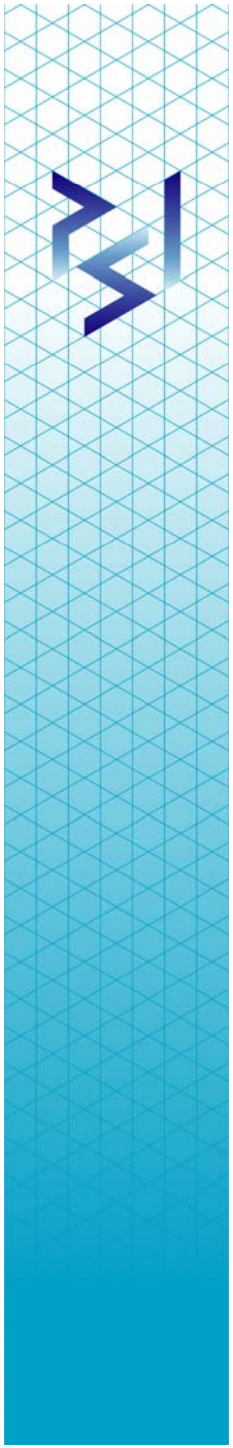
次世代スーパーコンピュータの設計開発を支援するシステム性能評価環境PSI-SIM

柴村, 英智
九州システム情報技術研究所

出版情報：SLRC プレゼンテーション, 2007-07-25
バージョン：
権利関係：



次世代スーパーコンピュータの設計開発を支援するシステム性能評価環境PSI-SIM

PSI-Project 柴村英智

（財）九州システム情報技術研究所

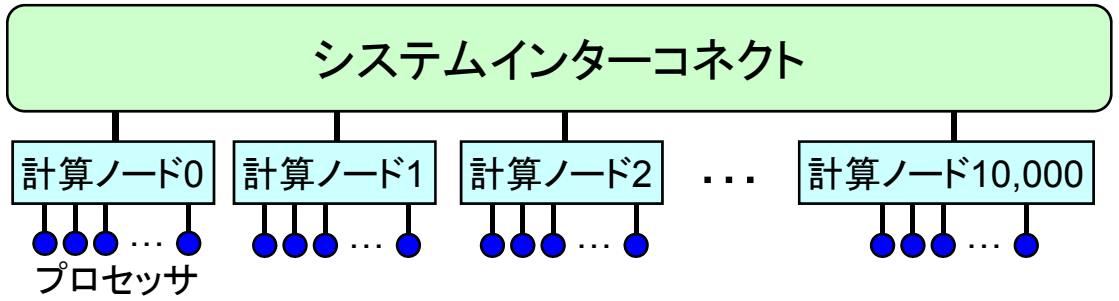
shibamura@isit.or.jp

PSI Project

- **P**etascale **S**ystem **I**nterconnect Project
  - 文部科学省「次世代IT基盤構築のための研究開発」、研究開発領域「将来のスーパー コンピューティングのための要素技術の研究開発」(H17-H19)
    - ➠ 研究開発課題「ペタスケール・システム インターコネクト技術の開発」

http://www.psi-project.jp/

- スーパーコンピュータの計算ノードを相互結合する システムインターコネクトの技術開発プロジェクト

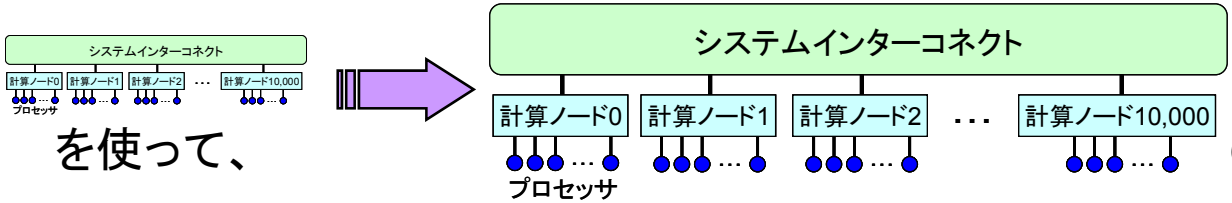


PSIプロジェクトにおけるミッション

- 実効性能 1P FLOPSの実現を目標とする3つの技術開発
  - 超高速光パケットスイッチの実現を目指した物理層技術
  - MPIから物理層までを通したインターコネクト全体の高機能化、高性能化技術
  - ペタフロップス級マシンの振舞いをシミュレーション可能とする統合型システム性能評価技術

本研究では！

- 「メモリ・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予測を可能にする大規模シミュレーション技術の確立



