# 「速い・易い・巧い」を提供する次世代スーパーコンピュータの性能予測技術の開発

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, 2006-07-19. 九州大学システムLSI研究センター
バージョン：
権利関係：

# 「速い・易い・巧い」を提供する次世代スーパーコンピュータの性能予測技術の開発

PSI-Project 柴村英智†

† (財)九州システム情報技術研究所

shibamura@isit.or.jp

# What's 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
  - 数十万規模の高速計算ノードを持つペタフロップス級の次世代スーパーコンピュータシステムを念頭
  - 現状のシステムよりもコスト対性能比で10倍以上を目指す
  - 高性能化、高機能化、低コスト化を同時に達成

# PSIプロジェクトにおけるミッション

- 実効性能 1Pflop/s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予測を可能にする大規模シミュレーション技術の確立

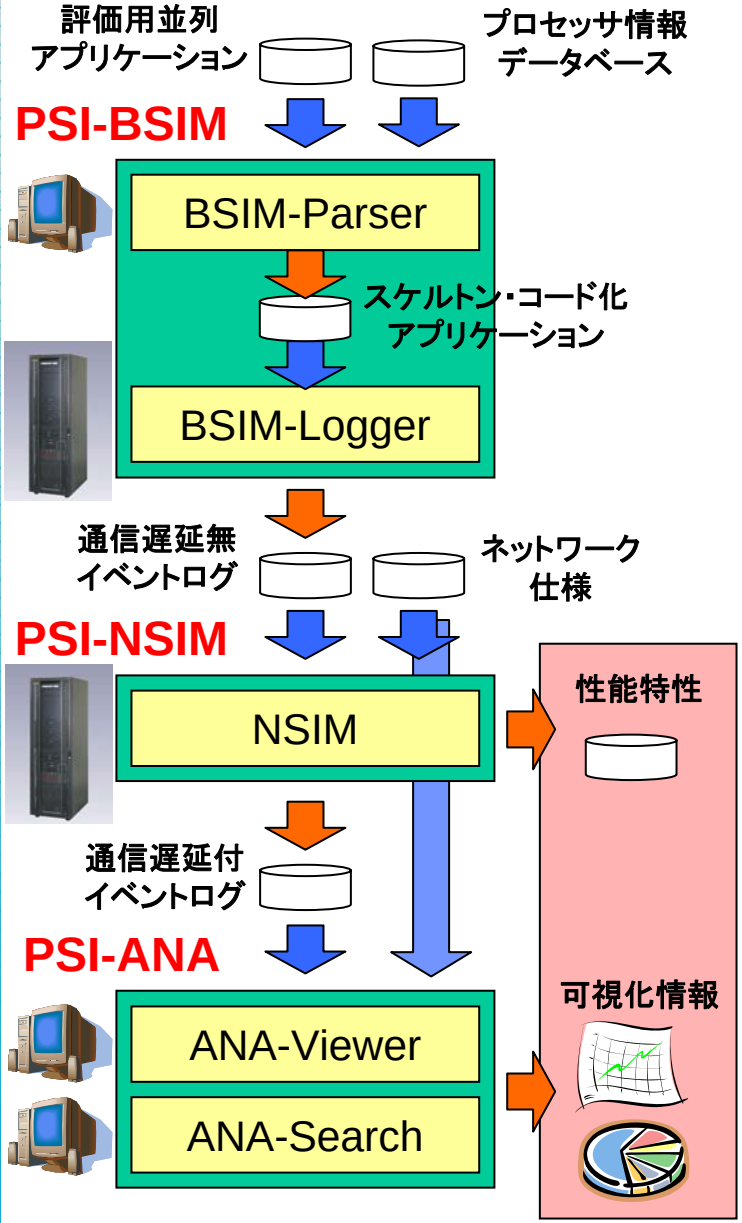
# 性能予測技術の必要性

- 次世代スーパーコンピュータの設計開発支援
  - 未知なるシステムアーキテクチャの設計空間を探索
  - 利用可能な要素技術に則した評価
- システムソフトウェアやアプリケーションの設計開発支援
  - アプリケーションの実行完了時間の予測
  - 高速化・最適化に向けた最新機構の有効性検証
- 性能をスポイルするボトルネックポイントの発見
  - ハードウェアとソフトウェアの両側面からアプローチ
  - 様々なアプリケーションの性能をバランス良く引き出す

