Vol.2009-OS-112 No.16
2009/8/6

# オーバレイネットワークにおける論理リンクの通信遅延変動に関する一考察

浅原 理人[†1]　河野 健二[†1]

PlanetLab や Emulab といった，オーバレイネットワークシステムの動作をインターネットと同様の環境下で評価できる環境が広く利用されている．しかし，既存の評価環境はインターネット上で非決定的に発生したエンドホスト間リンクの特性変動を記録し再現するシステムではない．そのためインターネットで現実に発生した状況下で繰り返し試験を行うことは難しいといえる．我々は現在，エンドホスト間リンクの特性を記録する基盤を構築中である．本論文ではこの試みの先駆けとして行った，通信遅延の変動に関する調査結果を報告する．PlanetLab 上の約 17 万のエンドホスト間経路で通信遅延の変動を記録した．分析の結果次の 3 点を確認した．1 点目はエンドホスト間のリンク特性の変動は大きく 4 種類に分けられることである．2 点目は通信遅延の変動とホストの負荷の変動との相関は極めて弱いことである．3 点目は同一ネットワーク内で観測した通信遅延の変動は類似性が高いことである．

# Analysis of the Latency Fluctuations on the Logical Links of Overlay Networks

Masato Asahara [†1] and Kenji Kono[†1]

To evaluate overlay network systems under an Internet-like environment, Internet testbeds and Internet emulation environments have been widely used. PlanetLab and Emulab exemplify them. Unfortunately, these traditional testbeds cannot help us repeatedly test our overlay networks under the same condition of the Internet. This is because these testbeds do not record and replay the fluctuations of the network conditions on end-to-end paths. We are developing a recording system for the fluctuations of the network conditions on end-to-end paths. In this paper we report the preliminary analysis of the latency fluctuations on 17 thousand end-to-end paths between PlanetLab nodes. Our analysis suggests that 1) the latency fluctuations on end-to-end paths are classified into four patterns, 2) there is little correlation between the latency fluctuations and the load of a node, 3) the latency fluctuation patterns on end-to-end paths from the same network are similar to each other.

## 1. はじめに

インターネット上で動作することを想定した大規模なアプリケーション層オーバレイネットワークシステムを研究開発する機会が多くなっている．その中には既に実運用され実際にサービスを提供しているものも多い．例えば著明なコンテンツ配信システムとして Akamai[1] や BitTorrent[2] がある．オーバレイネットワークシステムは運用段階において不具合が発見され改良が行われることがしばしばある．例えば通信遅延を予測する要素技術であるネットワーク座標系[3),4)] をオーバレイネットワークシステムに組み込んだ際，通信遅延の揺らぎがネットワーク座標系の予測精度に大きく影響を与えることが指摘されている[5)]．従ってオーバレイネットワークを構築する際に下層となるネットワークの状態を現実の状態に似せて試験を行うことが，オーバレイネットワークシステムの開発にとって重要であるといえる．

オーバレイネットワークシステムの開発や評価を支援するために，インターネットで発生する現象を試験に反映する様々な試みが行われている．ひとつは PlanetLab[6)] のような，インターネット上に分散配置された計算機群によって，インターネット上での実践的な試験を可能にするものがある．別のアプローチとして Modelnet[7)] や Emulab[8)] のような，インターネットをモデル化することでインターネットの状態をエミュレートしシステムの動作検証を行うものがある．また iPlane[9)] などのような，インターネットにおけるエンドホスト間のパケット転送経路を記録する基盤も提案されている．これらのシステムを用いることによって，インターネットの特徴を考慮したオーバレイネットワークシステムの動作検証を行うことが可能になってきている．

しかし，これらのシステムはエンドホスト間の論理リンクにおける非決定的な特性変動の記録や再現までは達成していない．インターネットではエンドホスト間の論理リンクの通信遅延やスループット，パケットロス率といった特性が短時間のうちに非決定的に変動することが知られている．このような特性の変動が実運用の際システムに影響を与えることがある[5)]．アルゴリズムの設計もしくは実装上の不具合は，一般に同一条件下で再試験することが原因の特定や修正の助けとなる場合が多い．インターネットで非決定的に発生した論理リンクの特性変動を記録することができれば，記録した時間帯のインターネットの状態を再現