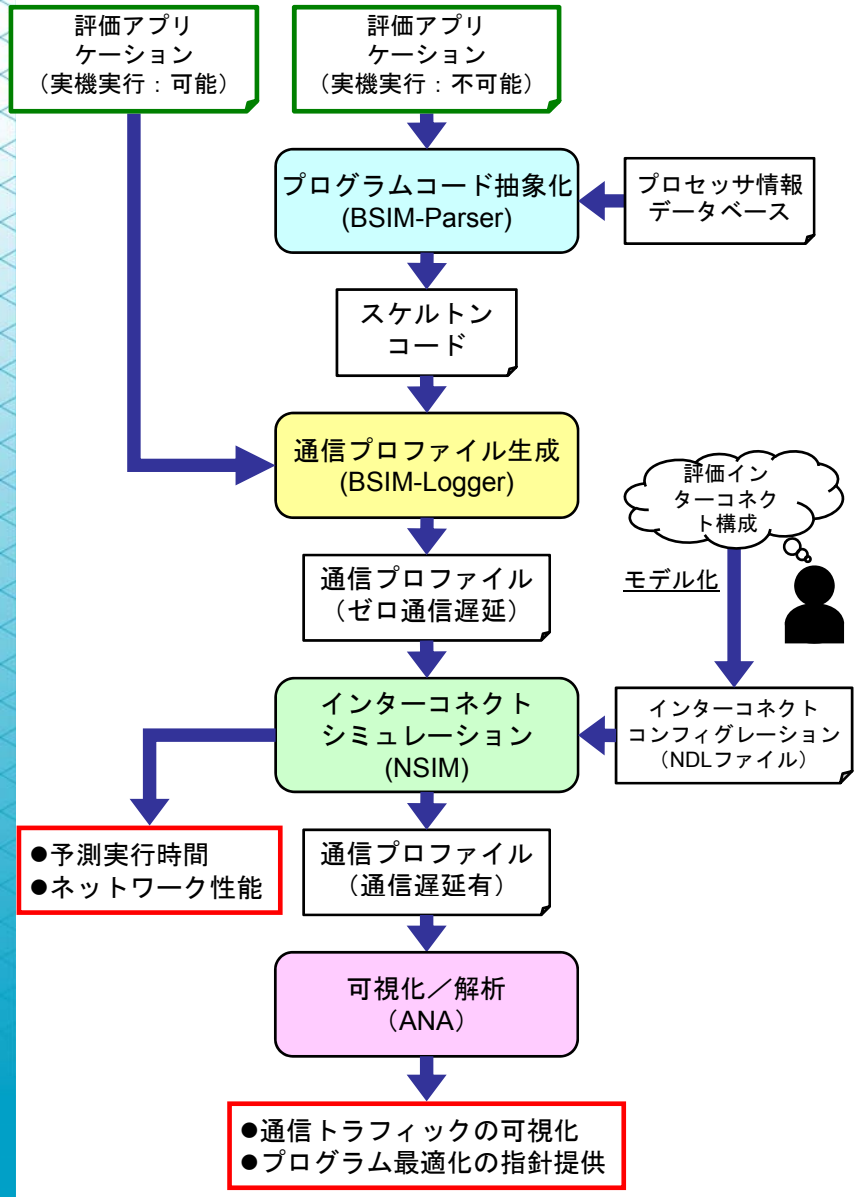
を使って、

の性能予測

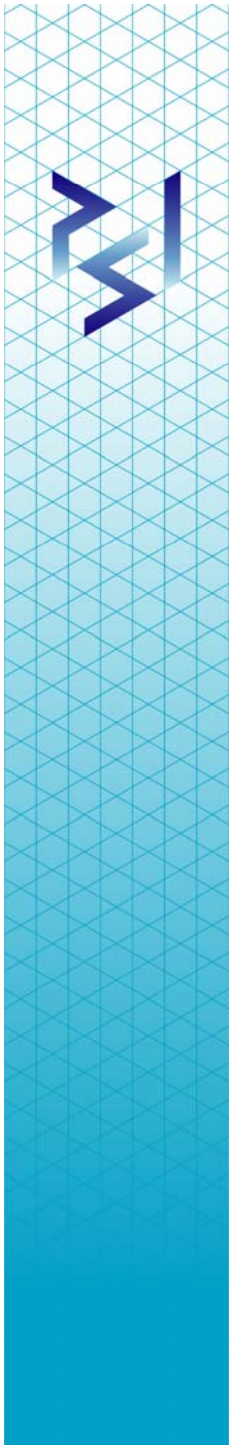
本研究の目的

- 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
- 性能評価環境(PSI-SIM)を構築
  - 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
- PSI-SIMの特徴
  - 数千から数万プロセッサを持つ大規模システムでも
    実用時間内でシミュレーションを完了 …(速い！)
  - 様々なシステムアーキテクチャに容易に対応できるよう、
    スケーラブルかつ高い柔軟性を持つ …(易い！)
  - スケルトン・コード実行と呼ぶプログラムコード抽象化技術を用いて、様々な評価項目を精度良く見積もる …(巧い！)

