M-041

ファットツリー型データセンタネットワークにおける障害箇所推定ルールを用いた自律的経路制御方式の検討

Autonomous Routing Algorithm using Fault Location Estimation for Fat-tree-based Data-center Networks

石川 さゆり† 坂田 匡通† 小川 祐紀雄†
Sayuri Ishikawa Masayuki Sakata Yukio Ogawa

1．まえがき

近年，クラウドコンピューティング／クラウドサービスが社会に浸透し，データセンタ(DC)の大規模化が進んでいる[1]．また，MapReduce 等の大規模分散処理アプリケーションの導入が進み，サーバ間通信が増加している[2]．このような中，DC では大規模・高性能なネットワークが求められるとともに[3]，構成変更や拡張にスケーラブルに追従し，通信状態の変化に対しても安定した性能を提供するネットワークが必要となる．

これに対し，我々のグループでは，多数の小型スイッチにて Fat-tree 型[4]のネットワークを構成し， スモールスタート/スケールアウトを可能とする DC 向けネットワークを検討している．小型スイッチを用いてスモールスタートすることにより，システム構築時の初期投資を大幅に抑えることができる．さらに，システムの構成変更や拡張に応じて，小型スイッチの並列化台数を増やすことで，システム規模をスケーラブルに拡大することが可能となる．これにより，無駄のない投資が実現できる．

しかし，サーバ数百台規模の大規模システムに適用した場合，数十台規模の多数のスイッチにより構成されるため，障害時の経路制御が複雑になるという課題がある．

本報告では，Box 型スイッチを多数利用して構成した Fat-tree 型ネットワークにおいて，障害が発生した際の対策の検討結果を報告する．

2．Fat-tree 型データセンタネットワーク

我々は，価格性能比の高いボックス型スイッチを複数用いて Fat-tree 型のネットワークを構成し，スモールスタート/スケールアウトを実現させる方式を提案している．本章では，Fat-tree 型データセンタネットワークの構成と，障害時における課題を示す．

2.1 基本構成と経路制御方式

Fat-tree のネットワーク構成を図 1 に示す．図に示す上段スイッチと下段スイッチの間はメッシュで接続し(上段スイッチは，全ての下段スイッチと接続する)，上段スイッチ同士，下段スイッチ同士は接続しない．ここでは，上段スイッチをファブリックスイッチ(F-SW)，下段スイッチをポートスイッチ(P-SW)とし，F-SW 及び P-SW から構成するネットワークをスケーラブルネットワーク(SCNW)と定義する．SCNW では，P-SW が負荷分散を担い，F-SW が中継(フォーワーディング)を担当する．従来，このような複数経路を持つ構成の経路制御には，STP(Spanning Tree Protocol)を利用することが一般的であったが，STP は，複数ある経路のうち，1 経路以外の通信をブロックするという仕様であった(マルチパス接続ではない)．これに対し，SCNW では，サーバ等のエンドノードと接続する P-SW から，F-SW へのフレーム転送を，MAC アドレスや IP アドレス等に基づく経路分散で構成し，マルチパスを実現している．

SCNW の，フレーム転送について説明する．全ての P-SW は，上記分散方法に従いフレームを転送し，全ての F-SW は，経路の学習に基づきフレームを転送する．例えば，図 1 に示すように，Server1 から Sever3 への通信において(往路)，Server1 が送信するフレームは，P-SW1 から F-SW1 へ転送される．このとき，F-SW1 では Sever1 の接続経路を FDB(Filtering Database)に学習する．次に，F-SW1 は，自身の FDB を参照しフレームの転送を行う．宛先が FDB に未登録の場合，F-SW1 は，残りの全ての経路にフレームを送信する(フラッディング)．また，Server3 から Sever1 への通信において(復路)，Server3 が送信するフレームを受信した P-SW3 では，往路と同一の中継スイッチ(F-SW1)にフレームを転送する．この同一中継スイッチへの転送(シンメトリックルーティング)は全 P-SW が同一の分散ルールを持つ事により実現する．F-SW1 では，Server3 の接続経路を FDB に学習し，既に学習済みの FDB を参照し，P-SW1 にフレームを転送する．