# 本研究の目的

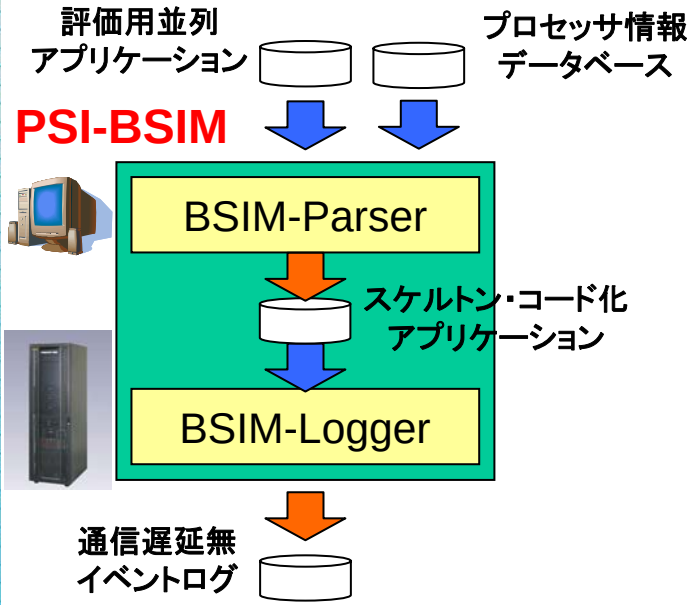
- 次世代スーパーコンピュータの設計開発に向けたシステム性能予測技術の開発
- 性能評価環境(PSI-SIM)を構築
  - コンピュータシミュレーションによる性能見積ツールキット
  - 高機能な検索機能を備えた可視化・解析ツールキット
- PSI-SIMの特徴
  - 数千から数万プロセッサを持つ大規模システムでも実用時間内でシミュレーションを完了 … Speedy（速い）!
  - 様々なシステムアーキテクチャに容易に対応できるよう、スケーラブルかつ高い柔軟性を持つ … Simple（易い）!
  - スケルトン・コード実行と呼ぶ擬似実行技術を用いて、様々な評価項目を高精度で見積もる … Skillful（巧い）!

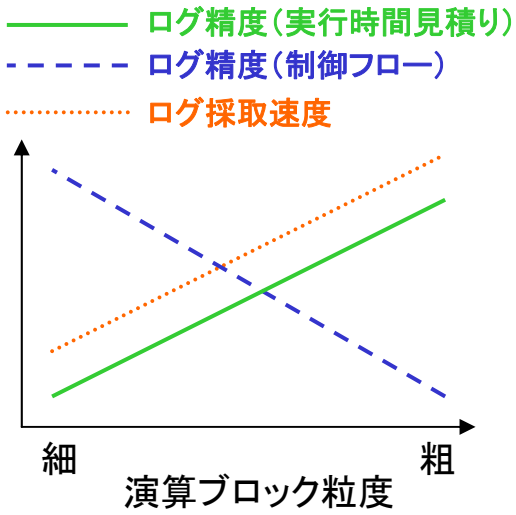
# 性能評価環境PSI-SIMの構成




- **PSI-BSIM (Base SIMulator)**
  - アプリケーションの構造解析
  - 演算ブロックの見積実行時間の封入
  - 擬似実行のためのスケルトン・コード生成
  - 擬似実行（並列処理）による理想的なネットワーク環境を想定したイベントログ生成
- **PSI-NSIM (Network SIMulator)**
  - 大規模なイベントログを利用し、様々なネットワーク仕様に基いた柔軟なネットワークシミュレーション
  - 並列処理による高速なシミュレーション
  - 開発対象とするシステムの様々な性能特性と通信遅延付イベントログを出力
- **PSI-ANA (Adv. Network Analyzer)**
  - ペタスケール級の大規模ログの俯瞰に優れた可視化機能
  - 従来にない高機能なサーチエンジン

