

汎用品を用いてスーパーコンピュータ のネットワークを作る

鯉渕 道紘

国立情報学研究所

アーキテクチャ科学研究系

イーサネット

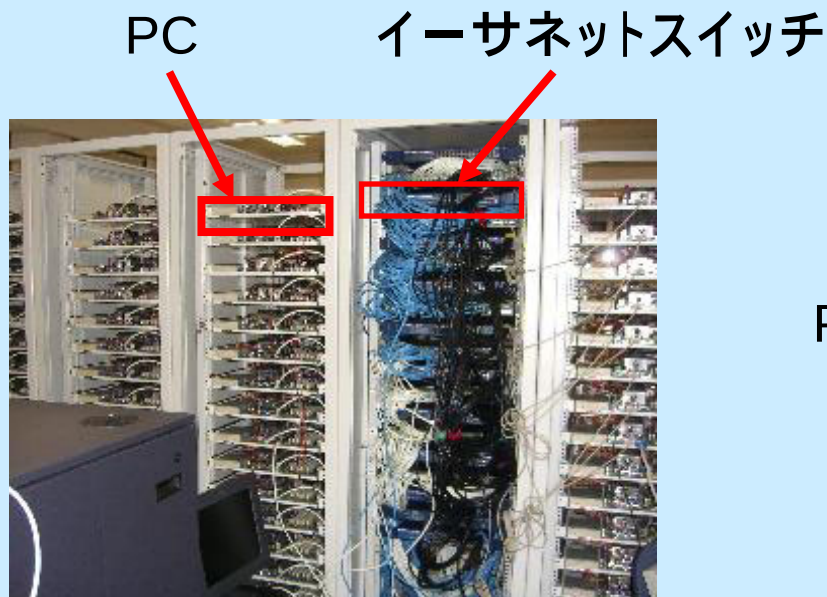
- 世界中の家庭、オフィスで一般的に使われているLAN(ローカルエリアネットワーク)
 - インターネットへの接続に利用(ADSL, B-Flets, CATV)
- 10M/100M/1G/10G/100Gbpsの帯域
- 数千円～数万円で購入可能



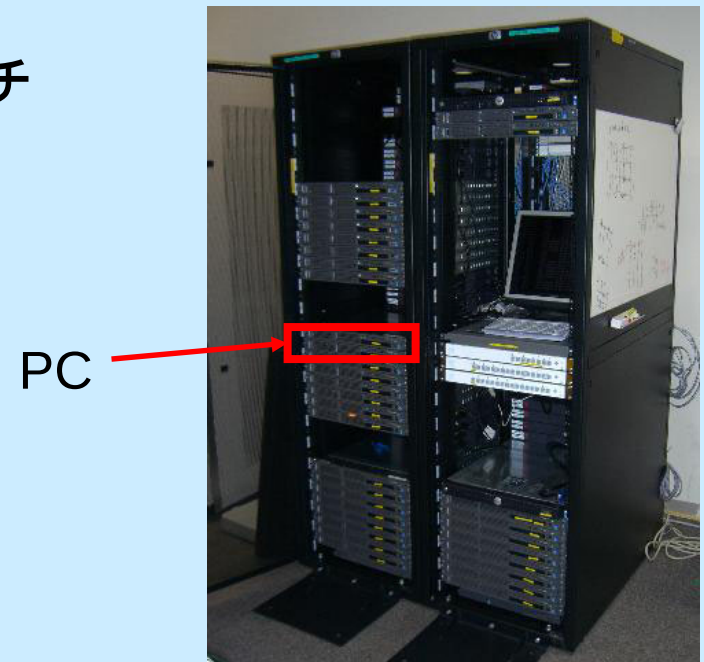
GbEスイッチ

PCクラスタ

- 安価な汎用の製品で構成した大規模計算機
 - パーソナルコンピュータ(PC) + イーサネットが一般的
 - バンド幅 (GbE, 10GE, 100GE) は専用NWに匹敵
- 目指せ！スーパーコンピューティング



