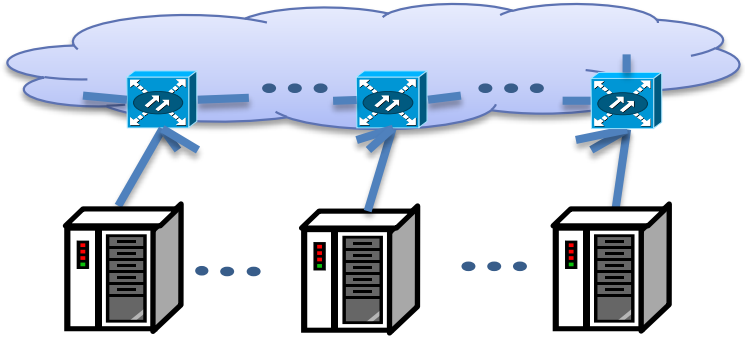
# 光電子融合型パケットルータを用いたデータセンターネットワークの設計

大阪大学 大学院情報科学研究科
大下 裕一
村田 正幸

# データセンターネットワーク

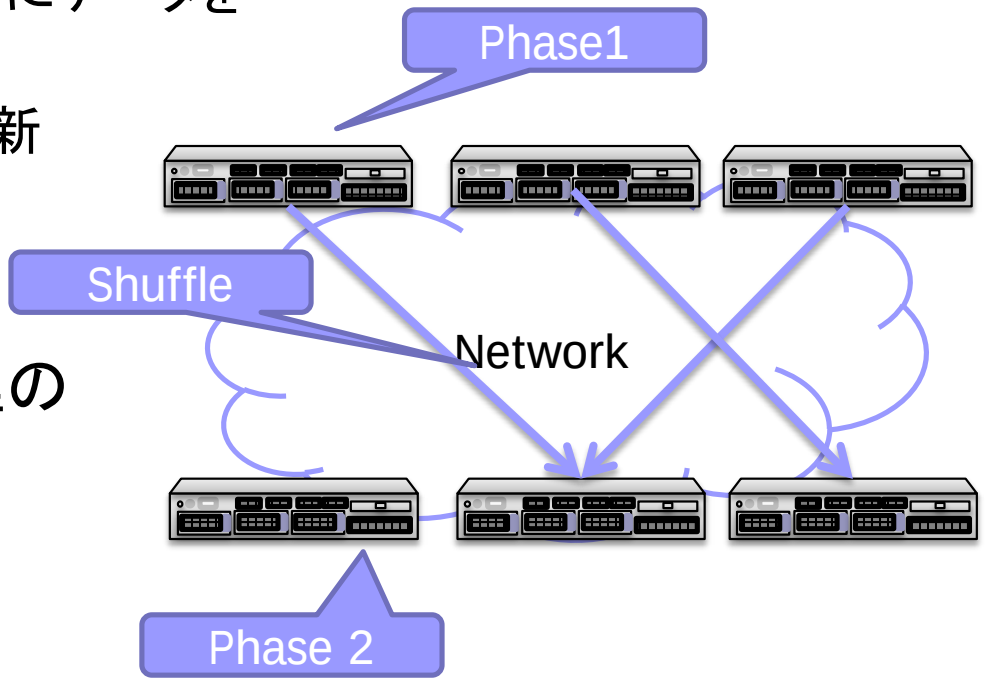
- 多数のサーバーとサーバー間のネットワークで構成
  - サーバーで連携をとることにより、多量のデータを処理
  - 一か所のデータセンターで数万台のサーバーを接続するものも設置

- ネットワークが性能に大きな影響を与える
  - ネットワークの遅延や帯域不足がサーバー間の連携を阻害



# データセンター上のアプリケーションの一例

- データセンター内では、サーバー間の連携によって多量のデータを処理
  - 例：サーチエンジンのバックグラウンド
  - Phase 1:　収集したWEBのキーワードを識別
  - Shuffle: 対応するサーバーにデータを送る
  - Phase 2:データベースを更新

