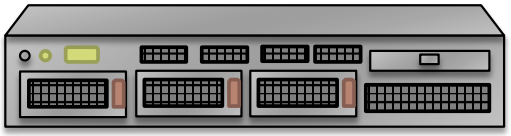
次世代データセンターネットワーク

大阪大学　大学院情報科学研究科　大下裕一

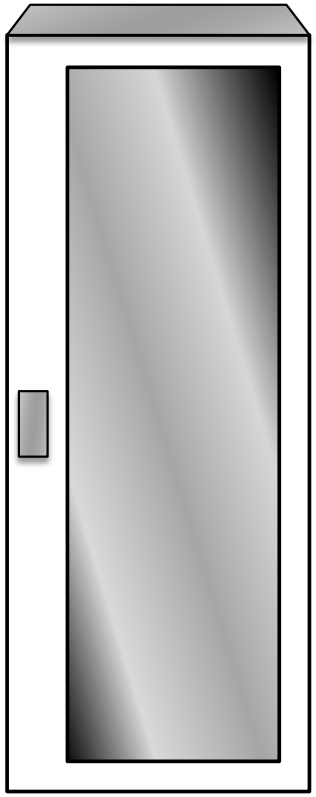
データセンターとは

- データセンターの構成
  - 多数のサーバをサーバラックに収容
  - サーバラックをネットワークで束ねてクラスタを構成
- データセンターの使われ方
  - クラウドサービスをホストする環境として
  - 一つのデータセンターが1台のコンピュータであるかのような使われ方も
    - データセンター内のサーバが連携して一つの大きな処理を行う
    - データセンター内の複数のサーバで分散して、一つの大きなデータを保持
- データセンターの規模
  - コンテナ型データセンター（＜サーバ数千台）
  - 大規模データセンター（数万台規模）（Google、Facebook等）

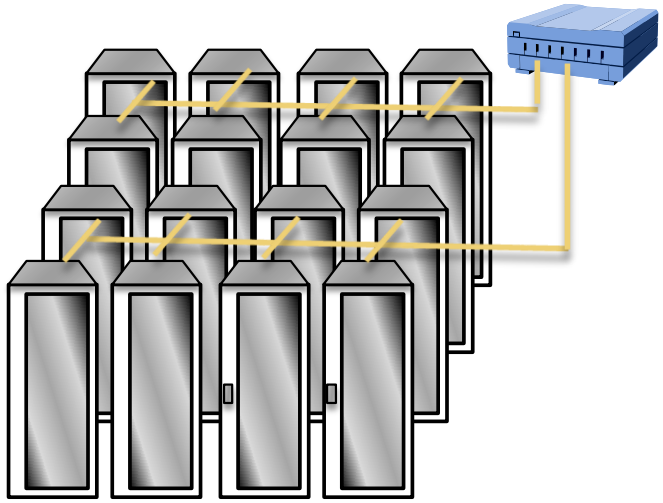
データセンターのイメージ図



サーバ
- CPU
- DRAM
- ストレージ

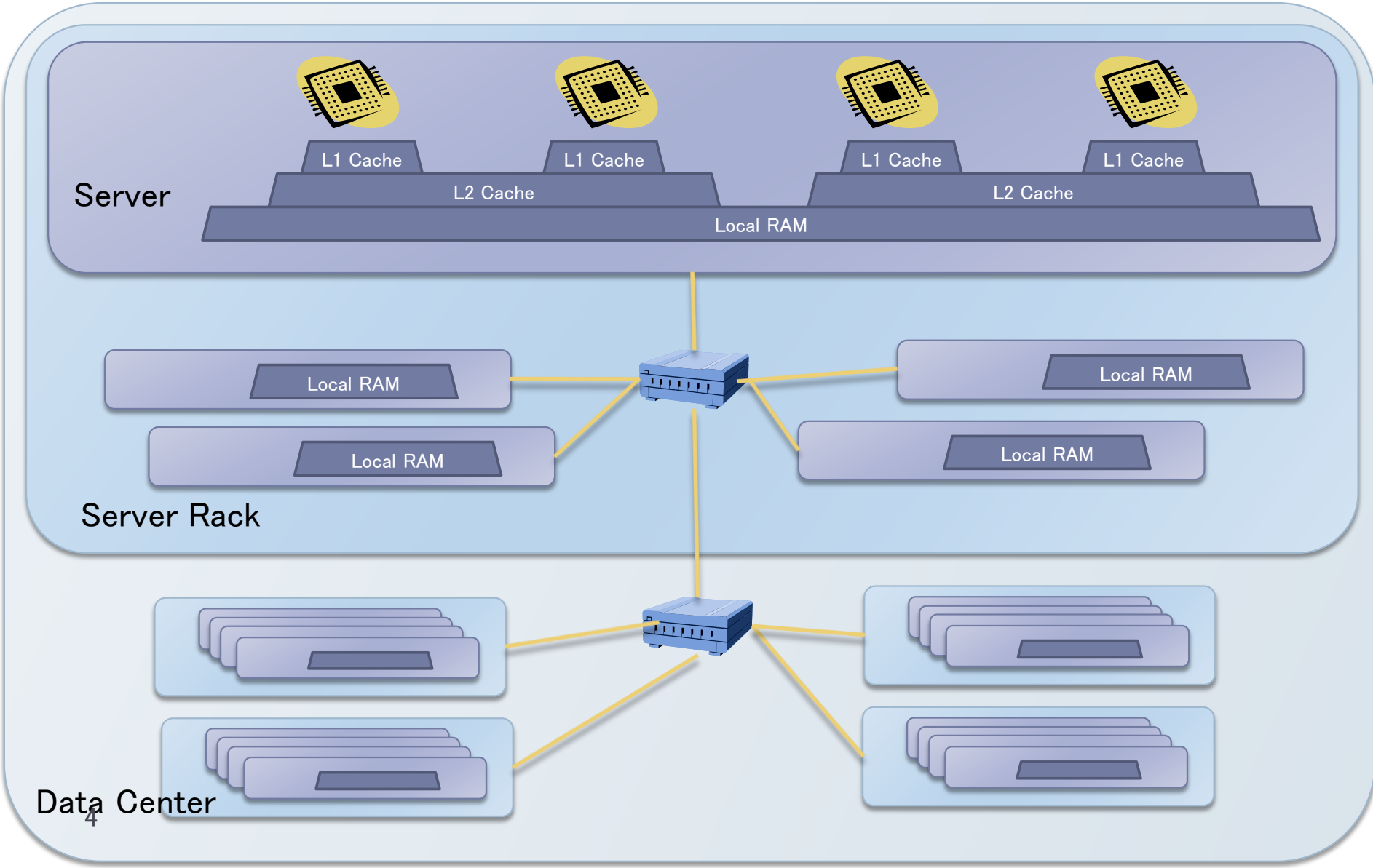


サーバラック
- サーバ×40–80台
- ラックスイッチ



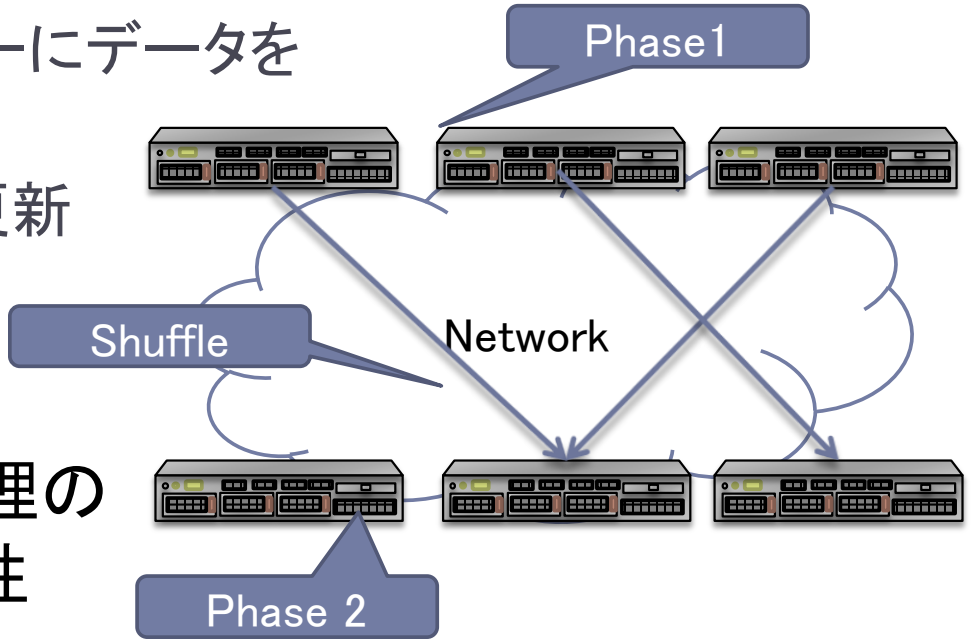
データセンタ
- 多くのサーバラック
- サーバラック間のネットワーク

データセンターをコンピュータとして見ると



データセンターにおける大規模データの処理

- 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[1]
  - 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

[1] J. Dean, and S. Ghemawat, “MapReduce: simplified data processing on large clusters,” Communications of the ACM, 2008.

データセンターの処理におけるボトルネック

- アプリケーション依存ではあるものの。。
  - ディスクアクセスがボトルネック
  - CPUがボトルネック

- ネットワーク自体もボトルネックになりうる
  - 帯域不足は処理に必要なデータを他から得るのにかかる時間を増大させる
  - 他のサーバのRAMを利用することにより、大規模なRAM空間を確保することの提案も[2]

[2] J. Ousterhout et. al. , “The Case for RAMClouds: Scalable High-Performance Storage Entirely in DRAM,” SIGOPS Operating Systems Review, Dec 2009

データセンターにおけるネットワーク

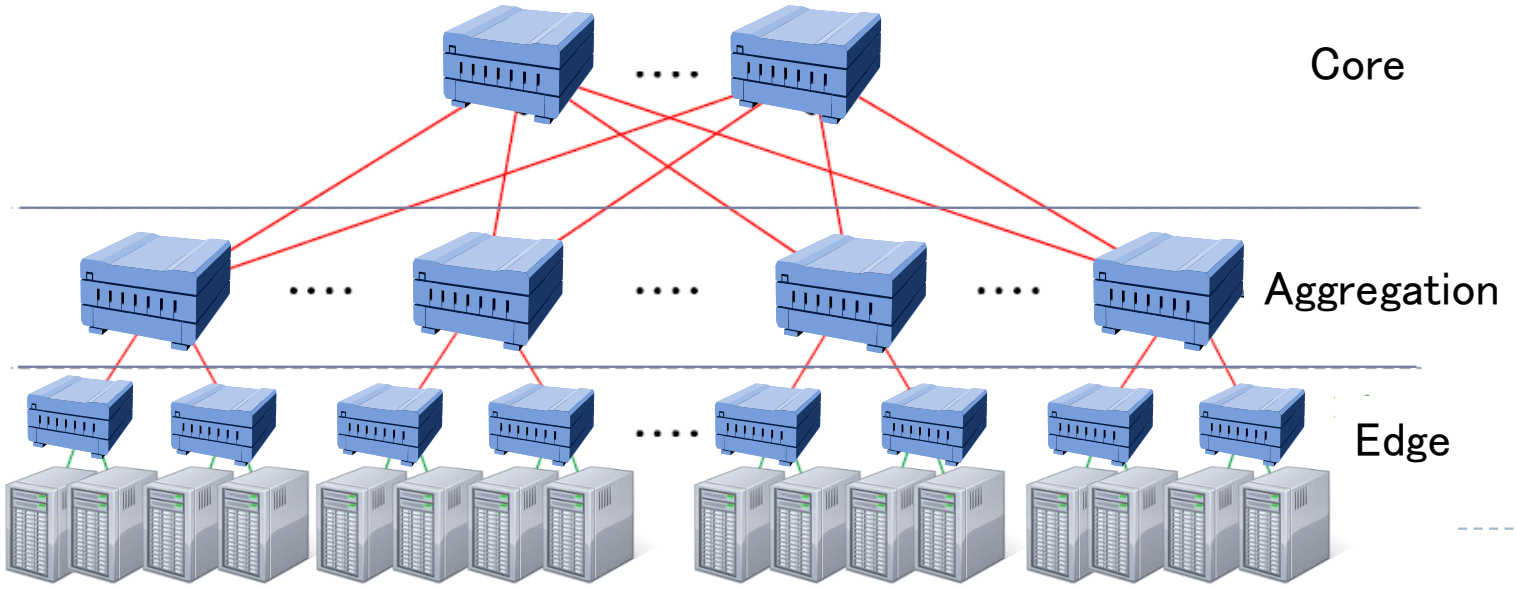
- データセンターの重要な位置づけを担う
  - データセンターを1台のコンピュータと見立てると、ネットワークはバスに相当
    - サーバ間で連携してデータを処理するため、データセンタ内部の通信が盛ん
    - ネットワークが十分な帯域を確保できないと、データセンター全体の性能劣化
      - 必要な情報を取得するまでの時間の増大→処理時間の増大

データセンターネットワークへの要求

- スケーラビリティの確保
  - 近年大規模化したデータセンターネットワークに対応して数十万台～百万台のサーバを接続できることが必要
- 通信性能の確保
  - サーバ間で連携して動作するアプリケーションの性能要求を満たすのに十分な広帯域・低遅延の通信をサーバ間に提供できることが必要
- 耐故障性の確保
  - 大規模システムなので、いずれかの箇所で故障が発生するという確率は高い。
  - 故障が発生しても、データセンターのサービスを維持することが必要
- 消費電力
  - サーバの低消費電力・高効率化が進むにつれ、ネットワークがデータセンター全体にしめる消費電力の割合も増加
  - ネットワーク自体も低消費電力になることが必要
- 安価であること