同志社大学

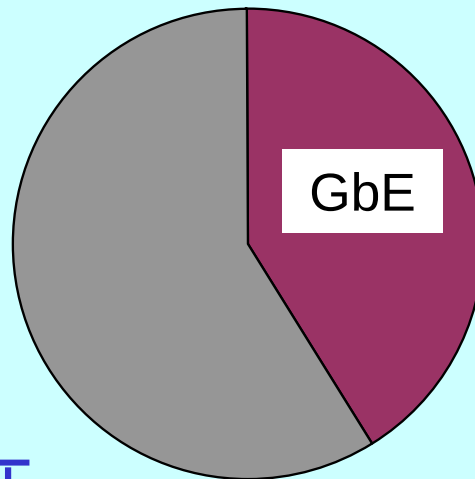


NII

スーパーコンピューティングの可能性

- インタコネク트에イーサネットを用いたPCクラスタ
 - 専用ネットワーク (SAN) よりも安価
 - バンド幅 (GbE, 10GE, 100GE) は SAN に匹敵

トップ500世界ランキングのスーパーコンピュータの内訳



2007年6月現在

(<http://www.top500.org/stats/list/29/conn/>)

1つのシステムに何台のスイッチが必要か？

- 並列システム構築の目安：1Gflops 1 Gbps
- Intel Core2 Duo E6700 2.66GHzだと15Gflops
 - ノードスイッチ間は、複数リンクが必要
 - 10GEスイッチ：数十ポート、GbEスイッチ：数百ポート
- バランスの取れたHPCシステムを作るためには、多数のスイッチが必要！！

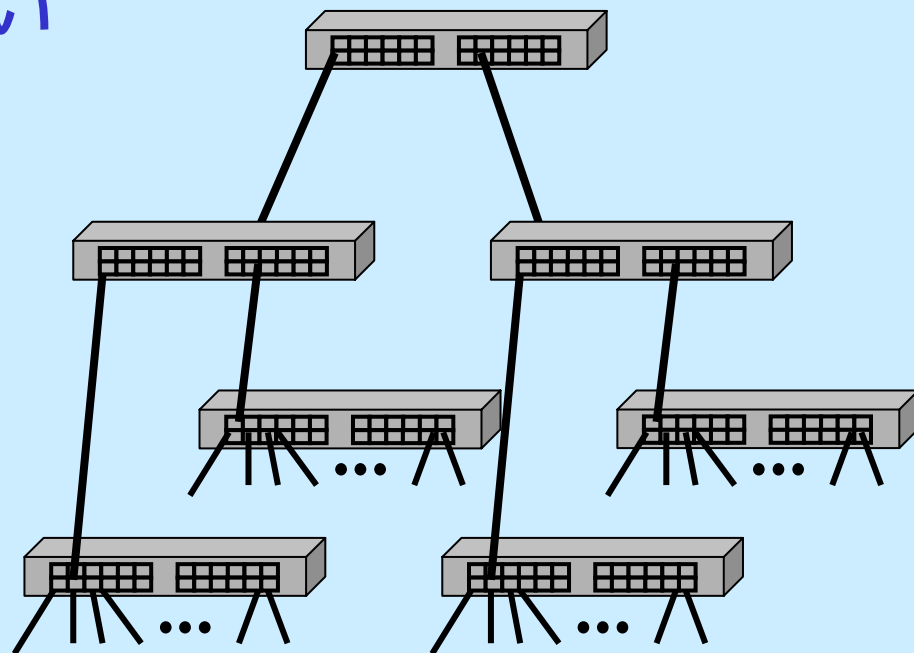


現状の問題点

- 超並列計算機は3Dトーラスなどの密結合トポロジ
- でも、イーサネットのトポロジは木構造のみ
 - ループ構造は不可
 - スパニングツリープロトコル (STP)
 - スケーラビリティに乏しい
 - リンク集約化も、