# PSI-BSIM



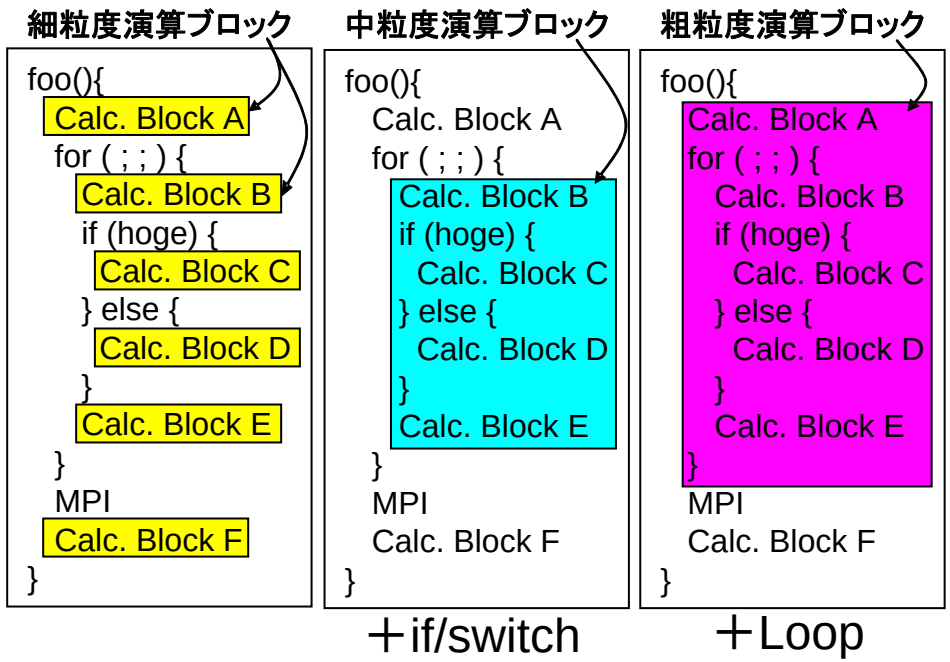

- 評価用並列アプリケーションの制御フローを維持したスケルトン・コードの生成（BSIM-Parser）
  - 演算と制御・通信の分離
  - 命令ブロックから演算ブロックを抽出
  - 出力コードへの見積実行時間の埋め込み
- 通信履歴を含むイベントログの生成（BSIM-Logger）
  - 実機（クラスタ計算機）によるスケルトン・コード化されたアプリケーションの擬似実行
  - 通信遅延時間0（理想的なネットワーク環境）のイベントログを出力
  - イベントの依存関係を保持

# 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）

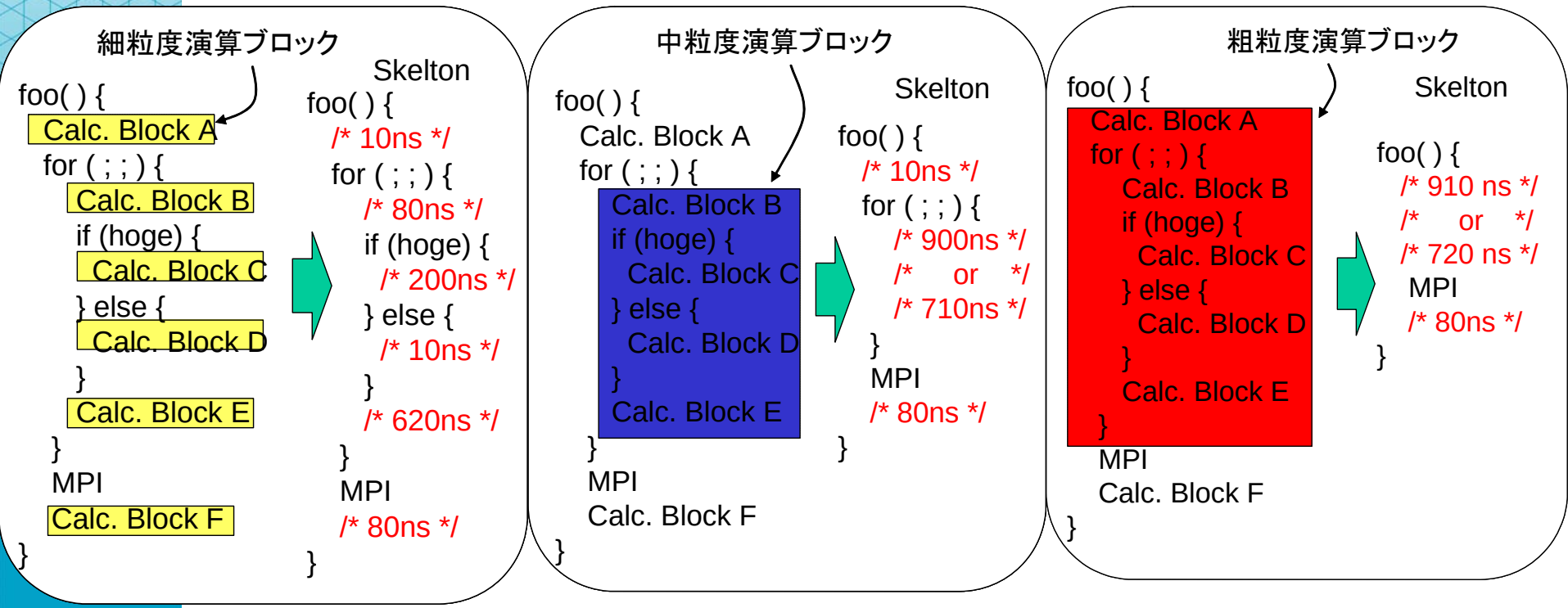
```
subroutine transpose_x_y_global(d, ..xout)
            :
call mpi_alltoall(xin, … commslice2, ierr)
```

