インターネットにおける トランスポート層アーキテクチャに関する研究

長谷川 剛 (ごう)
大阪大学サイバーメディアセンター

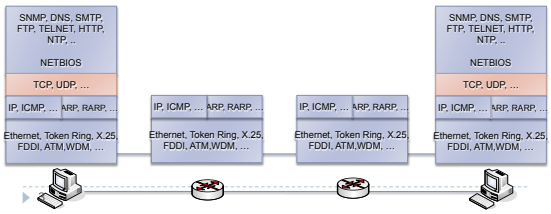
目次

- TCP輻輳制御機構の概要
- 問題点と解決策に関する研究例
- 近年の研究内容
  - トランスポート層プロトコルの改善による無線LANの消費電力削減
  - エンド間ネットワークパス上の複数区間の利用可能帯域同時計測

2

インターネットのプロトコルスタック

- 第4層以上はエンドホストのみに存在
  - アプリケーション処理、輻輳制御、…
  - ネットワーク内はIP (第3層) までの処理
- トランスポート層はネットワークをブラックボックスと扱う
  - エンド端末間制御



TCP (Transmission Control Protocol)

- トランスポート層プロトコル
  - ネットワーク技術の違いを吸収して、アプリケーション層にネットワークの切り口を見せる
  - ネットワーク層が相手にデータを確実に送り届けてくれるとは限らないという前提の上で、確実なデータ転送を実現する
    - ただし、遅延時間やスループットは保証しない
  - その結果、TCPを用いるアプリケーションはエンド端末間のネットワークの特性を気にする必要がなくなる
    - プロトコル階層化のメリット

4

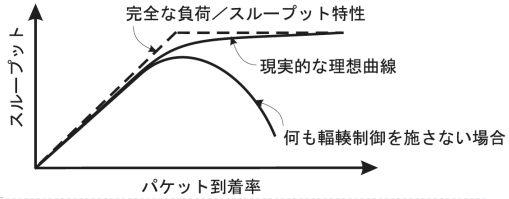
フロー制御と輻輳制御

- フロー制御
  - 送受信ホストの性能の違いなどを考慮した制御
- 輻輳制御
  - ネットワークの混雑(輻輳)を回避・解消するための制御
- 共に、送信ホストからのデータ転送速度を調整することで実現される

5

輻輳制御の重要性

- ネットワークへのパケット流入量が増えると、ネットワークスループットは劇的に低下する
- 輻輳制御は、輻輳発生時の性能低下を防止するために必要



6

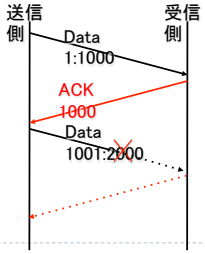
輻輳制御の3つのステップ

- 輻輳の監視
  - TCPはエンドホストでのみ動作するため、ネットワーク内の輻輳状況を直接監視できない
  - エンド端末間で得られる情報から推定
- 輻輳発生の通知
  - 監視同様、ネットワーク内からの明示的な輻輳通知は期待できない
- 輻輳制御の実行
  - 送信端末のデータ転送速度の調整

7

輻輳の監視

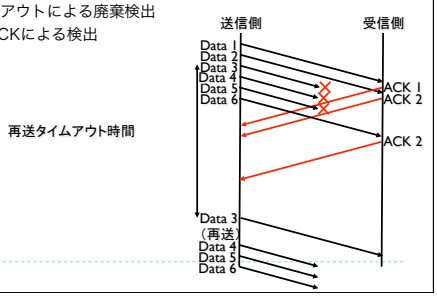
- 送出したパケットのACKが返ってくるか否かで判断
  - 返ってくる=ネットワークに余裕がある=輻輳していない
  - 返ってこない=パケット廃棄がネットワーク内で発生=輻輳発生



8

輻輳の通知

- ネットワークからの明示的な通知は行われない
  - 送信側端末が自身でパケット廃棄を監視する
    - タイムアウトによる廃棄検出
    - 重複ACKによる検出



9

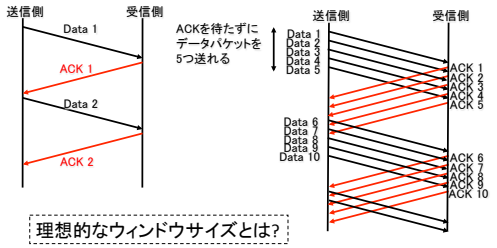
輻輳制御の実行

- 輻輳発生を検出したら、自身のデータ転送速度を低下させる
  - ウィンドウサイズを小さくする
- 輻輳が発生していなければ、ウィンドウサイズを徐々に大きくする
  - ウィンドウ型輻輳制御

10

ウィンドウ型フロー制御

ウィンドウサイズ=1　　ウィンドウサイズ=5



理想的なウィンドウサイズとは?

