社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.

［招待講演］データセンターネットワークの研究動向

大下　裕一† 　村田　正幸†

† 大阪大学 大学院情報科学研究科
〒 565-0871 大阪府吹田市山田丘 1-5
E-mail: †{y-ohsita,murata}@ist.osaka-u.ac.jp

あらまし　近年、大規模なデータセンターの建設が進んでいる。データセンター内では、サーバが連携して多量のデータの処理を行っている。そのため、サーバ間を結ぶデータセンター内のネットワークは、データセンターの処理性能に大きな影響を与える。そこで、データセンターネットワークに関する様々な研究が進められている。本発表では、データセンターネットワークに関する研究のうち、主にネットワーク構成、経路制御、低消費電力化のための制御手法に関する研究動向について紹介する。

キーワード　データセンタネットワーク、トポロジ、経路制御、消費電力

[Invited Talk]Trends in Research on Data Center Networks

Yuichi OHSITA† and Masayuki MURATA†

† Graduate School of Information Science and Technology, Osaka University
1-5, Yamadaoka, Suita, Osaka, 565-0871, Japan
E-mail: †{y-ohsita,murata}@ist.osaka-u.ac.jp

Abstract　In recent years, large data centers with tens of thousands of servers have been built to handle very large amounts of data generated by various online applications. In a data center, servers handle such very large amounts of data by communicating with each other via the network within the data center. Thus, the network within the data center has large impacts on the performance of the data center. In this presentation, we survey researches on the data center networks. Especially, we focus on the topology of the data center network, routing in the data center netwrok, and the control methods to achieve low energy consumption.

Key words　Data Center Network, Topology, Routing, Energy Consumption

データセンターの現状

- 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
  - 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
    - 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
    - データセンター内のサーバ間のRAMを束ねて大きなRAMを持つコンピュータを構成するプロジェクトも存在
- データセンターの規模
  - 大規模化している
  - 1か所のデータセンターに数十万台のサーバを配備することも。

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
- データセンターの大規模化に伴って、複雑化
  - サーバ台数が少ない場合は、少数のスイッチで全サーバラックを接続するのみでもよかったが…
  - 数十万台以上のサーバを効率的に収容するネットワークが求められる

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問題点

- スケーラビリティの確保
  - 機器数を増やそうとすると、以下の問題が生じ、他の要求との両立が困難
    - ポート数の大きなスイッチの導入→消費電力の増大
    - 階層数の増加→遅延の増大・機器増による消費電力の増大
- 通信性能の確保
  - Oversubscriptionによる通信帯域の制限
  - Store and Forwardスイッチを多段通過することにより遅延が大
- 耐故障性の確保
  - Coreスイッチの故障が大きな性能劣化をもたらす
- 消費電力
  - Coreスイッチに高性能なスイッチ（消費電力大）を用いる必要がある

データセンター内のトラヒック状況

- 現状：サーバラック内の通信が多い
  - 現状のプログラムが、ラック内の通信を有効に使うように組まれているため。
  - ただし、データセンター内の処理が複雑化、データが大規模化、データセンターが大規模化するにつれ、サーバラック内によく使うデータを固めて配置することは困難になってくる

  - **ネットワーク内のどのサーバとも広帯域・低遅延の通信ができる環境が望まれている**
- 現状：データセンター内のトラヒックは著しく変動する
  - 秒オーダーで流れるトラヒックの傾向が著しく変わる

  - **特定のトラヒック状況にネットワークを最適化するということは困難**

データセンターネットワークにおける研究課題

- ネットワーク構成
  - 安価・サーバ間の帯域の確保・サーバ間を低遅延で接続・耐故障性に優れるという要求を満たすネットワーク構成の提案
- 経路制御
  - ルーティング情報の交換によらず、ネットワークの構造を利用した経路制御手法の提案
  - 突発的な環境変動が発生しても対応可能な経路制御手法の提案
- 省電力化のための制御
  - ネットワーク機器のON/OFFの制御による省電力化手法の提案

