スケーラブルスイッチDynateraのトラフィックセンサ

天海 良治, 釘本 健司, 清水 奨
NTT 未来ねっと研究所

概 要

スケーラブルスイッチDynateraは, 市販ネットワークスイッチに光スイッチを組み合わせたスイッチクラスタで, 負荷の高いリンクを検出して, そこを通過するトラフィックを光スイッチ経由のショートカットパスに回すことで, 全体のスループットを確保する. ここでは, ソフトもハードも与えられたままの不可侵リソースである市販スイッチとのstruggleについて述べる.

1 はじめに

ネットワークトラフィックの増加を支えるには, 回線の増強とともに, データを交換するスイッチの高速化, 大容量化が必要である[1].

最近の高性能スイッチは, データ交換のノンブロッキング性を追求している. ノンブロッキング性とは, スイッチのポート間で転送パスが同時に複数設定でき, スイッチのバックプレーン (スイッチファブリック) や転送用ハードのリソース不足によって転送が止まる (ブロックされる) ことがないことをいう. ノンブロッキング性を確保するには, 負荷が集中するバックプレーンに (ポート数 * ポート最高速度) 以上の速度が必要であり, この部分の開発は年々困難となっている. 逆に, バックプレーン速度によってスイッチ全体の性能が押えられるので, このようなモノリシックな構成 (図1) では, ポートの増設や増速のためにはスイッチ全体のアップグレードが必要となる.

だが, スイッチのすべてのポートが最高速度で常にデータをやりとりしているわけではない. 通信の多い時間, 少ない時間といった時間的偏在, 通信の多いポート, 少ないポートといった空間的偏在が存在する. また, 時間的偏在にも長期的, 短期的なものがある. 例えば, 深夜は通信が少なくなる, 週末は少ない, 時々バースト的に通信が増大する, といったことがある.

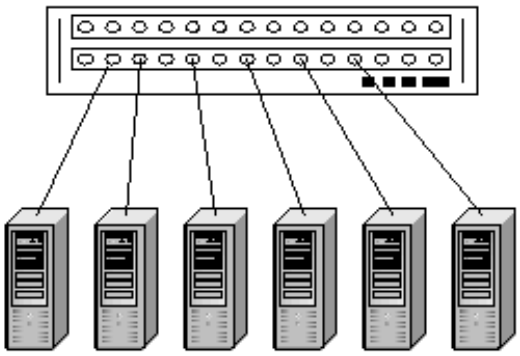

図 1: モノリシックスイッチ

我々の開発しているDynatera[1]は, このようなトラフィックの偏在を前提として, 複数の独立したスイッチ (スイッチモジュール) をインターリンクで接続したスイッチクラスタの構成をとる. インターリンクはスイッチの通常のポートに接続される. さらに, 動的に接続を変更できる光スイッチを経由するインターリンクを用意して, トラフィックの多い通信ペアを光スイッチ経由のショートカットパスに誘導する制御を行なう (図2). このアプローチでは, スイッチ全体でのノンブロッキング性を保証することはできないが, 光スイッチを制御して転送停止時間を少なくすることは可能である. 我々は, これをレアブロッキング性と呼んでいる. 現在, プロトタイプで動作検証を行なっている.

以下, Dynateraの構成, 特にスイッチ制御用データの収集部分, トラフィックセンサについて述べる.

[1] ここでは, レイヤ2スイッチ (ブリッジ) とレイヤ3スイッチ (IPルータ) の両方を含んで単にスイッチと呼ぶ

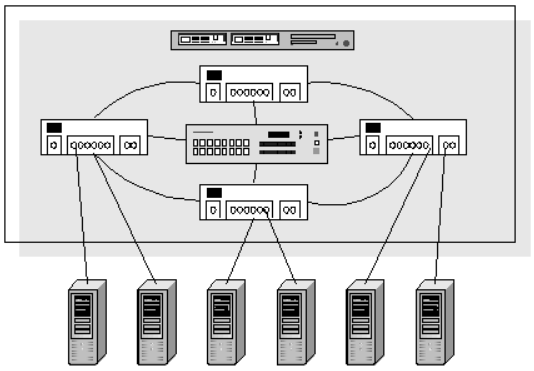

図 2: Dynatera の概念図. 市販スイッチ間を光スイッチで接続

2 Dynatera の構成

プロトタイプでは, クラスタのスイッチモジュールには, 市販のギガビットが 8 ポート装備されているスイッチを使用している. 光スイッチにはシングルモード光の入力 8 ポート, 出力 8 ポートの光インタフェースがあり, 各々の入力を 8 つの出力のうちの任意の 1 つに接続することができる. ただし, 入力を複数の出力に分岐させたり, 複数入力を 1 つの出力に接続することはできない. 接続の制御は RS-232C シリアルインタフェースで行なう.