†1 慶應義塾大学
Keio University

することが容易になる．これにより，実際に起きた特定の条件下で発生する不具合の修正をより効率よく行えることが期待できる．

我々は現在，インターネット上のエンドホスト間における論理リンク特性のスナップショット取得基盤を構築中である．ここでいうスナップショットとは，ある時刻におけるエンドホスト間リンクの特性値の集合である．このスナップショットを連続して取得することで，ある時間帯のインターネットの状態を記録しておくことができる．このスナップショットと既存のエミュレーション環境を併用すれば，記録した時間帯のインターネットの状態を任意の回数再現することができるようになる．本研究では最初の試みとして，任意のエンドホスト間における通信遅延のスナップショットを取得する基盤の構築を目指す．

本論文ではこの試みの先駆けとして行った，通信遅延の振る舞いに関する調査結果を報告する．より精度が高くかつ副作用の小さいスナップショット取得システムを構築するために，まずは取得の対象となる事象の分析を行った．PlanetLab で稼働する 514 台のノードを用い，約 17 万のエンドホスト間で 9000 万以上の ICMP パケットを送受信し，通信遅延の変動を記録した．分析の結果次の 3 点を確認した．1 点目はエンドホスト間のリンク特性の変動は大きく 4 種類に分けられることである．2 点目は通信遅延の変動とホストの負荷の変動との相関は極めて弱いことである．3 点目は同一ネットワーク内にあるホストは同一の傾向を示す場合が多いことである．また，この分析結果から設計上のアイディアを 2 点導出した．1 点目は，同一ネットワーク内のホスト間で協調し，非決定的に発生する通信遅延の異常状態を記録するというものである．2 点目は，統計手法を用いて通信遅延の異常状態を検出し，動的に記録の頻度を高めるというものである．

本論文の構成は以下の通りである．2 章では関連研究について述べる．3 章では本スナップショット取得システムが達成すべき要件と仮定する条件について述べる．4 章では調査した通信遅延の変動の分析結果について述べ，5 章ではその結果から導き出した設計上のアイディアをまとめる．最後に，6 章で本論文をまとめる．

## 2. 関 連 研 究

インターネット上での試験を可能にするテストベッドが提供されている．PlanetLab[6] などはインターネット上で稼働しているので，実際にインターネット上で発生するエンドホスト間論理リンクの特性変動を反映したオーバレイネットワークシステムの動作検証が行える．しかし，実運用されているインターネット上で直接試験を行うため，同じ条件を再度整えることは事実上不可能である．

インターネットをシミュレートもしくはエミュレートする評価環境が提案されている．ns-2[10] ではリンクやパケットキューといった，ネットワークを構成する要素を組み合わせて擬似的なネットワークを構築し，パケットの転送をシミュレートする．また，Emulab[8] や ModelNet[7] ではインターネットをモデル化しエミュレートすることで，実装したオーバレイネットワークシステムを擬似的なインターネット上で動作させ検証することができる．これらネットワークシミュレータやネットワークエミュレータは詳細に制御可能な仮想ネットワーク環境を構築し人工的に通信遅延を発生させるので，特定の条件を繰り返し再現することができる．しかし，ネットワークシミュレータやネットワークエミュレータは試験対象のネットワークをモデル化して表現するものであり，特定のインターネットの状況を記録するものではない．Flexlab[11] は，実際のインターネット上にパケットを送信することでホスト間のパケット転送に関する情報を取得し，その情報を元にエミュレーションを行う．Flexlab ではインターネットの状態を記録することは行わない．

インターネット上のトラフィックを再現する試みとして Swing[12] がある．Swing では，ある物理リンクのパケットトレースデータを元にして，記録した状況と同様のバックグラウンドトラフィックをエミュレートし擬似的に記録当時の状況を再現する．Swing は特定の物理リンクのバックグラウンドトラフィックを再現することを目的としているが，インターネットの状態を記録することは目的としていない．

インターネット全体の構造を記録する基盤として，iPlane[9] や iPlane Nano[13] がある．iPlane はインターネット上のパケット転送経路とその経路上の特性を記録し，オーバレイ上の論理リンクの特性を推測する環境を提供する．iPlane Nano は個々の物理リンクの情報を記録し，BGP などのルーティングプロトコルの特徴からパケットの転送経路を推測することで iPlane と同程度の精度を約 1000 分の 1 程度のデータサイズで提供する．これらの基盤は数時間から数日単位での変化を記録するものであり，分から秒単位で非決定的に発生する特性変化を記録する用途ではない．

このように，インターネットの状態を分から秒間隔で記録するシステムは我々が調査した範囲ではいまだ存在しない．本研究では新しい試みとして，インターネットの状態をこれまでより細かい間隔で記録できるシステムを構築する．このシステムが提供する情報と既存のエミュレーション環境を組み合わせることで，記録時のインターネットの状態を容易に再現できるようになる．

