

P2P モデルを用いたエンドノード間トラフィック測定*

豊野 剛† 仲山 昌宏† 杉浦 一徳†

慶應義塾大学 政策・メディア研究科† 慶應義塾大学 環境情報学部‡

Abstract

本研究では、P2P モデルを用いて、エンドノードのユーザからの分散したネットワーク測定環境を構築する。ネットワーク測定の手法は数多く存在するが、その多くは管理者が障害検知に用いるものや、中継ノードのトラフィックモニタリングを行う目的で設計されている。本研究では新たに P2P モデルを用いることにより、ドメインに捕らわれない広域のネットワーク測定手法を実現する。また、そのデータを一元的に管理することにより、既存の測定手法では行われていないネットワーク性能の相互把握やネットワークの動的な変化を捕らえることを可能にした。

1 はじめに

ネットワークの状態や動向を把握するために、さまざまなトラフィック測定手法が開発されてきた。特にネットワーク管理の観点でトラフィック測定は重要であり、現存するネットワークの測定手法の多くは、管理者が自ドメイン内の障害検知に用いるものや、中継ノードのトラフィックモニタリングを行う目的で設計されている。

しかしエンドノードからの一般的な利用形態を考えると、ドメイン内部の状態を把握するだけでは不十分であり、実際の利用形態により近似したトラフィック測定が必要となる。すなわち、

- ・ 上りと下りのデータフローの測定
- ・ ドメイン間にまたがったネットワークの性能測定
- ・ 時間の遷移によるネットワーク状態の変化の測定

などが上げられる。これらを測定する事は、いわゆるインターネット全体のトラフィックの状態や動向を把握することにも繋がる。

既存の測定手法では、CoralReef [1] などに代表される AS 単位のトラフィックプロファイラがドメイン間にまたがったネットワーク測定を行えるが、これらは管理者権限が必要であり、また運用コストが

大きく、エンドノードからの測定には適していない。エンドノードが容易に行える ping, traceroute などは、上りの経路のみしか測定することができない。

2 本研究の目的

本研究では、1 節で述べた、既存の測定手法に不足している以下の項目を満たすことで、ネットワーク間の性能の相互把握や、ネットワークの動的な変化を捉えることを目的とする。

1. 下りの測定情報の把握

ネットワークの的確な性能を把握するためには、上り(行き)の測定情報だけではなく下り(帰り)のトラフィックの測定も不可欠である。ユーザの利用形態を考えると、現在はまだアップストリームがリクエストで、ダウンストリームがデータであることが多く、このような通信だとダウンストリームの方がデータ量は多い。下りのトラフィックがネットワーク性能の要因となる。

これを実現するためには、他ネットワークを起点とした測定も可能でなければならない。

2. 時間の遷移によるネットワークの変化

エンドノードからの測定情報は、多くのものが一過性のもので記録して参照する事ができない。ネットワークの動的な変化を把握するために、時間遷移とともにネットワークの測定情報を記録する。

* "P2P Network Traffic Measurement by the End Node of view"
Tsuyoshi TOYONO† Masahiro NAKAYAMA† Kazunori SUGIURA†
E-mail:jr@sf.wide.ad.jp
Graduate School of Media and Governance, Keio University†
Faculty of Environmental Information, Keio University‡
Keio University Shonan Fujisawa Campus. 5322, Endo, Fujisawa, Kanagawa 252, Japan

3. 測定データの蓄積と共有

エンドノードが測定した、多地点からの測定されたデータを一元的に管理することにより、ドメイン内部に限らないネットワーク全体動向の把握や、相互の性能比較を可能にする。

本研究では、P2P モデルを用い、エンドノードのユーザからの分散したネットワーク測定環境を構築することでこれらの項目を実現する。これにより既存のシステムよりも低コストながら、広域なネットワーク測定を目指す。

3 設計

システムの概要を図1に示す。

本システムは ICQ や napster に代表される、いわゆる Hybrid P2P モデルと呼ばれる分散した環境で構築される。サーバは測定ノードのリストだけを保持しており、各ユーザは、協力してくれる測定ノードをサーバで検索した後、協力ノードとはサーバを介さず直接 peer to peer で測定を行えばよい。