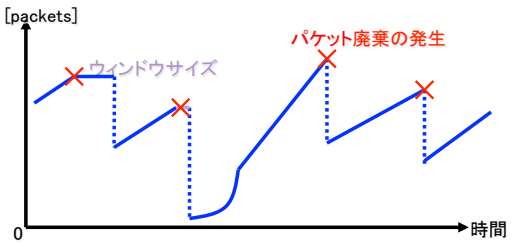
11

ウィンドウサイズと帯域遅延積

- ウィンドウサイズは、ネットワークの帯域遅延積と等しいことが理想的
  - 帯域と遅延の積=ネットワークに載せることができるデータ量
- 通信開始時点で帯域遅延積を知ることは不可能
  - 遅延時間、帯域の両方とも、通信を開始するまでわからない
  - 時間とともに大きく変動
- TCPのウィンドウ制御
  - パケット廃棄が発生しない限り、徐々に増加させる
  - パケット廃棄が発生すると、大きく減少させる

12

ウィンドウサイズと帯域遅延積（2）



13

TCP輻輳制御機構の特性

- TCP: 多様な環境において、安定的な通信を行うことができるロバスト性
- 個々のネットワーク環境における性能最適化は困難
  - アクセスネットワーク環境の進歩により、性能低下などの問題が顕在化しつつある
    - Ex) 高速・広帯域ネットワーク環境、高速無線ネットワーク環境、新たなアプリケーション特性、…

14 Go Hasegawa, Osaka Univ.

輻輳制御機構の重複

- TCP (第4層) では輻輳によるパケット廃棄の検出・再送を行っている
- 他の階層や、トンネリング環境における、複数の輻輳制御の相互作用の問題
  - TCPを用いたビデオストリーミングアプリケーション
    - TCPが原因でスループットが不必要に低下する
  - ATMのABRサービスクラス
    - 2つの階層における輻輳制御
  - VPNによるTCP over TCP問題
    - 外側と内側のフィードバックループの干渉
- 解決策
  - 慎重なパラメータチューニング
  - 一方の輻輳制御を停止させる (ex: TCP→UDPへ変更)

15

異なるバージョン間の公平性

- 異なるトランスポート層プロトコル間の不公平
  - UDP vs TCP
    - 輻輳に対して反応しないUDPが、TCPを負かす
  - 新しいTCP vs 旧来のTCP
    - インテリジェントな新しいTCPが、攻撃的な旧来のTCPに負ける
- ネットワーク内で第4層制御を行わないインターネットでは、本質的に避けることができない
  - エンド端末の「親切」に期待するしかない
  - 攻撃的なプロトコル同士の競争になり、共に性能低下
- 解決策
  - アプリケーションプロトコルによる制御 (TCP-friendly control over UDP)
  - 攻撃的な相手にも対応できるTCP (Compound TCP, CUBIC等)

16

新たなネットワーク環境への対応 (1)
無線ネットワーク

- TCP性能に影響を与える無線ネットワーク特性
  - 高いビットエラー率
    - パケット廃棄の発生 ≠ ネットワーク輻輳
  - 上りと下りの帯域共有
    - 混在環境における資源利用の不公平性
  - 通信路環境の時間的・空間的変動
    - リンク帯域、伝播遅延時間の大幅な変動
    - 非効率なデータ転送速度調整、delay-basedプロトコルの性能低下
- 解決策
  - ネットワークの分離(ネットワーク内で第4層処理)
  - TCPの高機能化
    - パケット廃棄原因の特定→輻輳制御のON/OFF切替