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 | × | ○ | ― | ― | ― |

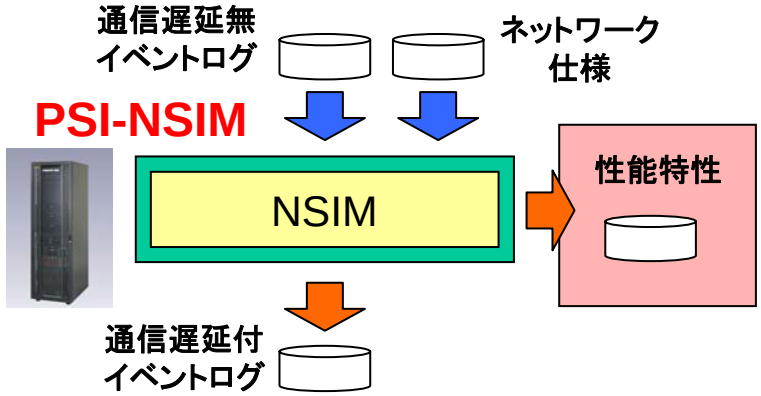
○：サポート　×：サポート困難　―：未サポート

# PSI-NSIM

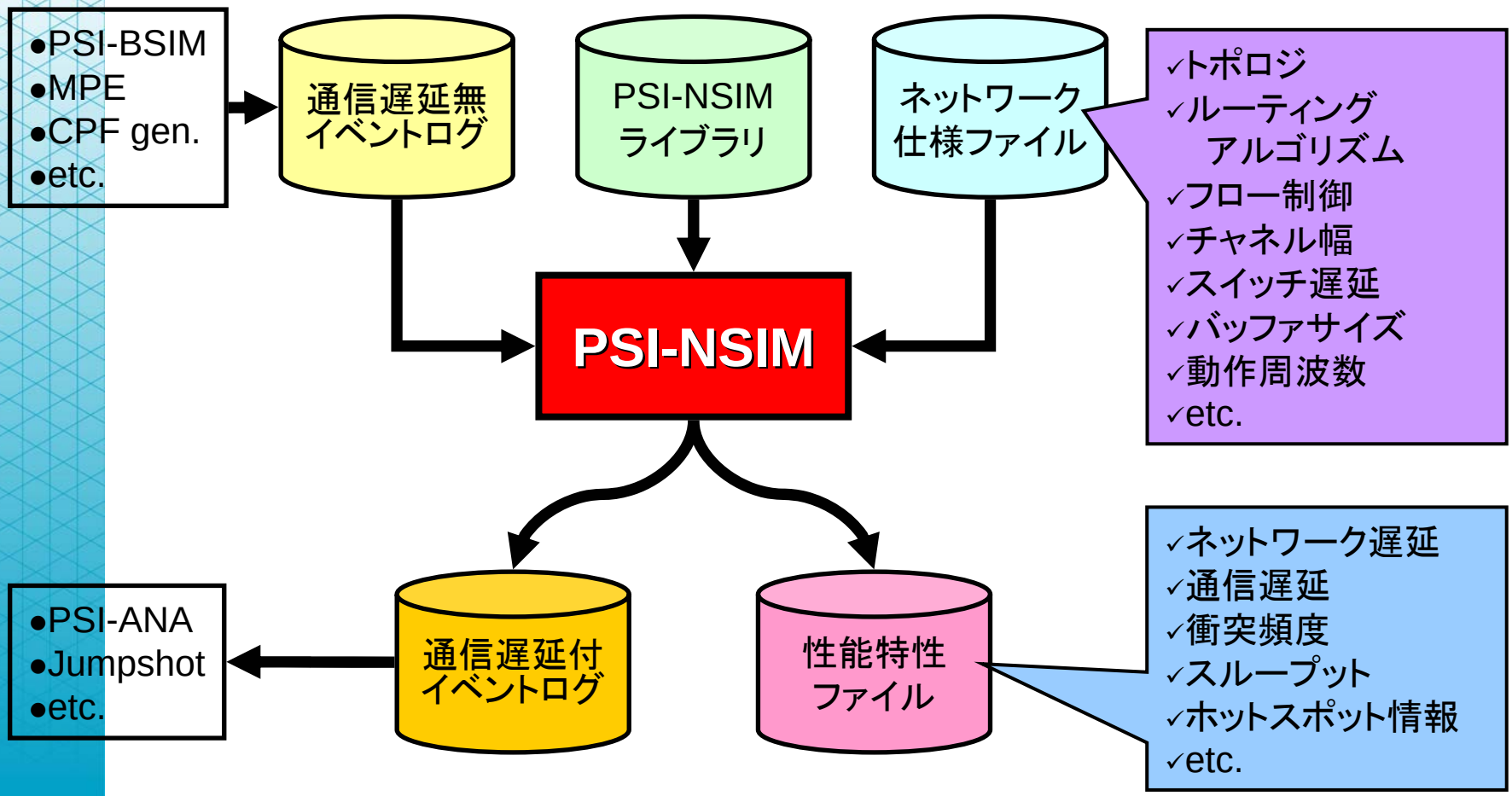
- ペタスケール級スパコンの相互結合網に対する要件を下記観点からの性能評価を通じて探る
  - アーキテクチャ
    - トポロジ、通信方式（フロー制御方式、ルーティング方式）
    - 集約通信、通信の動的最適化、優先度送受信などの専用機構の利用
  - ハードウェア実装
    - フリット・パケット長、各種バッファ量、レイテンシ、動作周波数
  - スケーラビリティ
    - ノード数、クラスタ数、パーティション形態
  - アプリケーション
    - 実践的なアプリケーションの通信パターンファイルの利用
    - ネットワークを高負荷状態にする人工的な通信パターンファイルの利用
- 中規模シミュレーションを基にした、大規模なシステムの性能予測をサポートする

# PSI-NSIMの特徴

- 数万ノードを結合する相互結合網の並列シミュレーション
  - ネットワーク仕様ファイルによる詳細かつ柔軟なシミュレーションパラメータの設定
  - コレクティブ通信機能などのインテリジェントなスイッチ仕様に対応
  - 種々のイベントログファイルを入力としてサポート
  - 遅延時間、転送処理能力、ホットスポットなどの多彩な評価項目を出力

