マルチメディア通信と分散処理 92－2
（1999．2．25）

メタモニタ：適応型ネットワークトラヒック観測機構

八木 哲　小倉 毅　川野哲生　丸山 充　高橋直久
NTT 未来ねっと研究所

本稿では，Gb/s クラスが実用水準に達してきた高速ネットワークの多様な運用管理のために，メタモニタと呼ぶ，比較的低い処理能力とシンプルな構成で，Gb/s クラスの高トラヒックを柔軟に観測する機構について述べる．メタモニタの特徴は次の通り．1) 複数の異なる性質のモニタで協調的に観測する．2) 観測負荷に応じて全数調査と標本調査を選択する．3) 観測要求や現在の観測結果などの状況に応じて各モニタの観測焦点を変更する．本稿では，メタモニタをモデル化するとともに，実現方法を考察し，我々が開発を進めている並列分散型高速通信スイッチ“COREswitch”上に構築したプロトタイプを用いて，実験，評価する．

META MONITOR:An adaptive network-traffic monitor

Satoru Yagi　Tsuyoshi Ogura　Tetsuo Kawano　Mitsuru Maruyama　Naohisa Takahashi
NTT Network Innovation Laboratories

A simple economical method called META MONITOR has been developed to monitor Gb/s network-traffic. META MONITOR can cooperate with any kind of monitor, select a sampling survey or a census survey according to a load of monitoring and change the focus of cooperating monitors dynamically according to a result of monitoring. This paper presents the META MONITOR model, describes its installation and the estimation the performance of a installed on a "COREswitch" a parallel distributed high-speed communication switch.

1　はじめに

ネットワークの詳細なトラヒック観測は，障害検出 (短期)，経路の制御や設計 (中期)，設備設計 (長期) など，多様なネットワーク管理のために有意である．これらの用途には，複数のノードの複数のポート (セグメント) を関連付けた観測を，リアルタイムで行う，柔軟な観測機構が望ましい．既存のネットワークでは，SNMP を基に，ネットワークノードをリアルタイムで観測する方法 (例[1]) が利用できる．遠隔ネットワークが対象の場合は，観測行為に起因するネットワークへの負荷を押えるため，観測結果を予め統計処理する方法も利用できる[2)3)]．また，より詳細な観測のために，パケットヘッダを蓄積し，バッチ的に観測するアプローチ[4)5)]がある．

一方，ネットワークの高速化は，Gb/s クラスが実用水準に達し，レイヤ 3 スイッチを例として，パケット転送処理のハードウェア化を促している．このような，転送処理のハードウェア化を必要とする，“高”トラヒックを対象にした柔軟な観測機構の実現には，処理能力不足や，高い処理能力を得るためのシステムの複雑化の問題がある．OC3MON[6)7)]は，汎用の PC と NIC を基盤とし，光スプリッタを介して OC-3 速度の回線を観測する，シンプルで低コストなシステムを実現している．しかし，Gb/s クラスを目標とする OC12MON や OC48MON では，バス速度等の制約から複数のポートの監視が困難である．また NMCVsystem[8)]では，独自の Gb/S クラスの回線インタフェース・チップ[9)]を基に，チッ

プの観測用ポートにマルチプロセッサ構成の処理装置を組み合わせたプローブを，ネットワーク上に複数配置することを想定している．しかし，多くの観測点が必要な場合には，処理能力の高い複雑なプローブを多数使用するため，システム規模が大きくなる．

これに対し，1) 負荷分散：複数の異なる性質のモニタで協調的に観測する，2) 負荷軽減：統計的な調査[10]の考えに従い観測負荷に応じて全数調査と標本調査を選択する，3) 処理能力の最適配置：観測要求や現在の観測結果などの状況に応じて各モニタの観測焦点を変更することにより，観測データ量の削減と観測能力の効率的利用を可能にし，比較的低い処理能力とシンプルな構成で，“高”トラヒックを柔軟に観測する手法を提案する．このような観測機構をメタモニタと呼ぶことにする．メタモニタは，トラヒックを観測するエージェントであり，観測焦点を指示するメタモニタ・マネージャに制御される．既に，我々が開発を進めている並列分散型高速通信スイッチ COREswitch[11] 上で簡単な実験を行ない，メタモニタ方式の有効性を確認している[12]．

