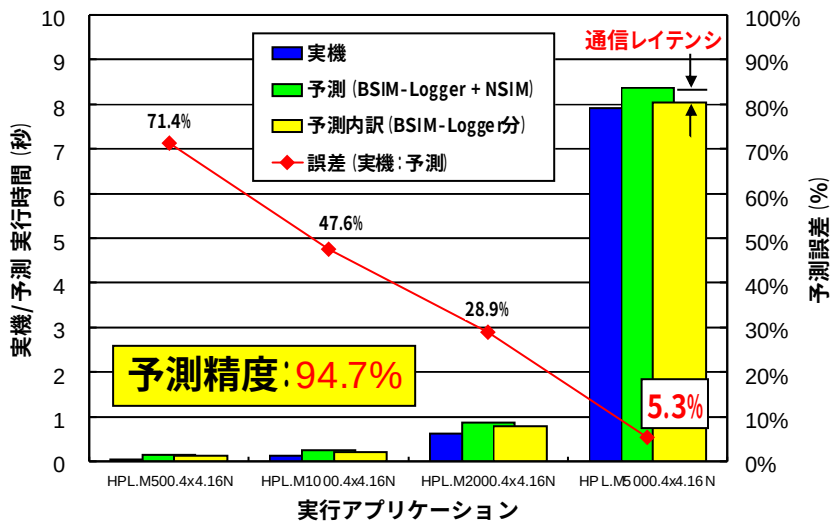
- **実用的な予測精度を有するか？**
- 評価アプリケーションの実機実行時間と、シミュレーションによる予測実行時間を比較する

実験内容

- **評価アプリケーションの実行時間の予測性能**
  1. **評価アプリケーション(HPL)を既存のクラスタシステムで実行**
     - **実際の実行時間を測定**
  2. **BSIM-Loggerを利用して、評価アプリケーションを理想ネットワーク環境下（ゼロ通信遅延時間）で実行した場合の通信プロファイルを生成**
  3. **NSIMを利用して、1.のクラスタシステムと同等のネットワーク（InfiBandスイッチによる単一段接続）をシミュレート**
     - CPU: Intel Xeon 3.0GHz、**ノード構成**：16ノード（1CPU/ノード）
     - InfiniBand**スイッチ**：1xLink DDR（4Gbps）、**スタートアップ遅延**：11.6μsec.（**実測、ポート間遅延を含む**）、**パケットペイロード**：1,024B、**パケットルーティング遅延**：100nsec.、3m**銅線ケーブルで接続**
     - **予測実行時間を算出**

実行時間の予測精度が得られる

実行時間の予測性能



評価アプリケーションの規模増加 ⇒ 予測精度が向上

まとめ

- 次世代スーパーコンピュータの設計開発に向けた**システム性能予測技術の開発**
  - システム性能評価環境（PSI-SIM）の開発
  - コンピュータシミュレーションによる性能見積ツールキット
  - 高機能な検索機能を備えた可視化・解析ツールキット
- **速い、易い、巧い** を提供
  - 高速な並列シミュレーション技術による実用時間内での評価
  - 評価ツールが兼備するスケーラブルかつ高い柔軟性
  - プログラムコード抽象化技術による高精度な性能予測
- 次世代スーパーコンピュータの設計開発のみならず、**アプリケーション開発時の強力な支援ツール**としても活用可能

プロジェクトスケジュール