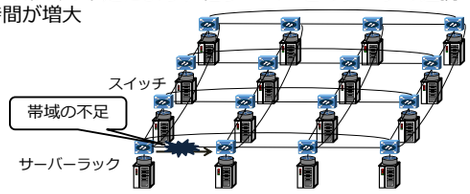
データセンターにおける輻輳回避のためのルーティング用論理トポロジ構築手法

下間　雄太（阪大）
大下　裕一（阪大）
村田　正幸（阪大）

データセンター

多数のサーバーとサーバー間を結ぶネットワークで構成

- サーバー間の連携により多量のデータを処理
  - 分散ファイルシステム、分散コンピューティングなど
- ネットワークが処理性能に与える影響大
  - サーバー間の帯域の不足により、他サーバーとのデータの連携にかかる時間が増大



データセンターネットワークにおける問題と本研究の目的

- トラヒック変動が頻繁に発生
- 故障が頻繁に発生

- 環境変動発生時でもサーバー間に十分な帯域を確保することが必要

研究の目的：
データセンターネットワークにおいて頻発する環境変動に対応してサーバー間に十分な帯域を確保する経路制御手法の確立

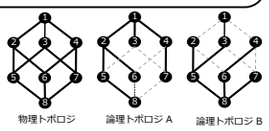
- トラヒック状況に応じて、通信サーバー間に十分な帯域を確保可能な経路を確立
- 頻繁な環境変動に瞬時に対応可能な自律分散型経路制御

環境変動に対応した自律的な経路制御手法

アプローチ

複数の論理トポロジを用いた自律的経路制御手法による輻輳箇所の迂回

- 論理トポロジを複数構成
  - 各論理トポロジは物理ネットワークの全ノードと一部のリンクから構成
  - いずれの論理トポロジを用いても全ノードに転送可能



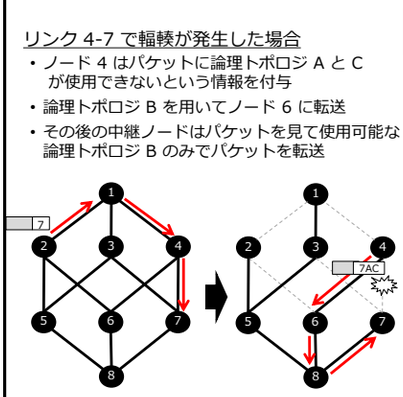
- 各ノードが論理トポロジを自律的に選択することにより、経路を決定
  - 輻輳箇所を迂回
  - 負荷分散の実現

各ノードにおける論理トポロジを用いた制御

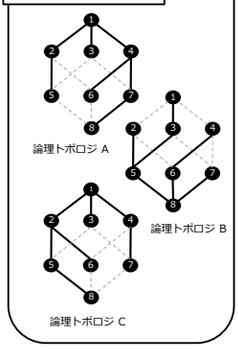
1. 利用可能な論理トポロジの集合を取得
   - ヘッダの参照により利用不可な論理トポロジを除外
   - 自身が接続しているリンクが利用不可な場合，当該リンクを含む論理トポロジを除外
     - 自身が検知した利用不可な論理トポロジの情報をパケットに付加
2. 候補論理トポロジの集合を取得
   - 利用可能な論理トポロジのうち，宛先までのホップ数が最短の論理トポロジ
3. 制御に用いる論理トポロジの選択
   - 次ホップのリンク負荷が最小の論理トポロジ

論理トポロジを用いた自律的な経路制御の動作例

リンク 4-7 で輻輳が発生した場合
- ノード 4 はパケットに論理トポロジ A と C が使用できないという情報を付与
- 論理トポロジ B を用いてノード 6 に転送
- その後の中継ノードはパケットを見て使用可能な論理トポロジ B のみでパケットを転送



ノードが保持する論理トポロジの集合



論理トポロジが満たすべき性質

1. 全ノード間に十分な迂回経路が存在
   - 輻輳が発生した場合にも、代替経路を用いた通信路を確保することが必要
2. 故障の影響が大きい機器は存在しない
   - 特定の機器の故障により、通信帯域が激減することは避けることが必要
3. 迂回経路上でトラヒックが集中しない
   - 迂回先でのトラヒック集中により、サーバー間に確保できる帯域が制限されることは望ましくない

要件を満たすような論理トポロジの集合を構成する手法を検討

論理トポロジの設計手順

1. 論理トポロジの候補の集合を生成

   論理トポロジ候補：
   物理トポロジ内のリンクのサブセットで構成された木構造の全パターン
   - 各論理トポロジに含まれるリンク数は最小
     →利用不可なリンクが生じた場合に、
     利用不可なリンクが各論理トポロジに含まれる確率は小
2. 論理トポロジを選択し、経路制御に用いる論理トポロジの集合へ追加
   - 各候補論理トポロジ追加時の論理トポロジ集合の適切性に関する指標を計算しもっとも良い論理トポロジを選択

