FX10における パケットペーシングを用いた アプリケーションの通信性能評価

柴村英智
<shibamura@isit.or.jp>
（公財）九州先端科学技術研究所
次世代スーパーコンピュータ開発支援室
JST CREST

どのようなトピックか？

- HPCアプリにおけるメッセージ通信
- インターコネクトでの通信混雑を解消・抑制し、アプリの実行を速くしたい
- 混雑が発生しないように、パケットの送信間隔を空けて（パケットペーシングして）送ろう

こんな風にしたい！

とあるスクランブル交差点

集団行動 by 日体大

パケット間に別のパケットを上手く(巧く)通し、
衝突の無いスムーズな通信を実現したい！

発表のながれ

- 研究の概要
- 背景
- パケットペーシング
- これまでにわかったこと
- 目的、方法、評価指標
- 評価実験
- まとめ
- 今後の課題

研究の概要

- これまでの研究では
  - HPCアプリの通信高速化

    ⇒ パケットペーシングに着目
  - ペーシングを施したアプリについて、「シミュレーション」による通信性能評価

    ⇒ パケットペーシングの効果を確認（SWoPP2010～）
- 本発表では
  - 「実機（FX10）」における通信性能評価

    ⇒ パケットペーシングの効果を確認

● ペーシングって効くの？
● 単に遅くなるだけじゃない？

● シミュレーションで効果が出てもねぇ。。。
● はやく実機で結果を示せ！

背景

- ポストペタ／エクサ コンピューティング時代に集団通信は使えるのか？
  - 大きな問題
    - 動かない（かも）：メモリ不足
    - 遅い（かも）：通信混雑、輻輳
  - 解決方法
    - 使わない ⇒ アプリケーション毎に通信アルゴリズムを考える
    - 速いものを選ぶ ⇒ 集団通信アルゴリズムの動的選択
    - なんとかしよう ⇒ 省メモリで高速な集団通信機構の開発
      - 省メモリ（実装）については別途研究
      - 集団通信の高速化に焦点

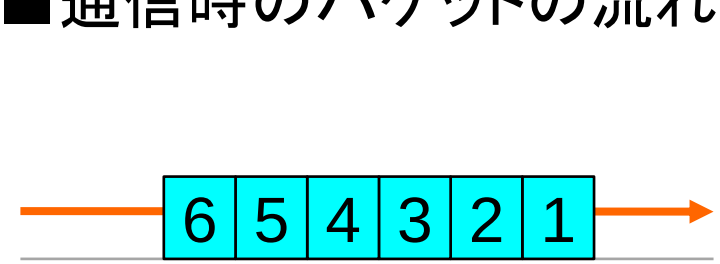
集団通信の高速化

- 高速化の方法
  - 新規に集団通信アルゴリズムを開発
    - システムとの親和性が重要
  - 適応ルーティングの利用
    - 集団通信は高トラフィック ⇒ 渋滞を迂回しても結局は渋滞
  - パケットペーシング
    - ネットワークへのパケット投入を抑制（Injection Rate Control）

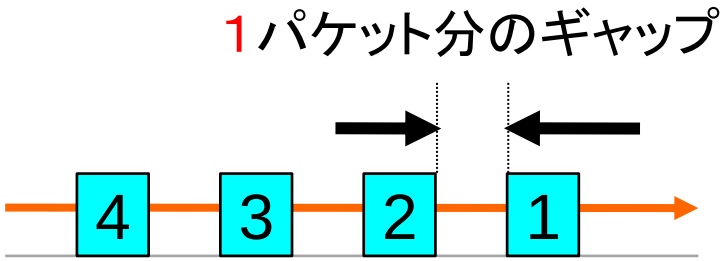
パケットペーシング

- 指定する間隔（パケット間ギャップ）を空けて、パケットをネットワークに送出

■通信時のパケットの流れ



パケットペーシング無し
（パケット間ギャップ＝0）



パケットペーシング有り
（パケット間ギャップ＝1）

通信効率を最適化（通信時間を最小化）するためには、適切なパケット間ギャップの導出がポイント

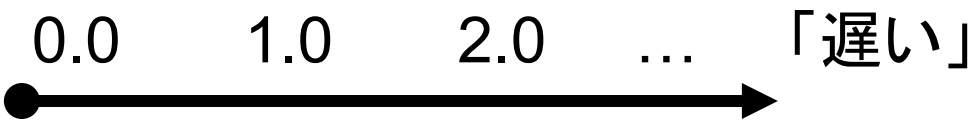
これまでにわかったこと

- パケットペーシングの有効性
  - 通信アルゴリズムに応じたペーシング効果
    - 通信アルゴリズムによっては通信帯域をほぼ最大に利用
  - メッセージ長やノード数によるペーシング効果の増加
    - ポストペタ時代の通信基盤技術となりうる
- パケットペーシングの弱点
  - 通信開始時刻のインバランス
    - 通信アルゴリズム、トポロジ、ノード数によって異なる
    - 場合によってはペーシングの効果が損なわれることも