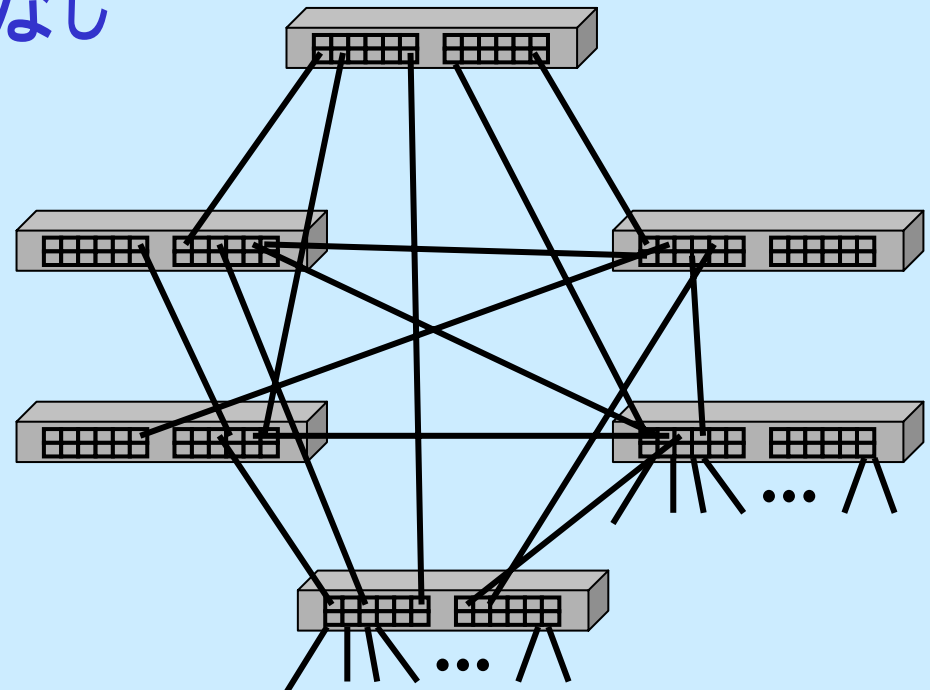


Blue Gene/L



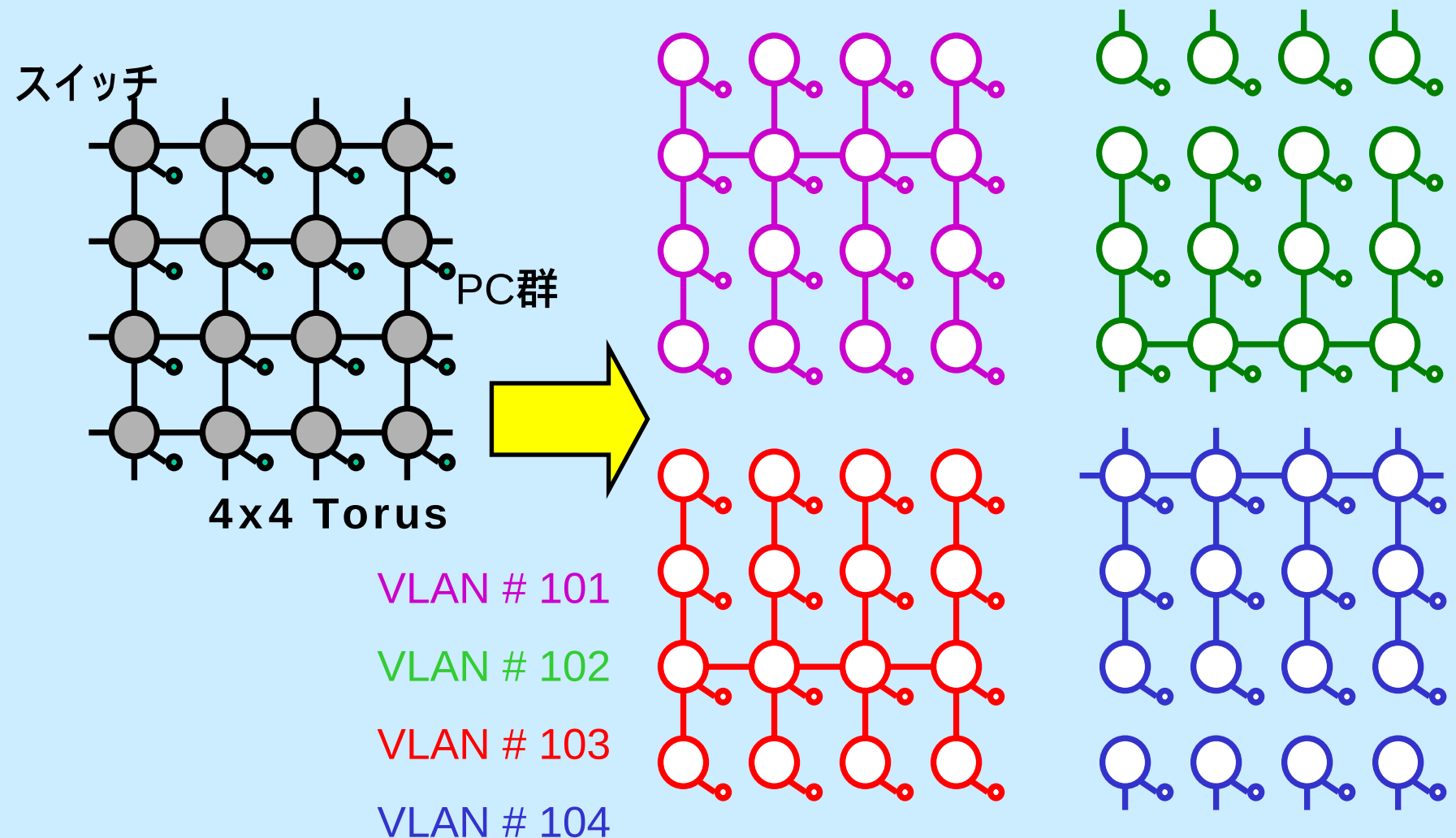
本研究の目的

- あらゆる大規模トポロジを既存の**商用イーサネットスイッチの機能を利用して実現する。**
 - VLAN(virtual LAN)を利用
 - VLANルーティング法 [工藤ら, 2004]
 - ホスト側は変更の必要なし (VLANに未対応でOK)