## 3. 本記録システムの要件と仮定する条件

本章では，本記録システムが達成すべき要件と，本記録システムが仮定する条件について述べる．

### 3.1 本記録システムの要件

本記録システムは次に挙げる 4 点の達成を目指す．

- **高い精度**．数分から数秒単位で非決定的に変動する経路特性を記録できなければならない．インターネット上のエンドホスト間経路の特性は様々な要因によって変動する．ここでは TCP や IP といったアプリケーション層より下位のプロトコルによる制御の影響や，物理的な障害による影響のことを指す．例えばパケットの輻輳制御や障害点回避，インターネットサービスプロバイダ間の契約条件で定められた Autonomous Systems (AS) 間の複雑なパケット転送ポリシーなどが要因としてある．本記録システムはこういった要因で変化したエンドホスト間経路の特性を記録できなければならない．
- **副作用が小さい**．本記録システムがネットワークに与える影響をできうる限り小さくする必要がある．本記録システムが送受信するパケットが，ネットワーク上のルータやセキュリティシステムの動作を変更する要因となってはならない．ルータの観点でいえば，例えばルータの処理パケットキューが本システムの送信するパケットで溢れてしまうようなことがあってはならない．またセキュリティの観点では，セキュリティシステムが攻撃とみなすような条件を本記録システムが満たしてはならない．例えばファイアウォールサービスを提供する機器のひとつである Netscreen[14] の標準設定では，特定のホストから毎秒 1000 以上のパケットが到着すると攻撃とみなす．よって，本記録システムは計測用に送受信するパケットの数をできうる限り少なくする必要がある．
- **対象が広範囲**．インターネット上に存在するあらゆるエンドホスト間の経路を記録できる構造でなければならない．そのためには，途中の経路上に存在するルータへ新たな機能を導入することなく本記録システムを実現しなければならない．
- **高いスケーラビリティ**．インターネット規模の記録基盤を提供するには，増加するホスト数に対して本記録システムがスケールする必要がある．例えば BitTorrent[2] プロトコルを実装した Azureus[15] は百万台規模のクライアントが動作しているといわれている[5]．

### 3.2 本記録システムが仮定する条件

本記録システムでは次の 2 点を仮定する．まず，本記録システムはアプリケーション層で構築されたオーバレイネットワークを対象とする．例えば Peer-to-Peer ファイル共有システム[2),16)] や Peer-to-Peer Content Distribution Networks[17)]，オーバレイマルチキャストシステム[18)]，分散ハッシュ表[19)–22)] などが対象である．ルータ間の協調動作の検証といった，アプリケーション層より下位のレイヤで構築されるオーバレイネットワークシステムは対象としない．

2 点目に，本記録システムはエンドホスト上からエンドホスト間経路の特性を観測し記録する．オーバレイネットワークの構成要素となりうるエンドホスト上にプローブを導入し，エンドホスト間の経路特性を計測する．これは，PlanetLab などの既存のテストベッド環境や，BitTorrent ノードなどのボランティアで参加しているノードを利用することで実現可能である．

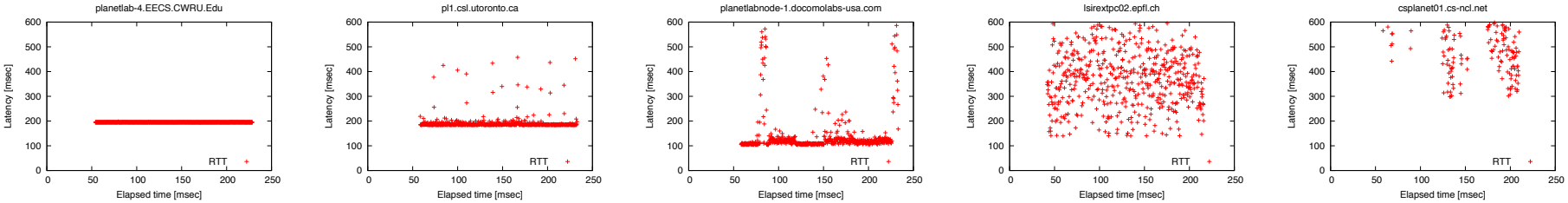
## 4. 通信遅延の振る舞い

本章では，本記録システムを構築する前に行った調査について報告する．3 章で述べたように，本記録システムは非決定的に発生するエンドホスト間の論理リンクの特性変動を記録する一方で，ネットワークに与える副作用を小さくするために計測用のパケットの送受信量を小さくしなければならない．そのために本記録システムは，論理リンクの特性変動の特徴を捉え，特性の変動に応じて記録頻度を変更するような構造でなければならないと考えられる．