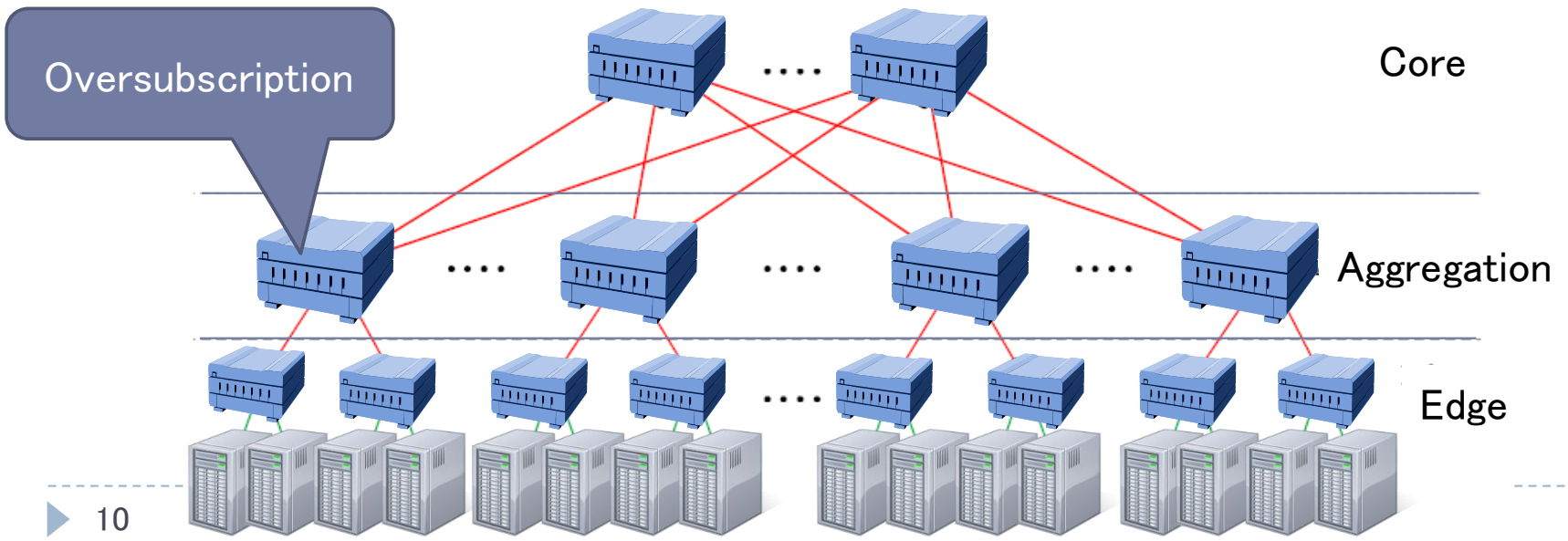
従来型データセンターの構造

- 3階層の木構造
  - Edge、Aggregation、Core
- Oversubscriptionにより設置コストを低減
  - Oversubscription：各スイッチにおいて、下位スイッチからのリンクの総帯域よりも小さな帯域のリンクのみを用いて上位スイッチと接続すること



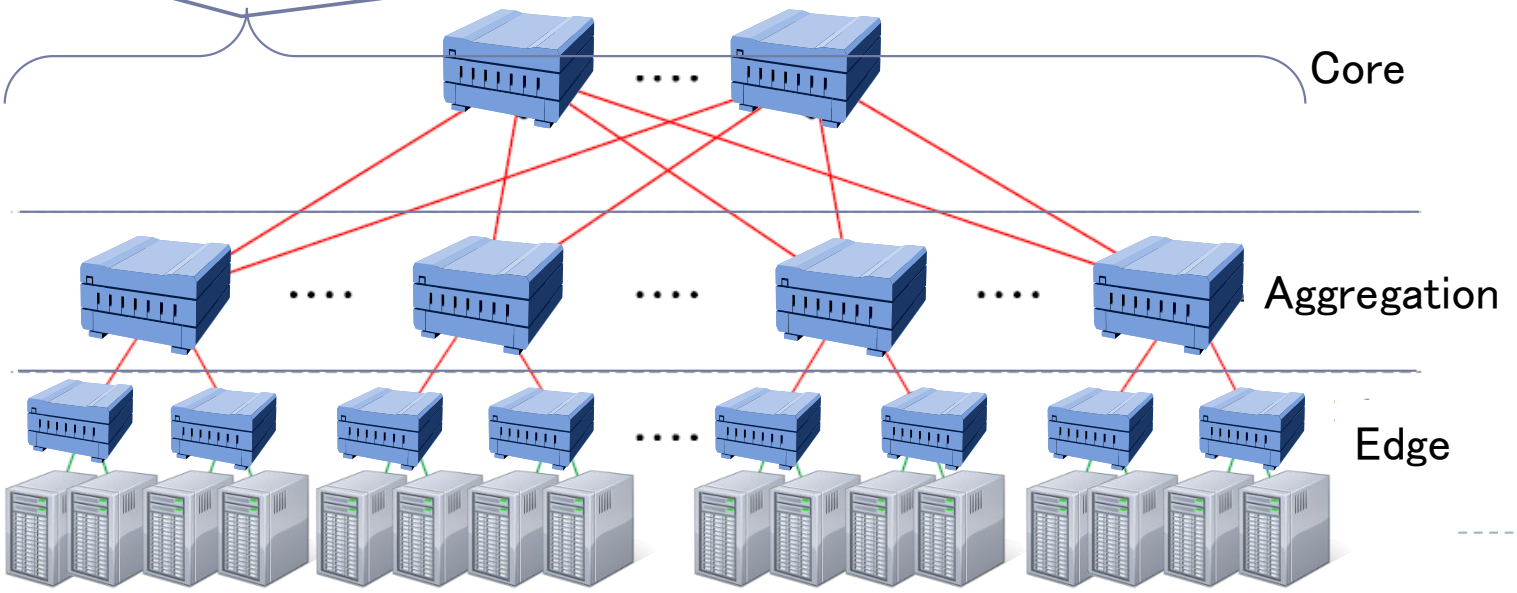
従来型データセンターの構造

- 性能の問題
  - Oversubscriptionにより、サーバ間に確保される帯域が限定



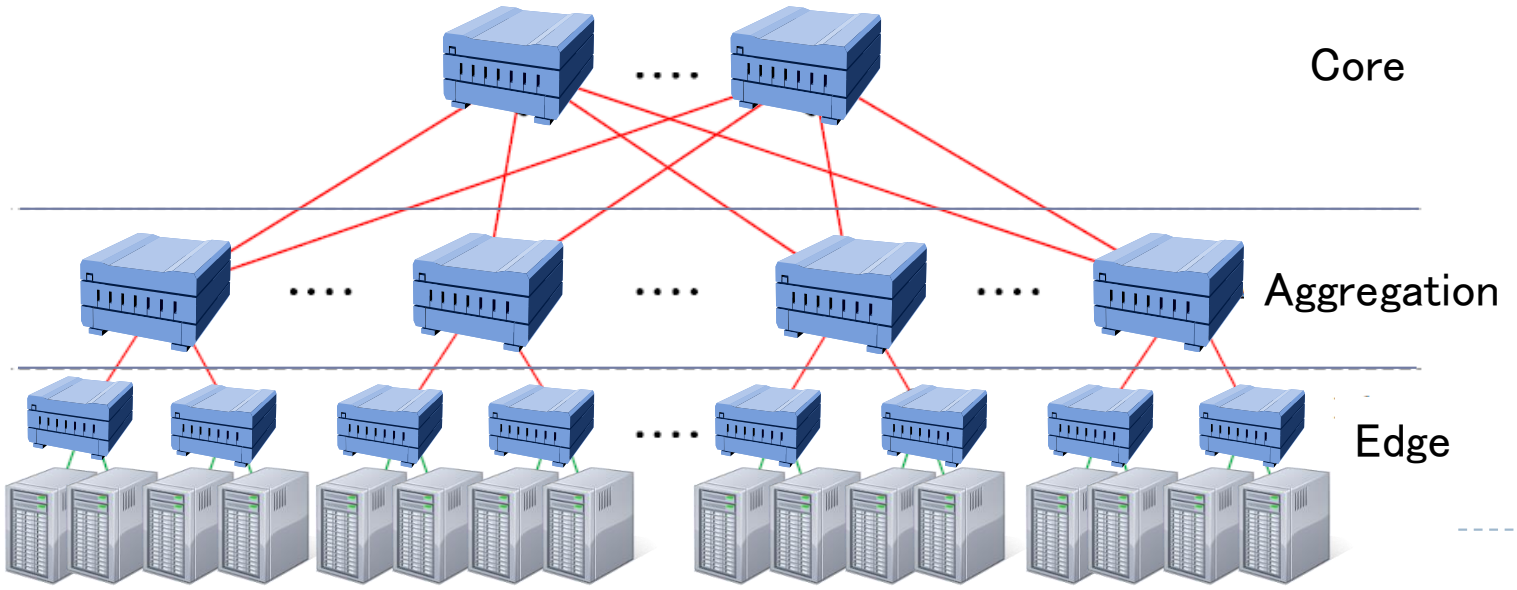
従来型データセンターの構造

- スケーラビリティの問題
  - 十分な性能を確保したまま、大規模化するのは困難
    - ポート数の大きなスイッチは消費電力が大きい
    - 木構造の階層数を増やすとスイッチ数が増える



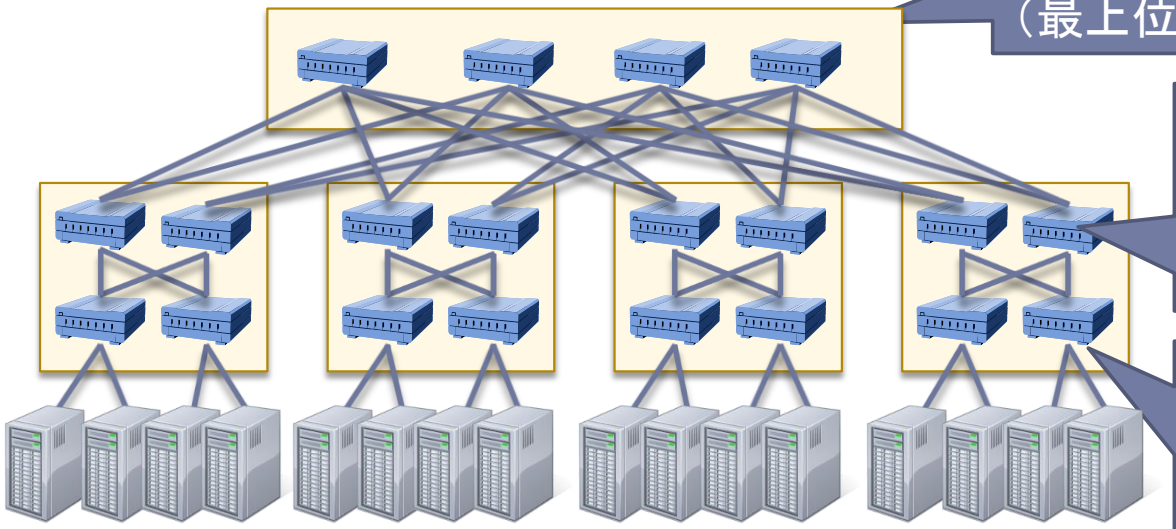
従来型データセンターの構造

- 消費電力の問題
  - コアに大規模なスイッチが必要＝消費電力大



新たなデータセンターの構造　例１

- 小規模なスイッチのみで十分な帯域を確保できる構造[3]



最上位層のスイッチのポート数と同じ数のPODを接続
（最上位のスイッチは、全PODに得接続）

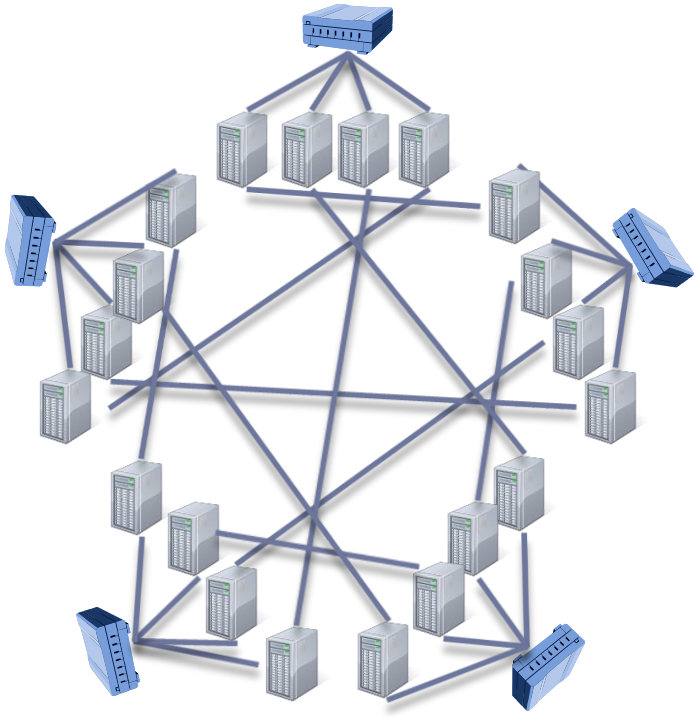
各スイッチの上位との接続と下位との接続のリンク数が同じ
＝Oversubscriptionなし

各POD内でButterflyトポロジを構築することにより、ポート数の多いスイッチをポート数の小さいスイッチから構築

[3] M. Al-Fares, A. Loukissas, and A. Vahdat, “A Scalable, Commodity Data Center Network Architecture,” in Proceedings of SIGCOMM, 2008.