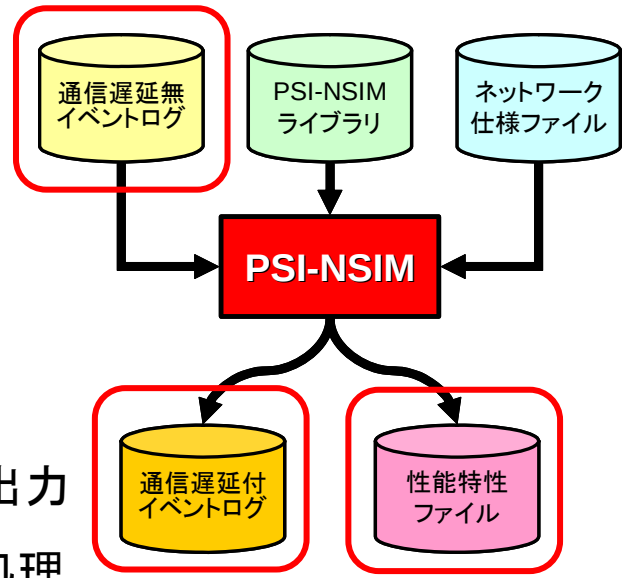


# PSI-NSIMのシミュレーションフロー



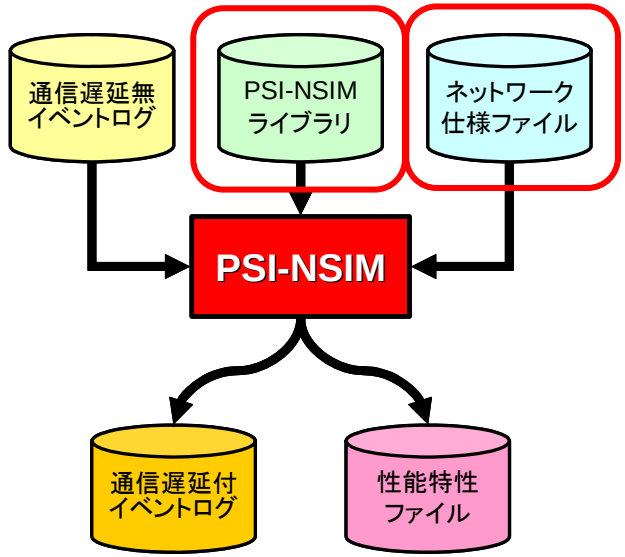
# PSI-NSIMの入出力ファイル(1)

- イベントログファイル
  - 通信遅延無イベントログ(入力)
    - ネットワーク遅延ゼロ (non-timed)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からの出力ログ)
    - 典型的な通信は、通信パターンファイル生成プログラム(CPF Generator)によって生成した出力ログも利用可能
  - 通信遅延有イベントログ(出力)
    - ネットワーク遅延を考慮した通信ログ
    - 可視化・解析ツールに入力可能な形式
- 性能特性ファイル
  - 各通信メッセージについて、送受信時刻、送受信ノード、ネットワークレイテンシ、通信レイテンシ、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レイテンシ、通信処理能力、ホットスポットなどに解析・算出に利用



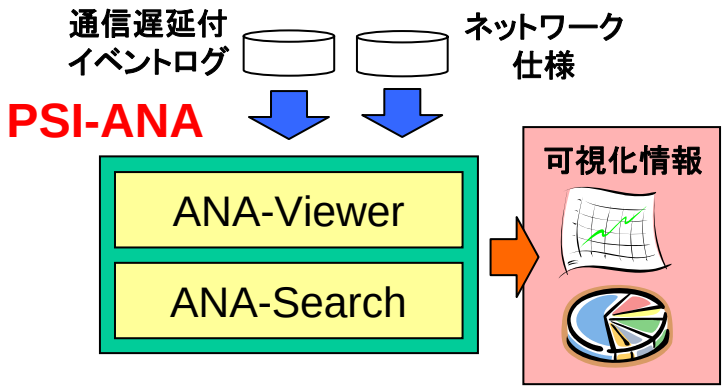
# PSI-NSIMの入出力ファイル(2)

- PSI-NSIMライブラリ
  - シミュレーションで多用される設定内容やネットワーク仕様ファイルでサポートが難しい機能をモジュール化し、ライブラリとして用意
    - トポロジ
    - ルーティングアルゴリズム
    - インテリジェント・スイッチの振舞いなど
- ネットワーク仕様ファイル
  - シミュレーション対象のモデル化や各種パラメータを設定するためのファイル
  - 設定項目
    - トポロジ、ルーティングアルゴリズム、フロー制御方式、チャネル幅、スイッチ等のレイテンシ、バッファサイズ、動作周波数、入出力ファイル名、など
  - 可読性の高い簡易言語による記述



# PSI-ANA

- ネットワーク状態（通信の混雑状況など）の高速な解析と可視化機能を提供
- 超巨大ログファイルの解析をサポート
- 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