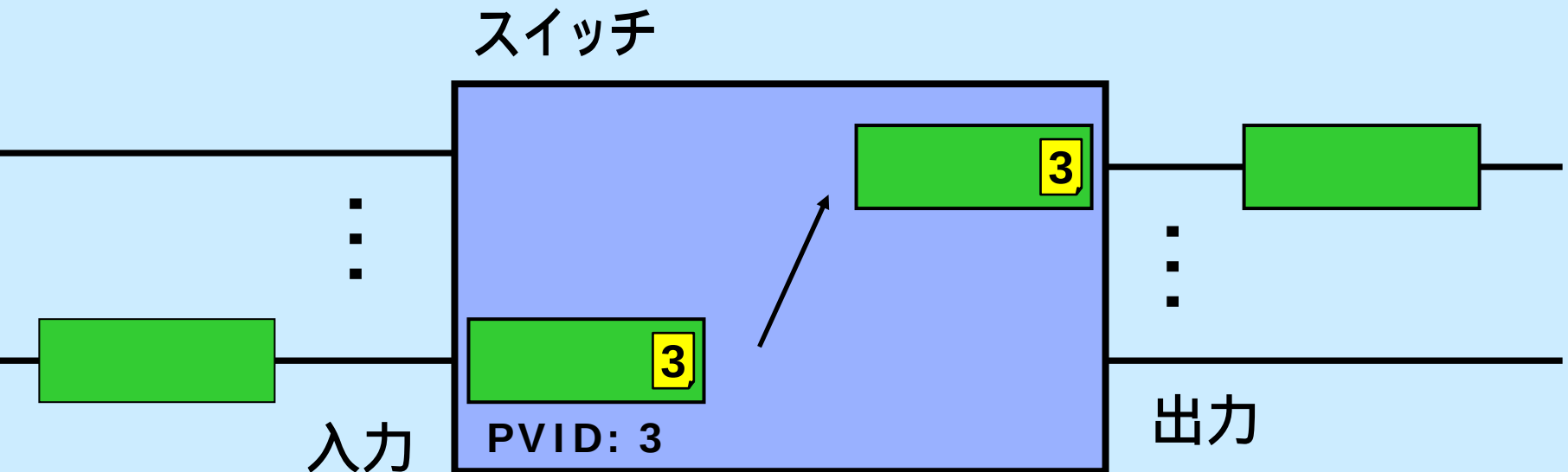
VLANルーティング法

- 複数のツリーを組み合わせてスパコンと同じトポロジ、ルーティングを実現



イーサネットスイッチの動作

- ホストではなく、スイッチでVLANタグ付けを行う
 - フレーム入力時に、PVIDタグを付加
 - フレーム出力時に、VLANタグを除去
- ホスト側のソフトウェアに対する変更は不要