本稿では，エージェントであるメタモニタについて述べる．先ず，前記の三つの方針を基に，メタモニタをモデル化する (2 章)．次に，メタモニタの実現法を検討する．実現上の要件は，a) 統計的な調査を可能にする標本抽出機構と，b) 観測焦点の変更を可能にする，ネットワークノードとの I/F である．a) について，統計的な考えをそのまま具現化する“直接的な実現方法”対して，実現の容易さを重視した，タイムスロット型標本抽出法と呼ぶ“簡易的な実現方法”を提案する (3.1 章)．b) について，I/F の要件をまとめる (3.2 章)．また，COREswitch 上に実現した実験用プロトタイプを示し (4 章)，プロトタイプを用いた実験と，実験結果の評価を行なう (5 章)．最後に，本稿の内容をまとめ，今後の課題を示す (6 章)．

2 メタモニタ

比較的低い処理能力とシンプルな構成で，“高”トラヒックを柔軟に観測するための方針を示す．

1) 負荷分散： 観測対象に選んだトラヒックの特徴を複数のパラメータに分解し，性質の異なる複数の**要素モニタ**と呼ぶモニタで各パラメータを観測する．**要素モニタ・コントローラ**と呼ぶ要素モニタの制御部で，各観測結果からトラヒックの特徴を求める．

2) 負荷軽減： 要素モニタは，統計的な調査の考えに基づき，観測処理が軽いパラメータの観測には全数調査を適用し，観測処理が重いパラメータの観測には標本調査を適用する．

3) 処理能力の最適配置： a) 注目して観測したいポイント，b) 現在の観測結果，c) 観測に必要な負荷，d) 要素モニタの処理能力に基づいて，**メタモニタ・マネージャ**と呼ぶ**メタモニタ**の制御部が出した指示に従い，各**要素モニタ**は観測焦点を適応的に変化させる．例えば，複数ポートの概観から特定ポートの詳細へ観測焦点を絞り，精度を上げて調査する．

このように，メタモニタ・マネージャから指示された，観測対象トラヒック (母集団) と，アドレスやプロトコルなどのパラメータ (母集団変数) に対して，要素モニタは，全数調査と標本調査を選択的に用いて調査を行なう．要素モニタ・コントローラは，個々の要素モニタの調査結果から観測対象トラヒック (母集団) の様子を推定し，更に複数の調査結果を組み合わせ，より複雑な条件での，観測対象トラヒック (母集団) の様子を求める．これを単位時間ごとに行ない，時間変化を観測する．要素モニタと要素モニタ・コントローラが持つ機能を定義し，メタモニタをモデル化する．

[定義 1] 観測機能 (要素モニタ)
ポートの集合 P からの入力パケットを確率 $s_p(0 < s_p \leq 1,\ 0 < s_p < 1$ の時は標本調査，$s_p = 1$ の時は全数調査) で抽出し，抽出したパケットのうち，条件 c を満たすパケットの数を期間 s_t ごとに集計する．P を**観測対象ポート**，c を**ストリーム識別条件**，s_t を**抽出期間**，s_p を**抽出確率**，集計した値の列 R_e を**観測結果**と呼び，**観測機能** m_e を $R_e = m_e(P, c, s_t, s_p)$ で表す．□

観測機構は，指定ポートからの入力パケットを母集団として，指定確率でサンプリングし，標本を作成する．更に，指定条件に従って内訳を求め，指定時間毎に集計する．

[定義 2] 推定機能 (要素モニタ・コントローラ)
観測結果 R_e の要素であるパケット数に対し，**推定関数**と呼ぶ関数 f_c を適用して得られたパケット数の列 R_c を，**推定した観測結果**と呼び，**推定機能** m_c を，$R_c = m_c(R_e, f_c)$ で表す．□

推定機構は，観測機構が行なった全数／標本調査の結果から，母集団を推定する．例えば，観測機構で母集団の 1/n の大きさの標本を作成し，調査したのであれば，集計結果を n 倍する．