そこで我々は本記録システムの詳細な設計を行う前に，インターネットにおけるエンドホスト間の経路特性の変動について調査した．今回の調査では通信遅延の変動に焦点を絞って行った．その結果，通信遅延の変動に関してシステム設計上有用と思われる 3 つの特徴を確認した．1 点目は，エンドホスト間の経路特性の変動が 4 種類のパターンに分類できることである．2 点目は，通信遅延の変動とエンドホストの負荷の変動との相関が極めて弱いことである．3 点目は，同じネットワークからの経路では似たような通信遅延の変動が観測できることである．以下，この調査の詳細について述べる．

本調査は 514 台の PlanetLab ノードを利用して行った．計測は 2009 年 3 月 15 日から 16 日の 27 時間にかけて行った．計測には Fedora 標準の ping コマンドを用い，1 つの送信先ホストあたり 0.3 秒間隔で 180 秒間 ICMP パケットを同時間帯に送信しラウンドトリップタイムを計測した．この計測を 514 台のホストに対して順次送信先ホストになるように繰り返した．ただし，計測の途中で制御が不能になったノードがある．計測不能になったノードに対してはそれまでに完了した計測データのみを取得している．この調査では

Vol.2009-OS-112 No.16
2009/8/6




(a) planetlab-4.EECS.CWRU.Edu to planetlab2.sfc.wide.ad.jp　(b) pl1.csl.utoronto.ca to planetlab2.sfc.wide.ad.jp　(c) planetlabnode-1.docomolabs-usa.com to planetlab2.sfc.wide.ad.jp　(d) lsirextpc02.epfl.ch to planetlab4.cse.nd.edu　(e) csplanet01.cs-ncl.net to planetlab2.sfc.wide.ad.jp

**図 1**　4 種類の通信遅延変動の例．(a) は通信遅延が安定している場合，(b) は通常は安定しているが突発的に通信遅延が増大する場合，(c) は一定の時間帯に通信遅延が増大もしくは減少する場合，(d)，(e) は常に不安定な場合の例である．

169,804 のエンドホスト間経路に対してのべ 97,500,532 回 ICMP パケットを送信した．

### 4.1　変動の種類

調査の結果，通信遅延の変動の様子は大きく 4 種類に分けられることがわかった．4 種類とは，1) 通信遅延の変動が常に一定の場合，2) 通信遅延が突発的に増大する場合，3) 一定時間通信遅延が増加もしくは減少することが不定期に発生する場合，4) 常に通信遅延が不安定な変動を示す場合である．

図 1 は 4 種類の通信遅延変動の例を示す．図 1 (a) は通信遅延が常に一定の値を示す場合の例である．極めて安定した通信経路が確保されている場合このような分布を示す．(a) の状態の場合計測不能回数と観測値の標準偏差 $\sigma$ が極めて小さい値を示す傾向がある．例えば調査したエンドホスト間経路では計測不能回数が 0 回であり，標準偏差は 0.2 ミリ秒 であった

図 1 (b) は突発的に増大した通信遅延が観測される場合の例である．(b) の場合では，大部分の観測では (a) と同様に一定の通信遅延値が得られるが，不規則に数倍から数十倍に増大した通信遅延値が観測されることがある．(b) の場合，標準偏差は (a) と同様に小さいが，外れ値と中央値との差が極めて大きくなる．図 1 (b) の場合，中央値が 185.5 ミリ秒，標準偏差が 32.5 ミリ秒であり，250.5 ミリ秒以上の計測値 18 個を外れ値とみなした．外れ値と中央値との差の平均は 175.2 ミリ秒であった．こうした現象が発生する原因は 2 点考えられる．1 点目は，経路中のルータの負荷が一時的に上昇したためにパケットの転送遅延が発生した影響である．2 点目は，BGP 等のルーティングプロトコルによる一時的なパケット転送経路の変更による影響である．

図 1 (c) は一定時間通信遅延が増加もしくは減少することが不定期に発生する場合の例である．図 1 (c) の場合，例えば 79 秒経過時から 7 秒間，他の時間帯と比べて通信遅延が増大している．(c) のような状態とは計測値の分布から見ると，あるスロットで時分割したときにスロットごとの中央値が他と比べて大きく異なっているものがある場合と見ることができる．