新たなデータセンターの構造　例２

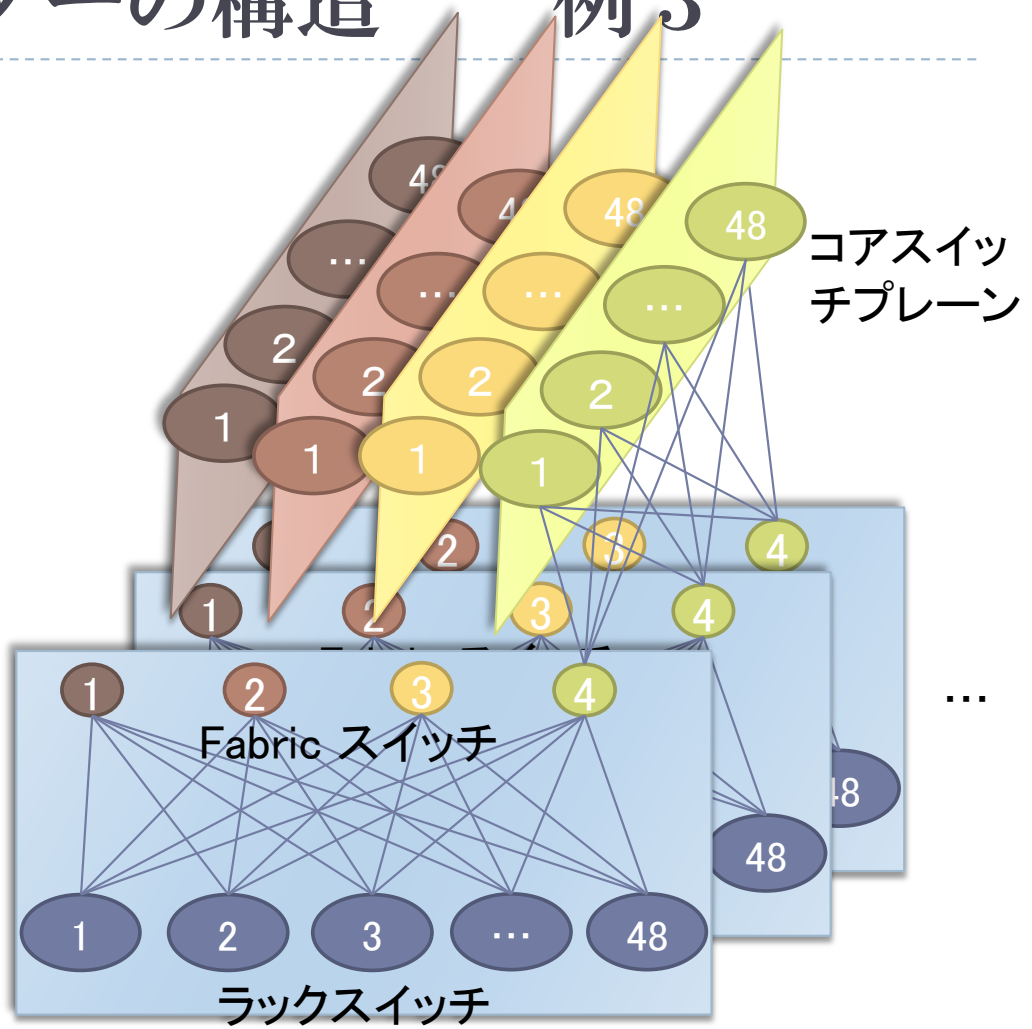
- Dcell[4]
  サーバ同士の接続により、少ないスイッチ数で大規模データセンターを構築
  - 構築方法：階層的に構築
    - $DCell_0$
      - 電気パケットスイッチの全ポートにサーバを接続
    - $DCell_k$（k階層目のDCell）
      - 異なる$Dcell_{k-1}$に属するサーバ同士を直接接続
        - 接続の際にはサーバのIDと$DCell_{k-1}$のIDを基準として用いる



[4] C. Guo, H. Wu, K. Tan, L. Shi, Y. Zhang, and S. Lu, “Dcell: a scalable and fault-tolerant network structure for data centers,” in Proceedings of SIGCOMM, 2008.

新たなデータセンターの構造　例３

- Facebookの新データセンターの構造[5]
  - 階層的な構造で大規模なデータセンターを構成
    - PODを48台のコアスイッチで構成：POD単位でデータセンターを拡充
    - Oversubscriptionなしでサーバラック間を接続



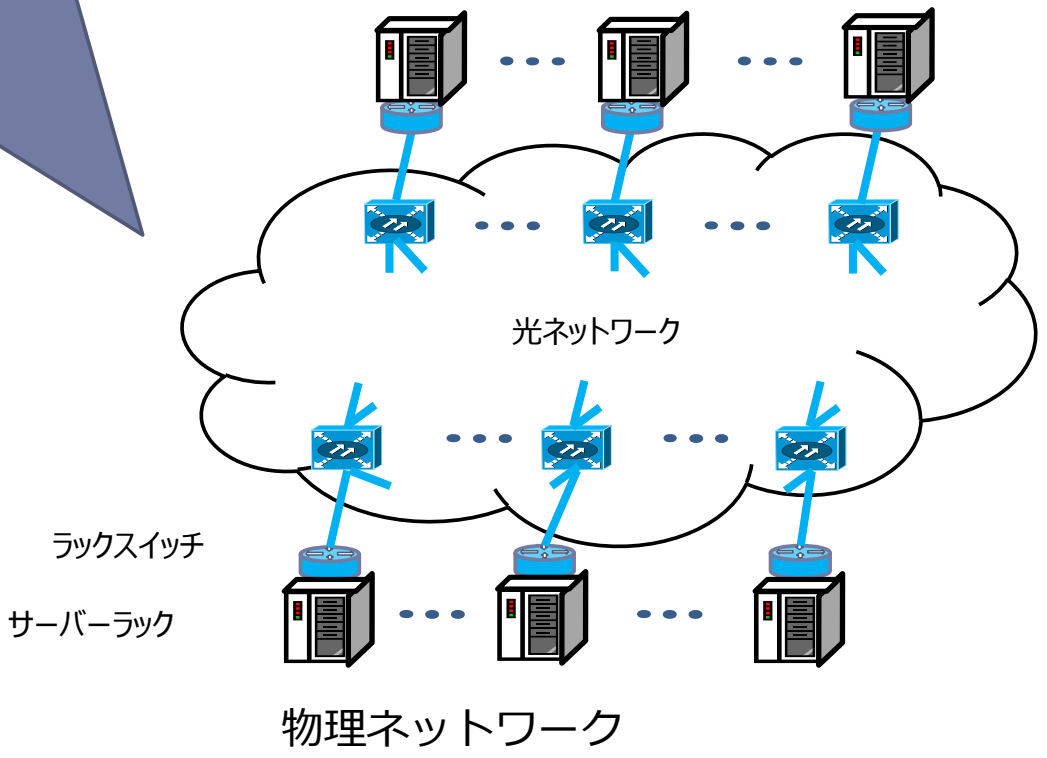
[5] A. Andreyev, “Introducing data center fabric, the next-generation Facebook data center network,” 2014 available on https://code.facebook.com/posts/360346274145943/introducing-data-center-fabric-the-next-generation-facebook-data-center-network/

データセンターにおける光ネットワークの活用

通信量が大きくなるデータセンターのコアネットワークに光通信技術を導入

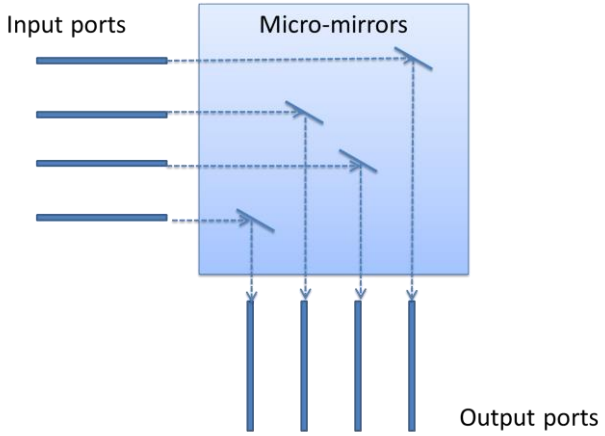
以下を達成

- 低消費電力:
  - 電気の処理を行うよりも、低消費電力なネットワークを構成可能
- 広帯域・低遅延:
  - 光スイッチの広帯域・低遅延性を生かした接続

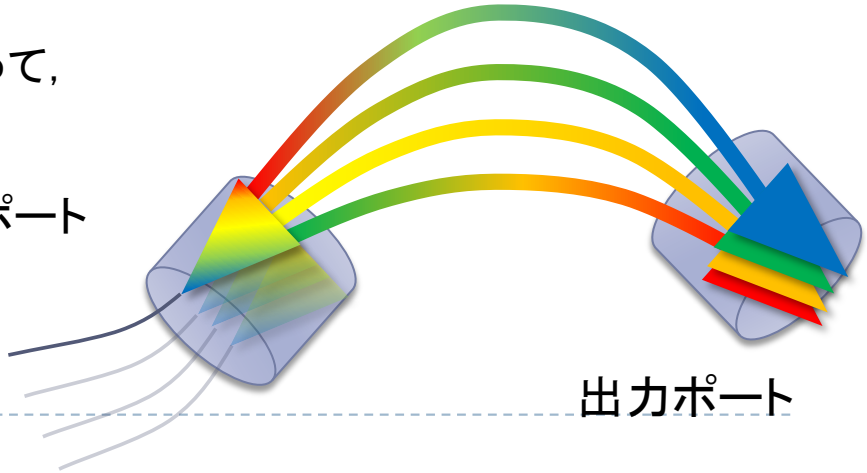


データセンターネットワークへの導入が検討されている光通信技術

- 光回線交換スイッチの導入
  - MEMS光スイッチ
    - 鏡の角度の調整で、各入力ポートと直結させる出力ポートを設定
    - 大規模なスイッチを構成可能

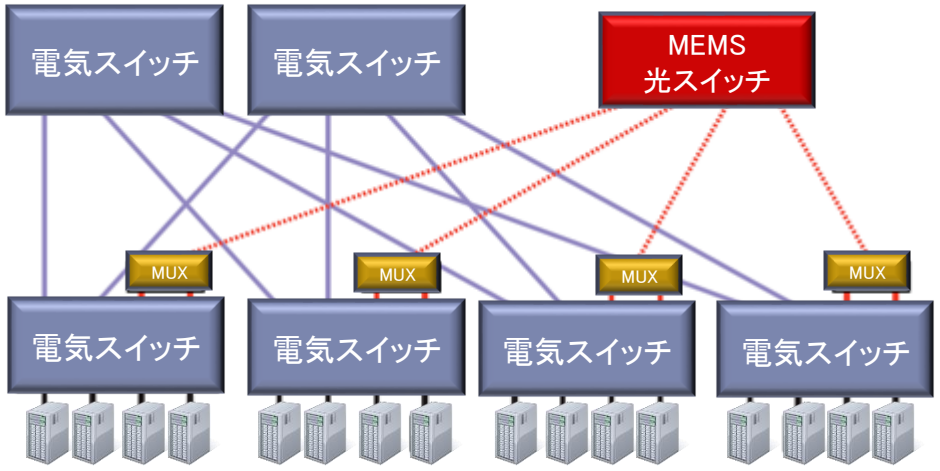


- 光パケットスイッチの開発
  - 光パケットのスイッチングする手法
    - アレイ導波路回折格子（AWG）を用いた構成
      - 長さの異なる複数の導波路群を使って，波長ごとに光を取り出す分光器
      - 入力波長を変えることにより、出口ポートを変えるスイッチングに利用



光回線交換スイッチを用いた構成　例１

- Helios[6]
  - コアスイッチに電気パケットスイッチと光回線交換スイッチを配置
  - トラヒックが多い地点間を接続するように光回線交換スイッチを設定
  - それ以外の地点間のトラヒックはコアの電気パケッットスイッチを経由して転送



[6] N. Farrington, G. Porter, S. Radhakrishnan, H. H. Bazzaz, V. Subramanya, Y. Fainman, G. Papen, and A. Vahdat, “Helios: a hybrid electrical/optical switch architecture for modular data centers,” in Proceedings of SIGCOMM 2010

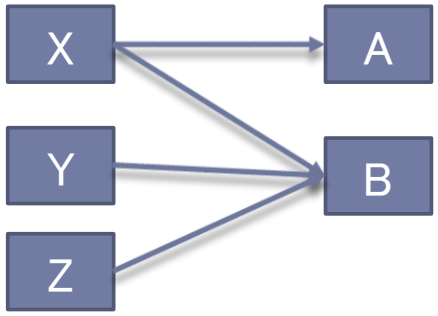
Heliosにおける制御

1. トラヒック需要の収集
2. 実際のトラヒック需要の推測
   - 観測されたトラヒック≠実際のトラヒック需要
     - 現在のネットワークのボトルネックにより抑えられているかもしれない
3. 推測されたトラヒック需要に合わせてMEMS光スイッチを設定

実際のトラヒック需要の推測[7]

各ノードの入出力リンクの帯域を均等にシェアするように実際のトラヒック需要を推測

- 例:
  - 各ノードの入出力リンクの帯域は1
  - X→A、X→B、Y→B、Z→Bの通信が発生
  - いずれの通信も観測されたトラヒック量がボトルネックリンクを通過



X、Y、Zの出力リンクの帯域をシェアするように需要を予測

X→A	1/2
X→B	1/2
Y→B	1
Z→B	1

A、Bの入力リンクの帯域をシェアするように需要を補正

X→A	1/2
X→B	1/3
Y→B	1/3
Z→B	1/3

X、Y、Zの出力リンクの帯域をシェアするように需要を補正

X→A	2/3
X→B	1/3
Y→B	1/3
Z→B	1/3

[7] M. Al-Fares, S. Radhakrishnan, B. Raghavan, N. Huang, and A. Vahdat, "Hedera: Dynamic Flow Scheduling for Data Center Networks," in Proceedings of NSDI, 2010