[定義 3] 導出機能 (要素モニタ・コシトローラ)
n 個の**推定した観測結果** $R_ci(i = 0,1,2,,,n)$ を要素とする集合 R_p に対し，R_ci の要素であるパケット数に**導出関数**と呼ぶ関数 f_m を適用して得たパケット数の列 R_m を，**導出した観測結果** R_m と呼び，**導出機能** m_m を $R_m = m_m(R_p, f_m)$ で表す．□

導出機能は，複数の推定機構の調査結果を組み合せ，より複雑な条件下での母集団の様子を求める．

[定義 4] *メタモニタ*
n 個の要素モニタを持つメタモニタは，個々の**観測機能**を $m_{ei}(i = 0,1,2,,,n)$，**推定機能**を $m_{ci}(i = 0,1,2,,,n)$，**導出機能**を m_m とすれば，以下のように表す．

- R_{ei}=$m_{ei}(P_i, c_i, s_{ti}, s_{pi})$ $(i = 0,1,2,,,n)$
- R_{ci}=$m_{ci}(R_{ei}, f_{ci})$ $(i = 0,1,2,,,n)$
- R_m=$m_m(R_p, f_m)$, $R_p \ni R_{ci}$ $(i = 0,1,2,,,n)$

□

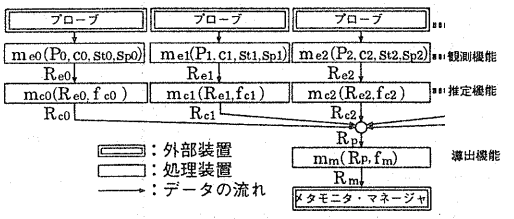

図 1: メタモニタのモデル

3 実現法

3.1 標本抽出

標本抽出の“直接的な実現方法”は，例えば，内部のデータ転送量の減少を考慮し，プローブ部で乱数を用いて，確率的にサンプリングすればよい (観測機構の標本抽出部をプローブに移設した形態)．しかし，1) プローブに高速の乱数生成機能が必要，2) 複数のポートを観測する場合，各プローブと観測機構間の経路を共有すれば，統計多重されたパケットはブロッキングされる可能性があり，経路を独立させれば，複雑な機構になるなど，メタモニタの基本方針に反した複雑化の問題がある．これに対し，実現性の高い“簡易的な実現方法”として，抽出期間を抽出確率に応じた複数のタイムスロットに区切り，一つのタイムスロットを一つのポートに割り当て，サンプリングする方法が考えられる．しかしこの方法では，割り当てられたタイムスロットの間に，トラヒック変動の山や谷があった場合には，これを観測できず，平滑化された結果が出る (図 2)．

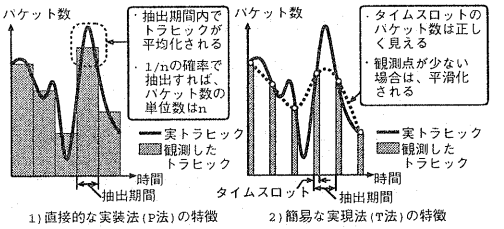

図 2: 標本抽出の実現方法の比較

“直接的な実現方法”を確率型標本抽出法 (P 法)，“簡易的な実現方法”をタイムスロット型標本抽出法 (T 法) と呼ぶことにし，以降では，平滑化の問題がある T 法を，実用的に使用するための方法を考察する．

- 観測機能の水準：抽出期間内の全てのタイムスロットで，観測負荷の軽い，観測対象ポートからの全入力パケット数を観測し，タイムスロットを単位時間として，累積入力パケット数と相関係数を求めれば，入力パケット数のバースト的な変化 (平滑化の発生) が推測できる．この時，タイマ割り込みの増加などの性能的問題が許せば，抽出期間

を小さくして対応できる．複数のストリームの入力パケット数が，その増減を相殺し合い，累積入力パケット数が一様に増加する場合は，この方法では推定できない．ただし，あるストリームのあるタイムスロットにおける誤差 e は，トラヒックが一様に増加した場合のパケット数を s_t，観測した観測対象ポートからの全入力パケット数を a_t とすれば，以下の値で押えられる．

$$0 \leqq e \leqq \max(s_t\ ,\ a_t\text{-}s_t)$$

- 推定機能の水準：割り当てられたタイムスロット以外の抽出期間のパケット数の情報を補う方法として，タイムスロットを単位時間とし，x 軸を時間，y 軸を単位時間当たりのパケット数とした観測点に対して，補完曲線を求め，抽出期間で積分する方法が考えられる．ただし，補完曲線を求めるために多くの観測点を用いると，リアルタイム性と処理量の点で問題が生じる可能性がある．