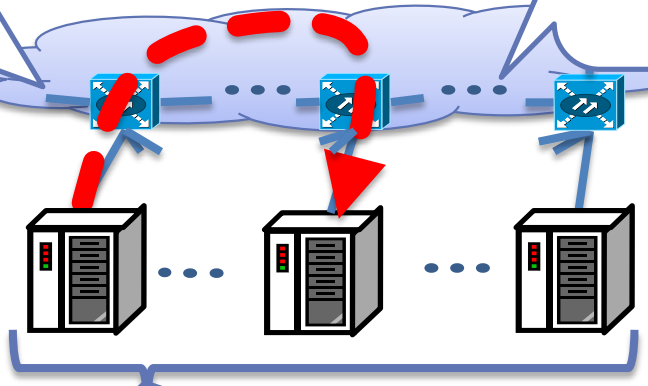



- サーバー間の転送が処理のボトルネックになる可能性

# データセンターネットワークの要求

- サーバー間を低遅延で接続
  - アプリケーションの性能の確保

- 低消費電力
  - 大規模ネットワークにおいても消費電力を抑制



- 多数のサーバの接続
  - 近年構築されている大規模データセンタの同程度の規模までの拡張性は必要

低消費電力性と性能の両立は電気スイッチのみのネットワークでは達成困難

⇩

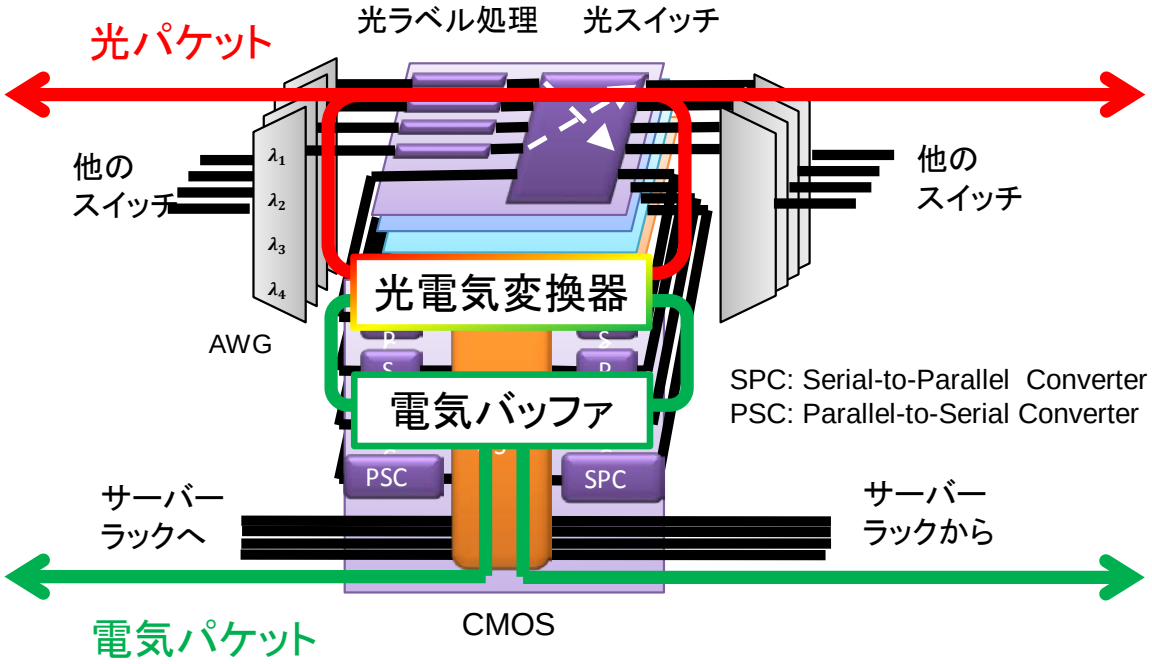
光通信技術の活用

# データセンターネットワークへの光技術の適用

- 光回線交換スイッチ＋電気パケットスイッチ
  - 広帯域の通信を光パスで収容
  - 回線交換スイッチの切り替え
    時間が問題となる場合も
- 全光パケットスイッチ
  - 中でも、バッファーレス
    なスイッチの研究が盛ん
    - ToRスイッチでの
      バッファを活用
  - ただし、タイミングの
    集中制御が必要
- 光電子融合型パケットスイッチ
  - 光スイッチ・電気バッファの組み合わせ
  - パケットの衝突がなければ、光信号のまま中継
  - パケットの衝突が起きたら、電気バッファを活用