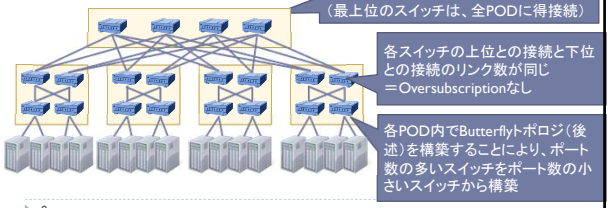
- ISPネットワークにおける研究と異なる点
  - 配線にかかる制約が緩く、単一管理者のネットワークのため、自由な構成・プロトコルを採用可能
  - 単一管理者が管理するノード数が著しく大きい
  - トラヒック変動が激しく、トラヒック観測・予測に基づいた制御が困難

既存電気パケットスイッチを用いたネットワーク構成

- ポート数が少なく帯域が小さいスイッチを用いて大規模なデータセンターを構築
  - ＝従来型ネットワークにおけるCoreルータの排除
    - メリット：ポート数が小さく帯域が小さいスイッチは、帯域あたりの消費電力(Watt/Gbps)が小さく、安価
- 電気スイッチのみを用いた構成
  - ポート数の少ないスイッチのみを用いて、広帯域の通信を提供できるようなネットワーク構造
- サーバ間にもリンクを構築する構成
  - 複数Network Interface Card (NIC)を持つサーバを用いて、サーバ間を接続
  - サーバ間の接続を効率的に利用することにより、広帯域の通信を提供しつつ、消費電力のさらなる削減

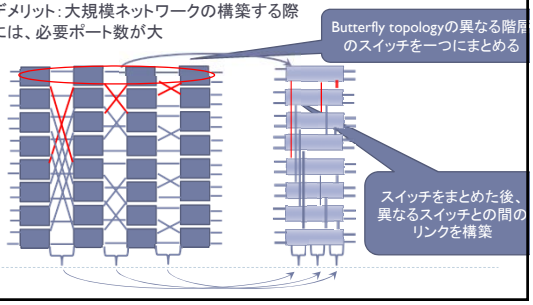
電気スイッチのみを用いた構成(1)　FatTree

- 小規模なスイッチのみで十分な帯域を確保できる構造
  - メリット：大きなスイッチを使わないことにより、コストを抑え、消費電力を抑えることが可能
  - デメリット：より大規模なデータセンターを構築するには、ホップ数を増大させるか、大容量のスイッチが必要



電気スイッチのみを用いた構成(2) Flattened Butterfly

- Butterfly構造をフラット化することによりホップ数を削減
  - メリット：Butterfly構造より、少ない消費電力で多くのサーバを収容可能（機器数削減のため）
  - デメリット：大規模ネットワークの構築する際には、必要ポート数が大

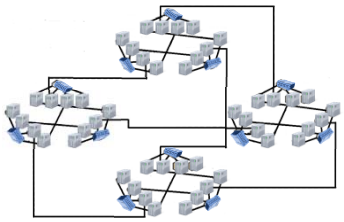


サーバ間にもリンクを構築する構成(1) DCell

- 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
  - メリット：数十万台以上のサーバを安価に接続可能
  - デメリット：通信帯域が狭い。遅延大

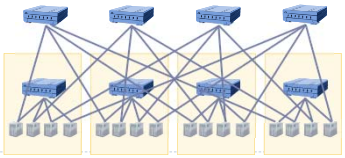
サーバ間にもリンクを構築する構成(2) FiConn

- Dcellからさらに各サーバのポート数を減らした構造
  - 一般的なNIC2枚のサーバで構築可能
  - 構築方法：階層的に構築（Dcellと同様）
    - **各層の接続において、未接続のポートが存在するサーバのうち半分しか用いない**
      =より上位の階層の接続のために未使用のポートを残しておく
  - メリット：一般的な機器のみで構成可能
  - デメリット：通信帯域が狭い。遅延大



サーバ間にもリンクを構築する構成(3) BCube

- スイッチを介して他のサーバと接続
  - 構築方法：階層的に構築
    - 最下位層：スイッチの全ポートにサーバを接続
    - 上位層：スイッチの全ポートに、下位層で異なるBCubeに属していたサーバを接続
  - メリット：Dcellよりも広い帯域を確保
  - デメリット：収容サーバ数を増やすためには、サーバのポート数を増やす必要がある

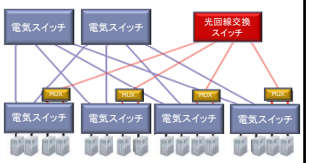


光通信技術を用いたネットワーク構成

- 光スイッチを用いることにより、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
- 構成
  - 光回線交換スイッチを用いた構成
  - 光パケットスイッチを用いた構成
- 課題
  - 回線交換スイッチ
    - パス切り替えに時間がかかる→短期的な変動に追随が難しい
  - 光パケットスイッチ
    - デバイス自体が研究対象