以上に示すように，2 階層の Fat-Tree 型のネットワークにおいて，多数のマルチパスにフレームを分散転送しつつ，シンメトリックルーティングにより効率的な転送を実現している．これにより，規模に依らず，SCNW 全体での負荷分散と低レイテンシが実現可能となる．

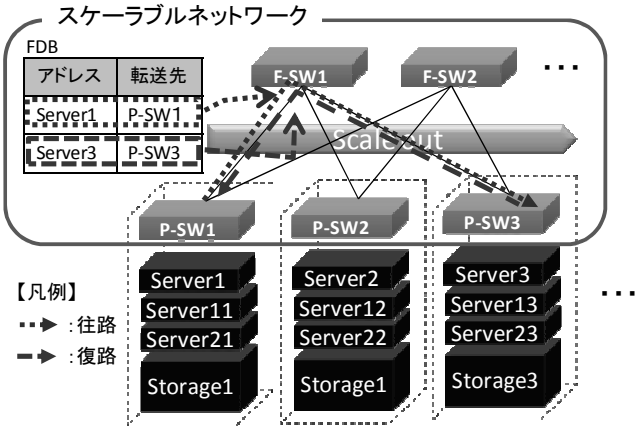

図 1 スケーラブルネットワークの構成

†(株)日立製作所 横浜研究所
Yokohama Research Laboratory, Hitachi, Ltd.

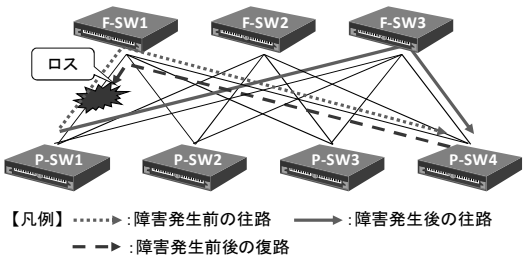

図 2　シンメトリックルーティング崩壊時の問題

2.2　SCNW における障害発生時の問題

SCNW では，構成と転送方法を限定していることで，ネットワーク内で障害が発生した場合，障害発生箇所へ接続する経路を通るフレームの転送を，他の経路に振り替えることにより，迅速な障害の回避を実現する．しかし，経路制御はシンメトリックルーティングを前提としているため，その前提が崩れた場合への対応が必要になる．例えば，図 2 に示すように，F-SW1 と P-SW1 の間のリンクにて障害が発生した場合，障害発生リンクと接続する P-SW においてのみ，経路変更が実施されると，シンメトリックルーティングが崩壊し，F-SW における FDB の学習が完結しないため，フレームのロスが発生する．

2.3　課題

本節では，SCNW の構成に基づく課題と，性能に関する課題について説明する．

2.3.1　SCNW の構成に基づく課題

前節に示した問題を解決するためには，SCNW 内のどの箇所で障害が発生した場合においても，シンメトリックルーティングを成立させることが必要となる．従って，各スイッチが，シンメトリックルーティングを成立させることが 1 つ目の課題となる．課題を解決するためには，各スイッチは，以下(1)～(3)に示す場合において，それぞれの経路制御を実施する必要がある．