ただし、これらはシミュレーションでのお話

目的、方法、評価指標

- 目的
  - 実機におけるパケットペーシングの有効性を実証
- 方法
  - パケットペーシング可能なシステムを用いて、パケット間ギャップを変えた通信プログラムを実行
- 評価指標
  - ペーシング無しの通信時間を基準(1.0)とした、ペーシング有りの通信時間の比 ⇒ ペーシング効果



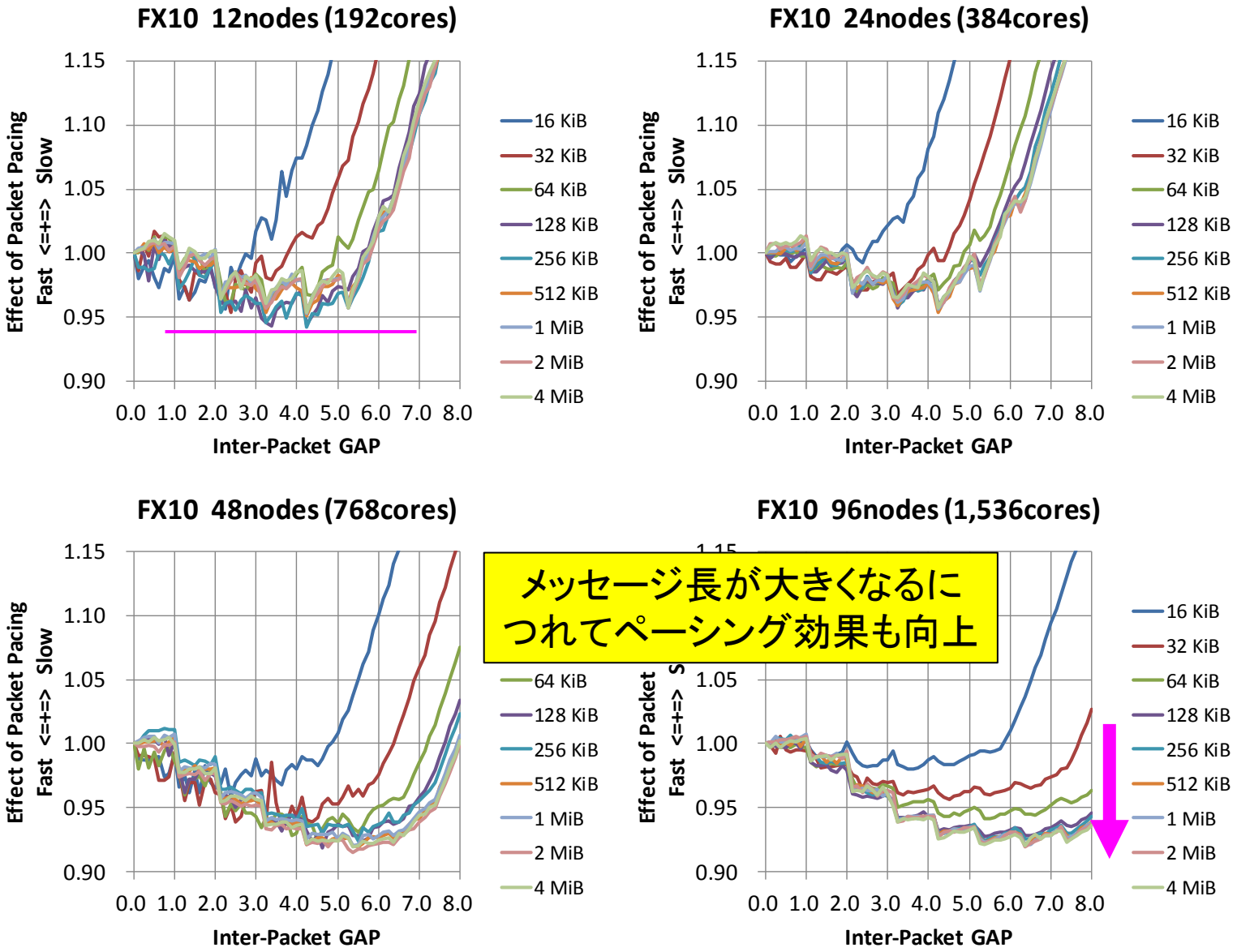
評価実験

- 実験環境
  - パケット転送間隔を変更可能なシステム
  - 富士通 PRIMEHPC FX10
    @九州大学情報基盤研究開発センター
- 実験1：ランダムリング通信
  - HPC Challengeより
  - ランダムに選んだランクの並びでリングを形成し、