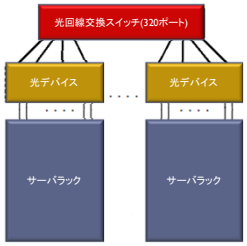
光回線交換スイッチを用いた構成(1)　Helios

- 構成
  - 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光パスで接続
  - それ以外の地点間のトラヒックはコアの電気パケットスイッチを経由して転送
- メリット
  - 少ないコストで広帯域・低遅延の通信経路を確保
- デメリット
  - パスの切り替え時間
    - 頻繁にトラヒック状況が変わる場合に対応不可



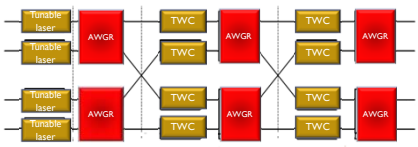
光回線交換スイッチを用いた構成(2)　Proteus

- 構成
  - 全サーバラックを大きな光回線交換スイッチに接続
  - 光パスで論理トポロジを構築
  - 全通信は論理トポロジを経由して通信
- メリット
  - 少ないコストで広帯域・低遅延の通信経路を確保
- デメリット
  - パスの切り替え時間
    - 頻繁にトラヒック状況が変わる場合に対応不可



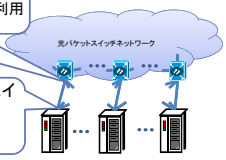
光パケットスイッチを用いた構成

- データセンター向け大規模光パケットスイッチを開発
  - 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
  - デメリット：
    - 一台の光パケットスイッチで接続する構成は耐故障性が弱い
    - ネットワーク規模が大規模化・トラヒック量が増大すると集中制御が困難に



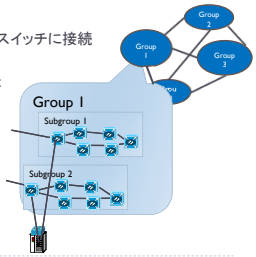
【発表者らの研究グループによる研究の紹介】光パケットスイッチを用いたネットワーク構成

- 光パケットスイッチを複数用いた効率的なネットワーク構造の検討
  - 概要
    - 光パケットスイッチをコアに配置
    - 1台の光パケットスイッチに複数台のサーバラックを接続
      - 光パケットスイッチの広帯域を効率的に利用するため
    - 各サーバラックも複数台の光パケットスイッチに接続
      - 故障発生時の接続性の確保のため



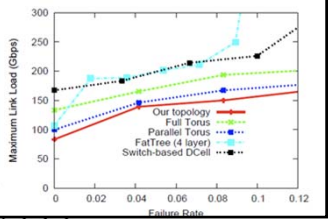
【発表者らの研究グループによる研究の紹介】光パケットスイッチを用いたネットワーク構成

- 光パケットスイッチネットワークをグループに分割
  - 各サーバラックからは、同一グループの複数の光パケットスイッチに接続
  - 同一グループに接続することで、各サーバラックからの配線長が長くなることを防止
- 各グループ内をサブグループに分割
  - 各サーバラックは異なるサブグループの光スイッチに接続
  - 各サブグループ間にリンクはなし
    - 同一のサーバラックが接続しているスイッチ間はネットワーク上近くである必要なし
- グループ内・グループ間のリンク数は輻輳が発生しないようにパラメータ調整



【発表者らの研究グループによる研究の紹介】評価

- 評価指標
  - 最大リンク使用率
    - トラヒック：各サーバラックはランダムに選択した特定のサーバラックとのみ通信
    - 故障：ランダムに発生
- 評価結果
  - 故障発生時も、提案手法が最もリンク使用率が低い
    - 必要な箇所にリンクを重点的に構築する構造になっているから