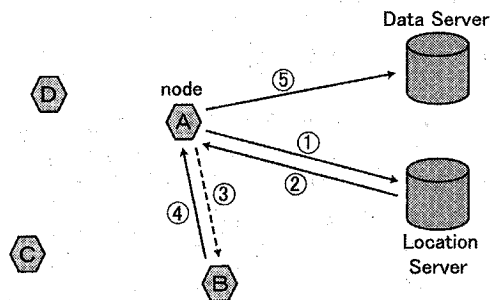


図1: システム概要図

以下にユーザ（ノードA）が測定を行うまでのシーケンスを示す。

1. 接続通知、および P2P ノードの検索クエリを送信
2. ロケーションサーバはノードをリストに登録し、また検索結果を返信
3. 協力ノード B に対して測定、および測定リクエストを送信
4. 協力ノード B は測定リクエストに基づき測定を行い、測定データを元ノード A に返信

5. 元ノード A は測定した全てのデータをデータサーバに送信

このように、測定対象側のノード B と協力することで、下りの経路の把握や、時間同期など様々なメリットが生まれる。

また、各エンドノードが行った測定のデータはデータベースに送信され、他のエンドノードが行った測定データとともに一元的に管理される。ネットワーク的に分散した地点の測定データを、統一したフォーマットで蓄積することで、後にネットワークの時間遷移やネットワーク全体の動向を解析することが可能となる。

例えばユーザ A が行ったトポロジ測定結果、同様に B, C が行ったトポロジ測定結果があるとする、これらを一つのデータベースにマージすることにより、ネットワーク全体のトポロジの様子が把握できる。（図2）

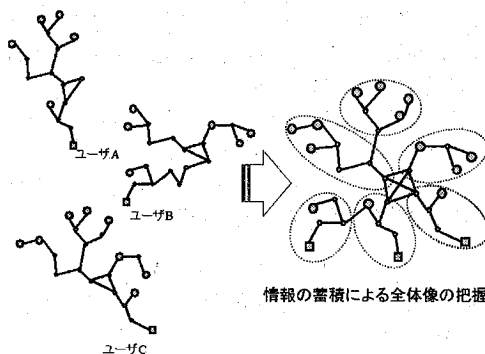


図2: 各ノードのデータを一元管理すると効果的な例：トポロジ

本システムでは、同様に自分が所属しない他ドメインネットワークからの測定を透過的に実現した。図3は他ドメインネットワークからの測定の概要を示したものである。

以下にユーザ（ノードA）がノードB-C間の測定を行うまでのシーケンスを示す。

1. 接続通知、および2つのP2Pノードの検索クエリを送信
2. ロケーションサーバはノードをリストに登録し、また検索結果を返信
3. 協力ノードBに対してB-C間の測定リクエストを送信

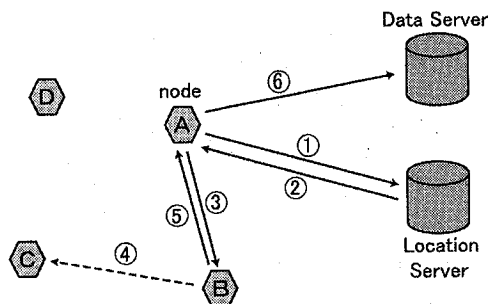


図 3: 他ネットワークからの測定

4. 協力ノード B はリクエストに基づき、C に対して測定、および測定リクエストを送信。C は測定データを B に返信
5. ノード B は、B-C 間の測定結果を元ノード A に送信
6. 元ノード A は測定した全てのデータをデータサーバに送信

ユーザ (ノード A) はロケーションサーバに任意の 2 つのノードを問い合わせる。合致する対象ノードが 2 つあれば、その 2 地点間の測定を行う。この場合、測定的一方が自ノードである必要はない。

このように Hybrid P2P 型の環境を用いることで、任意の 2 地点の測定を実現できる。

3.1 測定ノード

測定ノードは、ネットワークに接続されるとまず規定のロケーションサーバに接続通知メッセージを送信する。また、測定を行う際には、ロケーションサーバで測定ノードを検索する。実際の測定リクエストには以下のフォーマットを用いる。(図 4)

3.2 サーバ

3.2.1 ロケーションサーバ

ロケーションサーバは現在アクティブな測定ノードのリストだけを保持する。ロケーションサーバの役割は 2 つである。ひとつは測定ノードの接続通知を受け、そのノードをリストに加える。もうひとつ

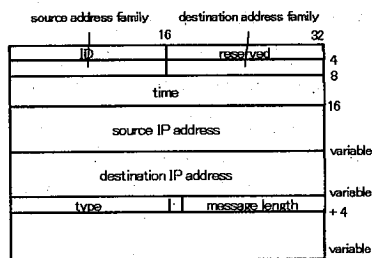


図 4: 測定リクエスト・データフォーマット

はノードの検索を受けて、条件を満たす協力ノードのリストを返答することである。

ロケーションサーバと測定ノードの応答は図 5 のデータフォーマットを介して行われる。また、このフォーマットで用いられるディスクリプタは表 1 のようになっている。