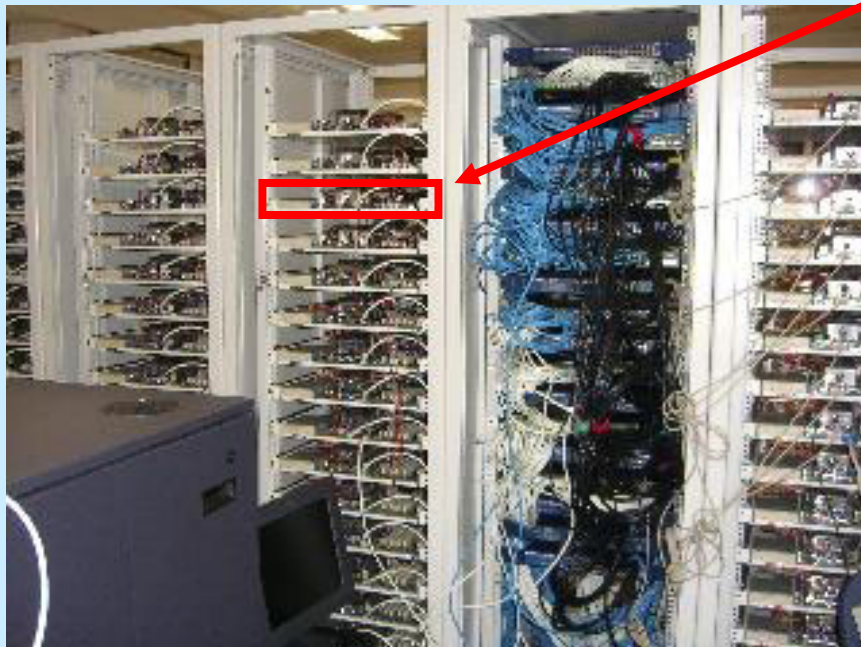
スパコンの実現性: VLAN数

	単純な割当て	局所性利用
Fat ツリー (u,d,r)	u^r	u
メッシュ (k-ary n-cube)	k^{n-1}	n
トーラス (k-ary n-cube)	$2k^{n-1}$	$3n$