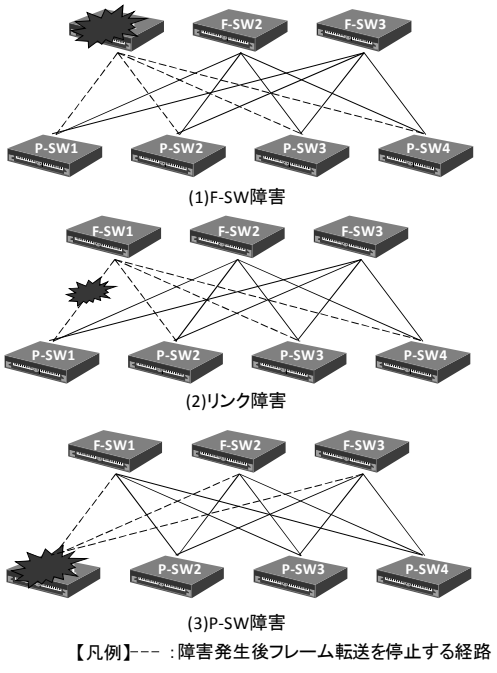

図 3　障害発生時の経路制御

(1)F-SW1 に障害が発生した場合

各 P-SW は，障害の F-SW1 への経路を，経路分散から除外し，全ての P-SW での経路分散を統一させる．

(2)F-SW1～P-SW1 間のリンク障害が発生した場合

前節にて示したように，シンメトリックルーティングを成立させるため，障害発生リンクに接続する F-SW1 を経由するフレームの転送を停止する．従って，各 P-SW は，F-SW1 への経路を，経路分散から除外し，全ての P-SW での経路分散を統一させる．

(3)P-SW1 に障害が発生した場合

各 F-SW は，障害の P-SW1 への経路を，経路分散から除外する．P-SW 側の経路変更はない．

2.3.2　性能に関する課題

DC の規模が拡大するにつれ，ネットワークにおける装置数の削減，ケーブルの簡素化，それに伴う装置コストや運用コストの軽減のニーズが高まり[5]，LAN/SAN 統合の動きが活発化している[6]．SCNW においても，このような動きに対応するため，LAN/SAN 統合の適用を考えると，SAN では，ファイバチャネル(FC)ノード(ex.ストレージ)の接続要件から，625msec 以内での経路切り替えが必要となる[7]．そのため，SCNW における経路切り替えを，625msec 以内で実現することを 2 つ目の課題とする．

3.　障害箇所推定ルールを用いた自律的経路制御方式

本章では，前章にて示した課題を解決するための，障害時箇所推定ルールを用いた自律的経路制御方式を提案する．以下，提案方式に関して説明する．

3.1　検知方法

障害が発生した際に経路制御を実施するためは，障害が発生したことを検知する必要がある．はじめに障害検知方法を説明する．

SCNW では，F-SW 同士又は P-SW 同士は非接続であるため，経路分散の役割を担う P-SW のうち，障害が発生したリンクに直接接続していない P-SW は，障害が発生したことを知り得ない．従って，P-SW は，障害のリンクに接続する F-SW から通知を受け，障害の発生を認識する．

(1)F-SW 障害や(3)P-SW 障害の場合は，スイッチが，ポートのリンクダウン状態を監視することで検知を行う．

(2)リンク障害の場合は，障害が発生したリンクに接続する P-SW 及び F-SW のみが障害を検知する．障害を検知した F-SW は，P-SW に対して障害の通知を実施することにより，全ての P-SW が障害を検知する．

また，(2)リンク障害は 2 種類考えられ，(2-1)リンクダウンと，(2-2)片断線とがある．(2-1)リンクダウンの検知は，ポートのリンクダウン状態を監視することにより実施する．一方，(2-2)片断線の検知は，各スイッチ間で Keep alive を実施し，Keep Alive を受け取らないことにより検知する．

3.2　障害箇所推定ルール

P-SW に対して障害を通知する F-SW は，(2)リンク障害であるか，(3)P-SW 障害であるかを判別することができない．そのため，経路制御を実施する各 P-SW は，障害発生箇所を推測し，障害箇所に応じた適切な経路制御を実施する必要がある．