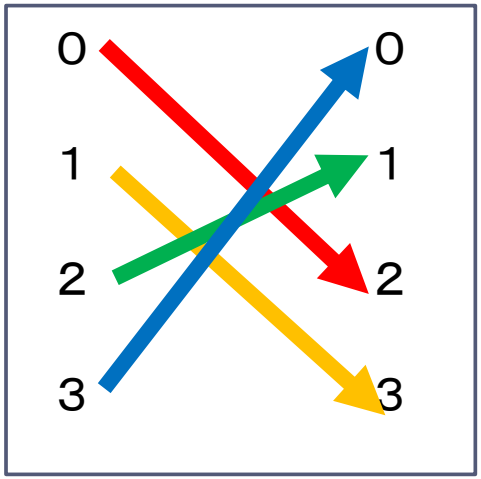
HeliosにおけるMEMS光スイッチの設定

- 最大重みマッチング問題を解くことにより求める
  - 入力PODと出力PODのマッチングを二分グラフで表す
  - 二分グラフの辺の重みは、当該POD間の需要

収容できる需要が最大化されるようなMEMS光スイッチの設定が得られる

	0	1	2	3
0	**0**	1	**3**	1
1	4	0	1	3
2	1	3	0	4
3	**3**	2	1	0

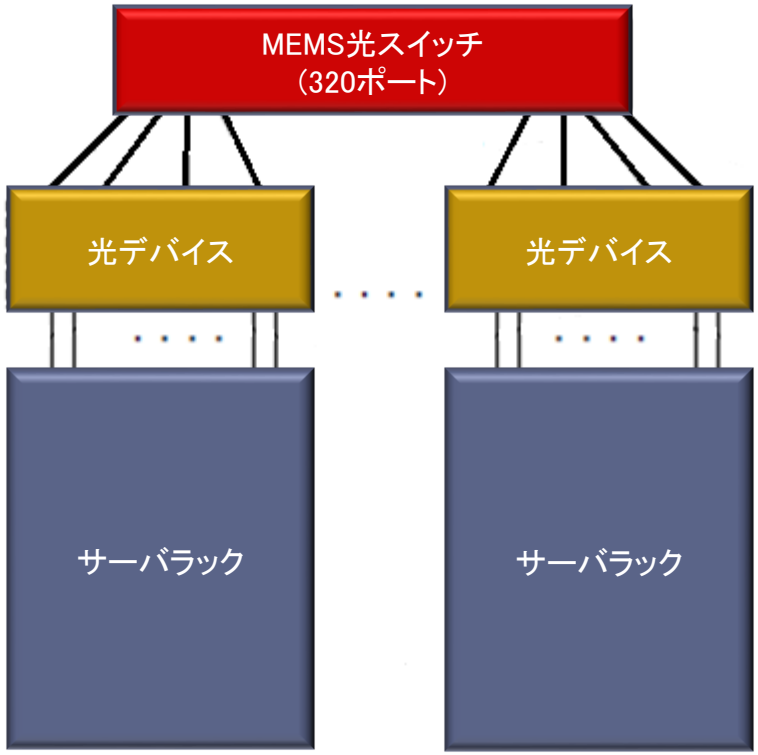
需要行列



光スイッチ設定の解

光回線交換スイッチを用いた構成　例２

- Proteus[8]
  - 全サーバラックを大きな光回線交換スイッチに接続
  - 光パスで仮想ネットワークを構築
  - 全通信は仮想ネットワークを経由して通信



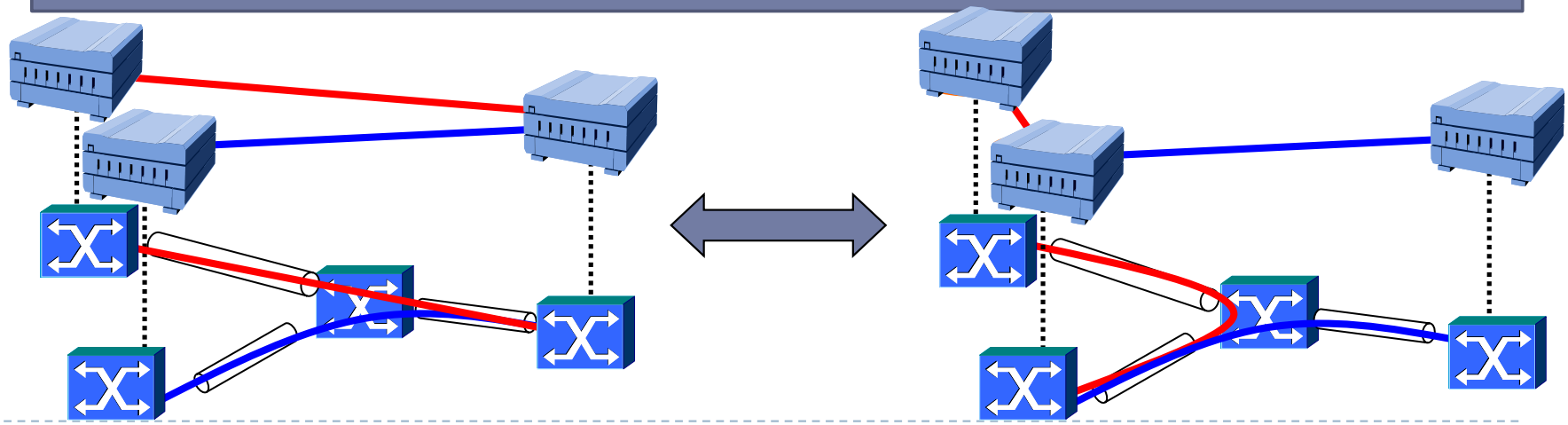
[8] A. Singla, A. Singh, K. Ramachandran, L. Xu, and Y. Zhang, “Proteus: A Topology Malleable Data Center Network,” in Proceedings of HOTNETS, 2010

光回線交換スイッチを用いた データセンターネットワーク内仮想ネットワーク

- 光回線交換スイッチの設定により、電気スイッチ間に光パスを設定

- 構築されたパスは、電気スイッチからは物理的な配線として見える

光回線交換スイッチの設定の変更により、電気スイッチのネットワークの接続構成が変更可能



光回線交換スイッチ上仮想ネットワークの制御によるネットワーク低消費電力化

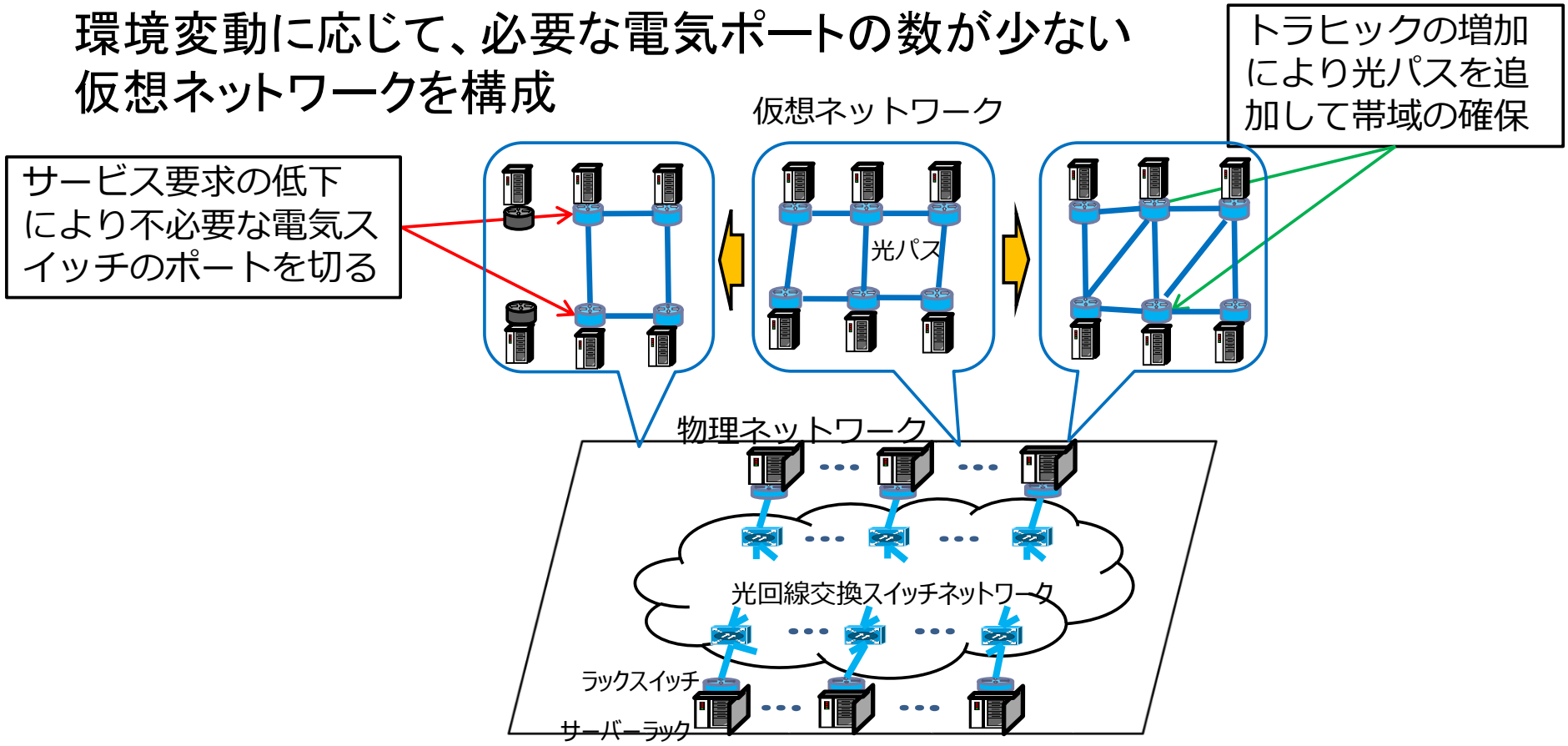
- 電気スイッチの消費電力は大きい

以下を達成するように仮想ネットワークを制御

- 不要な電気スイッチは電源断
- 不要な電気スイッチポートも電源断

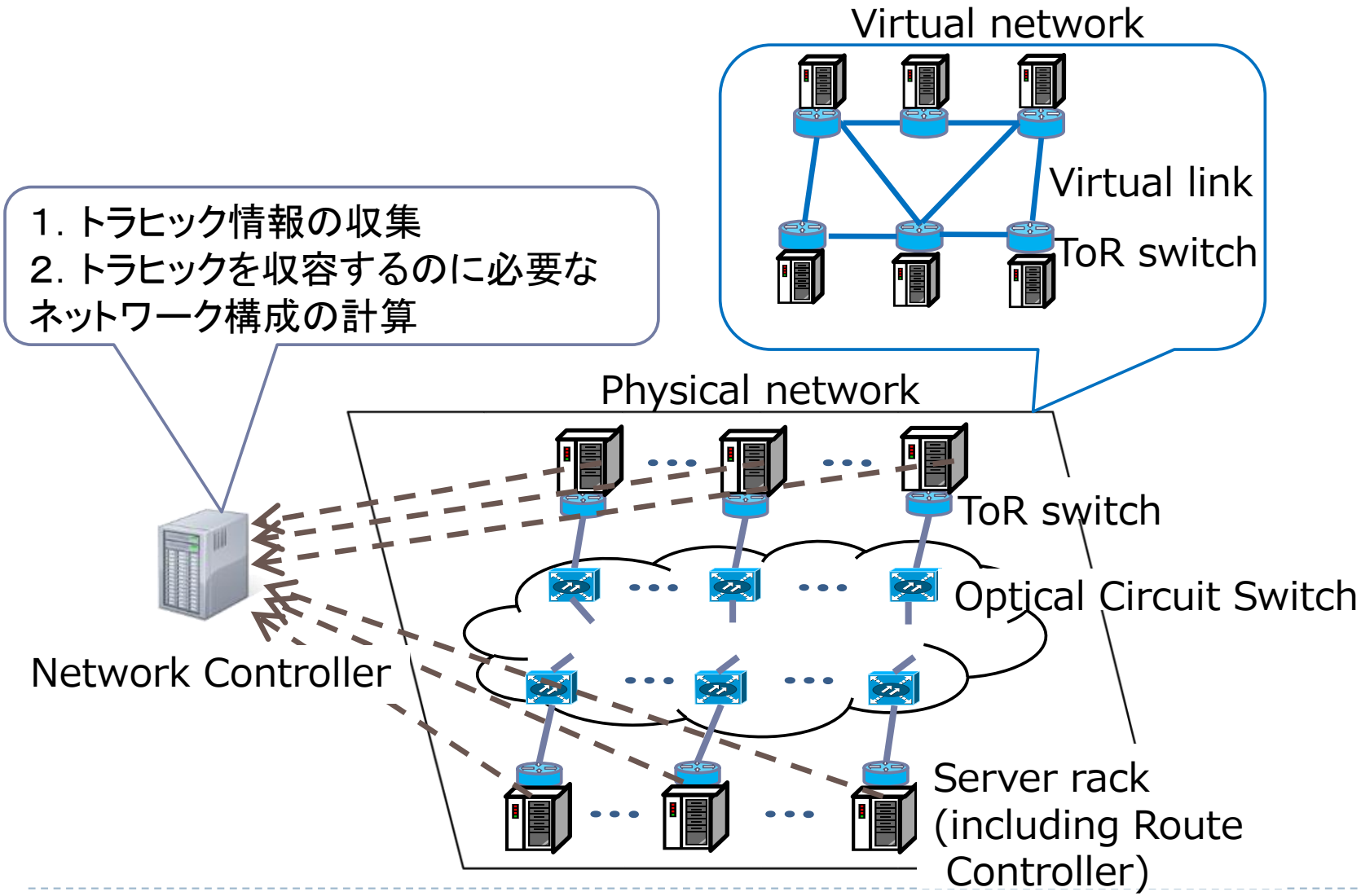
光回線交換スイッチ上仮想ネットワークの制御による ネットワーク低消費電力化[9]