スイッチモジュールと光スイッチの接続トポロジは, (図 2) のように, スイッチモジュールをリング状に配置してショートカットパスの生成に光スイッチを使用するのが一例であるが, 他にもスイッチモジュールを直列に配置する, ツリー形状をとる, といった選択肢がある. 実際には, Ether スイッチをリング状に接続するとパケットの無限転送が発生するので, トポロジには制限がある.

Dynatera 制御ソフトウェアは主にトラフィックセンサ, トポロジプランナ, トラフィックアクチュエータから構成される. センサは, Dynatera を通過するトラフィックを計測し, プランナのための基礎資料を提供する. 詳細は後述する. プランナは, センサから提供されたクラスタ内のトラフィック状況からトラフィック状況モデルを構築し, データのブロッキングを抑え, スループットを保つような構成計画を立て, 必要ならアクチュエータに指令を出して Dynatera 内のトポロジを変更する. 判断の時間間隔は, 長期的なものと短期的なものとに分けている. 長期的判断は, 例えば 1 日を単位として, それまでの統計情報を元にある程度のトラフィック予測をして Dynatera のスループットを保つ. 短期的判断は, 不定期に発生するバーストトラフィックに反応して, インターリンクでのブロッキング発生を回避することを目的とする. アクチュエータは, プランナの指示に基づき, 光スイッチを操作してトポロジ変更を行なったり, スイッチモジュール内の MAC アドレス表などの操作をして実際のトラフィックの誘導を実施する. 操作タイミングはセンサにもフィードバックして, トラフィック変動がアクチュエータの動作によるものであることを伝えている.

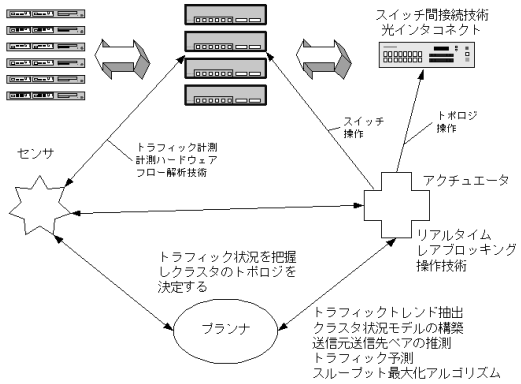

図 3: ソフトウェア構成

3 Dynatera のトラフィックセンサ

センサでどのくらいの情報が得られるかによって, Dynatera の動作は大きく影響をうける. Dynatera で最も有用な情報は, Dynatera 内を通過するデータストリームの MAC アドレスペアごとのトラフィック (マイクロフロー) である. これが実時間で計測できれば, プランナがスイッチモジュール間でトラフィックのバランスをとったりバーストトラフィックを迂回させることも比較的容易である.

フロー計測には, 直接計測と間接計測の 2 つがある. 直接計測は, スイッチのポートごとにパケットキャプチャ装置を挿入してすべてのパケットを取り込み, データを比べることでフローを見付ける. だが, たとえ TCP ヘッダ部までを取り込むとしても, 数 10 の GbE ポートからのデータは膨大なものとなり, これからフロー

を得るのは現実的ではない．また，キャプチャ装置のコストも大きなものとなる．間接計測は，データとして比較的容易に得られるスイッチのポートごとの転送オクテット数やパケット数から，トラフィックの送受ペアの推測やディレイを推測するものである．パケット到着についてモデルを仮定して，実測データに推測値を近づけていくくり返しアルゴリズムを使用する．ただ，Dynatera に応用できるかどうか，収束の速度や計算量について検証中である．

現在，Dynatera のセンサは間接計測を行なっている．スイッチモジュールや接続されているユーザマシンには手を入れずに，どこまで必要な情報が得られるかを試している．もともと市販スイッチは，固定化された固いリソースで手を入れる余地はほとんどない．また機種によって保持している情報も異なる．それでも，SNMP[2] エージェントはほぼ装備されているので，同じインタフェースで情報を得ることは可能である．SNMP で得られる情報のうち Dynatera で利用するのはポートごとの転送オクテット数である．ただ，転送オクテット数でも機種によって，無符号 64 ビットの値 [3] まで得られるものと，無符号 32 ビットまでのものがある．1GBps のポートが最高速度で転送すれば，32 ビットのオクテットカウンタは 50 秒たたずにあふれてしまう[2]．さらに，SNMP で得られる情報がある程度の時間がたたないと更新されない機種も存在する．連続してデータが転送されていても，SNMP 経由で観測すると断続的に送られているように見えてしまう．複数のポートの送受関係を調べたいときに大きな問題になった．

Dynatera では，くり返しアルゴリズムによる送受ペア推測はまだ検証中であって実装していない．プロトタイプでは，次のようにしてペアを検出している．まず，スイッチのポート単位のトラフィックペアを求める．