- 導出機能の水準：観測対象ポートの総入力パケット数を数えるモニタを用意し，総入力パケット数の内訳として，推定機能が求めた各ストリームの入力パケット数の値を適用すれば，各ストリームの入力パケット数の合計が保証できる．

3.2 ネットワークノードとのI/F

ネットワークノードとのI/Fの要件を以下にまとめる．観測焦点を自由に変更するためのプローブの要件 (1,2) と，正しく観測するためのプローブと観測機構間の転送経路の要件 (3,4) は，以下の通りである．また，これらの機構はシンプルであることも重要である．

1. 任意のポートを等しく対象にできる．
2. プローブエフィクトが小さい．
3. パケットが破棄されない．
4. 通常の転送経路と同程度の遅延．

4 実験用プロトタイプ

プラットフォームとなるCOREswitch[11]を図3に示す．IFP と CIF は，XSW と C-bus に対してシンメトリであり，柔軟な構成が取れる．

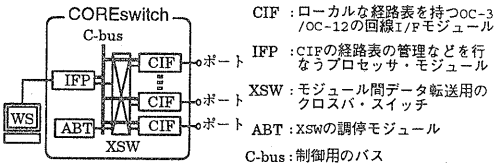

図 3: COREswitch

本プロトタイプのメタモニタは，IFP 上の3つの要素モニタと，WS 上の要素モニタ・コントローラからなり，指定したポートからの入力パケットの中で，指定したIP アドレスを宛先に持つパケット数を求める．T 法の動作を基本とするが，P 法の動作も限定的に可能である．メタモニタの各機能について示す．

- プローブ：XSW と ABT による One-shot 型マルチキャスト機能[13]を用い，IFP にパケットを取り込む．設定は CIF の経路表のエントリにあるフラグの操作で行ない，任意のポートを対象にできる．動作はユニキャストと比較して百数十 nsec 程度遅いだけであり，プローブエフィクトは小さい．また，同時に複数のポートを対象としなければ，XSW-IFP 間でブロッキングは生じない．

- 観測機能：T 法の動作の場合，IFP は定期的に CIF の経路表を更新し，割り当てられたタイムスロットの時にパケットを取り込む．この時，経路表の更新時間が問題になるが，本プロトタイプでは，使用中のエントリのみを更新することで回避している．実用システムでは，各エントリ共通の取り込み先フラグを用意し，経路検索の FPGA に，このフラグと各エントリのフラグとの OR を取る，単純な機構を追加することで解決できる．P 法の動作の場合，一つのポートのパケットを全て IFP に取り込み，乱数[14]を用いて指定された確率でパケットを選択する．各要素モニタの観測機能を以下に示す．

パケット数モニタ (要素モニタ 0)：
CIF の入力パケットのカウンタを IFP で定期的にポーリングし，抽出期間 s_t ごとの入力パケット数を計測する．全数調査のモニタ．

– $R_{e0} = m_{e0}(P, \text{全パケット}, s_t, 1)$

パケット種別モニタ (要素モニタ 1,2)：
IFP に取り込んだ標本となるパケットを，宛先 IP アドレスで分類し，抽出期間 s_t ごとのパケット数として計数する．標本調査のモニタ．

- $R_{e1} = m_{e1}(P,$ 宛先 IP =指定 IP, $s_t, s_p)$
- $R_{e2} = m_{e2}(P,$ 宛先 IP ≠指定 IP, $s_t, s_p)$