環境変動に応じて、必要な電気ポートの数が少ない仮想ネットワークを構成



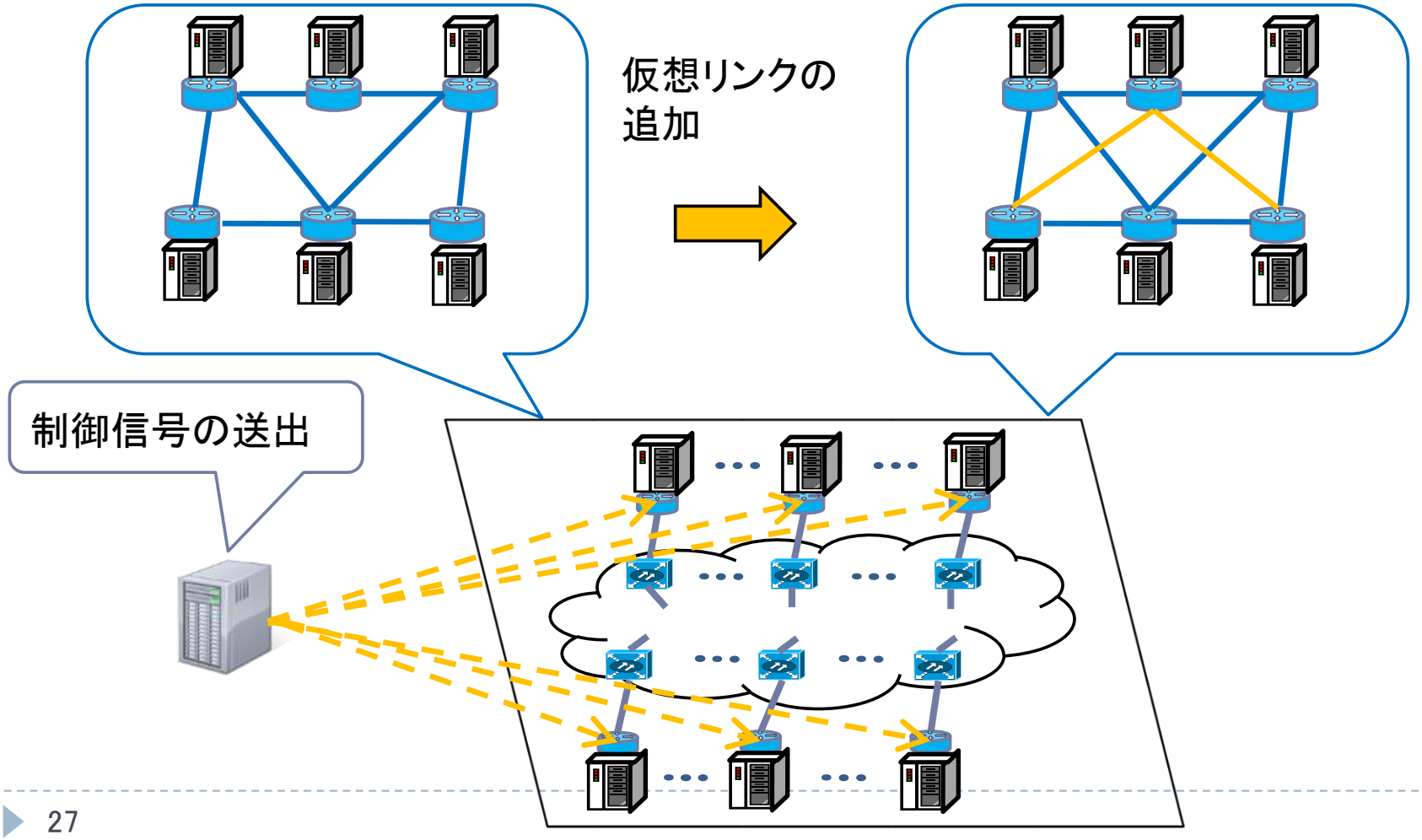
[9] Y. Tarutani, Y. Ohsita, and M. Murata, “Virtual Network Reconfiguration for Reducing Energy Consumption in Optical Data Centers,” IEEE/OSA Journal of Optical Communications and Networking, 2014.