定期的に，SNMP でもってすべてのスイッチモジュールの全ポートの転送オクテット数を取得する．トラフィックの小さなポートは無視する．これを除いた 2 つのポートの単位時間あたり転送量 (転送速度) が (ほぼ) 一致しているものを検出する．TCP ストリームであれば，データの流れと逆方向に ACK が流れているので，両方を検出することでフローを決める．さらに，転送速度の変化量の一致を検出してトラフィックの重なりもある程度検出している．ポートの単位でトラフィックペアが検出できればスイッチの MAC アドレス表の MAC アドレス–ポート番号のデータを突き合わせることでマイクロフローを得ることができる．

スイッチ内の MAC アドレス表を得るインタフェースは統一されていない．telnet でスイッチにログインして，コマンドライン インタフェースで表示させたり，WEB で設定できるスイッチであれば，表を表示させる URL を送るといった方法にならざるを得ない．表の形式もさまざまなので，機種ごとにデータの正規化が必要となっている．

現在の方法では，ある程度の時間持続するバーストトラフィックの検出は可能であるが，長期的判断に使用できるような細かいデータは得られない．

4 課題

Dynatera の課題は多く残っている．センサの精度向上もその 1 つだが，これ以外に，長期判断におけるトラフィック予測，スイッチモジュールを多数接続したときのスケーラビリティの確保，大きなマルチキャストトラフィックの誘導などが大きな課題である．

5 関連研究/技術

パケットの直接観測の例:
ベル研では，TCP の振舞いや統計情報を得るために回線をながれるパケットをすべて取り込んで解析をしている [6]．Duffield らは，ルータにハードを追加し，1 つのパケット全体を DSP でハッシュしてラベルを付け，パケットがどのルートを通過したかを解析している [7]．

間接観測の例:
人間の内部を非破壊に観測する Computerized tomography から言葉を得た，Network Tomograpy[8] と呼ばれる観測手法が報告されている． パケット到着がポアソン分布に従うとの仮定のもと，計測したスイッチのポート転送カウンタの値から EM (Expectation–Maximization) アルゴリズムなどで Origin-Destination Matrix を推測する例がある．

スタッカブルスイッチ/ポートトランキング
スイッチのポート数を増やすために，スイッチ同士を専用ケーブルで結んで 1 つのスイッチとできるスタッ

[2]MRTG[4] は標準では 5 分間隔でデータを得る．32 ビットカウンタのギガビットスイッチの MRTG グラフはまったく意味がない

カブルスイッチは市販されている. ポート間を結ぶリンクを束ねて 1 つのリンクとして動作させて障害対策や通信量を確保する機能も多くのスイッチに備わっている. だが, これらはトポロジは固定である. 3Com XRN (eXpandable Resilient Networking) [9] も複数のスイッチをインターリンクで結んで分散管理し, 高いスループットと冗長性, 可用性を提供しようとしているが, 動的なトポロジ変更までは提供していない.

Juniper T シリーズ [10] には, 320Gbps ものスイッチファブリックをもったスイッチをシームレスに結合する機能がある. あくまでもノンブロッキング性を追求しているため, スイッチの結合には大量の光ファイバを必要とする.

参考文献

[1] 清水 奨, 釘本 健司, 天海 良治, 村上 健一郎, スケーラブルな知的制御スイッチクラスタ Dynatera, マルチメディア, 分散, 協調とモバイル (DICOMO 2002) シンポジウム, pp. 463–466, 2002.

[2] J.Case, M.Fedor, M.Schoffstall, J.Davin, A Simple Network Management Protocol (SNMP), STD0015, RFC1157, May 1990.

[3] K.McCloghrie, F.Kastenholz, The Interfaces Group MIB, RFC2863, June 2000.

[4] MRTG: Multi Router Traffic Grapher http://mrtg.hdl.com/mrtg.html

[5] R.W.Wolff, Poisson Arrivals See Time Averages, Operations Research, Vol.30, No.2, pp. 223–231, 1982.

[6] J.Cao, W.S.Cleveland, D.X.Sun, S-Net: A Software System for Analyzing Packet Header Databases, Proc. Passive and Active Measurement, pp. 34–44, 2002.

[7] N.G.Duffield, M.Grossglauser, Trajectory Sampling for Direct Traffic Observation, IEEE/ACM Transactions on Networking, Vol.9, No.3, pp. 280–292, 2001.

[8] M.Coates, A.Hero, R.Nowak, B.Yu: Internet Tomography, IEEE Signal Processing Magazine, May 2002.

[9] 3Com, Introduction to XRN Technology, http://www.3com.com/other/pdfs/legacy/en_US/xrn_intro_whitepaper.pdf

[10] Juniper Networks, The Essential Core, http://www.juniper.net/solutions/sol_prof/351006.pdf