指標

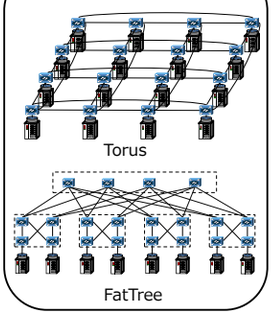
1. 十分な迂回経路数：サーバー間の迂回経路数
2. 機器の故障の影響：平均ノード媒介中心性の最大値
   $B_n^{node} = \left|\frac{1}{G}\right| \sum_{g \in G} \left(\sum_{s,d \in S} \frac{R_g^{node}(s,n,d)}{R_g(s,d)} \right)$
3. 故障迂回時の輻輳可能性：平均リンク媒介中心性の最大値
   $B_l^{link} = \left|\frac{1}{G}\right| \sum_{g \in G} \left(\sum_{s,d \in S} \frac{R_g^{link}(s,l,d)}{R_g(s,d)} \right)$

シミュレーション評価

評価環境

- 評価対象トポロジ
  - 4 ポートのスイッチ 16 台で構築された Torus
  - 4 ポートのスイッチ 20 台で構築された FatTree
- 比較対象経路制御
  - Equal Cost Multi-Path (ECMP) [1]
- 各リンクの帯域
  - 10 Gbps
- トラヒックの発生方法
  - 10 % のサーバーラック間で発生
    - 空き帯域がある限りトラヒックの送信量を増加させていく

評価対象トポロジ



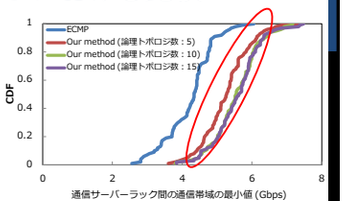
評価指標

- 通信サーバーラック間の通信帯域の最小値 (Gbps)
- 通信サーバーラック間の通信帯域の合計値 (Gbps)

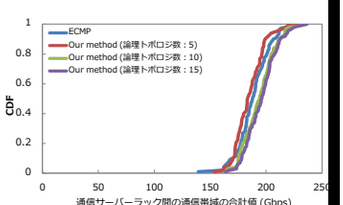
[1] C. HOPPS, "Analysis of an Equal-Cost Multi-Path Algorithm," RFC 2992, Internet Engineering Task Force, Nov. 2000. http://tools.ietf.org/html/rfc2992.

評価結果：Torus における比較

- ECMP より提案手法の方が通信帯域の最小値が大きい
- 少量の論理トポロジでも ECMP より多く通信帯域の最小値を確保できている

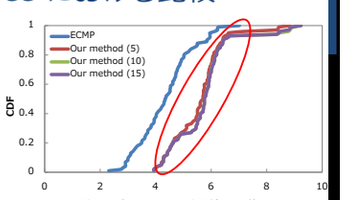


- 合計値も同様に ECMP より提案手法の方が大きい

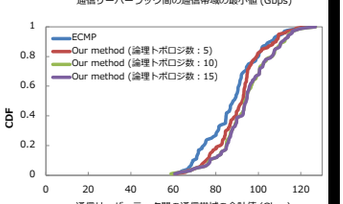


評価結果：FatTree における比較

- ECMP より提案手法の方が通信帯域の最小値が大きい
- 少量の論理トポロジでも ECMP より多く通信帯域の最小値を確保できている

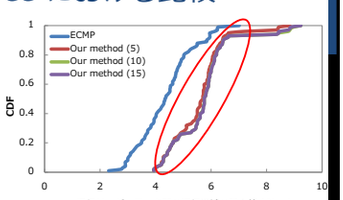


- 合計値も同様に ECMP より提案手法の方が大きい

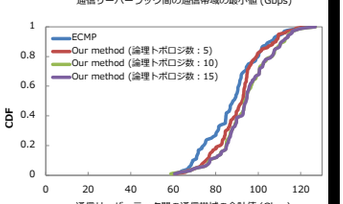


評価結果：FatTree における比較

- ECMP より提案手法の方が通信帯域の最小値が大きい
- 少量の論理トポロジでも ECMP より多く通信帯域の最小値を確保できている



- 合計値も同様に ECMP より提案手法の方が大きい



トポロジに関係なく提案手法の方が ECMP よりも大きな通信帯域を確保できている

まとめと今後の課題

まとめ

- 論理トポロジを用いた自律的な経路制御手法および論理トポロジの構築手法を提案
- Torus において提案手法で ECMP より通信帯域を多く確保できることを確認
- FatTree においても Torus と同様，提案手法で ECMP よりも通信帯域を多く確保できることを確認

今後の課題

- 規模の大きいネットワークでの評価
- 故障が発生した場合の評価
  - ノード故障
  - リンク故障