17 Go Hasegawa, Osaka Univ.

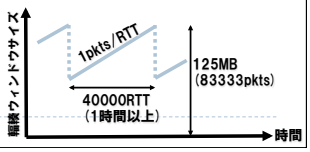
新たなネットワーク環境への対応 (2)
高速・高遅延ネットワーク

- 高遅延(Long)、高帯域(Fat)ネットワーク
  - インターネットの大規模化、ネットワーク技術の進歩
- 帯域遅延積が巨大化
  - Ex) 10Gbps, 100msec で125MB
    - 83333パケットに相当(パケットサイズ：1500B)
- この帯域遅延積を使い切るには、
  - 送受信端末に巨大なバッファが必要
  - 輻輳ウィンドウサイズの増減アルゴリズムがよくない
    - 1パケット／RTTの増加
    - パケット廃棄時に半分に減少

18

新たなネットワーク環境への対応 (3)
高速・高遅延ネットワーク

- 従来のTCP(Reno)で、10Gbps, 100msecのネットワークを使い切るには、
  - 2 x 10^{-10} 以下のパケット廃棄率が必要
    - 現在の光ファイバ技術でも不可能
    - リンクレイヤでのエラー回復制御で可能になるかもしれない
  - パケット廃棄が発生すると、スループットが回復するまでに40000 RTTs(1時間以上)必要
  - 現実的には10Gbpsは達成不可能
- 解決策
  - コネクションの並列化
  - 輻輳制御の高度化
    - Comound TCP
    - CUBIC
    - …



19

新たなネットワーク環境への対応 (4)
その他: 新たなアプリケーションによる影響

- 例: データセンタネットワーク
  - 今までにないトラヒック特性
    - MapReduceなど
  - 安価な機器を用いたネットワーク構築
    - コアネットワークのルータ等に比べて資源が乏しい
  - 起こらないと考えられていた、LAN内での輻輳が発生
    - 今までのWANでの輻輳を前提としたTCP輻輳制御機能がうまく機能しない
- 解決策
  - データセンタ専用の輻輳制御機構

20

トランスポート層プロトコルの改善による無線LANの消費電力削減

本研究の一部は、NICT委託研究
「新世代ネットワークを支えるネットワーク仮想化基盤技術の研究開発(課題ウ)」の支援による。

研究の背景

- 小型の無線端末を利用したインターネットアクセスが一般的になってきた
  - ノートPC，タブレット PC やスマートフォンなど
- 無線端末は通常バッテリ駆動である
- 無線端末の消費電力の 10% から 50% を無線通信が占めている

無線端末の駆動時間を長期化するには，無線通信の省電力化が重要

Atheros Communications, “Power consumption and energy efficiency comparisons of WLAN products,” Atheros White Papers, May 2003.

22

無線 LAN 環境における省電力化

無線 NIC の省電力化

無線 NIC の消費電力

消費電力は約 1/10 に削減

品名 (発表年)	送信	受信	アイドル	スリープ
Atheros AR5004 (2003年)	1.4 W	0.9 W	0.8 W	0.16 W
Atheros AR6002 (2007年)	0.8 W	0.5 W	0.05 W	0.002 W

消費電力は約 1/2 に削減

Wistron NeWeb Corp., “CM9: WLAN 802.11 a/b/g mini PCI … available at microcom.us/CM9.pdf.
Silex, “SX-SDCAG 802.11a/b/g SD … http://www. silexamerica.com/products/data sheets/sx-sdcag br…

MAC プロトコルレベルにおける省電力化

IEEE 802.11 Power Saving Mode (PSM) によって省電力化が可能

省電力化が可能である一方，ネットワーク性能を損ねる場合がある

- スループットの低下
- エンド間遅延の増加

23

効果的に省電力を行うためには?