    隣接するプロセス同士で1対1通信を行う
- 実験2：全対全通信
  - Pairwise exchangeアルゴリズム
  - MPICHやOpenMPIなどの多くのMPI実装で採用

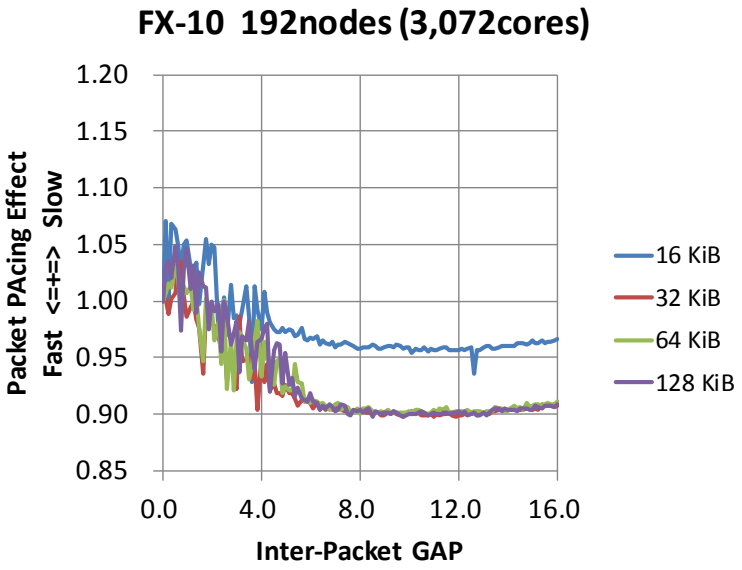
実験1：ランダムリング通信

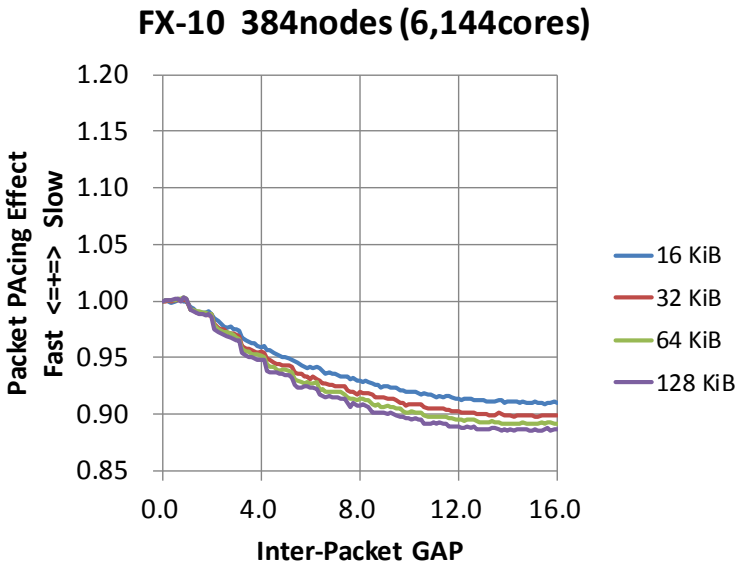
- MPIライブラリ（MPI_～）で記述
- ノード数：12，24，48，96，192，384，768
- プロセス数：ノード数×16コア
- パケット間ギャップ：0（ペーシング無）～16.0
  - ジョブオプションで指定
- メッセージサイズ：16KiB～4MiB
  - パケットヘッダサイズ含む