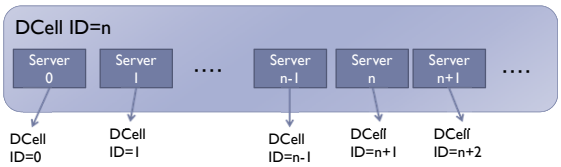


経路制御の研究の概要

- 大規模なネットワークに対応可能な経路制御手法の研究
  - 大規模なため、Link State Routing等では、以下の問題が発生
    - 故障等のトポロジ変動後の収束に時間がかかる
    - メッセージ交換量が大きい
  - アプローチ：トポロジ構造を利用
    - データセンタのトポロジは構造的に構築される
    - トポロジ構造を反映したノードのアドレスと、そのアドレスを用いてルーティング
- 予測不可能なトラヒック変動に対応する手法の研究
  - 輻輳の発生がデータセンタのパフォーマンス低下を招く
  - トラヒック観測に基づかずトラヒックを負荷分散させる手法

トポロジ構造を利用したルーティングの例 (Dcell Routing)

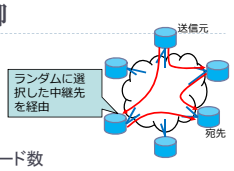
- Dcellの構造
  - 各Dcell内のサーバのIDから、そのサーバが接続している他のDcellのIDがわかる

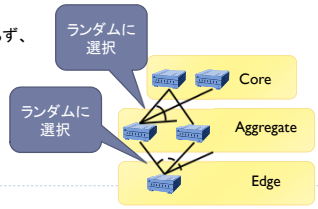


- ルーティングの方法
  - 宛先サーバが所属するDcell宛へのリンクを持つサーバに転送することを繰り返す

負荷分散のための経路制御

- Valiant Load Balancing
  - 宛先によらずランダムに経由する中継先を選択することにより負荷分散
  - 各ノードを流れるトラヒック量＝各ノードが送信するトラヒック量／ノード数
- Valiant Load Balancingをデータセンターに応用
  - メリット
    - トラヒックの傾向によらず、十分な帯域を確保
  - デメリット
    - 余分な資源消費
    - ホップ数の増大
    - 制御粒度の問題



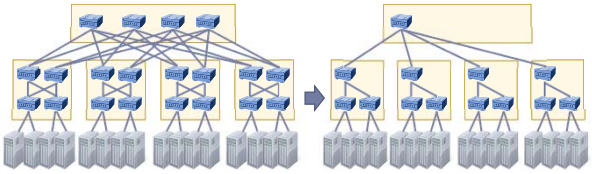


省電力化のための制御手法

- 前提
  - データセンターの使用率は常に100%ではない
- 基本的なアイデア
  - 現在の使用率に合わせて適切な消費電力にする
    - インタフェースの通信速度を下げる
    - 不要な機器の電源をOFFにする
    - より少ない機器でサービスを提供可能な仮想ネットワーク構成へ移行する

省電力化のための制御　ElasticTree

- 従来型データセンターネットワークを対象
- トラヒック量に応じて、そのトラヒック量を収用することができる最小のネットワークのサブセットのみ電源を投入する制御により、消費電力を抑制

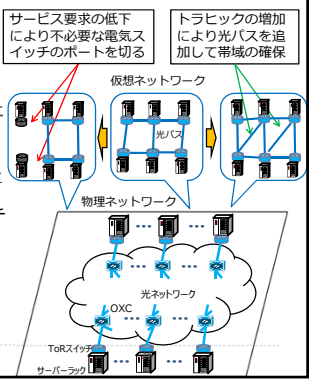


【発表者らの研究グループによる研究の紹介】省電力化のための仮想ネットワーク制御

想定するネットワーク構成
- 光ネットワークでコアを構成
- エッジにサーバラックの電気スイッチが接続
- エッジ間を光パスで接続することにより、仮想ネットワークを構成

光スイッチの消費電力は電気スイッチより著しく少ない
- 省電力化のためには電気スイッチの電力を抑えることが重要

- 電力削減の方針
  - 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

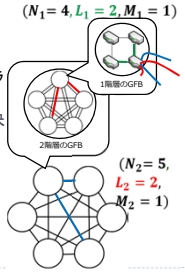


【発表者らの研究グループによる研究の紹介】仮想ネットワーク制御方法

- 提案手法の概要
  - パラメーターの調整によって経由トラヒック量、最大ホップ数、リンク数が調整できるネットワーク構造を提案
    - 階層的な構造
    - 各層のリンク数、ノード数がパラメータ
    - 各パラメータで構築された構造の各リンクのトラヒック量、最大ホップ数は容易に計算可能
  - 性能要件を満たすための構築パラメーター決定手法を提案