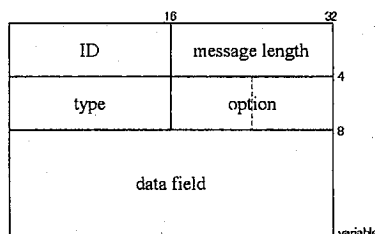


図 5: ロケーションサーバ・データフォーマット

表 1: ロケーションサーバのディスクリプタ

type	option (0-8)	option (9-16)	data field
Connect			
Close			
Search	0 IP addr/block	0 any protocol	search request
	1 AS number	1 IP v4 only	
	2 Domainname	2 IP v6 only	
Result	hit count		IP addr list
DS Search			Data Server IP addr

3.2.2 データサーバ

データサーバは各測定ノードが収集したデータを一元的に管理し、蓄積する。各測定ノードはロケーションサーバに type "DS Search" ディスクリプタを送信することによりデータサーバのアドレスを知

ることができる。データサーバと測定ノードの応答は図6のデータフォーマットを介して行われる。

また、ネットワークの時間遷移やネットワーク全体の動向を捉えるためのデータベースとして検索に基づき測定データを切り出す。このデータベース検索のためのインターフェースは実装に依存する。

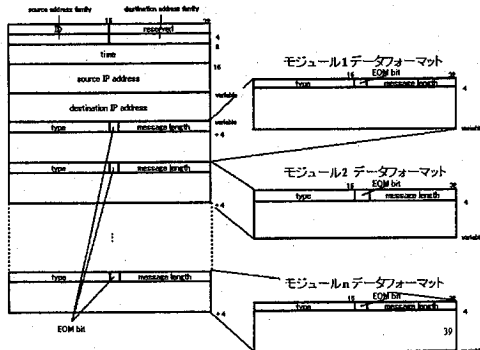


図 6: データサーバ・データフォーマット

4 実装

本システムは、以下の環境で実装を行った。(表2)

表 2: 実装環境

	クライアントノード	サーバ側
OS	FreeBSD4.4-REL	FreeBSD4.3-REL
Compiler	gcc version 2.95.3 20010315 (release)	同定
Database		PostgreSQL 7.2b2

4.1 測定ノードの実装

図7では測定ノードのシステムを示した。

測定ノードシステムは大きく分けてロケーションサーバと通信をする Node list manager, 測定を制御する Measurement controller, 最終的に測定したデータをユーザにアウトプットし、またデータサーバに送信する Data Sender に別れる。

Measurement controller により行われる測定項目は、各機能を独立したモジュールとして切り分けることにより、新しい測定手法にも柔軟に対応できるようにした。現在は表3のような測定モジュールを実装している。新しい手法で測定を開始する時も、モジュールを加えるだけで容易に測定項目を追加できる。

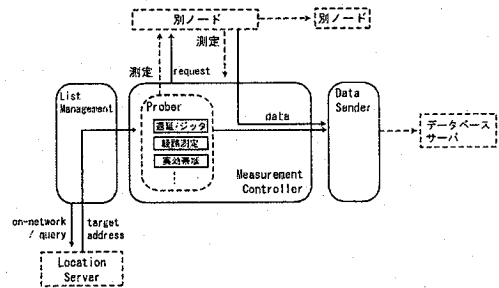


図 7: 各ノードシステム概念図

表 3: 測定モジュール

モジュール	パラメータ	オプション設定
遅延測定	上りの遅延, ジッタ, パケット喪失率	測定バケット数, バケットサイズ
途中経路	途中経路, RTT, パケット喪失率	測定バケット数, バケットサイズ
パケットペア	実効帯域	プロトコル, 測定バケット数, バケットサイズ
TCP スループット	TCP スループット	測定バケット数, バケットサイズ

各測定ノード間は SNTP(Simple Network Time Protocol) で同期をとり、誤差はミリ秒以内に押えられていると仮定している。SNTP が動作していないノードを用いた測定データはデータサーバに蓄積されない。

4.2 サーバの実装

4.2.1 ロケーションサーバの実装

ロケーションサーバは図5のフォーマットにより TCP で各測定ノードと通信し、図8のようなノードリストのデータを保持する。

接続通知があると RADB と DNS により AS number と Hostname を追加し、アドレスをノードリストに加える。登録されているノード数が一定値を超過しても現実装では生存確認応答は行わず、時間毎にタイムスタンプの古いノードを順に破棄していく。これによりノードリストは常に最新を保つ。