前述の通り，P-SW が障害を検知する状況は，(1)F-SW 障害と(2)リンク障害である．このときの，P-SW における経路制御処理は，いずれも，障害の F-SW，もしくは障害のリンクに接続する F-SW を，分散経路から除外することである(図 3(1)(2))．F-SW が障害を検知する状況は，(2)リンク障害と(3)P-SW 障害である．このときの，P-SW における経路制御処理は，一方は，障害のリンクに接続する F-SW を，分散経路から除外することであり(図 3(2))，他方は，何もしないことである(図 3(3))．従って，P-SW は，リンク障害であるか，P-SW 障害であるのかを瞬時に推測し，適した経路制御を実施する必要がある．

ここで，両者の障害時における状況を比較すると，(2)リンク障害の場合には，障害を検知する装置は単一の F-SW であることに対し，(3)P-SW 障害の場合には，障害を検知する装置は複数の F-SW であるという違いがある．そこで，P-SW は，障害を検知した F-SW の数，つまり，F-SW から障害通知を受信したポートの数に従い，リンク障害であるのか，P-SW 障害であるのかを判別するルールを用いることで，迅速な障害回避処理が可能となる．

3.3 障害箇所推定ルールを用いた自律的経路制御アルゴリズム

前節までに示した障害箇所推定ルールを用いた経路制御の内容を，図 4 を用いて説明する．P-SW は，①障害通知を受信した場合，②一定時間の間に障害通知を受信したポート数をカウントする．その後，③カウント数を元にしきい値判断を実施する．障害通知を単一のポートで受信した場合，障害を検知した F-SW は 1 台であることがわかるため，リンク障害である．従って，障害通知を受信したポート数がしきい値よりも小さい場合，リンク障害であると判断する．そして，各 P-SW は，フレーム転送経路から，障害通知を受信したポートを，除外する．このとき，しきい値を設けた理由は，リンク障害の複数箇所同時発生に対応するためである．一方，障害通知を複数のポートで受信した場合，障害を検知した F-SW は複数であることがわかるため，P-SW 障害である可能性が高い．従って，この場合，P-SW が実施する処理はない．

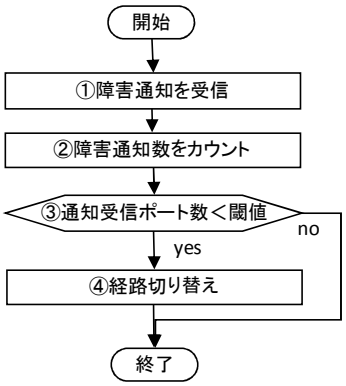

図 4　自律的経路制御アルゴリズム

4. 評価実験

提案方式に基づく経路制御ソフトウェアをレイヤ 2 スイッチ(実機)に搭載し，障害を発生させた際の経路制御処理に関して，評価を行った．障害箇所に応じた処理を実施し，かつ，目標値以内に経路切り替え完了しているかを確認した．

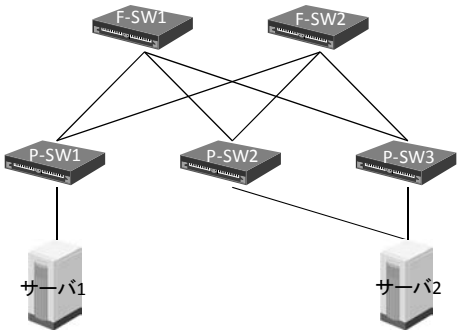

図 5　評価構成

4.1 条件

評価構成を図 5 に示す．F-SW2 台，P-SW3 台構成とし，各 P-SW から，2 台の F-SW に対して 1 本ずつリンクを接続した．さらに，サーバ 1 から，P-SW1 へ接続し，サーバ 2 から，P-SW2 と P-SW3 へと接続した．このとき，提案方式におけるサーバ 1～サーバ 2 間の通信は，P-SW1，F-SW1，P-SW3 を経由するものとした．発生させる障害の種類は，(1)F-SW1 の装置障害，(2-1)F-SW1 と P-SW1 の間のリンク障害(リンクダウン)，(2-2)F-SW1 と P-SW1 の間のリンク障害(片断線)，(3)P-SW3 の装置障害の 4 種類とした．スイッチの装置障害はスイッチの電源を落とすことで実施し，リンク障害はケーブルを抜くことで実施し，片断線は途中経路にスイッチを挟み，片経路のみパケットをフィルタリングすることで実施した．