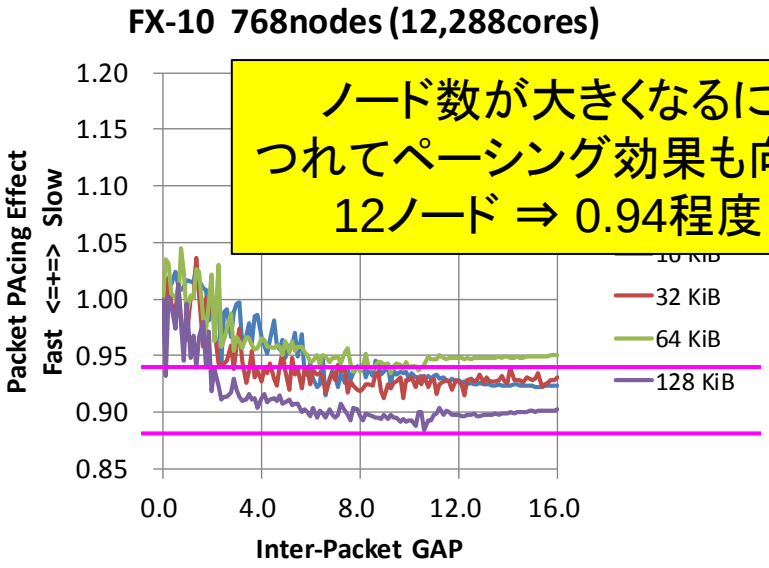
ランダムリング（12～96ノード）



ランダムリング（192～768ノード）







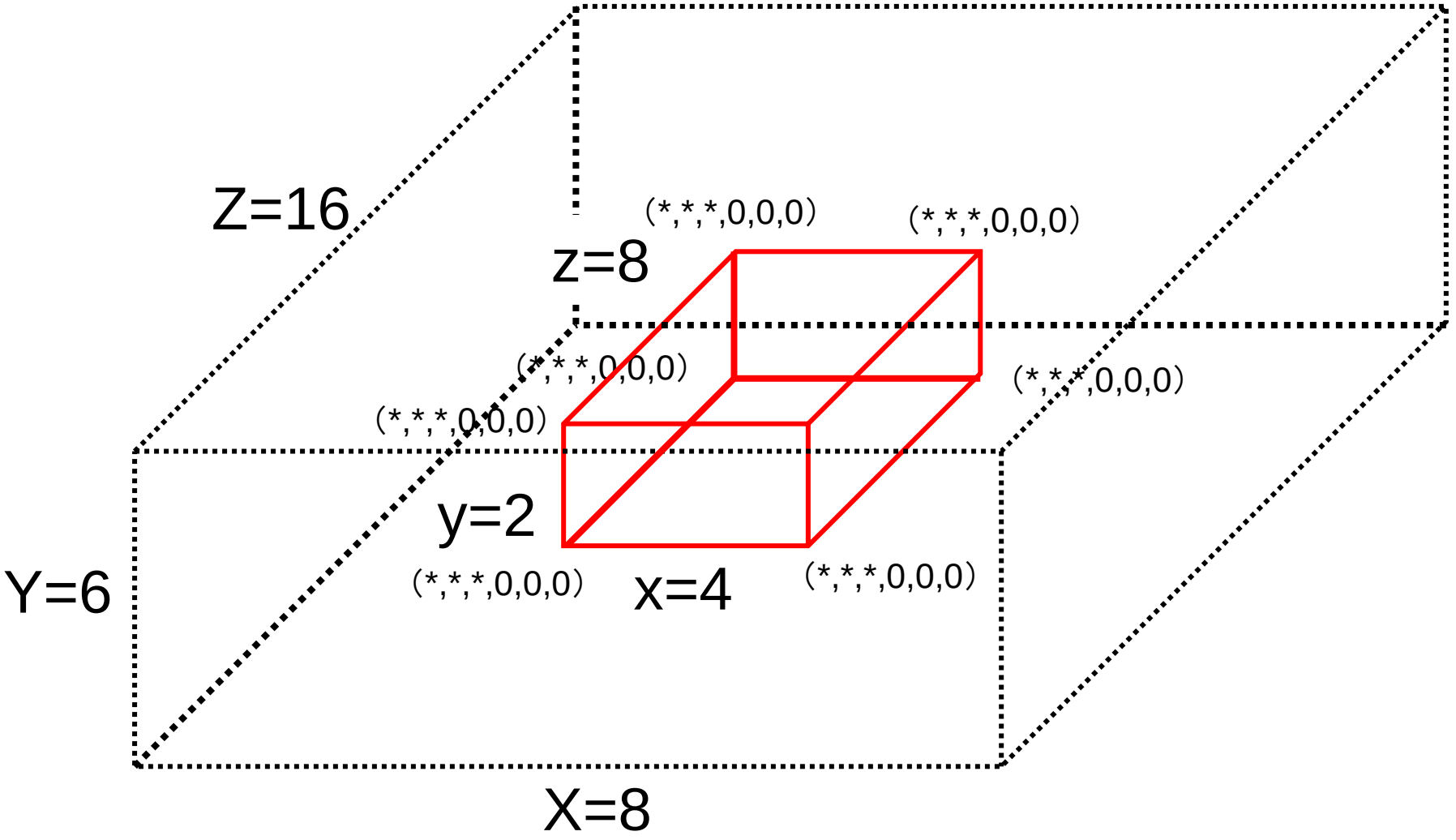
ノード数が大きくなるにつれてペーシング効果も向上
12ノード ⇒ 0.94程度

実機での
ペーシング効果を確認

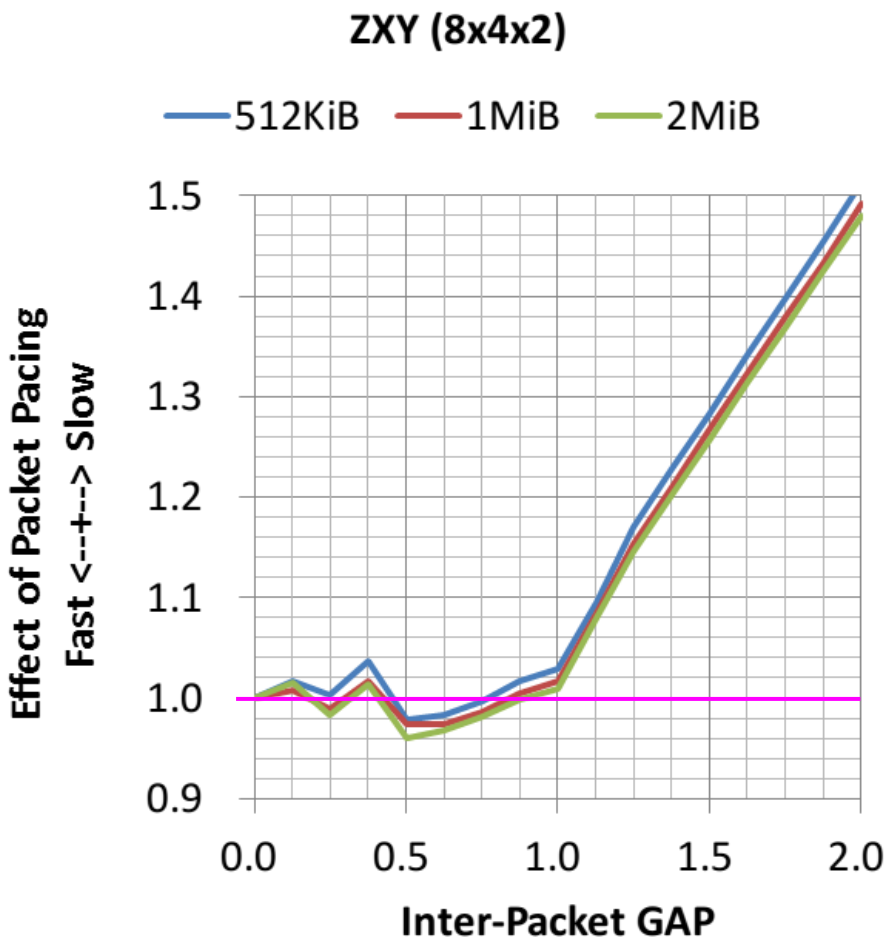