- 推定機能：処理の軽減とリアルタイム性に配慮したうえで，精度のために，少ない観測点から補完曲線を求める．ある抽出期間の推定のために，前後の二つずつの 4 点の観測結果から 3 次曲線を求め，台形積分を適用する．
  - $R_{c0} = R_{e0}$
  - $R_{c1} = m_{c1}(R_{e1},$ 3 次曲線と台形積分)
  - $R_{c2} = m_{c2}(R_{e2},$ 3 次曲線と台形積分)
- 導出機能：精度のために，**パケット種別モニタ**が観測した，標本抽出した入力パケットの内訳が，**パケット数モニタ**が観測した，全入力パケットの内訳を代表していると仮定し，元のトラヒックを導出する．
  - $R_m = m_m(\{R_{c0},R_{c1},R_{c2}\},$ 内訳を求める)
    $= [r_{c0}0\frac{r_{c1}0}{r_{c1}0+r_{c2}0}, r_{c0}1\frac{r_{c1}1}{r_{c1}1+r_{c2}1}, r_{c0}2\frac{r_{c1}2}{r_{c1}2+r_{c2}2}, ,,]$
    $r_{c0}i \in R_{c0}, r_{c1}i \in R_{c1}, r_{c2}i \in R_{c2}$ $(i$=0,1,2,,,)

5 実験と評価

COREswitch に 3 台の端末を接続し，端末 a から，端末 b と端末 c へ UDP でトラヒックを生成し，端末 a の継るポートを観測した．ポート数は最大 16 のため，抽出期間を 4 秒 (s_t=4 秒)，サンプリング時間を 1/4 秒 (s_p=1/16) とした．

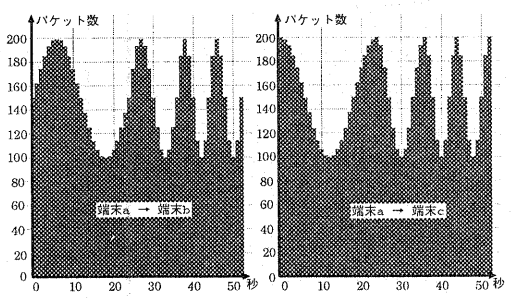

図 4: 生成したトラヒック

生成した二つのトラヒックは，周期を 24,12,8,8 秒と短くしながら，1 秒単位でパケット数を変化させた (図 4)．これを 4 秒で平均化して示す (図 5)．

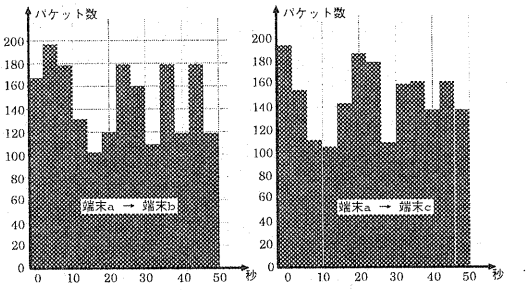

図 5: 平均化した結果

P 法による観測結果 (図 6) と図 5 を比べると，時間同期が取れていないため，少し時間軸をずらして平均した結果になっているが，近似性が分かる．

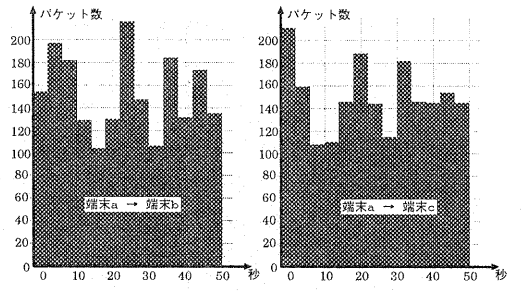

図 6: P 法による観測結果

T 法による観測結果 (図 7) と図 5 を比べると，変動周期が抽出期間の 3 倍から 2 倍程度まで短くなる，30 秒を経過した当たりから増減が鈍り，平滑化の発生が分かる．

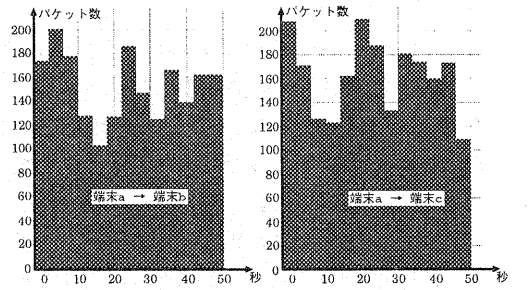

図 7: T 法による観測結果

各抽出期間について，タイムスロットを単位時間とし，単位時間ごとのポートからの累積入力パケット数とで相関係数を求めると(図8)，やはり30秒経過したあたりから値が小さくなり，平滑化を示唆している．このように平滑化の発生が分かる場合には，システムの性能が許す範囲で，抽出期間を短くして対応できる．