- トポロジの次数以内のVLAN数で実現可能
- スパコン並の大規模クラスターコネクが構成可能!
- VLAN技術による遅延、帯域の低下は極めて小さい

実装と評価

- 同志社大学の大規模PCクラスタにおいて運用へ
 - Top500世界ランキングにおいて93位(2003/11)
- NAS並列ベンチマーク



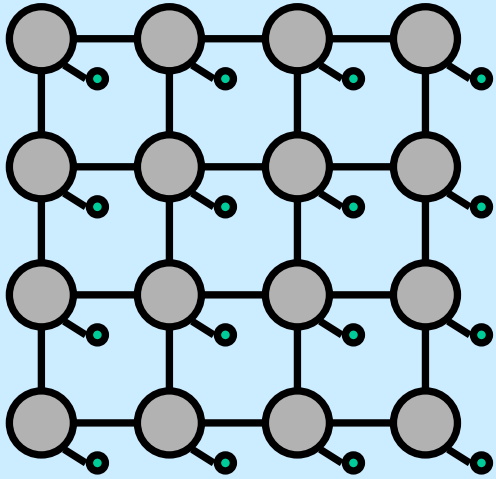
大規模PCクラスタ、同志社大学

PC

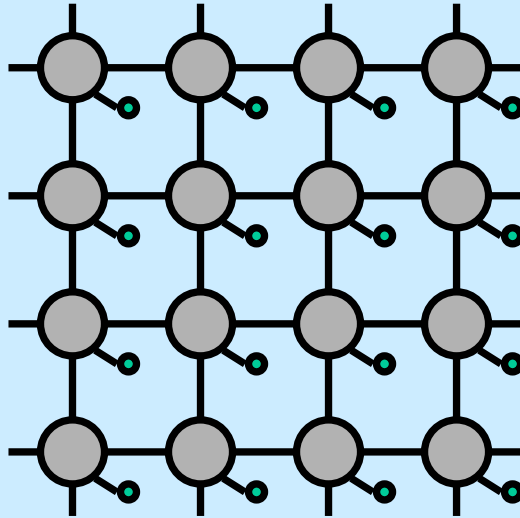


中規模テストベッド、NII

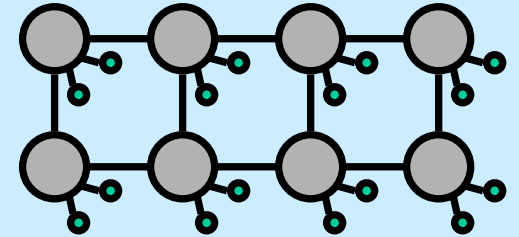