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

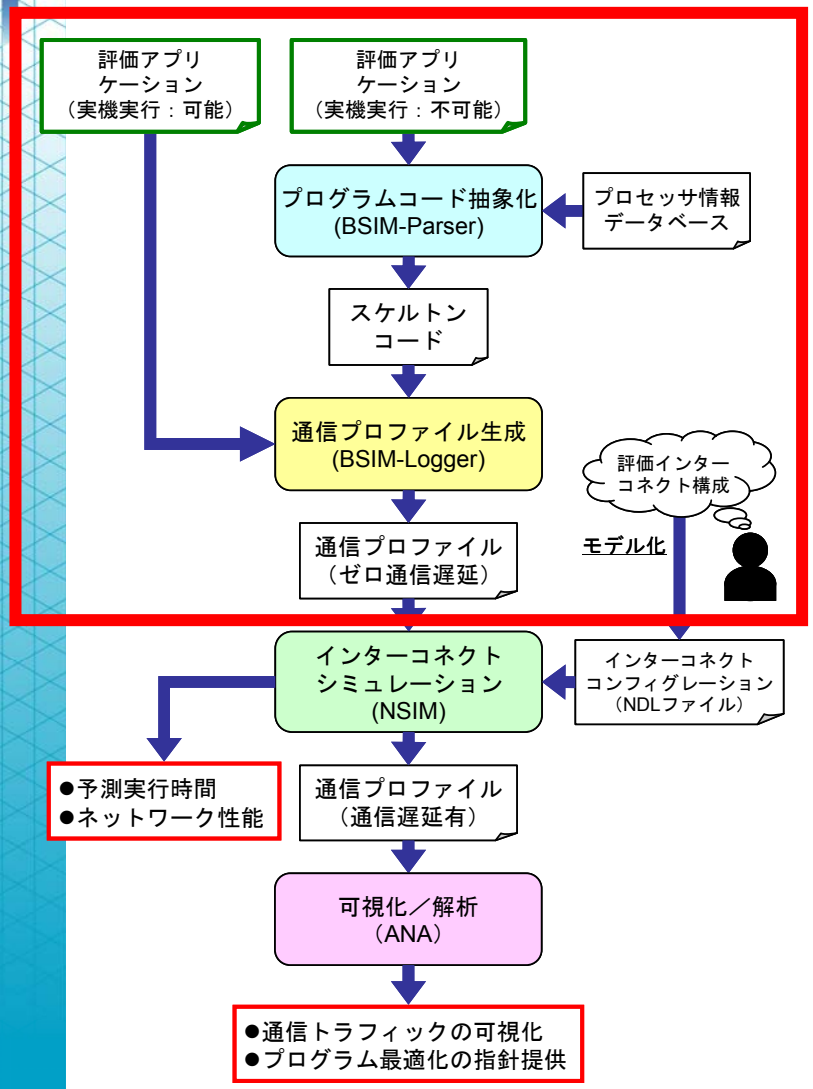


BSIM-Parser/Logger

プログラムコード抽象化技術に基づく

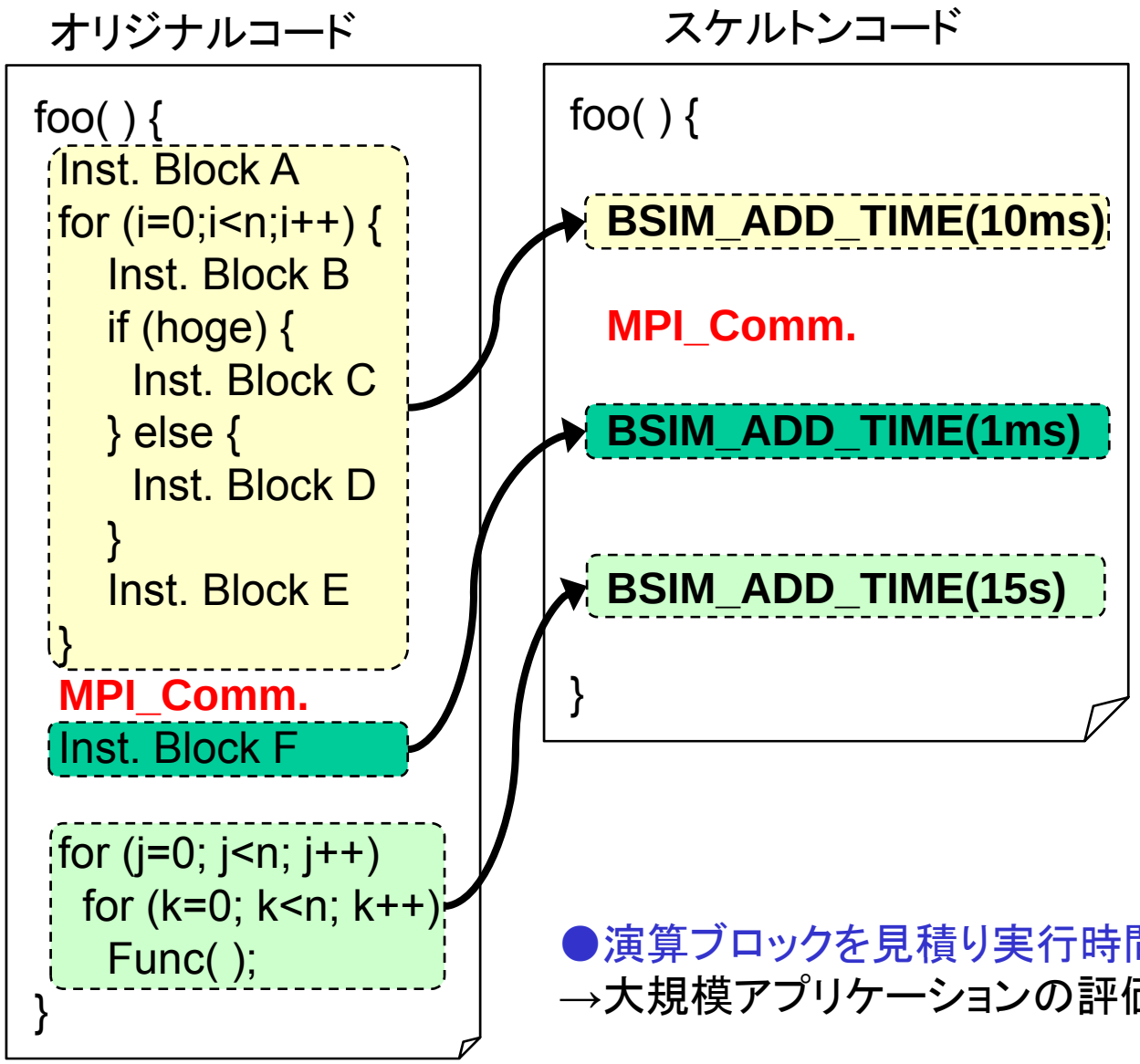
通信プロファイルの高速生成環境

BSIM-Parser/Logger

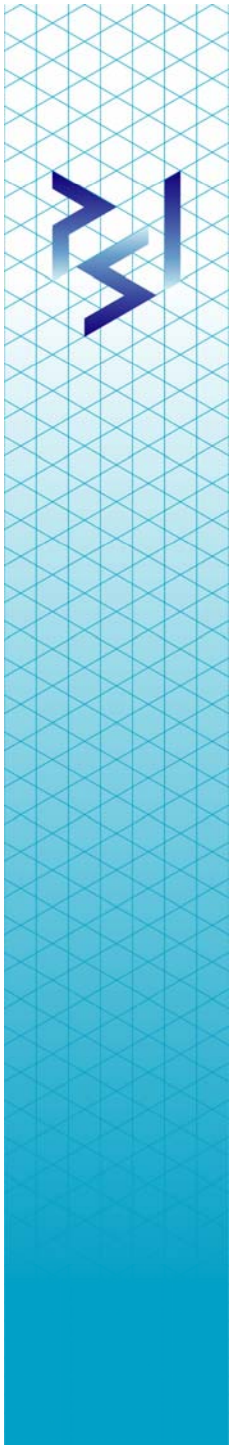


- 評価用アプリケーションの制御フローを維持したスケルトンコードの生成（BSIM-Parser）
  - 演算と制御・通信の分離
  - 命令ブロックから演算ブロックを抽出
  - 出力コードへの見積り実行時間の埋め込み
- 通信・計算処理の履歴を含む通信プロファイルの生成（BSIM-Logger）
  - 実機（クラスタ計算機）によるスケルトンコード化されたアプリケーションの擬似実行
  - 通信遅延時間0（理想的なネットワーク環境）の通信プロファイルを出力
  - 通信イベントの依存関係を保持