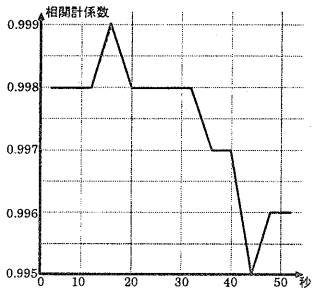

図 8: 抽出期間ごとの相関係数の変移

平滑化が分からない場合の簡単な実験を行なう．二つのストリームに対して，パケット数の合計を一定にし，割合(0-100%)と，その割合が継続する時間(0-8秒)を乱数[14]で変化させ，最大継続時間の50%,100%,200%の抽出期間(4,8,16秒)で160秒間観測する．この時，観測誤差が20%以内である抽出期間の，全抽出期間に占める割合を表1に示す．最大継続時間に対して抽出期間を小くすれば，多くの抽出期間で誤差を小さくできる．

表1. 平滑化の発生が分からない時の観測誤差

抽出期間の長さ(秒)	4	8	16
誤差20%以下の割合(%)	50	35	25

6 おわりに

本稿では，Gb/sクラスが実用水準に達してきた高速ネットワークの多様な運用管理のために，1)負荷分散：複数の異なる性質のモニタで協調的に観測する，2)負荷軽減：観測負荷に応じて全数調査と標本調査を選択する，3)処理能力の最適配置：観測要求や現在の観測結果などの状況に応じて各モニタの観測焦点を変更し，比較的低い処理能力とシンプルな構成で，柔軟に“高”トラヒックを観測する観測機構を示した．

今後，メタモニタのパラメータの記述方法，それらを制御するAPI，観測シナリオの記述方式など，適応的な観測のための制御機構を検討し，メタモニタ・マネージャを具体化する．これらの良否判定は，管理者やシステムが，ネットワークを観測する時の戦略に依存し，ドメイン知識によるところが大きい．

謝辞

日頃御指導いただくグループの皆様方に深謝します．

参考文献

1) MRTG(Multi Router Traffic Grapher), http://ee-staff.ethz.ch/~oetiker/webtools/mrtg/mrtg.html

2) S. Waldbusser, “Remote Network Monitoring Management Information Base”, RFC1757, Feb. 1995

3) S. Waldbusser, “Remote Network Monitoring Management Information Base Version2 Using SMIv2”, RFC2021, Jan. 1997

4) 串田高幸, “インターネットのTCPトラヒックの解析”, 情処学会研究会報告, 97-DSP-84-4, Sep., 1997

5) 小松原, 鈴木, 茂木, 三上, “インターネット・トラヒックの短期的特性の分析”, 情処学会研究会報告, 97-DSM-7-3, Oct., 1997

6) J. Apisdorf, K. claffy, K. Thompson, R. Wilder “OC3MON:Flexible, Affordable, High-Performance Statistics Collection”, Proc. INET97'

7) OC3MON,http://www.nlanr.net/NA/Oc3mon/

8) G. Parulkar,D. Schmidt, E. Kraemer, J. Turner, and A. Kantawala, “An Architecture for Monitoring, Visualization and Control of Gigabit Networks”, IEEE Network,Sep./Oct.,1997

9) Z.D.Dittia, J.R.Cox Jr, and G.M.Parulkar, “Design of the APCI:A High Performance ATM Host-Network Interface Chip”, Proc. IEEE INFOCOM'95, Apr.,1995

10) 山崎, 有馬, 片山, 他, “確率・統計入門”,1998, 実教出版

11) 高橋, 村上, 丸山, 八木, 小倉, 川野, “並列分散型高速通信スイッチCOREswitch”, 情処学会第56回全国大会論文集(3), Mar., 1998

12) 八木, 高橋, 丸山, 小倉, 川野, “高速ネットワーク向け統計的トラヒック観測機構の提案”, 情処学会第57回全国大会論文集(3), Oct., 1998

13) 小倉, 高橋, 丸山, 八木, 川野, “COREswitchにおけるマルチキャスト方式”, 情処学会第56回全国大会論文集(3), Mar., 1998

14) MersenneTwister,http://www.math.keio.ac.jp/matumoto/mt.html