光回線交換スイッチ上仮想ネットワークの制御によるネットワーク低消費電力化の手順



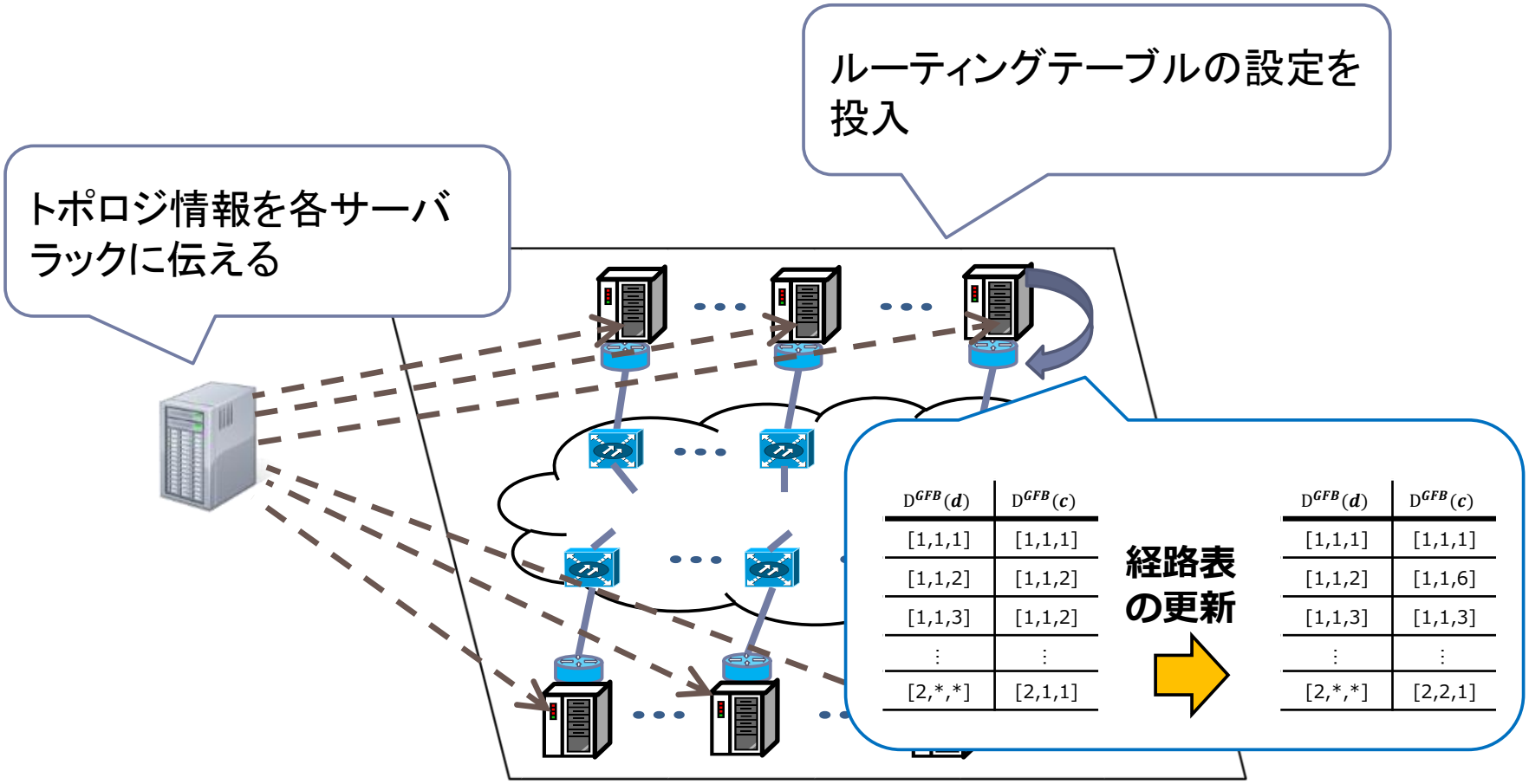
光回線交換スイッチ上仮想ネットワークの制御によるネットワーク低消費電力化の手順

3. リンクの追加



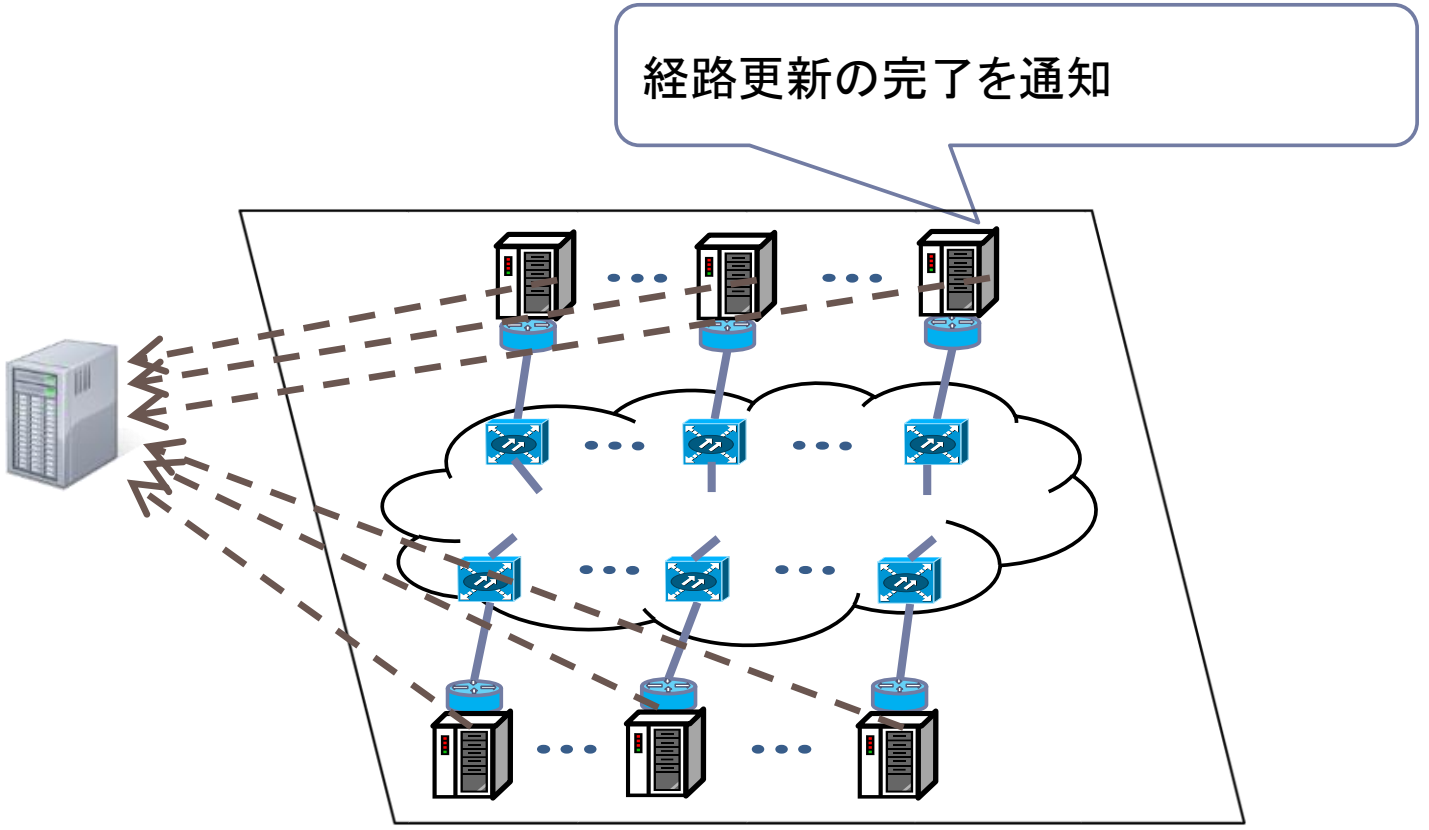
光回線交換スイッチ上仮想ネットワークの制御によるネットワーク低消費電力化の手順

4. 電気スイッチの経路の更新



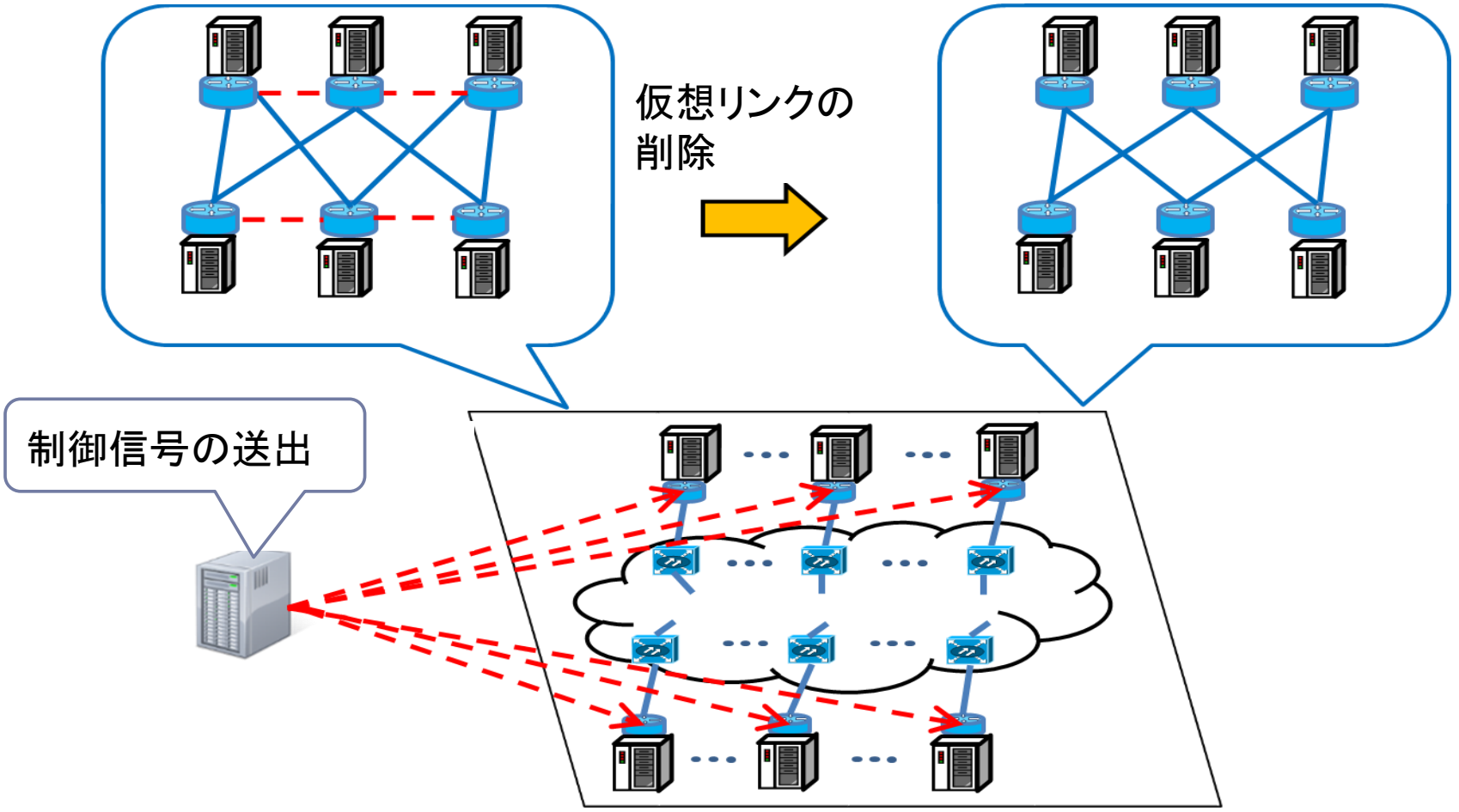
光回線交換スイッチ上仮想ネットワークの制御によるネットワーク低消費電力化の手順

5. 経路更新完了



光回線交換スイッチ上仮想ネットワークの制御によるネットワーク低消費電力化の手順

6. 不要な仮想リンクの削除



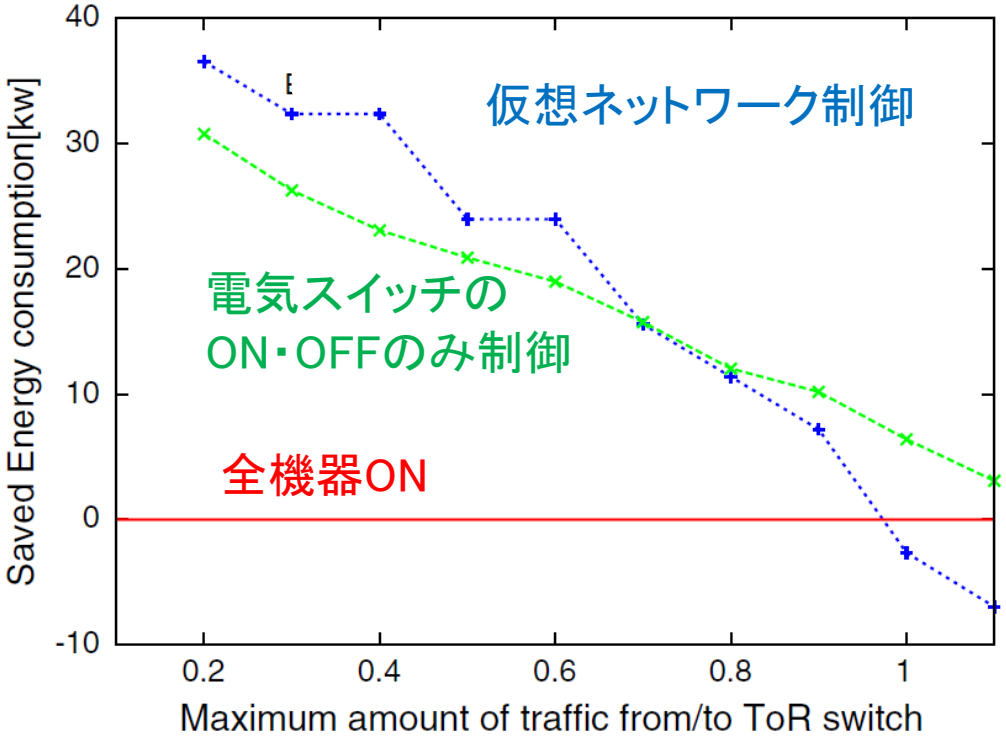
光回線交換スイッチ上仮想ネットワークの制御によるネットワーク低消費電力化の効果

- 全ネットワーク機器の電源をONにした場合と比較した消費電力の比較
- 仮想ネットワークを制御することにより、特にトラヒックの少ない時間帯では、消費電力を大きく削減可能

評価条件

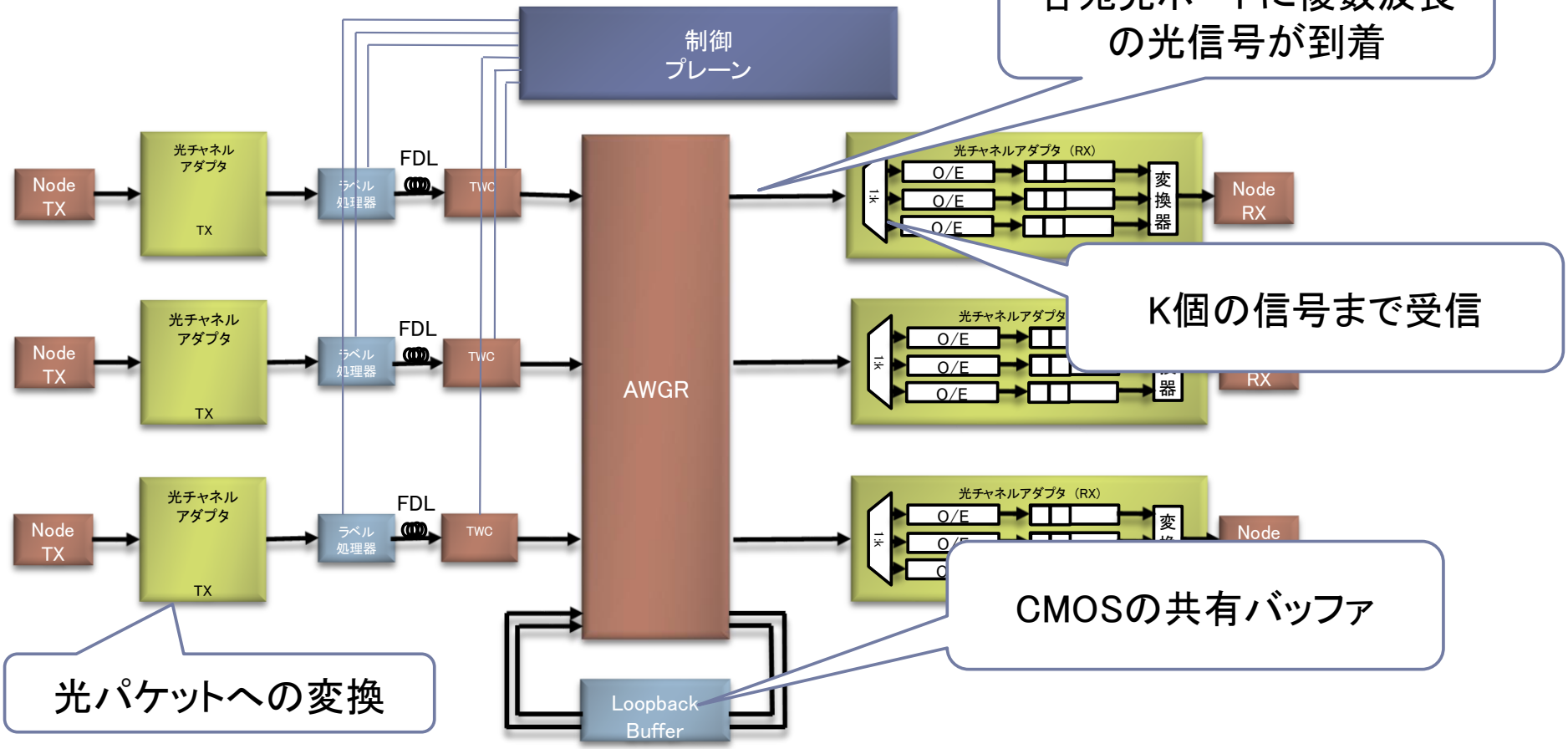
サーバラック数420台のネットワークで評価

電気スイッチネットワークは、最大通信量に合わせて設計



光パケットスイッチの研究　例１

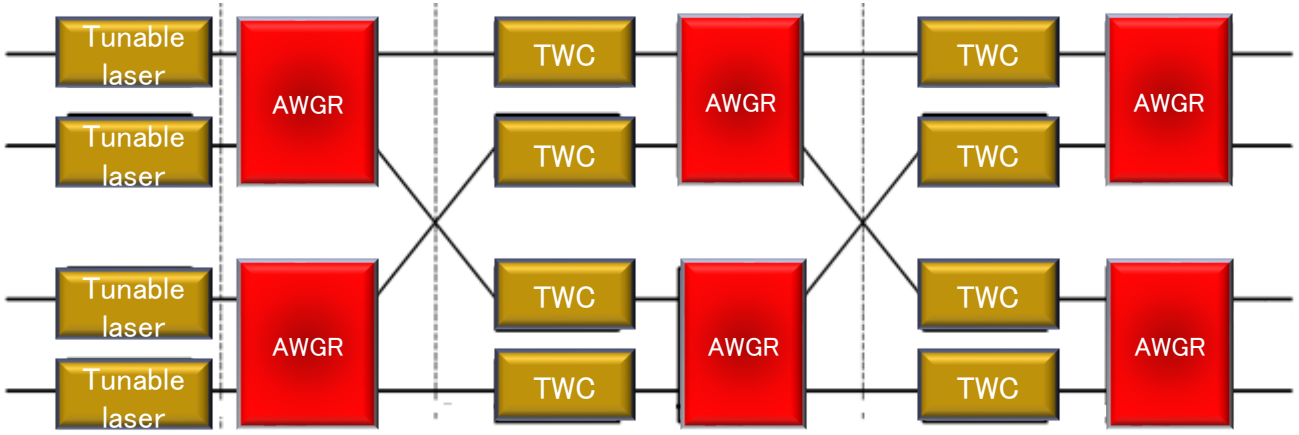
- LIONS[9]：AWGRを利用したパケットスイッチ



[9] Y. Yin, R. Proietti X. Ye, C. J. Nitta, V. Akella, and S. J. B. Yoo, “LIONS: An AWGR-Based Low-Latency Optical Switch for High-Performance Computing and Data Centers,” IEEE JOURNAL OF SELECTED TOPICS IN QUANTUM ELECTRONICS, 2013.

光パケットスイッチの研究　例２

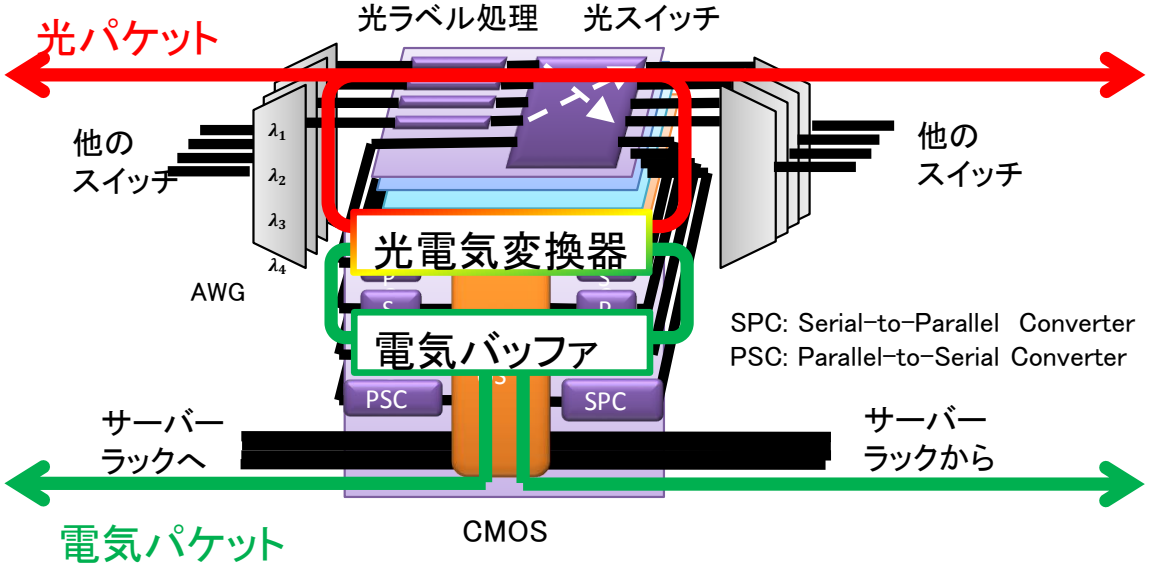
- AWGRを多段接続することによる、大規模光パケットスイッチ[10]
  - 構成
    - AWGRを多段で構成
    - AWGR：入射する波長によって出口ポートが決まる
    - TWC：入力波長を指定した波長に変換→変換することにより、経路を変えて制御
    - 内部にバッファはなく、集中制御により衝突を防止
      - バッファはラックスイッチ任せ



[10] H. J. Chao and K. Xi, “Bufferless Optical Clos Switches for Data Centers,” in Proceedings of OFC, 2011.