- いつ，どれくらいスリープするかがスリープ効率を左右する
- パケットの送受信タイミングはアプリケーションやトランスポート層プロトコルによって決定される

上位層プロトコルの挙動を考慮する必要がある

スリープ状態からアクティブ状態に遷移する際の消費電力は無視できない

- 瞬間的に送信時の消費電力を越える場合がある
- より消費電力の低いスリープを行うほど遷移に時間がかかる

上位層プロトコルの挙動をスリープしやすいようにする必要がある

24

研究の目的

無線ネットワークにおける
省電力トランスポートアーキテクチャの確立

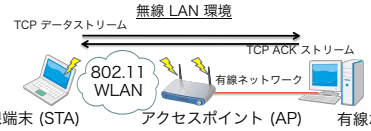
研究のアプローチ

1. 無線 LAN 環境における TCP の挙動を考慮した消費電力解析
   - 理論的な消費電力の下限を求める
   - 省電力を行う上で効果的なパラメータを見つける
   - バースト的なパケット送受信の効果を検証
2. 省電力トランスポート方式の提案
   - 複数のパケットをまとめて転送するバースト転送
3. 無線ネットワーク全般に対して適用できるエンドツーエンドな省電力トランスポート方式を考案

25

ネットワークモデルと仮定

無線 LAN 環境



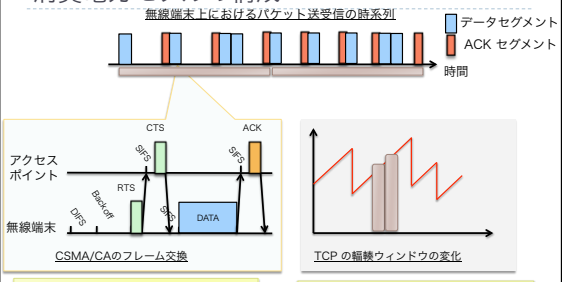
上り TCP データ転送をした場合に，無線端末 (STA) において消費される電力をモデル化

仮定

- バルクデータの転送を想定
- パケットの送受信のタイミングは TCP 輻輳制御にしたがい，既知である
- STA は RTS/CTS を利用し， AP は RTS/CTS を利用しない
- 無線区間ではフレームの衝突がなく，フレームが損失しない
- 有線ネットワーク上でデータセグメントが輻輳によって廃棄される．ACK セグメントは廃棄されない

消費電力モデルの構成

無線端末上におけるパケット送受信の時系列



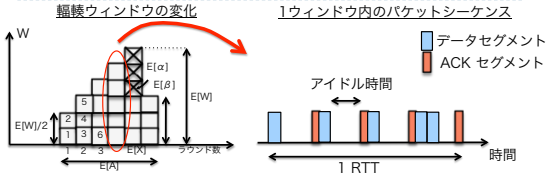
CSMA/CAのフレーム交換から
データフレームの送信/受信時の消費電力を計算

1 RTT 間に送受信されるパケット数は輻輳ウィンドウサイズによって決定

27

輻輳回避フェーズにおける消費電力

輻輳ウィンドウの変化　　1ウィンドウ内のパケットシーケンス



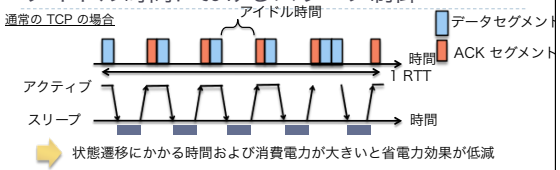
輻輳回避フェーズにおける消費電力

$$J^{TD} = \text{データセグメント送信時の消費電力} + \text{ACK セグメント受信時の消費電力} + \text{アイドル時の消費電力}$$
$$= E[Y]J^t + (E[Y] - E[W]/2)J^r + P^l\{E[A] - E[Y]T^t - (E[Y] - E[W]/2)T^r\}$$