実装したトポロジ



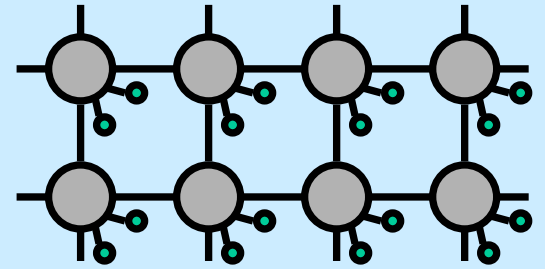
4x4 Mesh



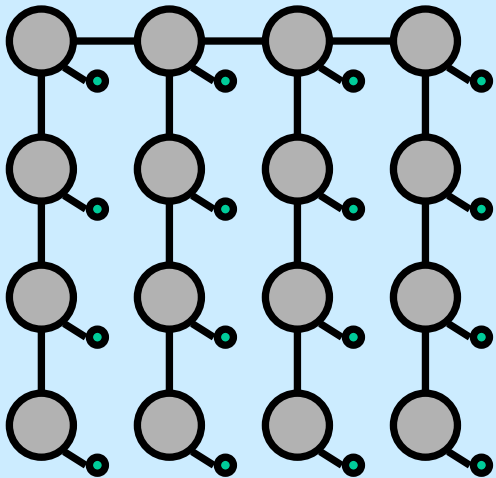
4x4 Torus



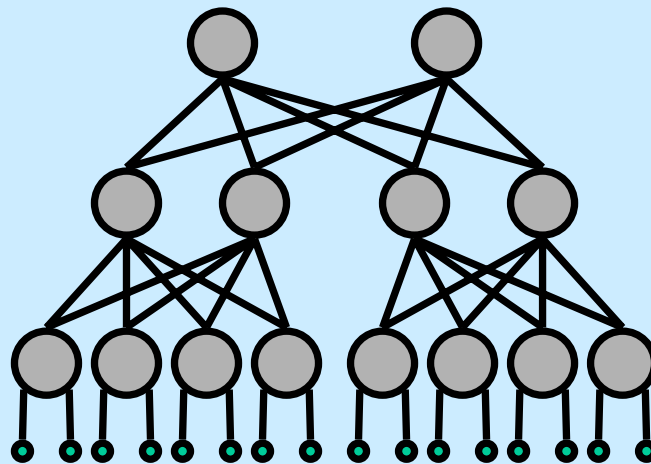
4x2 Mesh



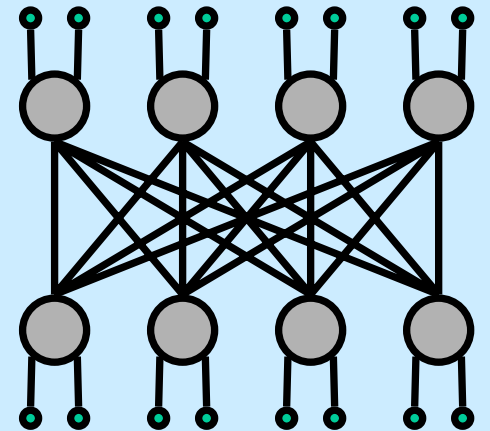
4x2 Torus



Tree

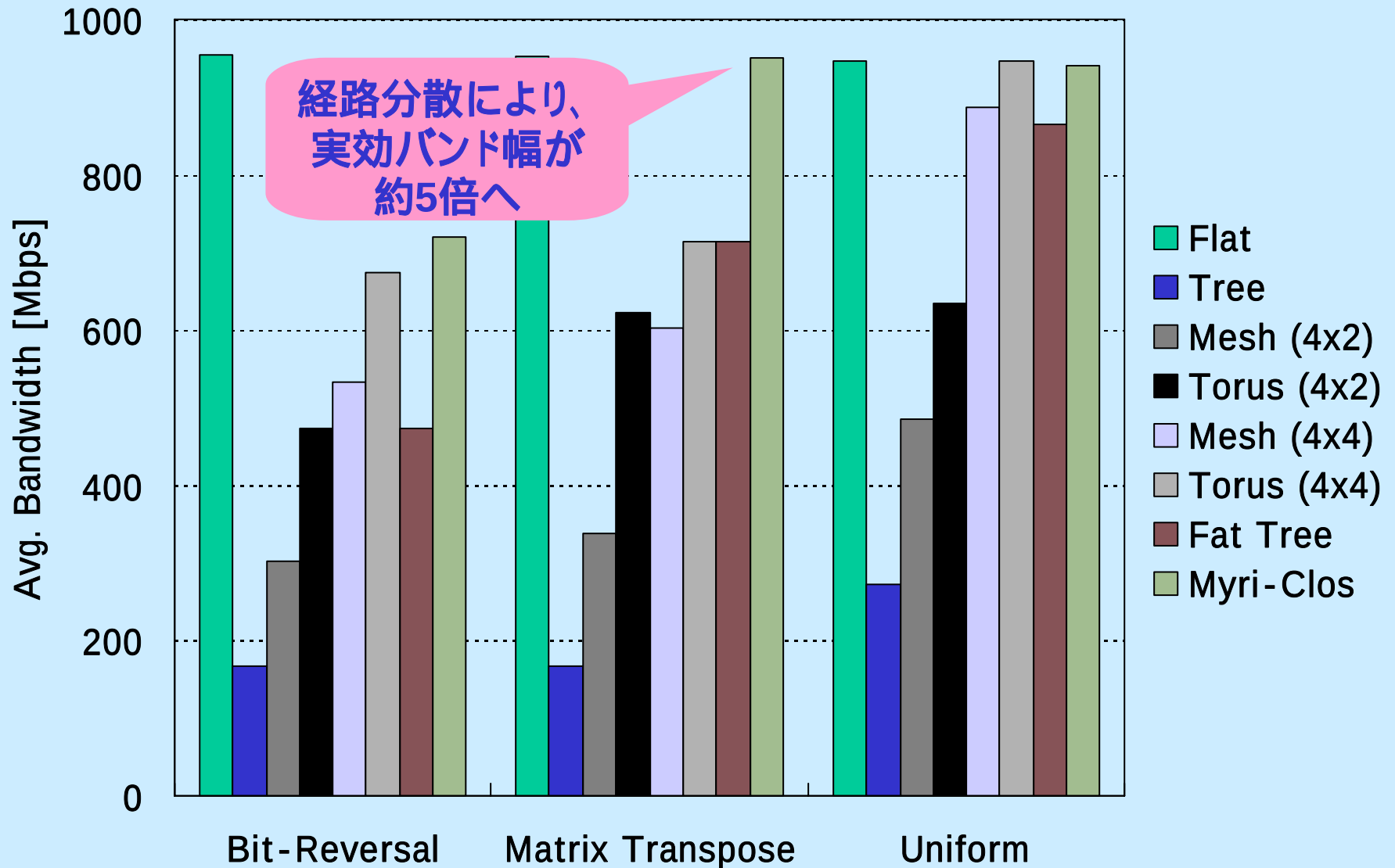


Fat Tree (2,4,2)