# 光電子融合型パケットスイッチ

- 光通信技術と電気技術を融合
  - 光ポートと電気ポートを持つ
    - 光ポートを用いて他のスイッチと接続
    - 電気ポートを用いてサーバラック内の電気スイッチと接続
  - 光パケットと電気パケットを変換するための変換器・電気バッファを持つ



# 光電子融合型パケットスイッチの特徴

- 光通信技術の高性能・低消費電力を維持したまま、光通信技術における制御や実現性の問題を解決
  - パケットの衝突が発生しない場合、光/電気変換が不要で、光パケットをそのまま中継可能

    → 低遅延・低消費電力で通信可能
  - パケットの衝突が発生する場合、電気バッファに一旦保存したのち、再度転送を試みることが可能

    → 大容量光バッファの実現やパケットの衝突回避の制御が不要

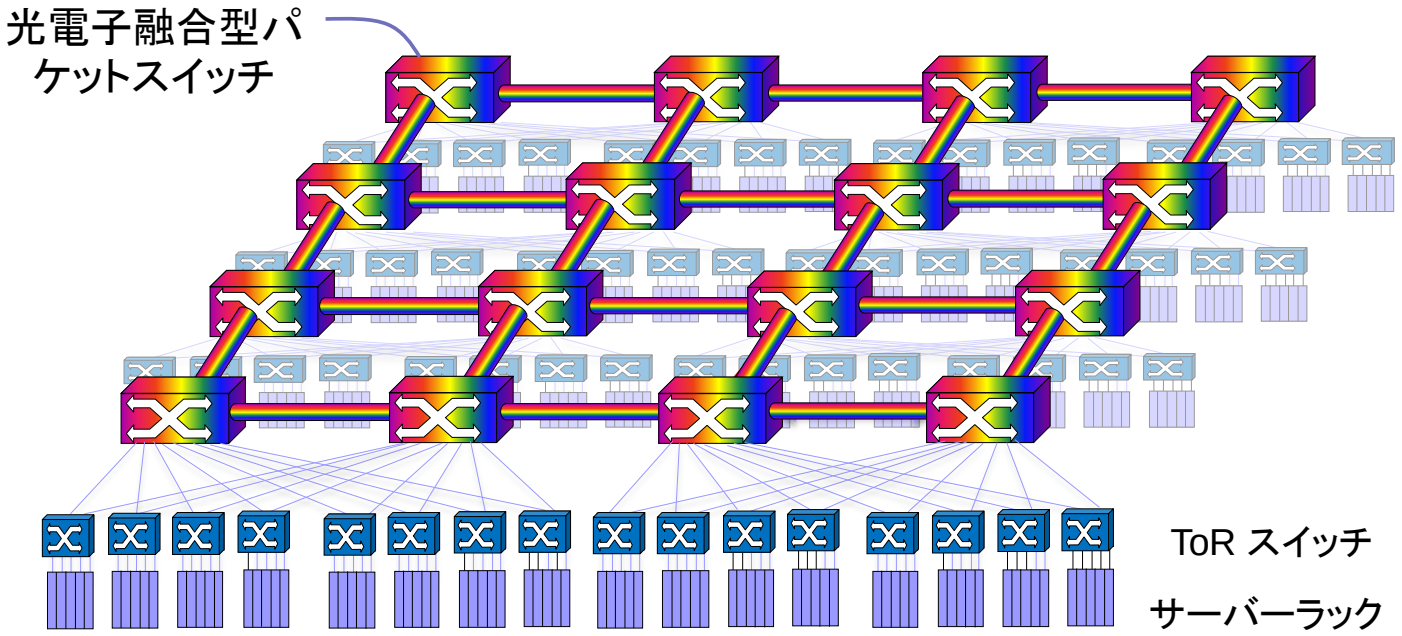
# 本研究の目的

- 光電子融合型パケットルータを用いたデータセンターネットワークに適した構造の構築

- 考慮にいれる点
  - 光電子融合型パケットルータでは、パケットの衝突を回避した経路を設定できることが低遅延なトラヒック収容に直結

# 光電子融合型パケットスイッチを用いたデータセンターネットワークの考え方

- データセンター内のコアネットワーク
  - 各光電子融合型パケットスイッチが多数のサーバラックからの通信を束ねて転送するネットワーク構造
  - 各サーバラックからは、複数の光電子融合型パケットルータに接続し、障害時にも通信経路を確保