また，経路切り替え時間の測定は，サーバ 1 からサーバ 2 へ，2msec 間隔で UDP パケットを送信し，途中経路にて障害を発生させ，ロスパケット数をカウントすることにより算出した．このとき，従来技術との比較として，①提案方式と，②Rapid Per-VLAN Spanning-Tree（R-PVST）における，経路切り替え時間を比較した．

4.2 経路制御時間の試算と Keep Alive 間隔の決定

提案方式において，障害発生から，経路切り替えの処理が完了するまでに要する時間を試算した．試算結果を図 6 に示す．図に示す t は，提案方式の Keep Alive 間隔となる．提案方式は，一定時間毎に障害の検知処理を実施しているため，1 回のループ処理で障害検知までに要する時間は，最大 t となる．

(1)F-SW 障害が発生した場合，障害を検知するスイッチは，P-SW となる．従って，P-SW が，(a)障害が発生してから障害を検知するまでの時間と，(b)障害を検知してから経路制御が完了するまでの時間，つまり(a)+(b)が，経路切り替えに要する時間となる． (3)P-SW 障害の場合は，障害を検知するスイッチは F-SW となるため，図の(c)＋(d)＋(h)が，経路切り替えに要する時間となる．(2-1)リンクダウンの場合は，F-SW が，リンクダウン箇所に接続していない P-SW に対して，障害発生の通知を実施し，P-SW が，経路切り替えの処理を実施するため，図の(c)＋(d)＋(e)＋(f)＋(g)が，経路切り替えまでに要する時間となる．(2-2)片断線の場合は，F-SW と P-SW 間で実施している Keep Alive の失敗を検知した後に，(2-1)リンクダウンと同様に，F-SW から P-SW への障害発生の通知を実施した後，P-SW において経路切り替えの処理を実施する．今回の実装では，片断線の検知タイミングを Keep Alive の 3 回失敗

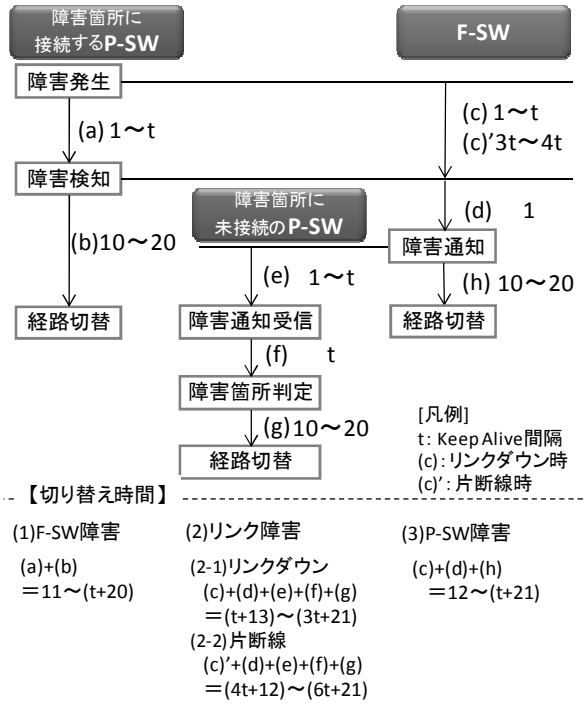

図 6　経路切り替え時間の試算(数値単位:msec)

時としたため，(c)'＋(d)＋(e)＋(f)＋(g)が，経路切り替えまでに要する時間となる．

従って，前述した 2 つ目の課題である 625msec 以内での経路切り替えを達成するためには，経路切り替えに要する時間が最大となる(2-2)片断線時の時間 (6t+21)＜625msec となればよい．そのため，t=100 以下が，この要件を満たす値となる．しかし，実運用においては，他処理の割り込みやトラフィックの混雑等が原因で，Keep Alive 間隔のずれが発生することを考慮し，今回は，Keep Alive 間隔を 50msec と設定した．