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

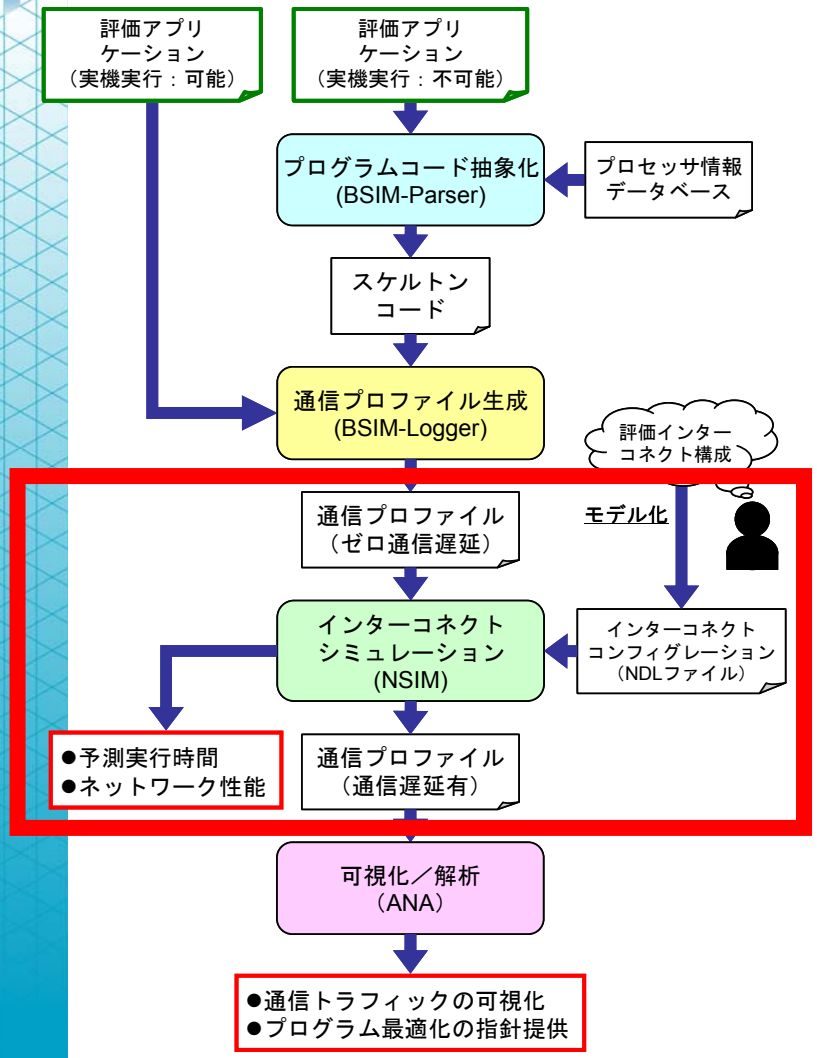


NSIM

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



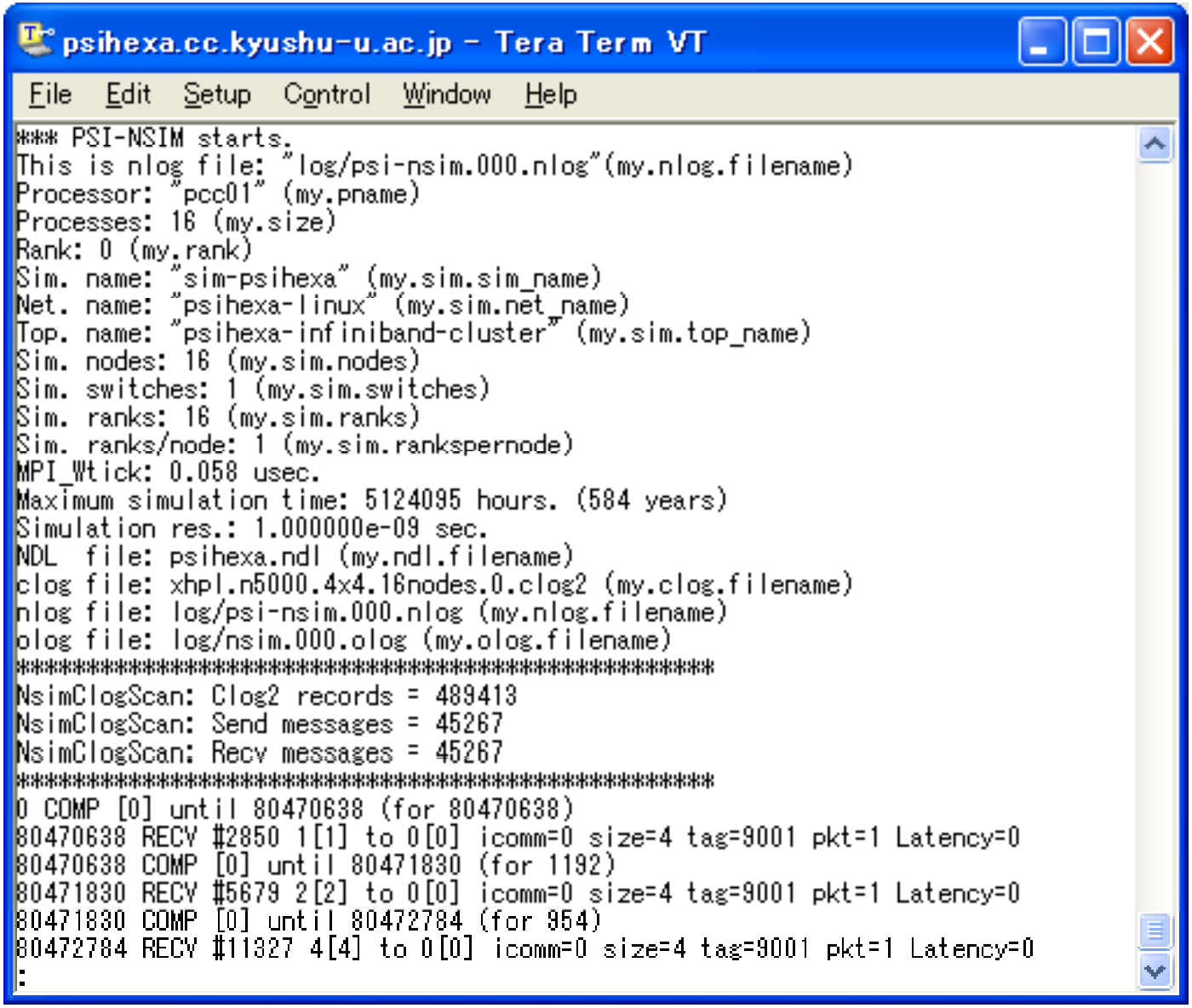
- 通信プロファイルに基づいたインターコネクトシミュレータ
  - 超大規模インターコネクトへの対応
  - 実用時間内におけるシミュレーションの完了を目的
  - 設計開発現場での実用性
    （シミュレーション解像度：１ナノ秒～）
  - 評価インターコネクトをコンフィグレーションファイルによってモデル化
  - ゼロ通信遅延時間の通信プロファイルを入力
  - 通信遅延時間を付加した通信プロファイルを出力 → PSI-ANAへ
  - 並列離散事象シミュレーション