# 光電子融合型パケットルータを用いたデータセンター向け経路制御

## 方針

データセンター内の光電子融合型パケットルータの台数は少ない

- 光電子融合型パケットルータ間の候補経路を**事前に計算**
  (最短ホップ＋nホップの経路まで候補として含めて計算)

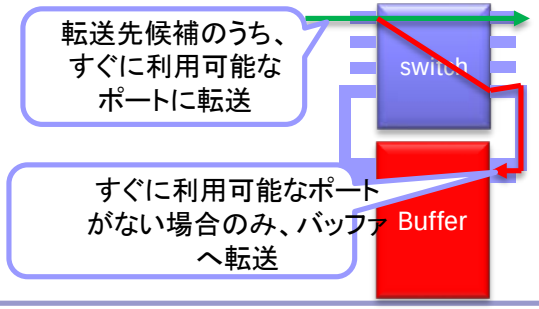
**光電子融合型パケットルータ**:
宛先光電子融合型パケットルータごとに**転送先候補を保持**し、それに従い転送

**サーバラック**:
宛先サーバラックが接続する光電子融合型パケットルータを宛先とした**カプセル化をしてパケット**を送出

| 宛先1 | 次ホップ候補1, 次ホップ候補2, ..... |
|---|---|
| 宛先2 | 次ホップ候補2, 次ホップ候補3, ..... |
| 宛先3 | 次ホップ候補3, 次ホップ候補4, ..... |
| ⋮ | ⋮ |