全対全通信

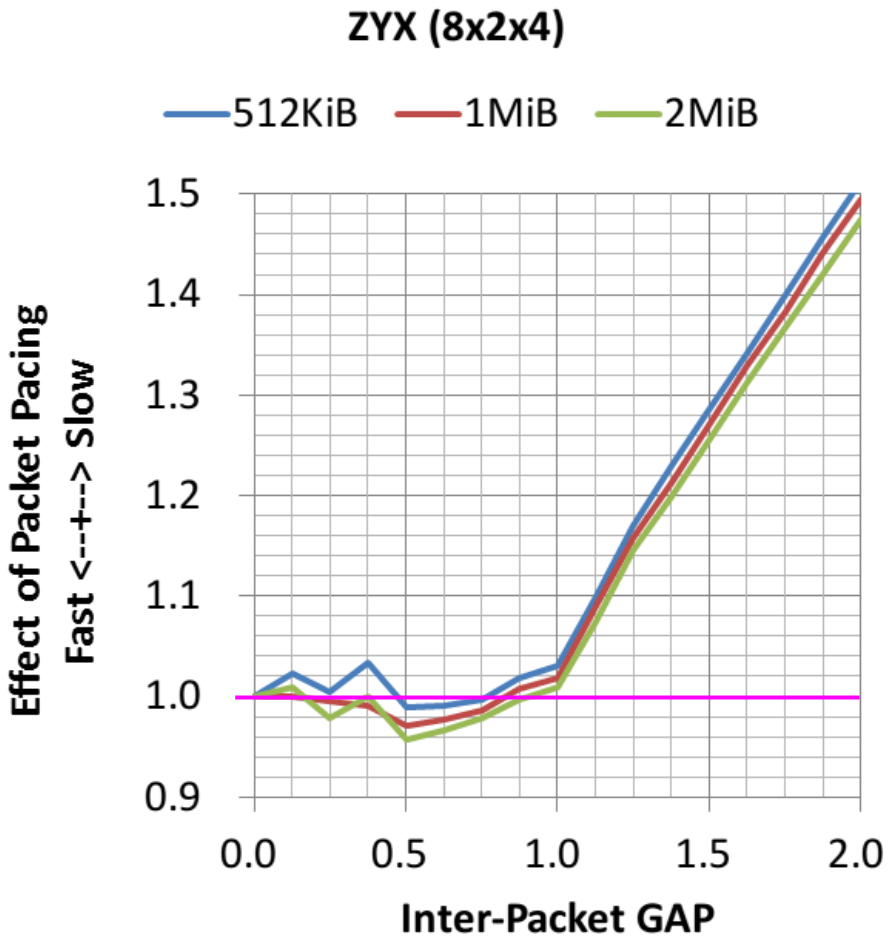
- Tofuライブラリで記述（RDMA put）
  - パケットペーシングの挙動を変える要因を排除
    - 要因：通信モード（高速型、省メモリ型）や通信方式（Eager、ランデブー）の切替
- ノード数：64／768
  - （X,Y,Z,0,0,0）のみ使用⇒ 物理3次元メッシュ・トーラス網（4,2,8）を構成（Z軸はメッシュ）
- プロセス数：64（1プロセス／ノード）
- パケット間ギャップ：0（ペーシング無）～2.0
- メッセージサイズ：512KiB、1MiB、2MiB
- ランクの割り当て方（3軸に割り当てる順番）：
  XYZ（4×2×8）、YXZ（2×4×8）、XZY（4×8×2）、
  YZX（2×8×4）、ZXY（8×4×2）、ZYX（8×2×4）

64ノード 3次元メッシュ／トーラス網



全対全通信(FX10)