光パケットスイッチの研究　例３

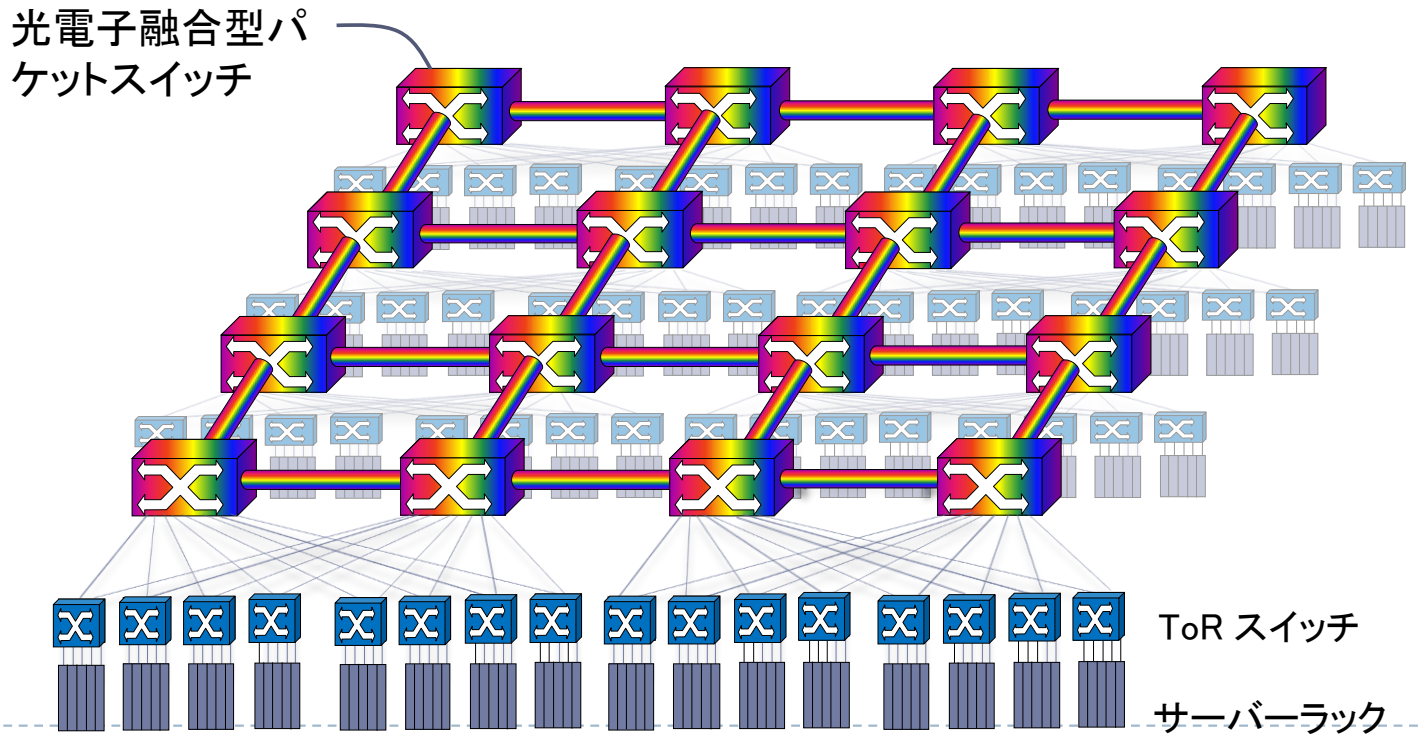
- 光電子融合型パケットスイッチ[11]
  - 光ポートと電気ポートを持つスイッチ
    - 光ポート：　多波長光パケットを他の光電子融合型パケットスイッチと交換
    - 電気ポート：　サーバラック内の電気スイッチと接続
      バッファに蓄えた後、多波長光パケットに変換して送出
    - 電気バッファは、衝突が発生した光パケットの保持にも利用



[11] S. A. Ibrahim, H. Ishikawa, T. Segawa, T. Nakahara, Y. Suzaki, and R. Takahashi, “100-Gb/s optical packet switching technologies for data center networks,” in proceedings of Photonics in Switching, 2014.

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

- 複数の光電子融合型パケットスイッチを接続し、データセンターのコアネットワークを形成
- 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集約して転送



光電子融合型パケットスイッチの特徴

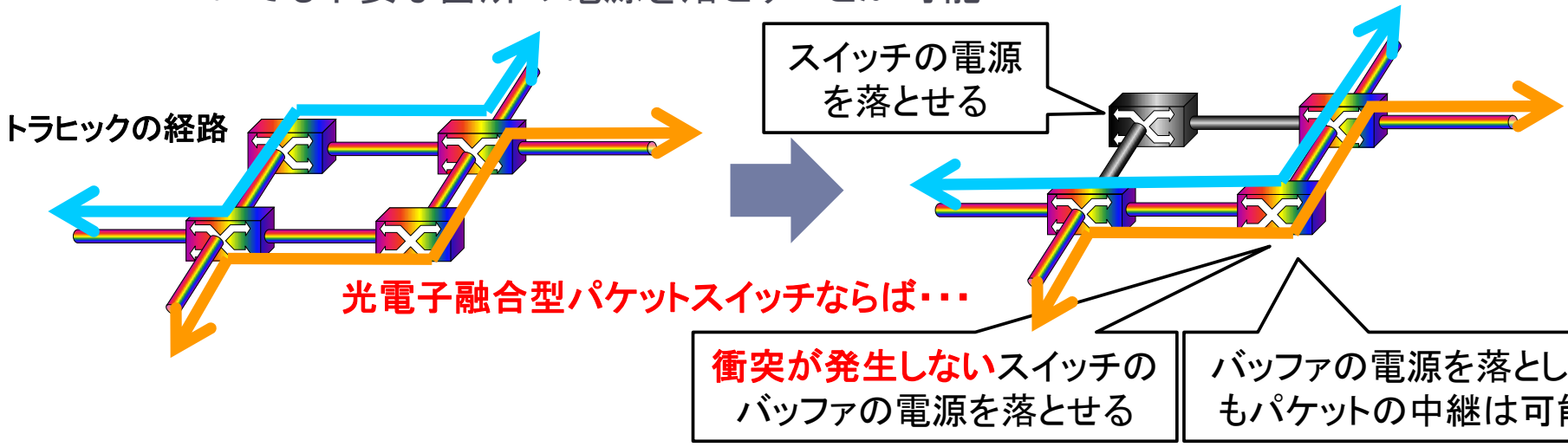
- 光通信技術の高性能・低消費電力を維持したまま、光通信技術における制御や実現性の問題を解決
  - パケットの衝突が発生しない場合、光/電気変換が不要で、光パケットをそのまま中継可能

    → 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

光電子融合型データセンターネットワークにおけるトラヒック経路選択による低消費電力化制御[12]

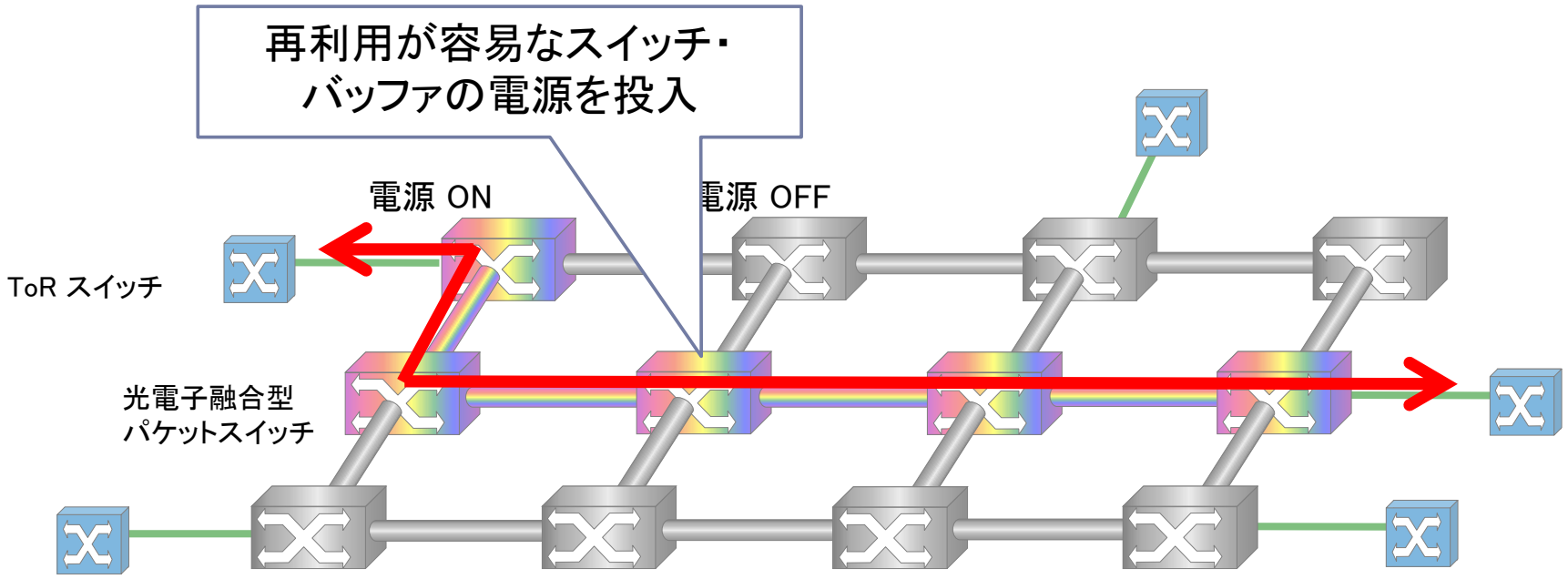
- 従来のネットワークの低消費電力化制御
  - 発生したトラヒックの経路を選択する際、必要な機器のみ電源を投入
    - IP スイッチの電源や、NIC のポート、光電気変換器など
- 光電子融合型パケットスイッチネットワークの低消費電力化制御
  - 発生したトラヒックの経路を選択する際、スイッチの電源のみではなく、**バッファ**についても不要な箇所の電源を落とすことが可能



[12] 西島孝通, 小泉佑揮, 大下裕一, 村田正幸, "光電子融合型パケットルーターを用いたデータセンターネットワーク向け低消費電力化制御の一検討," 電子情報通信学会技術研究報告(IN2013-179), 2014.

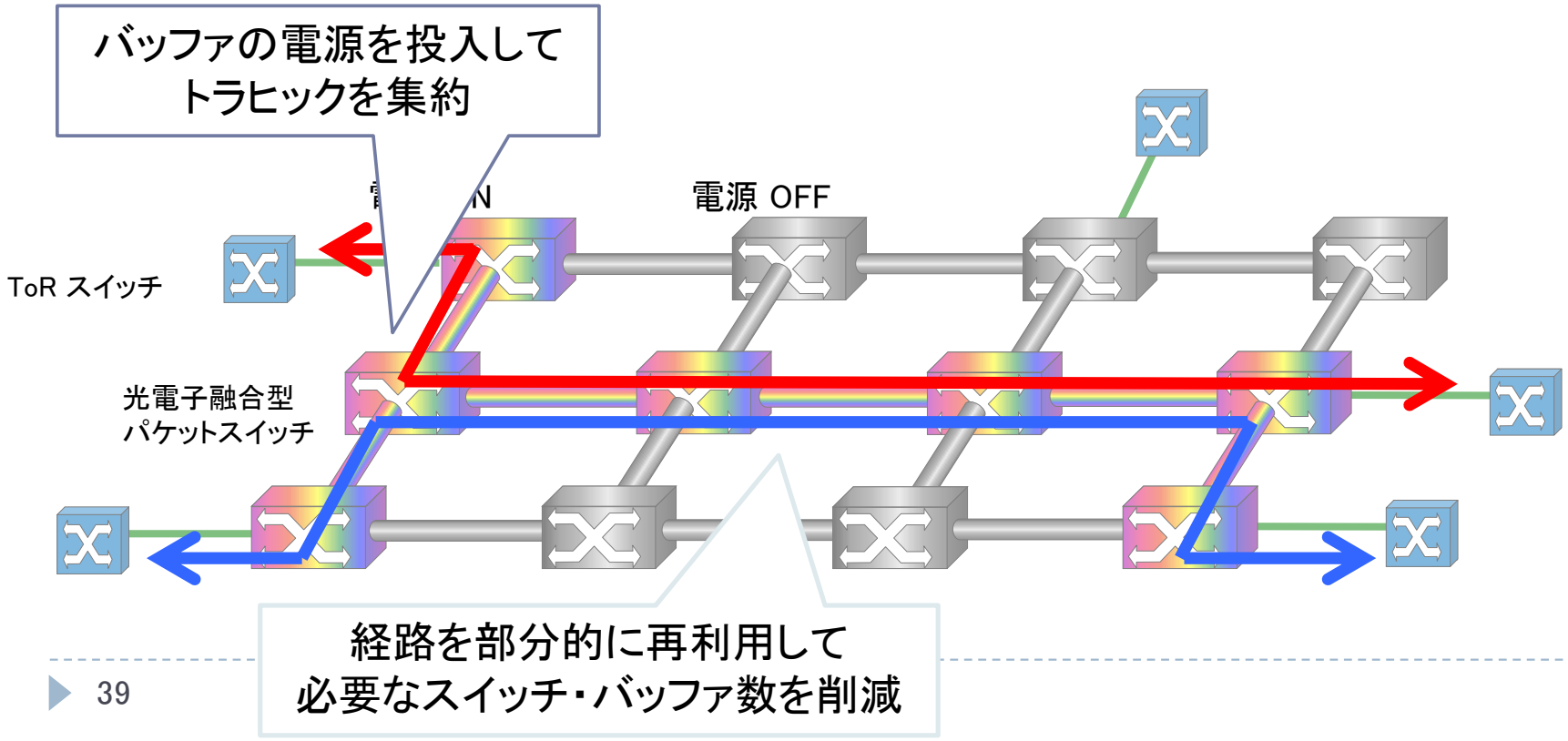
光電子融合型データセンターネットワークにおける低消費電力なトラヒック経路選択手法のアイデア