Generalized Flattened Butterfly (GFB)

$(N_1 = 4, L_1 = 2, M_1 = 1)$



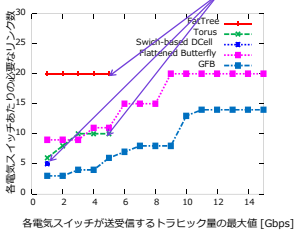
$(N_2 = 5, L_2 = 2, M_2 = 1)$

$(N_3 = 6, L_3 = 2, M_3 = 1)$

【発表者らの研究グループによる研究の紹介】仮想ネットワーク構造の評価

- 評価指標
  - 輻輳を発生させないトポロジを構成するのに必要なリンク数
    - リンク数が少ない
      ＝より多くの電気スイッチのポートの電源をOFF
- 提案手法で構築した仮想ネットワークトポロジー
  - 既存のトポロジーの半分のリンク数でトラヒックを収容可能
  - 必要な箇所にのみリンクを構築することにより、必要リンク数を抑えることが可能

これ以上のトラヒック量は収容不可



まとめ

- データセンターネットワークの研究課題
  - 適切なトポロジ構造
    - 安価に大規模なデータセンタネットワークを構築することに既存研究は主眼
    - 環境変動へ対応したルーティングの容易性も考慮したトポロジ構造は未検討
  - 環境変動への対応
    - データセンタネットワークの著しいトラヒック変動に効率的に対応する方法
      - 動的な経路最適化では間に合わないタイムスケールでのトラヒック変動が発生
      - 動的な経路最適化と局所的な輻輳判断の組み合わせが現実的な解？
  - 省電力化
    - 消費電力を考慮に入れた経路制御に関する研究
      - 進められているものの、大規模なデータセンタへの適用は未達成
  - マルチテナントデータセンタネットワーク
    - データセンタ内アプリケーションごとに仮想ネットワークを構築
  - 仮想ネットワークの動的制御・調停

謝辞:
本発表で紹介した発表者らの研究の一部は情報研究機構の委託研究によるものである。

参考文献

1. J. Ousterhout et. al. , "The Case for RAMClouds: Scalable High-Performance Storage Entirely in DRAM," SIGOPS Operating Systems Review, Dec 2009
2. D. Abts et al., "Energy proportional datacenter networks, " in Proc. the International Symposium on Computer Architecture, 2010
3. T. Benson et al. , "The Case for Fine-Grained Traffic Engineering in Data Centers" in Proc.INM/WREN 2010.
4. M. Al-Fares et al."A Scalable, Commodity Data Center Network Architecture" in Proc. SIGCOMM, Aug. 2008.
5. J. Kim et al., "Flattened Butterfly A Cost-Efficient. Topology for High-Radix Networks in Proc. of the International Symposium on Computer Architecture (ISCA), June 2007.
10. C. Guo, et al., "DCell: A scalable and fault-tolerant network structure for data center networks" in Proc. SIGCOMM, Aug. 2008.
11. D. Li, et al., "Scalable and Cost Effective Interconnection of Data-Center Servers Using Dual Server Ports," IEEE/ACM Transactions on Networking, Feb 2011.
12. C. Guo, et al. "BCube: A High Performance, Server-centric Network Architecture for Modular Data Centers" in Proc. SIGCOMM, Aug. 2009.
13. N Farrington et al. "Helios: A Hybrid Electrical/Optical Switch Architecture for Modular Data Centers," in Proc. SIGCOMM, Aug 2010.
14. Ankit Singlay et al. "Proteus: A Topology Malleable Data Center Network," in Proc. Hotnets, Oct 2010.
15. H. J. Cao et al. "Bufferless Optical Clos Switches for Data Centers," in Proc. OFC, Mar 2011.
16. Y. Ohsita et al., "Data Center Network Topologies Using Optical Packet Switches," in Proc. DCperf, Jun 2012.
17. A. Greenberg et al. "VL2: Scalable and flexible data center network" in Proc. SIGCOMM, Aug 2009.
18. B. Hellar et. al. "ElasticTree: Saving Energy in Data Center Networks," in Proc. NSDI, Apr. 2010.
19. 樺谷　他, "低消費電力を達成する大規模データセンタ内仮想ネットワーク制御手法," 電子情報通信学会フォトニックネットワーク研究会, Mar 2012.