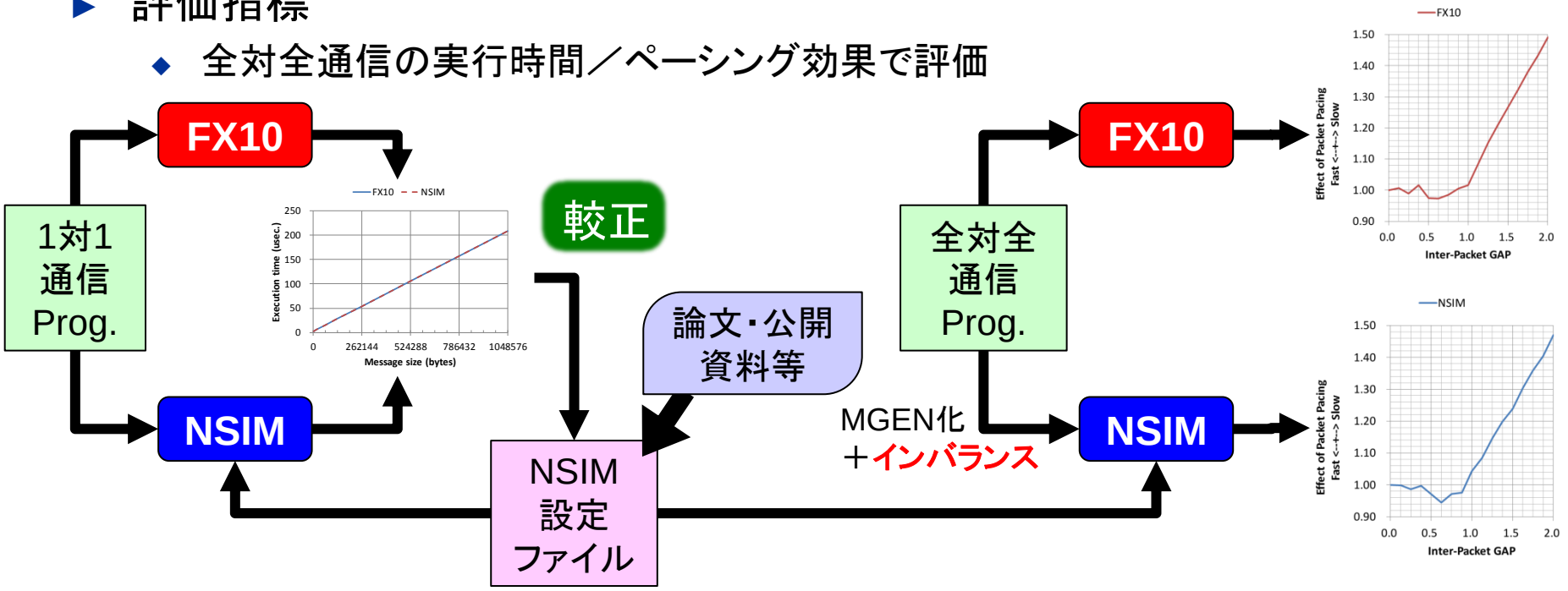




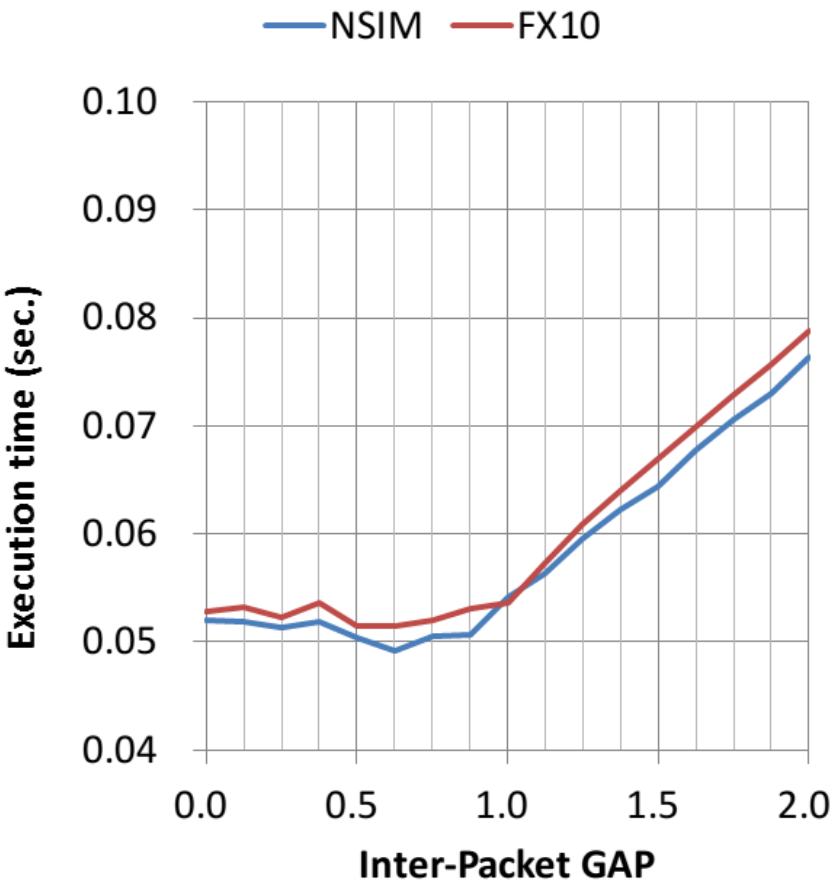
実機でのペーシング効果を確認

FX10とNSIMの評価結果を比較

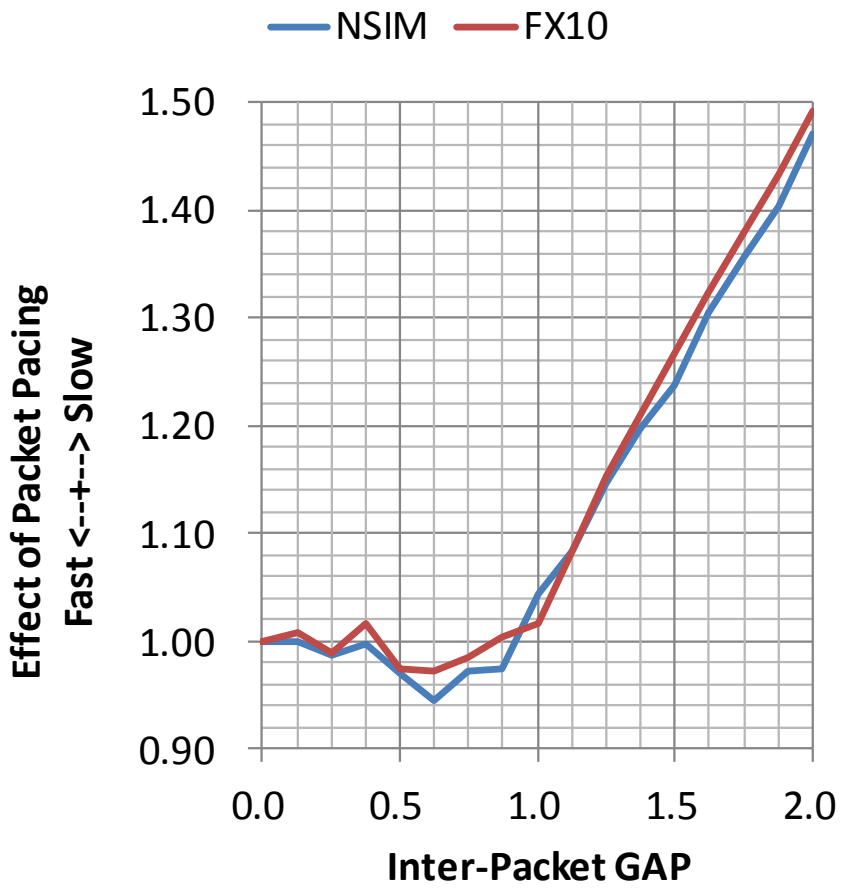
- 目的
  - NSIMのシミュレーション精度の確認
- 方法
  - 論文・公開資料等からインターコネクト仕様を設定
  - 実機1対1通信からの区間計測値等を利用して較正
  - NSIMに与える設定ファイルを精密に作成 ⇒ 全対全通信をFX10とNSIMで実行
- 評価指標
  - 全対全通信の実行時間／ペーシング効果で評価



FX10とNSIMの比較



実行時間



ペーシング効果

実機に近いシミュレーション精度を達成

考察

- ランダムリング
  - 実機におけるペーシングの有効性を確認
  - ノードサイズ、メッセージサイズが大きくなるとペーシングの効果も大きくなることも確認 ⇒ ポストペタに有効
- 全対全通信
  - 64ノードは規模が小さくペーシング効果が低い
    - シミュレーションでは8K、16Kノードを対象
    - X軸（=4）、Y軸（＝2）ではペーシングは効かない
    - Z軸（=8）で若干作用するが、PWXでは機会が少ない
  - 実機評価とシミュレーション評価で同等の挙動を確認

関連研究

- TCP輻輳制御
  - 輻輳ウインドウサイズの制御（RTTとメッセージ長で計算）
  - スロースタート、加算増加、高速回復
  - TCP Tahoe, Reno, etc.