協力ノードを検索する際には、type "Search" ディスクリプタが、AS 番号, IP address, Domain-name のいずれかの条件と共に送信されてくるので、longest match を行いより条件に近いノードをリストにして返信する。

ID	Time	Addr family	IP address	AS number	hostname
⋮	32	16	128	16	Variable(max 255 bytes)

図 8: ロケーションサーバ・データベース

4.2.2 データサーバの実装

データサーバは図 6 のフォーマットにより TCP で各測定ノードと通信する。このデータを図 9 のようにリレーショナルデータベースで保持する。メインデータベースが検索のキー ID を生成し、測定データは項目毎の各サブデータベースに ID と共に格納する。hash したキー ID を検索に用いることによって検索の高速化と追加される測定モジュールの拡張性に対応する。

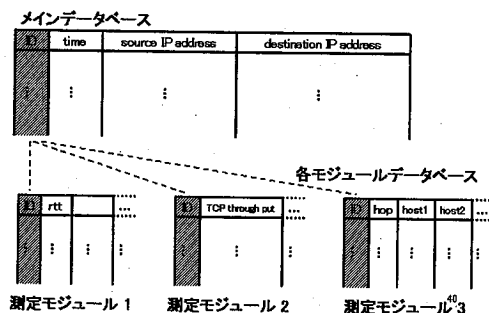


図 9: データサーバ・データベース

このデータサーバからデータを切り出すフロントエンドとして PHP を用いた。(図 10)

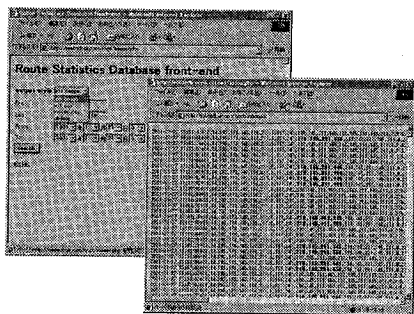


図 10: データサーバ・検索フロントエンド

5 評価

本システムを実際に運用し、2 節で述べた目的をもとに評価を行う。以下は本システムを 2001 年 10 月-12 月にかけて運用した収集した測定データを元にしたものである。

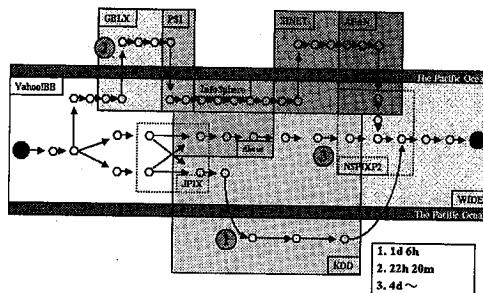


図 11: 同一 2 地点間の経路の振れ (A → B)

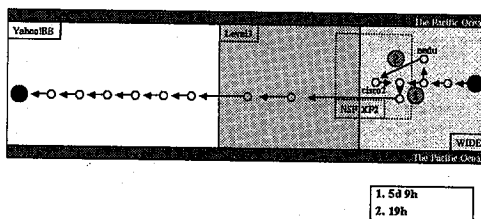


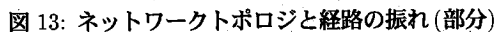
図 12: 同一 2 地点間の経路の振れ (A ← B)

図 11 と図 12 は共に同じ同一 2 地点間の上りと下りの経路情報をまとめ、AS 毎に区分したものである。時系列の経路の大きな変化の流れは、図中で 1 期、2 期、3 期とし、その期間を表に示した。

この図を見ると、同一 2 地点間でも、上りの経路と下りの経路が全く異なっていることが分かる。また、下りの経路は比較的安定しているのに比べ、上りの経路は測定期間中に大幅に変わっている。(図 11)

既存の測定手法では、上下のトラフィックを切り分けることができなかったため、このような状態を把握することが困難であった。

また、第 2 期 (図 11 中の経路 2) の 22 時間は経路的に非常に不安定で、2 度も日米国際線を往復している様子が伺える。時系列にデータを蓄積することで、このような状態の把握が容易となっている。



本システムの定性的評価を目的に沿ってまとまる。

- hybridP2P 型システムを構築することにより、
透過的に任意の他ドメイン間の測定を行えるよ
うになった。

また、多数地点からの測定情報の取得と、そのデータを動的に分析可能な基盤の構築を整えた。分散した各エンドノードの測定データを一元的に管理し、データベース化することによって、低いコスト

[7] Kevin Thompson, Gregory J. Miller, and Rick Wilder (MCI Telecommunications Corporation), "Wide-Area Internet Traffic Patterns and Characteristics", IEEE Network, November/December 1997, pp.10-23