- **必要最低限のバッファのみを用いてトラヒックを集約**し、電源の投入が必要なスイッチ・バッファの総消費電力を最小化
  - 多くのトラヒックで利用可能なスイッチおよびバッファのみ電源を投入



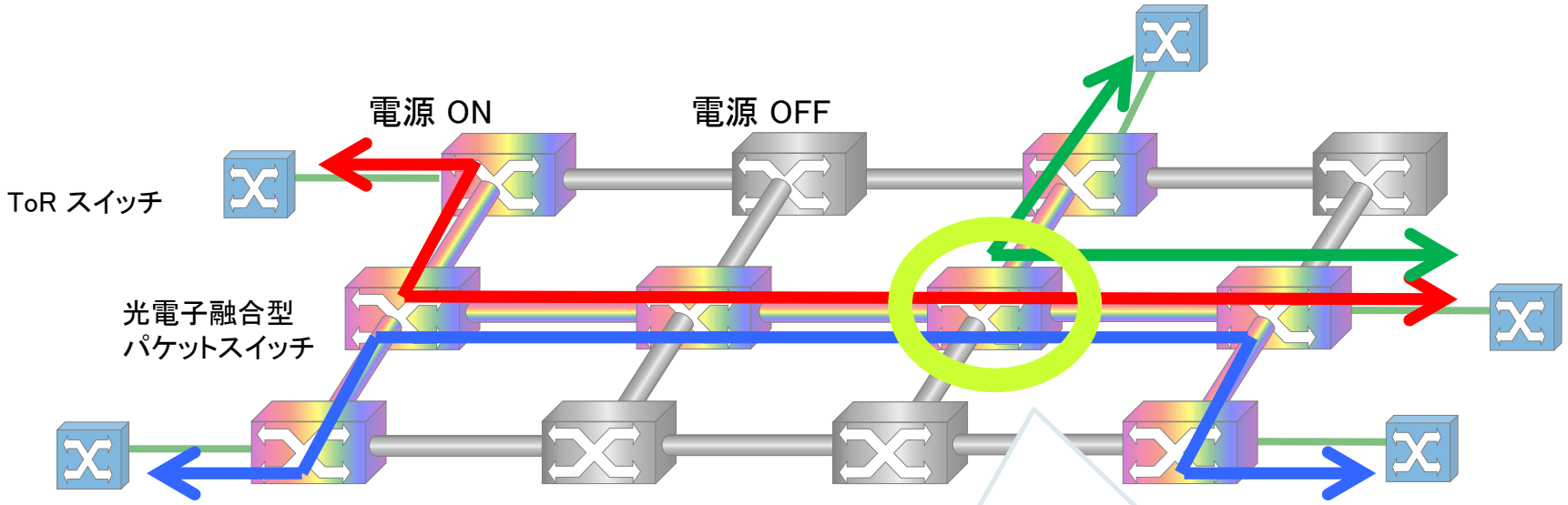
光電子融合型データセンターネットワークにおける低消費電力なトラヒック経路選択手法のアイデア

- **必要最低限のバッファのみを用いてトラヒックを集約**し、電源の投入が必要なスイッチ・バッファの総消費電力を最小化
  - ⇒ 多くのトラヒックで利用可能なスイッチおよびバッファのみ電源を投入



光電子融合型データセンターネットワークにおける低消費電力なトラヒック経路選択手法のアイデア

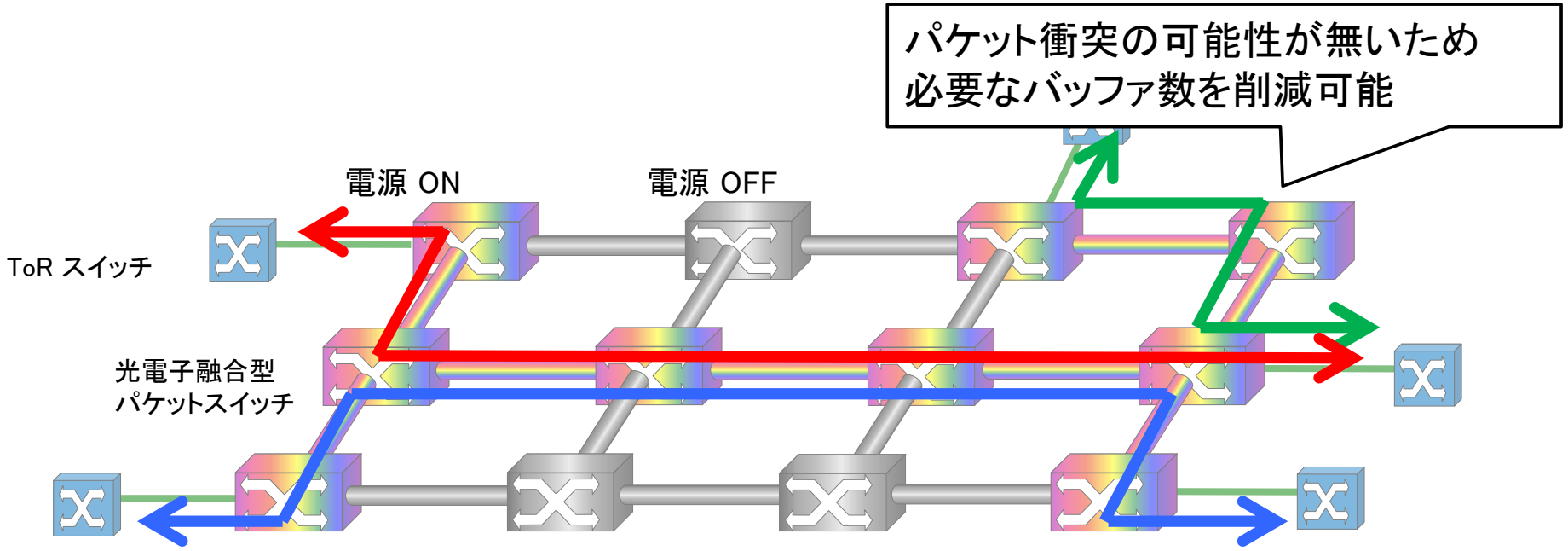
- **必要最低限のバッファのみを用いてトラヒックを集約**し、電源の投入が必要なスイッチ・バッファの総消費電力を最小化
  - 多くのトラヒックで利用可能なスイッチおよびバッファのみ電源を投入



光電子融合型データセンターネットワークにおける低消費電力なトラヒック経路選択手法のアイデア

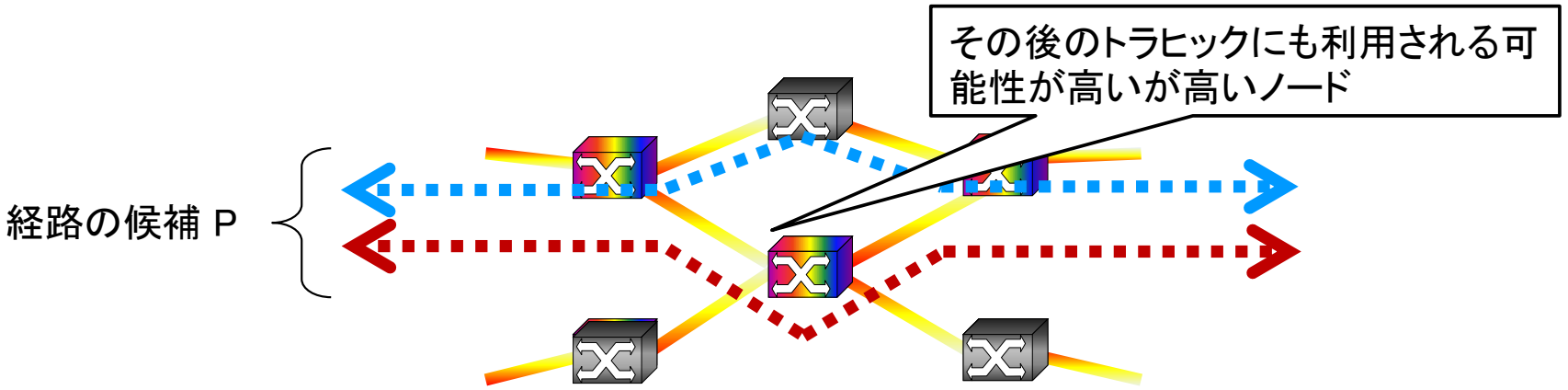
- **必要最低限のバッファのみを用いてトラヒックを集約**し、電源の投入が必要なスイッチ・バッファの総消費電力を最小化

  ➡ 多くのトラヒックで利用可能なスイッチおよびバッファのみ電源を投入



光電子融合型データセンターネットワークにおける低消費電力なトラヒック経路選択手法

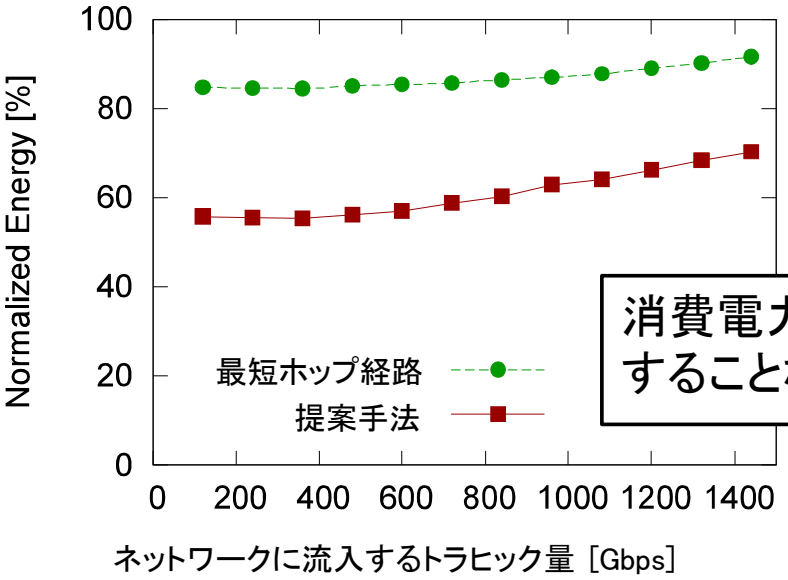
- 物理ホップ数が短いトラヒックから順に収容先の経路を確定
  - 短時間で経路を決めるため、最適化問題は用いない
- エンド間のトラヒックの経路決定の手順
  1. 性能要件を満たすエンド間の経路の候補 P を取得
  2. 経路の候補 P の内、新たに電源の投入の必要があるスイッチ・バッファの総消費電力が最小の候補を選択
  3. 最小の候補が複数ある場合、その後のトラヒックにも利用されやすい経路を選択



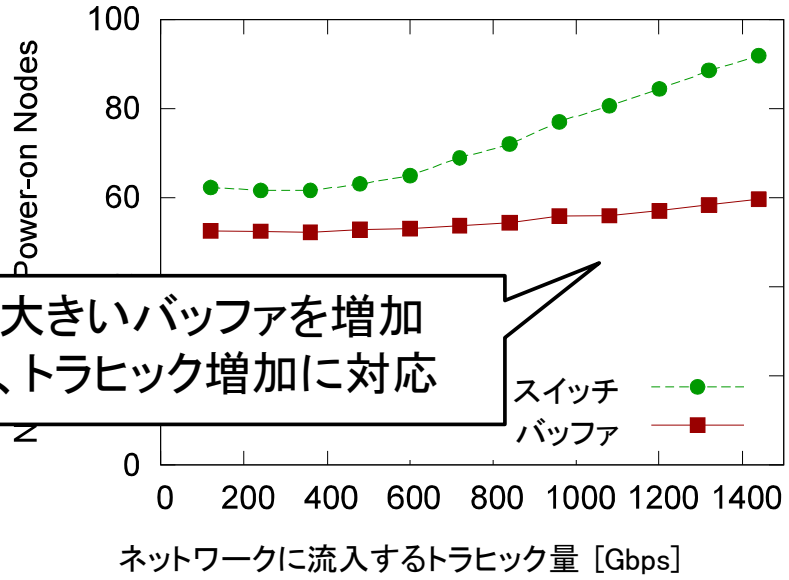
ネットワークに流入するトラヒック量の影響

- トラヒック量によらず消費電力を削減
  - トラヒック量が大きくなるにつれ削減効果が低下
  - 消費電力削減効果を高めるためには、トラヒック量が大きいエンド間を同一サーバラックに配置するなどのトラヒック量を削減する工夫が有効

消費電力削減効果：
全機器 ON に対する消費電力



提案手法により
電源が投入された機器数



消費電力の大きいバッファを増加することなく、トラヒック増加に対応

まとめ

- データセンターネットワークにおける光通信技術への期待
  - 従来： 光ファイバを介した中継のみ。スイッチングは電気。
    - 電気機器の消費電力の増大

  - 将来：
    - 光のままスイッチングするネットワークを導入→消費電力削減
      - 光パケット/光回線交換
    - 電気機器の利点と光ネットワークの利点を併せ持つネットワークへ

謝辞：
本講演で紹介した研究のうち、講演者らによる研究は情報研究機構の委託研究によるものである。