  - 関連論文多数
- ECN（Explicit Congestion Notification）の利用
  - 輻輳を検出すると送信先から送信元に輻輳発生を通知
  - HPC分野でECNを利用した輻輳回避の研究は少ない
    - IBにおけるECNを用いた輻輳回避（メラノックス社）

本研究で主張するところ

- パケットペーシング
  - 新規技術ではない
  - インターネットワーキングにおけるオープンな技術
  - 本研究ではクローズドなインターコネクトに適用
- 輻輳回避
  - 従来研究
    - 輻輳が発生してから対処
    - RTTやECNに対する反応を速くすることが難しい
      ⇒「パッシブ・ペーシング」
  - 本研究
    - 輻輳を発生させないように、パケット間ギャップを事前に算出
    - 積極的にペーシングを行う
      ⇒「アクティブ・ペーシング」
- 同様の研究は現在確認できていない

まとめ

- これまでの研究
  - 「シミュレーション」による通信性能評価
- 本発表では
  - 「実機（FX10）」における通信性能評価
    - ランダムリング通信と全対全通信を実行

⇒ 実機におけるパケットペーシングの効果を確認

⇒ メッセージ長やノード数によるペーシング効果の増加

今後の課題

- プログラムの「実行時」にホップ数に応じてパケット間ギャップを変えるMODペーシングの実機評価
  - シミュレーションでは効果大
- さらにノード数の大きなシステムで評価（したい）
  - 理研「京」
    - 応募が通るのか？ 通信のみなので調査シートが空白に。。。
  - 東大「FX10」
    - 有望か？