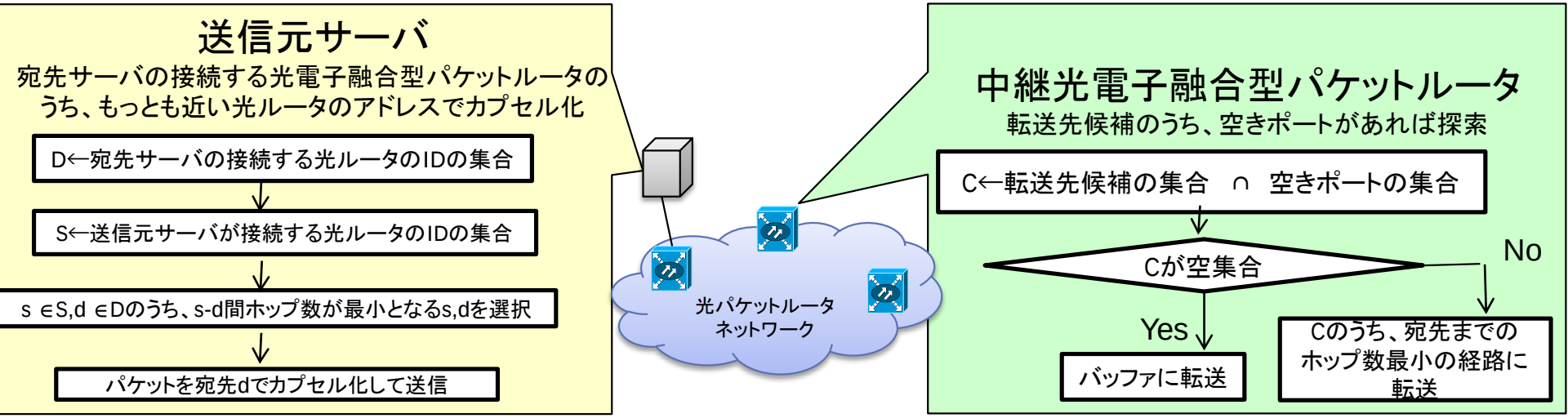
バッファを経由することによる遅延がバッファを経由しない場合と比較して大

- 最短ホップ経路以外にも利用可能な候補を複数保持
- 転送先候補のうち利用可能なポートがあれば転送
- 利用可能なポートがない場合のみバッファへ転送



# 光電子融合型パケットルータを用いたデータセンター向け経路制御

方法

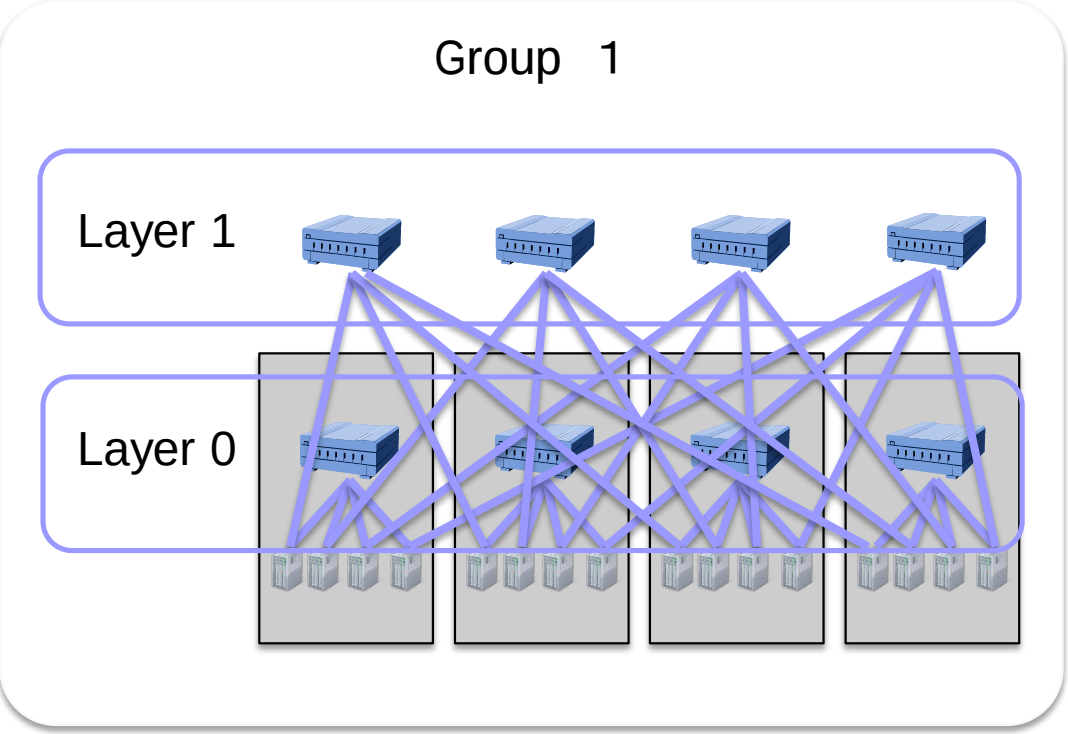


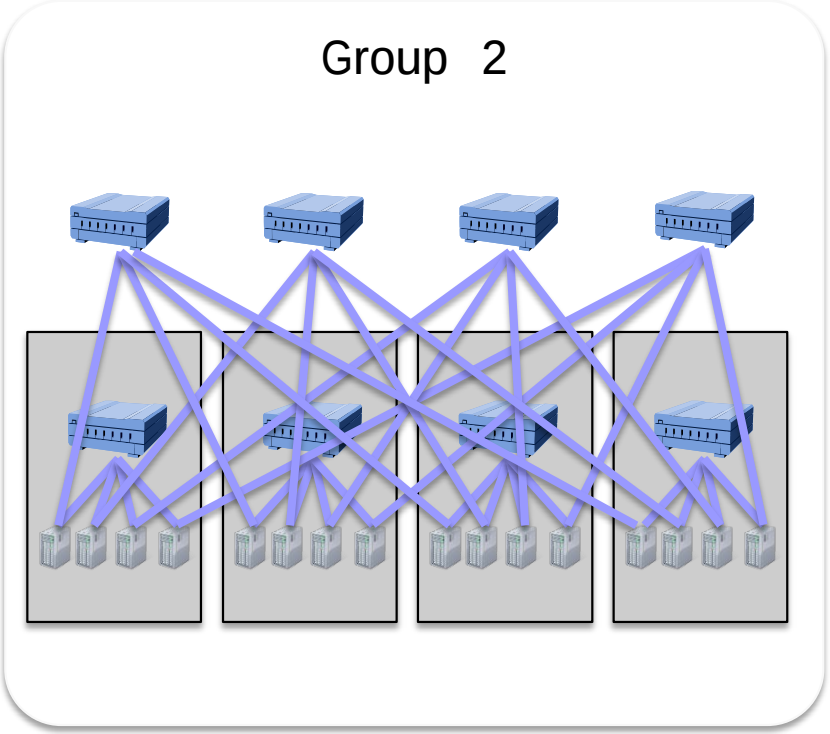
# 光電子融合型パケットルータを用いたデータセンターネットワークトポロジの構築方法