28

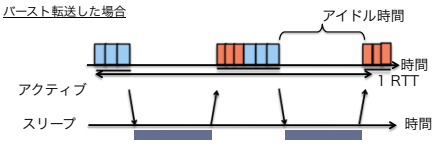
アイドル時間におけるスリープ制御

通常の TCP の場合



状態遷移にかかる時間および消費電力が大きいと省電力効果が低減

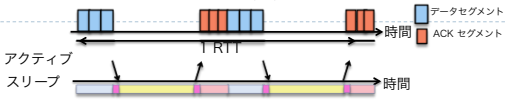
バースト転送した場合



アクティブ/スリープ状態間の状態遷移回数を減らすことにより，省電力効果が高まる

29

スリープした場合の消費電力



スリープ条件: アイドル時間 > アクティブ状態からスリープ状態への遷移時間 + スリープ状態からアクティブ状態への遷移時間

スリープする場合の輻輳回避フェーズにおける消費電力

$$J^{TD} = \text{データセグメント送信時の消費電力} + \text{ACK セグメント受信時の消費電力} + \text{スリープ時の消費電力} + \text{アクティブ/スリープ状態遷移時の消費電力} + \text{アイドル時の消費電力}$$
$$= E[Y]J^t + (E[Y] - E[W]/2)J^r + P^s E[T^s_{id}] + E[N^s_{id}](P^{as}T^{as} + P^{sa}T^{sa}) + P^l\{E[A] - E[Y]T^t - (E[Y] - E[W]/2)T^r - E[T^s_{id}] - E[N^s_{id}](T^{as} + T^{sa})\}$$

30

数値解析 - パラメータ設定

パラメータ設定

- データサイズ： 100 MB
- IEEE 802.11a
  - データレート: 54 Mbps
- データセグメントサイズ: 1500 バイト
- ACK セグメントサイズ： 40 バイト

無線 NIC の消費電力 [2]

送信	受信	アイドル	スリープ
1.4 W	0.9 W	0.8 W	0.16 W

スリープからの復帰にかかる時間は 1 ms，消費電力は送信時と同じとした

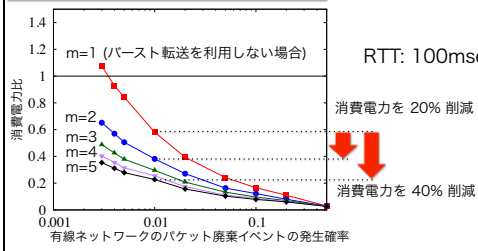
[2] Wistron NeWeb Corp., "CM9: WLAN 802.11 a/b/g mini-PCI Module." available at microcom.us/CM9.pdf.

31

解析結果 – 消費電力比

消費電力比 = スリープした場合の消費電力 / スリープしなかった場合の消費電力

m: バースト転送でまとめるパケット数



RTT: 100msec

消費電力を 20% 削減

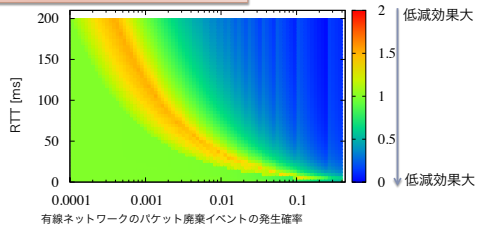
消費電力を 40% 削減

32

解析結果 – RTT とパケット廃棄イベントの発生確率を変化させた場合の消費電力比 (m=1)

消費電力比 = スリープした場合の消費電力 / スリープしなかった場合の消費電力

m: バースト転送でまとめるパケット数



RTT が大きく，パケット廃棄率が大きいほど省電力効果が高い

RTT が大きいとアイドル時間が多くなる ⇨ スリープ時間が長くなる

パケット廃棄率が大きいと平均的なウィンドウサイズが小さくなる

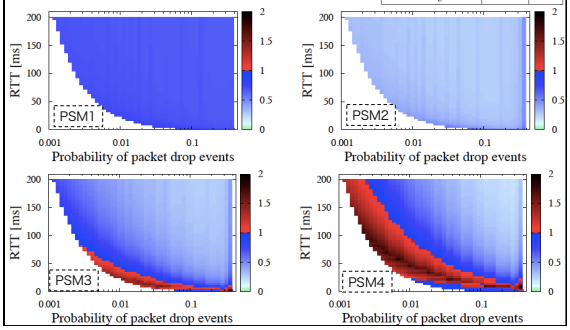
⇨ アクティブ/スリープ状態間の遷移回数が減る

33

省電力モードの違いの影響

表 1 PRISM1 のスリープモード

スリープモード	P^{st}	T^{wa}
Power Saving Mode 1	190 mW	1 μs
Power Saving Mode 2	70 mW	25 μs
Power Saving Mode 3	60 mW	2 ms
Power Saving Mode 4	30 mW	5 ms