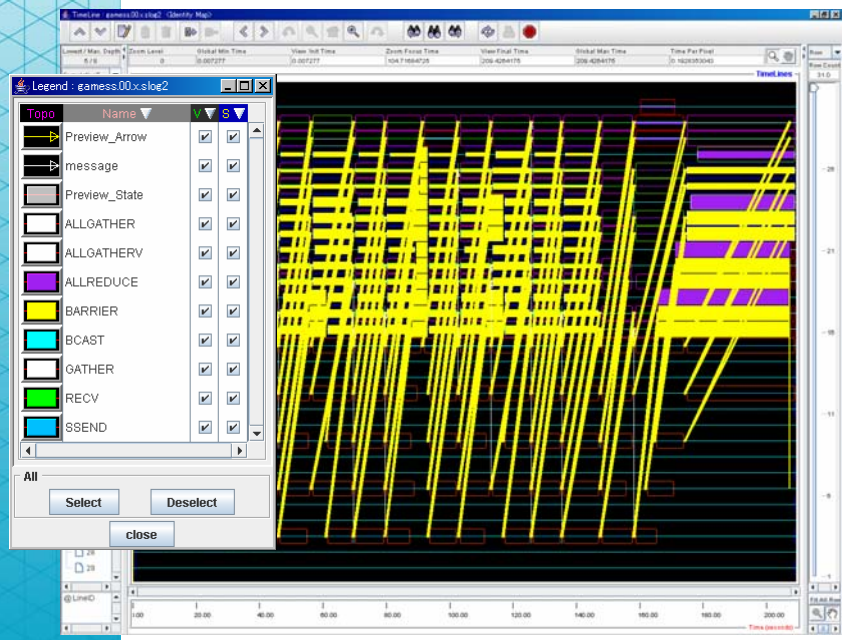
# PSI-ANAの各種機能

- 高機能ユーザインタフェース
  - マルチコミュニケーターやマルチランクのグループ化表示への対応
  - 観測性に優れた拡大・縮小・移動などの操作
  - 動画によるプロセス実行状況の再生・巻戻し・早送り
- 検索機能
  - 混雑具合の類似検索機能
  - クリティカルパス（ボトルネックポイント）の発見機能
- プログラミング支援機能
  - 送受信に時間がかかる理由の提示
  - 計算割り当ての不均衡を指摘
  - 高速化の可能性のある箇所を強調表示
  - ボトルネック箇所に対応するソースコードを表示
- その他
  - 分割ログファイルへの対応
  - ログファイルを時間軸方向／ランクごとに分割

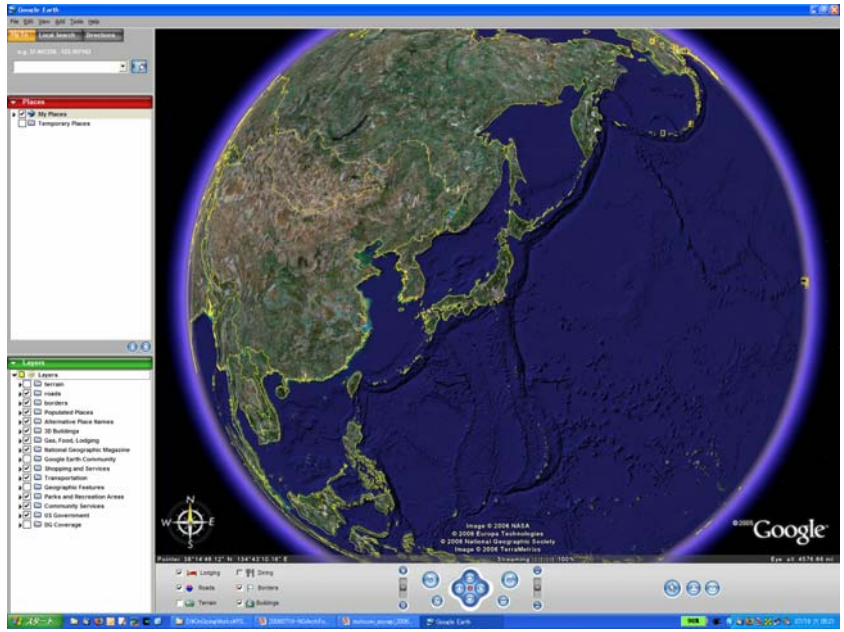
※検討項目を含む

# 魅力的なPSI-ANAの可視化機能

- 既存のViewer（Jumpshotなど）やGoogleEarthが持つ充実した俯瞰特性・操作性＋αを兼備




＋



# 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
  - 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
  - 評価ツール群が兼備する高い柔軟性
  - 擬似実行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