- 光電子融合型パケットルータとサーバラックの接続
  - 各サーバラックからより多くのサーバラックに1ホップで通信できるように接続
- 光電子融合型パケットルータ間の接続
  - 経路制御を模擬しながら、より多くのトラヒックを収容できるように接続

# 光電子融合型パケットルータとサーバラックの接続

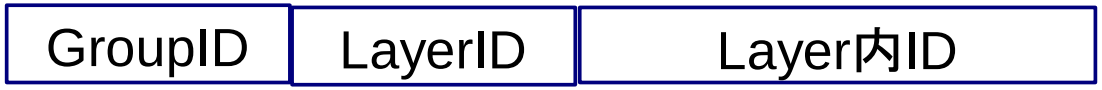
- 階層的な接続構造（Bcube)を模した接続構成
  - 各階層の光電子融合型パケットルータは、下位層で異なるグループに属するサーバラックと接続



# 光電子融合型パケットルータ間の接続手順

- 方針
  - 光電子融合型パケットルータにIDをふる




  - IDが0のサーバが接続している光電子融合型パケットルータの接続先のみを決める
  - IDが0のサーバの接続先ルータn1とn2が接続される場合、ルータn1'とn2'は以下の条件を満たす場合に接続

$$i_0^{\mathrm{SW}}(n_2') = \left(i_0^{\mathrm{SW}}(n_1') + \left(i_0^{\mathrm{SW}}(n_2) - i_0^{\mathrm{SW}}(n_1)\right)\right) \bmod K$$

$$i_1^{\mathrm{SW}}(n_2') = i_1^{\mathrm{SW}}(n_2)$$

$$i_i^{\mathrm{SW}}(n_2') = \left(i_i^{\mathrm{SW}}(n_1') + \left(i_i^{\mathrm{SW}}(n_2) - i_i^{\mathrm{SW}}(n_1)\right)\right) \bmod L^{\mathrm{SW\text{-}SV}}$$