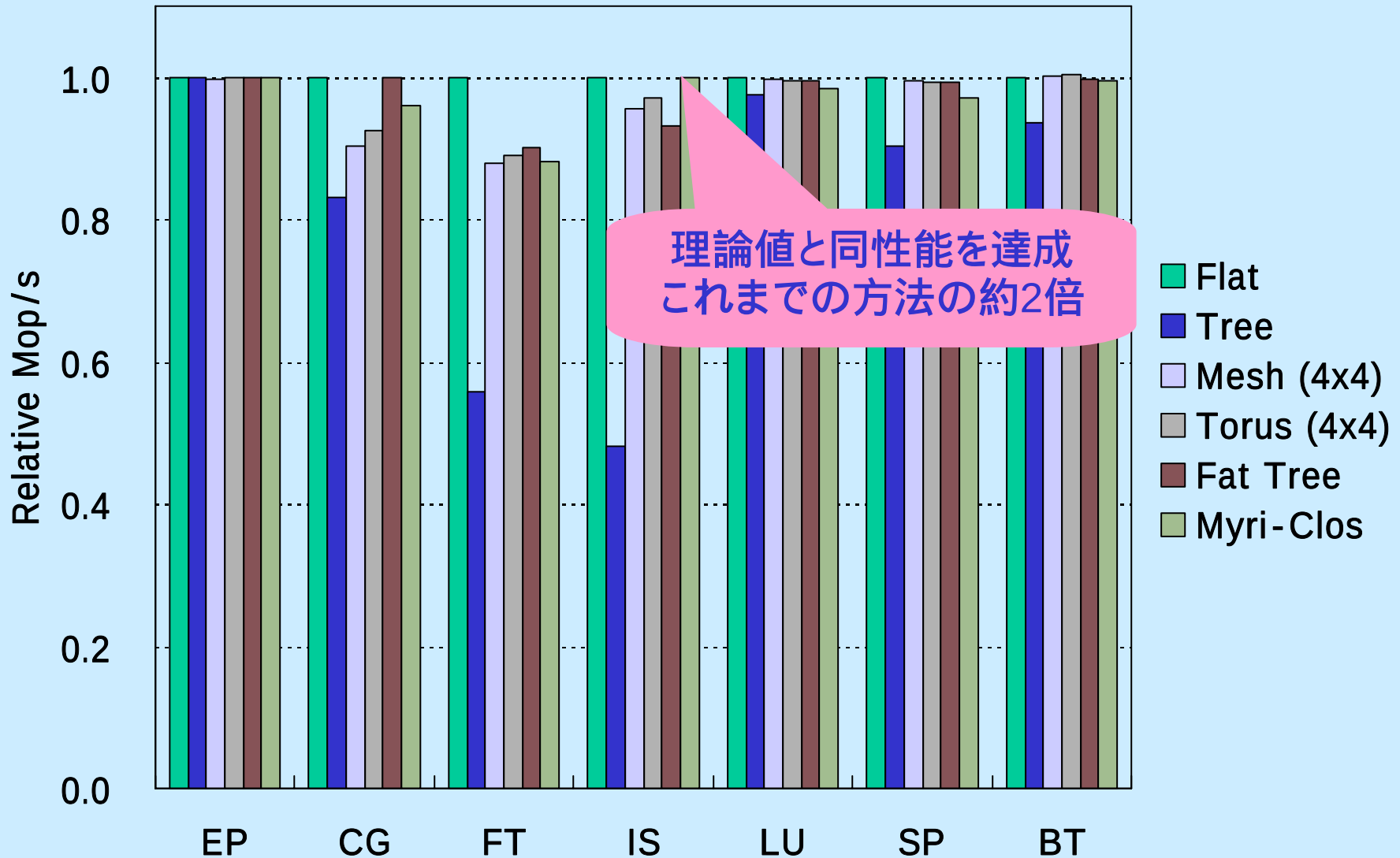


4x4 Myrinet-Clos

トラフィックパターンの結果



NAS並列ベンチマークの実行結果



まとめ

- PCクラスタはもはやスーパーコンピュータの王道！
- スーパーコンピュータの結合網と同じトポロジ、ルーティングをPCクラスタにおいて実現する手法を考案、実装、評価
- 既存のPCクラスタにそのまま適用可能
 - **すでに運用段階の技術であり、追加コスト0で性能向上**
 - 同志社大学大規模PCクラスタ(世界93位(2003/11), NIIテストベッド
- 世界中のPCクラスタは本技術の適用を検討する時

	これまで	提案手法
トポロジ	ツリー	トーラス、hypercube等
経路数	1	複数
VLAN数	1 (無)	数個程度