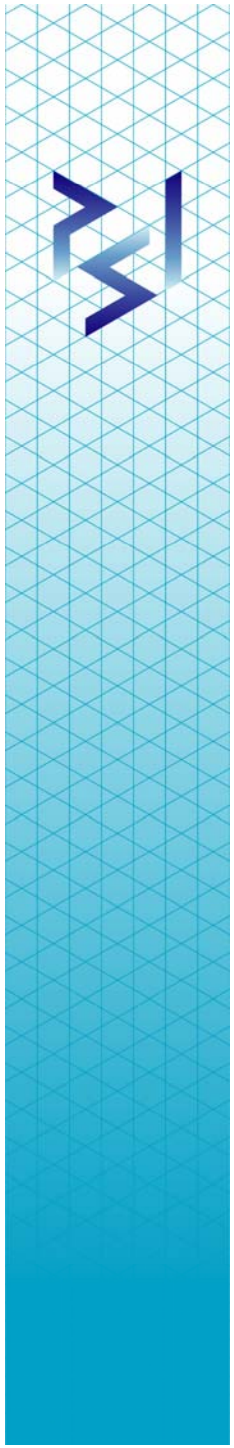
インターコネクト コンフィグレーション

```
; 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

通信遅延付プロファイル

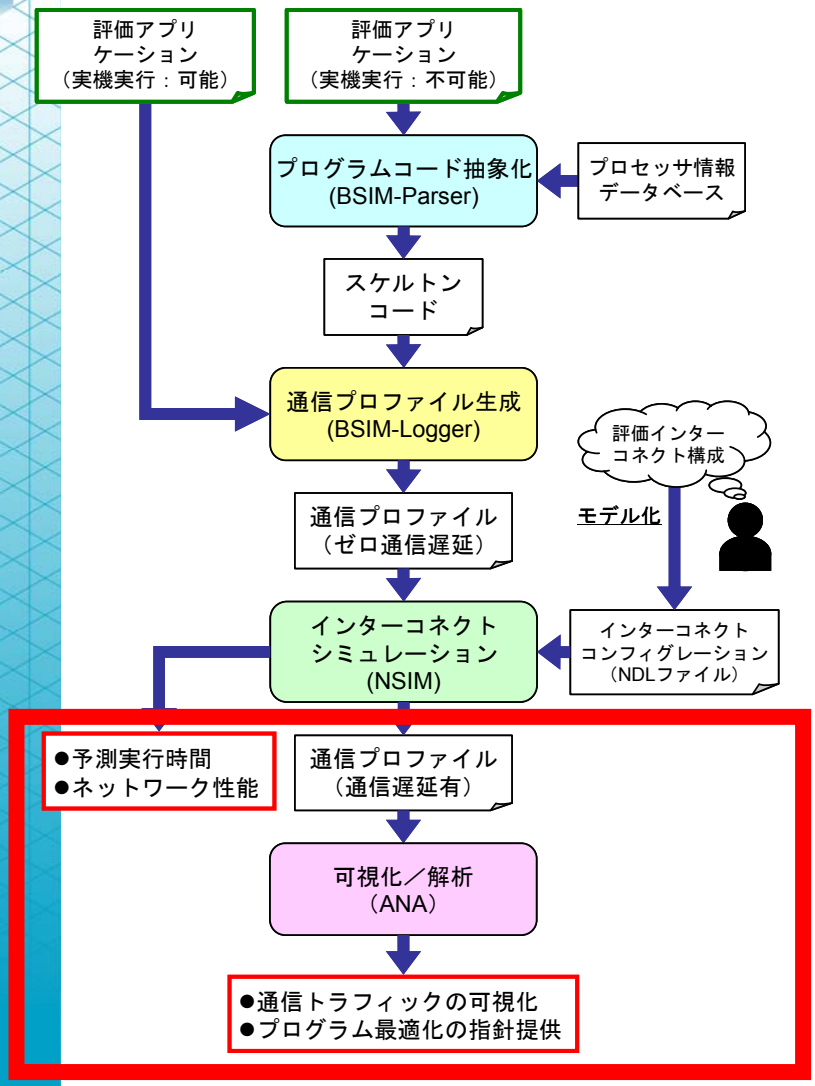




ANA

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

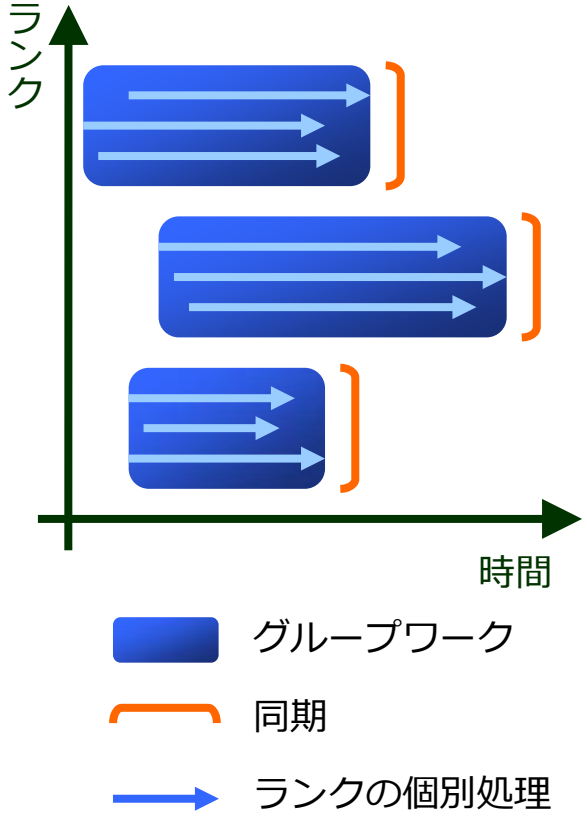
ANA



- 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

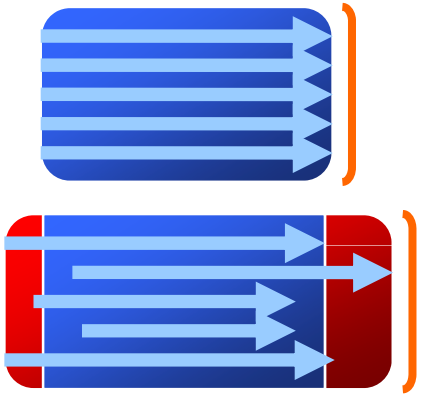


グループワーク

- ◆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

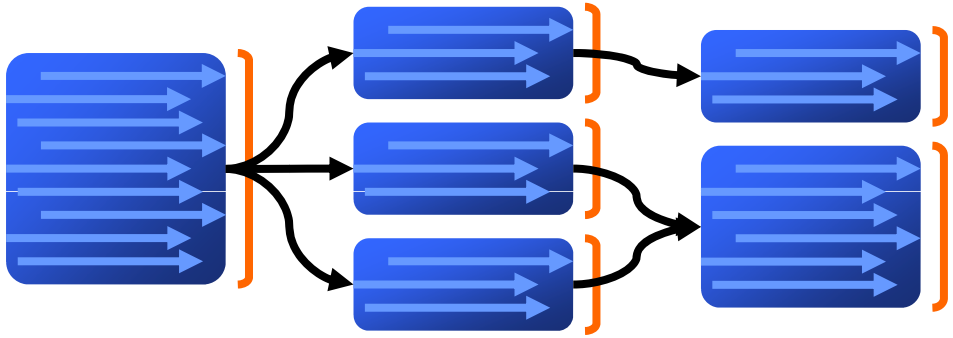
- すべてのワーカーが、同時に開始、同時に終了
- 待ち時間がない

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



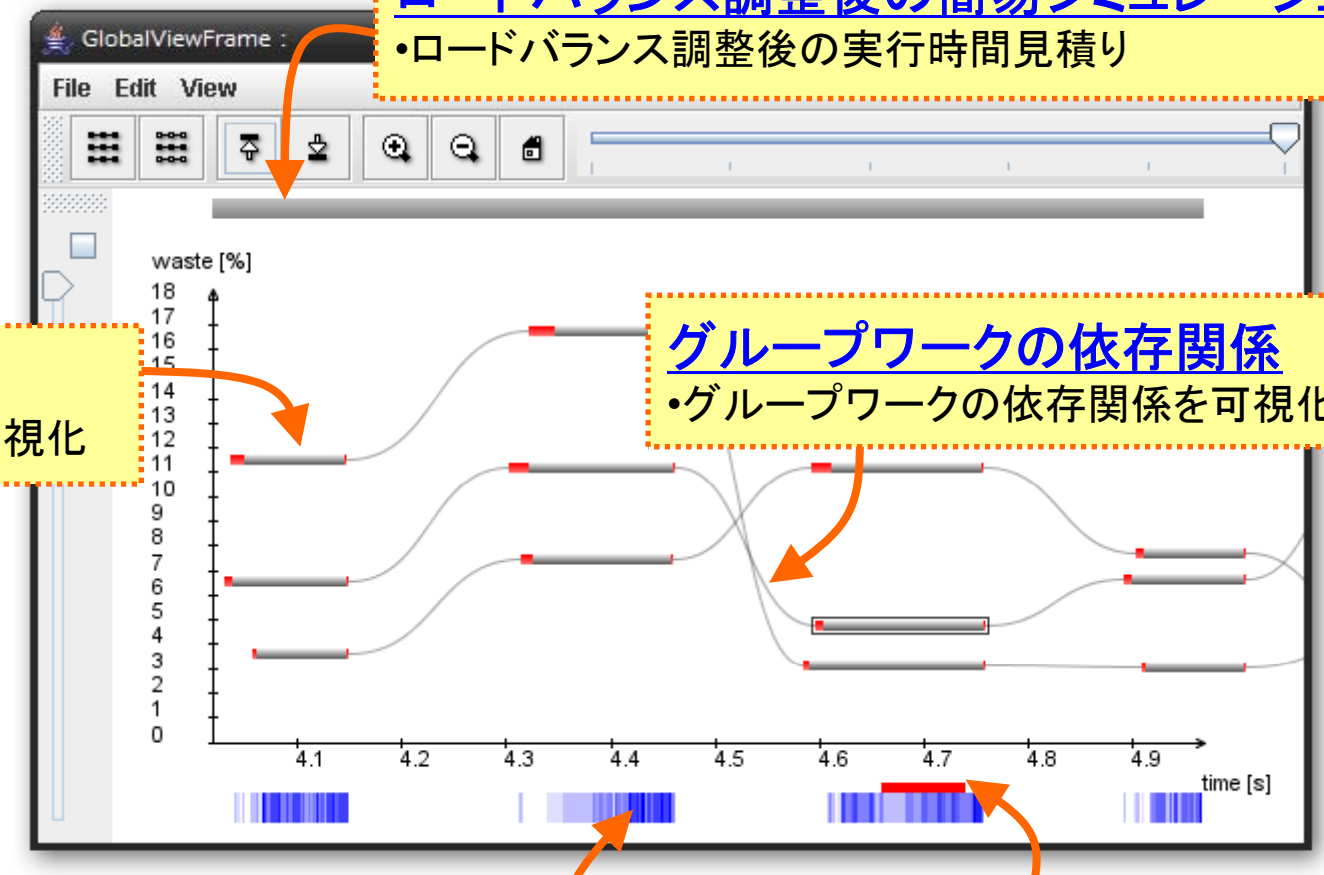
- グループワーク単位のロードバランス解析が容易
- 優先して改良すべきグループワークを把握できる
- クリティカルパスの早期発見に役立つ

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