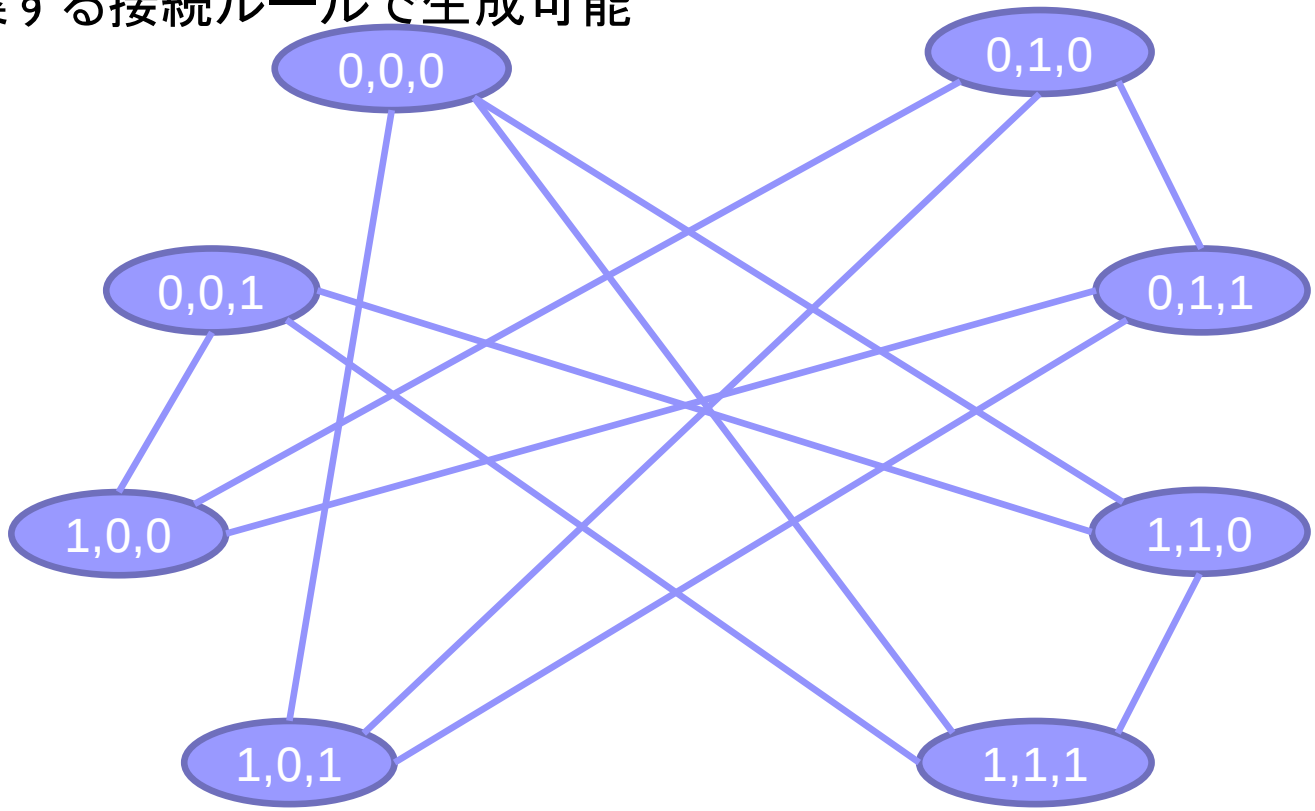
各光電子融合型ルータのトポロジ内における役割が同じであれば、規則的なトポロジになると考えられるため

# 光電子融合型パケットルータ間の接続手順

- IDが0のサーバが接続している光電子融合型パケットルータの接続先のみを決め方
  - 全候補を列挙し、経路制御を模擬した上で、各リンクに流れるトラヒック量（＝リンク負荷）を求める。
  - 以下の基準で接続構成を選ぶ
    - 最大リンク負荷がもっとも小さいネットワークトポロジを最適なトポロジとする。
    - 最大リンク負荷がもっとも小さいネットワークトポロジが複数ある場合は、そのうち、平均リンク負荷がもっとも小さいものを最適とする。
- 経路制御の模擬の方法
  - 全サーバラック間にトラヒックを生成
  - 各光電子融合型パケットルータで、宛先までの最短ホップ経路に属する転送先のうち、もっとも転送先までのリンクの負荷が低いリンクに転送