まとめと今後の課題

まとめ

- 無線 LAN 環境におけるTCP データ転送を行った場合の消費電力解析
  - 省電力化のためにパケットをまとめてバースト転送
- 構築した消費電力モデルを用いた数値解析
  - バースト転送を利用することで，単にスリープする場合より20 % から 40 % 消費電力を削減

今後の課題

- 消費電力モデルの改良
  - 無線区間におけるフレーム衝突の考慮
  - 複数コネクションの考慮
- 解析結果を用いた消費電力特性の解明
  - 消費電力低減の条件、複数の省電力モードの選択指針
- 具体的な省電力トランスポート方式の考案
  - パケット送受信タイミングの推定

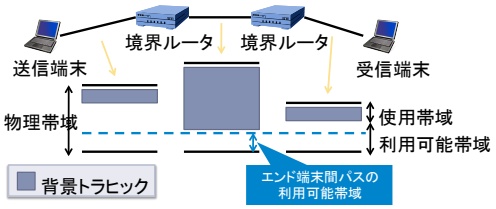
35

エンド間ネットワークパス上の
複数区間の利用可能帯域同時計測

本研究の一部は、NICT委託研究
「新世代ネットワークを支えるネットワーク仮想化基盤技術の研究開発(課題ウ)」の支援による。

研究背景

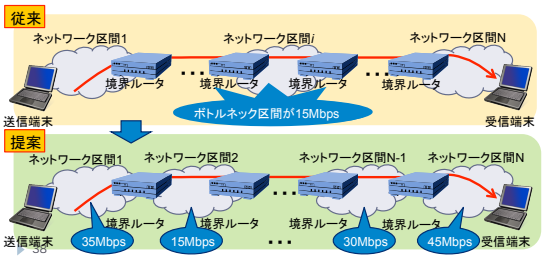
- エンド端末間パスの利用可能帯域計測
  - エンド端末間パスの利用可能帯域は，ボトルネックリンクの利用可能帯域により決定される



37

研究目的

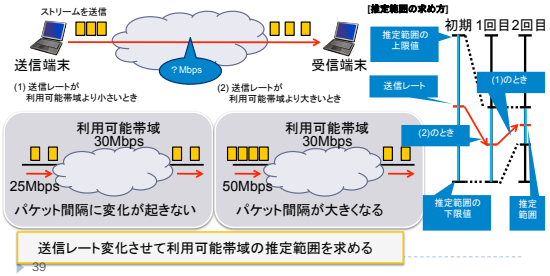
- エンド端末間パス上に存在する複数のネットワーク区間に対して，利用可能帯域を同時に計測する手法を提案



38

従来手法

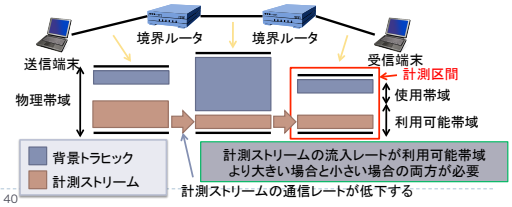
- Pathloadの計測原理



39

複数区間計測の問題点

- 計測区間における計測ストリームの流入レートを制御することができない
  - 計測パケットが計測区間に到着するまでに，背景トラヒックの影響を受ける
  - さらに，計測できない場合が生じる



40

複数区間の同時計測の可能性

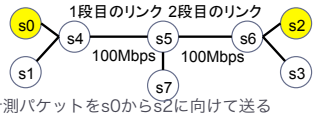
- 送信端末から高いレートで計測パケットを送ることにより，計測区間への流入レートを大きくすることができ，計測できる場合がある

シミュレーションにより計測可能であることを確認し，
複数区間の同時計測手法を提案する

41

シミュレーション設定

- 複数区間の同時計測可能性を調べるためのシミュレーションをns-2[6]により行った
  - ネットワークモデル



  - 計測パケットをs0からs2に向けて送る
  - 利用可能帯域の計測対象区間: 2段目のリンク
  - 計測区間の前後で連続したK_0個の計測パケットの平均通信レートをみる
  - 背景トラヒック: 以下を平均とする指数分布のパケット間隔
    - 1段目のリンク: X_1 [Mbps], 2段目のリンク: X_2 [Mbps]