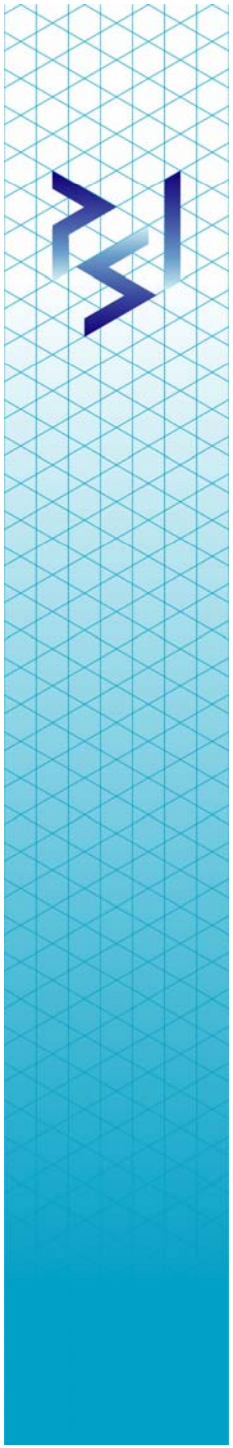
グループワーク
- ロードバランスの可視化

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

BSIM-LoggerとNSIMを用いて
「シミュレーション時間」と「実行時間の予測性能」
について調査

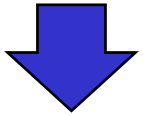
NSIMの評価実験

- 調査項目
  - 実験（1）：NSIMが要するシミュレーション時間
    - 実用時間内での性能評価が遂行可能か？
    - 理想ネットワーク環境下（ゼロ通信遅延時間）で実行し、シミュレーション時間の下限（最低必要時間）を測定
  - 実験（2）：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（1）

- NSIMが要するシミュレーション時間の調査

1. 問題サイズやプロセス数を変化させた通信プロファイルを生成
   - ゼロ通信遅延時間
2. 各通信プロファイルをNSIMで実行
   - 理想ネットワーク環境下を想定してシミュレート
   - NSIMの利用CPU数を1、2、4、8、16、32CPU と変化
   - 各NSIMのシミュレーション時間を測定
     （プロファイルの読み込み時間は含まない）



最低限のシミュレーション時間、並列化効率が得られる

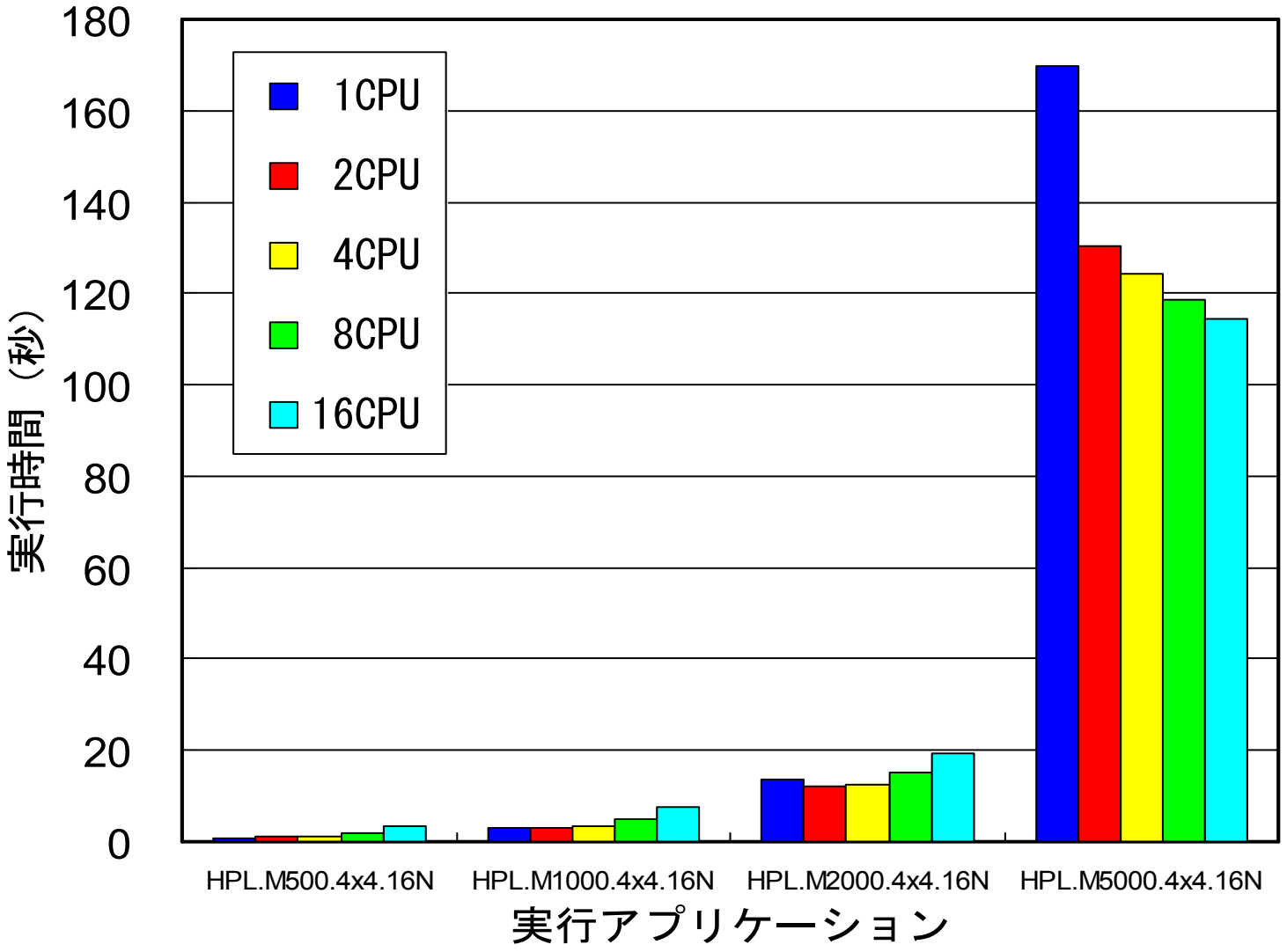
通信プロファイルの生成

- 評価アプリケーション：HPL