4.3　経路切り替え時間の評価結果

前述した 4 種類の障害における，経路切り替えの評価実験を行った．結果を以下に示す．

①提案方式(Keep Alive 間隔は 50msec)にて，4 種類の障害を発生させた場合の，経路切り替えを評価した．その結果，全ての障害時において，適した経路制御を実施することを確認した．

表 1 には，①提案方式と，マルチパスではないが，従来技術による経路切り替えの参考として②R-PVST とを用いて切り替えに要する時間と，障害からの復旧時に通信不能になる時間を計測した結果を示す．値は，それぞれ 5 回ずつ測定した平均値とし，単位は msec とする．①提案方式では，全ての障害発生時の経路切り替えにおいて目標値以内を達成し，②R-PVST と比較しても高速に経路切り替えを実施していることがわかり，提案方式の有効性が確認できた．また，障害からの復旧に関しては，①提案方式では，マルチパスを実現しているため，ロスが発生しないことを確認した．一方，②R-PVST では，経路が複数ある際，1 経路以外の経路をブロックするという仕様であるため，障害復旧時において，通信経路が切り替わる際，フレームロスが発生していることがわかった．

②R-PVST において，復旧時に通信断が発生する原因は，R-PVST の経路の変更において，ポートをブロックするタイミングと，FDB を更新するタイミングが合っていないためであると考えられる．また，片断線は想定外のためか，FDB が再更新されるまで復旧することはなかった．

表 1　障害時の経路切り替え時間と障害復旧時の通信断時間

条件 / 障害の種類		障害時の経路切り替え時間		復旧時の通信断時間
		①提案方式	②R-PVST	②R-PVST
(1)	F-SW 障害	50	735	534
(2-1)	リンクダウン	100	460	316
(2-2)	片断線	289	526	-(復旧せず)
(3)	P-SW 障害	224	274	36

また，①提案方式において，表 1 に示した値は，図 6 で試算した値に適していることを確認した．

5. まとめ

本稿では，DC 向け高スケール・低遅延な Fat-tree 型ラック間ネットワーク(SCNW)における，障害発生時の経路制御方式に関して検討を行い，以下を達成した．

- 各スイッチが障害箇所を推定し経路制御を実施
- マルチパス及びシンメトリックルーティングを保障
- Keep alive 間隔 50msec にて平均 289msec での経路切り替えを達成

今後の課題は，DC を想定した大規模構成にて検証を行い，本方式の有効性を確認することである．

参考文献

[1] IDC Japan, “国内データセンター規模別分布 2009 年の推定と 2010 年～2014 年の予測”, J10280101 (2010).
[2] T. Benson, A. Akella, and D.A. Maltz, “Network traffic characteristics of data centers in the wild”, in Proc. Internet Measurement Conference, 267-280 (2010).
[3] David Bernstein, Erik Ludvigson, “Networking Challenges and Resultant Approaches for Large Scale Cloud Construction”, 4-8, 136-142 (2009).
[4] Charles E. Leiserson, “Fat-trees - University networks for hardware-efficient supercomputing”, IEEE Transactions on Computers, Vol. C-34, 892-901 (1985).
[5] Loukas Paraschis, Sudhir Modali, “Data Center Transport in the Zettabyte IP Network”, Photonics Society Summer Topical Meeting Series 2010 IEEE, 19-21, 227 - 228 (2010).
[6] Reinemo. S., Skeie. T., Wadekar. M.K., “Ethernet For High-Performance Data Centers: on The New IEEE Datacenter Bridging Standards”, Micro, IEEE, 42-51 (2010).
[7] INCITS, “FIBER CHANNEL BACKBONE-5(FC-BB-5) REV 2.00”, 122 (2009).