42

シミュレーション結果

- 利用可能帯域が1段目: 40Mbps, 2段目: 50Mbpsの場合

流入レートと流出レートの関係が一定でない　計測困難

流出レートが流入レートより小さくなっている　計測可能

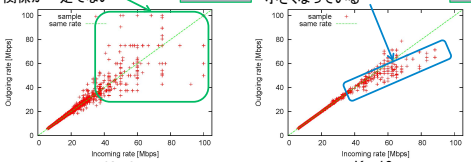


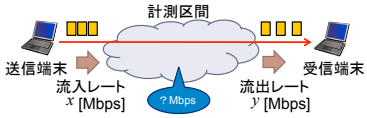
図 計測区間における流入レートと流出レートの関係

K_0 を大きくすることで計測が可能となる

43

提案手法 (1)

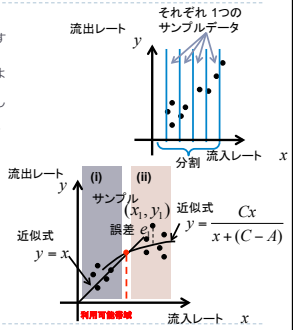
- 送信端末から様々なレートで計測パケットを送出し，各計測区間における計測パケットの流入レートと流出
  レートを用いて，利用可能帯域を推定する



44

提案手法 (2)

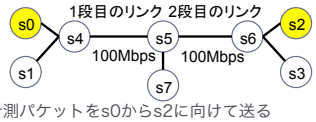
- 利用可能帯域の推定手順
  1. サンプルデータ (x_i, y_i) を取得する
  2. サンプルデータを流入レートにより分割し，分割されたサンプルデータに対して平均をとることで1つのサンプルデータとする
  3. 利用可能帯域を仮定し，モデルとの平均2乗誤差を算出する
  4. 3.で算出した平均2乗誤差が最小のときの利用可能帯域を推定値とする



45

性能評価: 評価環境

- ns-2を用いたシミュレーションにより，提案手法に対して計測精度の評価を行った
  - ネットワークモデル

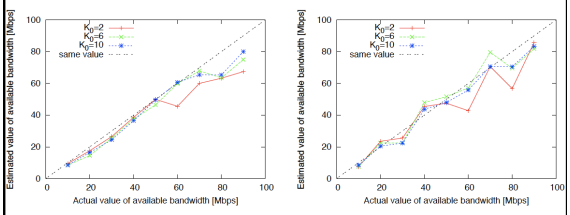


  - 計測パケットをs0からs2に向けて送る
  - 利用可能帯域の計測対象区間: 2段目のリンク
  - 計測区間の前後で連続したK_0個の計測パケットの平均通信レートをみる
  - 背景トラヒック: 以下を平均とする指数分布のパケット間隔
    - 1段目のリンク: 10, 20, ..., 90 [Mbps]
    - 2段目のリンク: 10, 20, ..., 90 [Mbps]

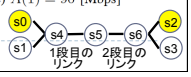
46

性能評価: 1段目の利用可能帯域が大きい場合

- 高い精度で計測が可能



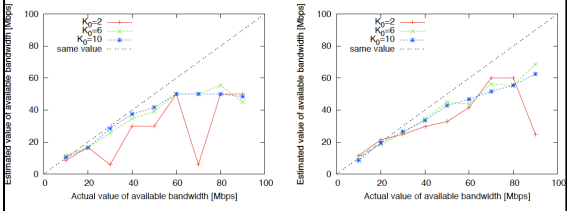
(d) $A(1) = 70$ [Mbps]　(e) $A(1) = 90$ [Mbps]



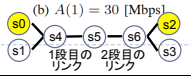
47

性能評価: 1段目の利用可能帯域が小さい場合

- 2段目の方が帯域が大きいと、精度が若干低下



(a) $A(1) = 10$ [Mbps]　(b) $A(1) = 30$ [Mbps]



48

まとめと今後の課題

- まとめ
  - エンド端末間パス上に存在する，複数ネットワーク区間の
    利用可能帯域を同時に計測する手法を提案
  - 受信端末に近いネットワーク区間の利用可能帯域が送信端末に近いネットワーク区間と比べ大きくなる場合でも計測可能であることを示した
- 今後の課題
  - 計測精度と計測負荷を考慮した計測レートの決定方法

49