  - 問題サイズ（N）：500、1000、2000、5000
  - プロセス数（PxQ）：4x4、16x16、32x32
  - ブロックサイズ：128
- BSIM-Loggerによる通信プロファイル生成
  - 上記パラメータの組合せから数種を選択
  - ゼロ通信遅延時間の通信プロファイルを生成

実験環境

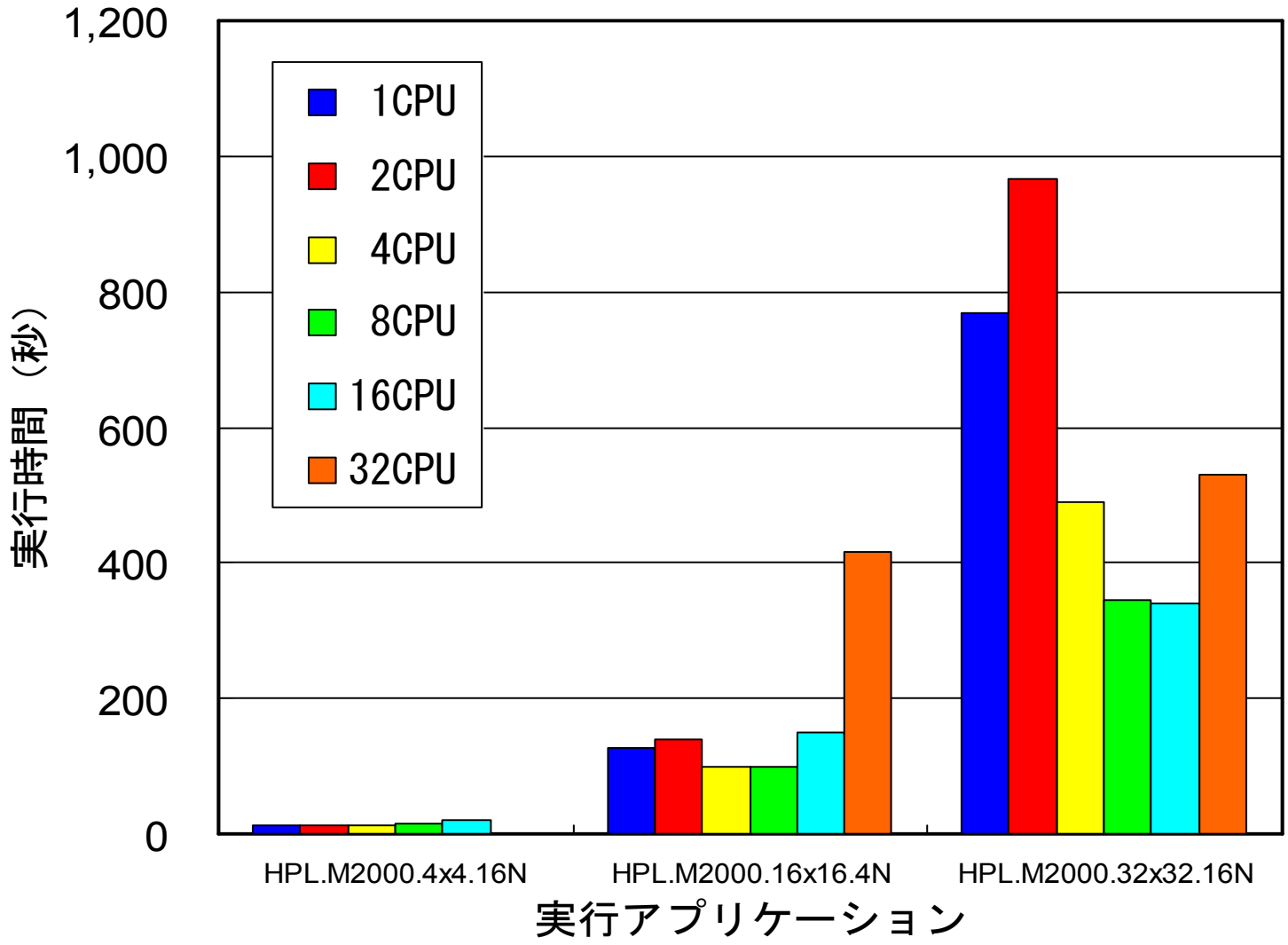
- 1CPU（デスクトップPC）
  - CPU: Intel Xeon 3.8GHz (EM64T)
  - Memory: 2GB
  - OS: Linux 2.6.20-1.2320.fc5
  - Compiler: GNU C Compiler ver.4.1.1
  - MPI: Mpich2-1.0.5p4
- 2CPU～（クラスタシステム）
  - CPU: Intel Xeon 3.0GHz (EM64T)
  - Memory: 7GB MEM
  - OS: RedHat Enterprise Linux AS rel.3 (Linux Kernel 2.4.21)
  - Compiler: Fujitsu Fortran&C compiler ver.5.0
  - MPI: Fujitsu MPI over Score
  - Network: InfiniBand (1xLink DDR)

シミュレーション時間（プロセス数固定：16）



評価アプリケーションの問題サイズ増加 ⇒ 並列処理効率が向上

シミュレーション時間（問題サイズ固定：2000）

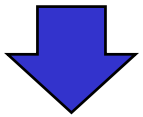