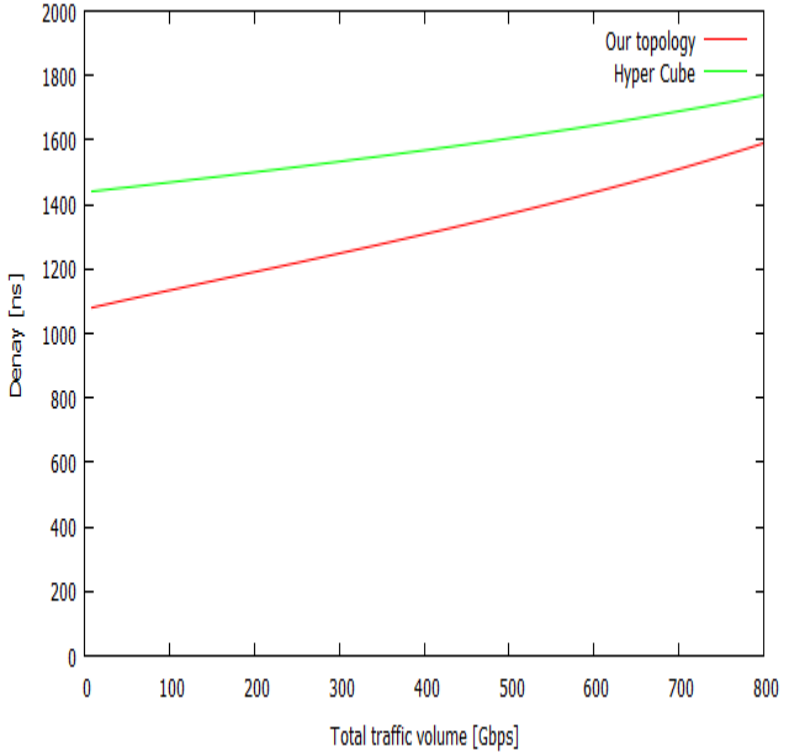
# 検証

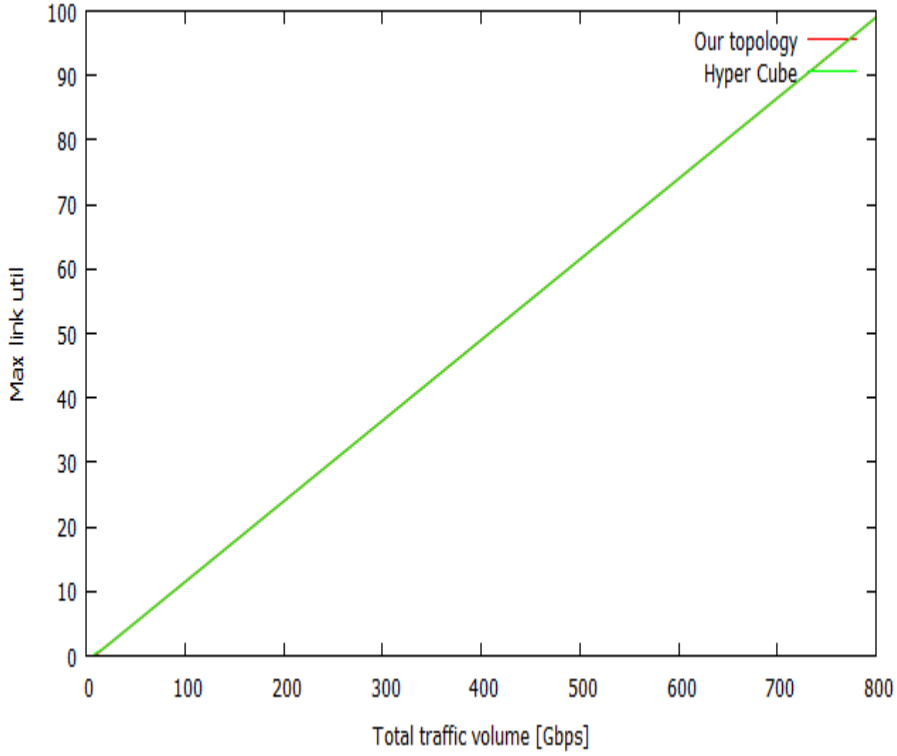
- 小規模なネットワーク（光パケットスイッチ８台、サーバーラック8台、サーバーラックからのリンク数2）の環境で、取りうる全構造のうち、もっともリンク負荷が低くなった構造
  - 本研究で提案する接続ルールで生成可能



# 遅延の比較

- HyperCube型トポロジと遅延・最大リンク使用率を比較
  - HyperCube型トポロジよりも低遅延でトラヒックを収容できていることが分かった。
  - サーバーラックからの複数のリンクを効率的に利用しているため。



# まとめ

- 光電子融合型パケットルータを効率的に用いたデータセンターネットワークトポロジ構成手法を提案

- 今後の予定
  - 段階的なノード追加による、大規模なデータセンター構築への適用について評価