評価アプリケーションのプロセス数増加 ⇒ 並列処理効率が向上

実験(2)

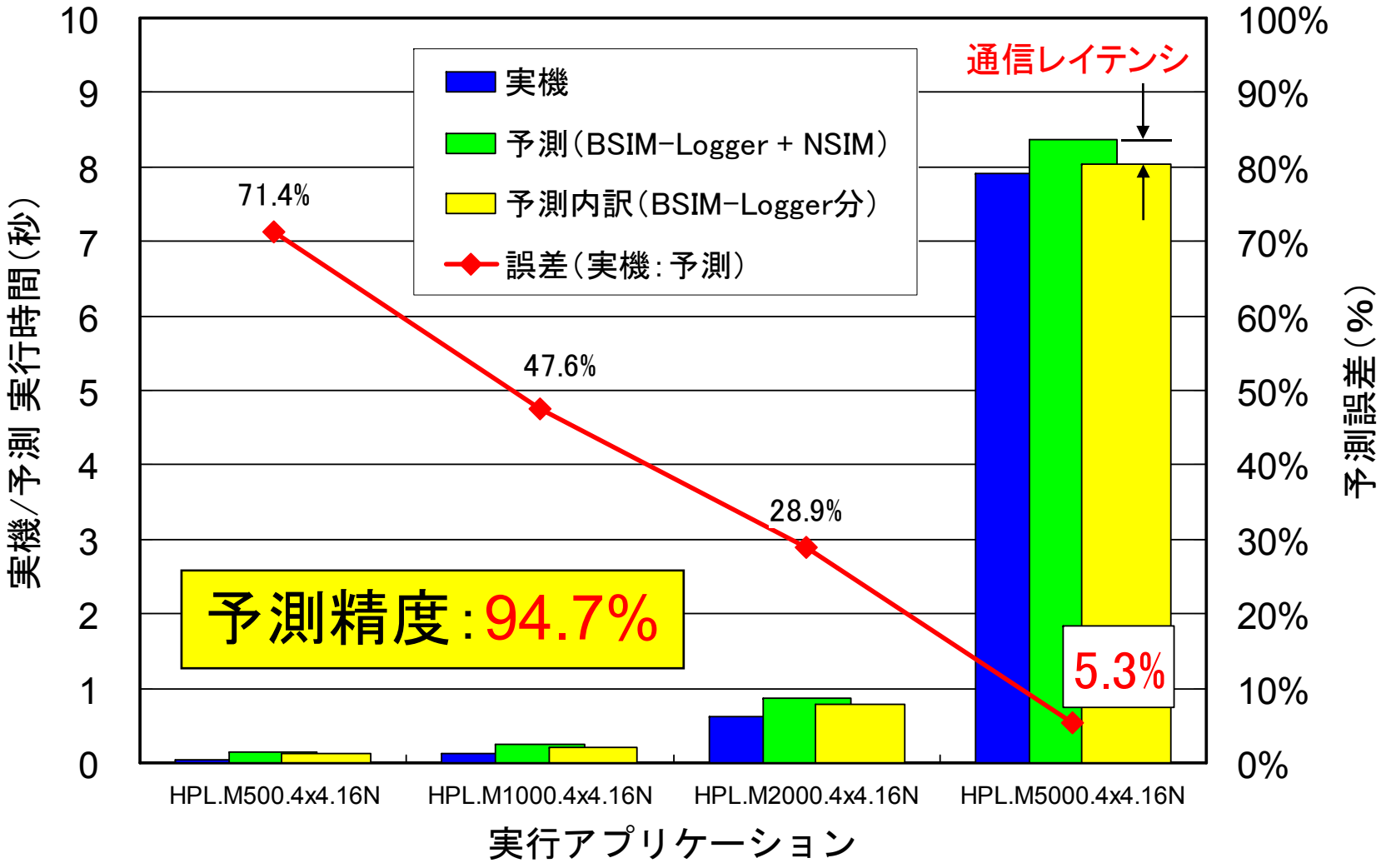
- 評価アプリケーションの実行時間の予測性能

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

ペタスケール級スパコンの評価に向けて

- 約100万回のMPI通信を行うアプリケーション(HPL.M2000.32x32.16N)のシミュレーション時間
  ⇒ 8CPUシステムで約5分強

- 超大規模アプリケーションを想定した見積り
  - HPL（並列数：262,144、問題サイズ：6,656K、ブロック数：128、MPI通信回数：1.75E+12回、通信プロファイル：325Tバイト）
    - 実効性能 10PFLOPS（CPUコア数：256K）のマシン
      ⇒ HPLの予測実行時間：約5時間
  - 現在のNSIMの性能を基にシミュレーションした場合
    ※ ただし、通信プロファイルの読込時間を除く
    - 並列効率 100% の場合
      - 1,024CPUのマシンでシミュレーション ⇒ 8.6 時間
      - 512CPUのマシンでシミュレーション ⇒ 17.1 時間
    - 並列効率 90% の場合
      - 1,024CPUのマシンでシミュレーション ⇒ 9.5 時間
      - 512CPUのマシンでシミュレーション ⇒ 19.0 時間

実用圏内

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