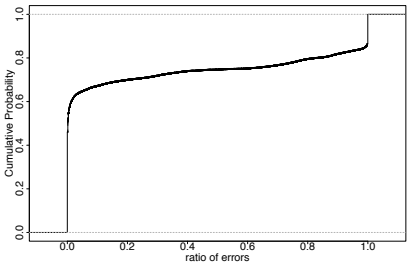



**図 2**　計測不能回数の割合の分布．

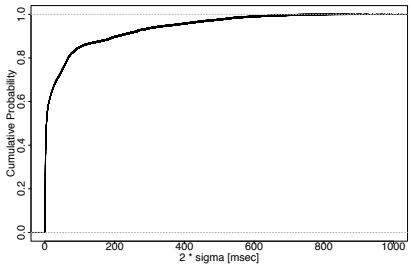



**図 3**　$2\sigma$ 値の分布．

図 1 (d) や (e) は通信遅延が不安定に変動する場合である．この種類に分類されるものの特徴は 2 点ある．ひとつは (d) のように，標準偏差が 100 ミリ秒オーダーと非常に大きいことである．もうひとつは (e) のように，計測不能であった回数が比較的大きいことである．

4 種の分類の根拠とした，測定不能数の割合，標準偏差，外れ値の割合および中央値と最低値の差の分布について調査した．図 2 は計測不能であった回数の割合を累積頻度グラフで示している．計測不能回数の割合が大きい経路は図 1 (e) のような，通信できない時間帯がしばしば発生するような経路だと考えられる．図 2 から，計測不能回数の割合がきわめて小

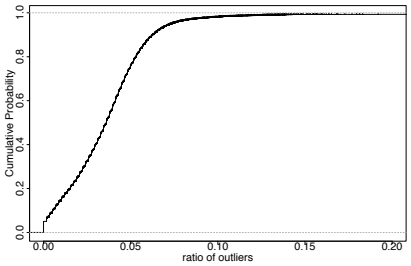



**図 4** 外れ値の割合の分布.

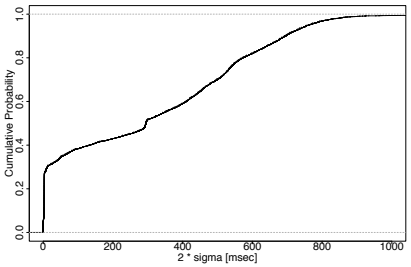



**図 5** 外れ値が 0 である経路の $2\sigma$ 値の分布.

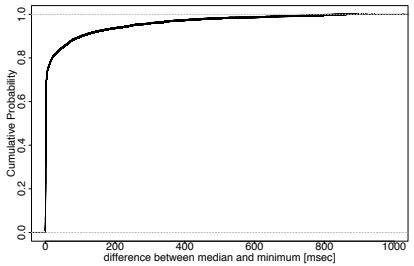



**図 6** 中央値と最低値の差の分布.

さい経路の割合が大きいことがわかる．全くエラーが発生しなかった経路が全体の 46.0 % を占めていた．また，エラーの発生割合が 1 % 未満であったものは全体の 58.3 %，10 % 未満のものは全体の 66.9 % であった．一方で，ほとんどが計測不能であった経路の割合も大きい．100 % 計測不能であった経路が全体の 13.5 % を占めていた．また，90 % 以上計測不能であった経路は全体の 18.3 % であった．このことから，計測不能になりやすい経路となりにくい経路とで極端に分かれる傾向にあることが読みとれる．

図 3 は $2\sigma$ 値の分布，すなわち 75 % 信頼区間[⋆1]の分布を表す．$2\sigma$ 値が小さいものは図 1 (a)，(b) のような通信遅延の変動が比較的小さい経路であるといえる．また，$2\sigma$ 値が大きいものは図 1 (d) のような通信遅延の変動が大きい経路であるといえる．図 1 が示す結果は，通信遅延の揺らぎが小さい経路が比較的多いことを示している．75 % 信頼区間が ±1 ミリ秒未満の経路が全体の 21.3 % を占めている．また，±10 ミリ秒未満の経路は全体の 50.3 % であった．一方で，75 % 信頼区間が ±100 ミリ秒以上の経路は全体の 13.0 %，±500 ミリ秒以上の経路は全体の 2.2 % であった．通信遅延の変動の多くは ±10 ミリ秒内であることから，記録システムではミリ秒オーダーの通信遅延の変動を記録できる必要があるといえる．また，数百ミリ秒のオーダーで変動する経路も確認できたことから，計測値を単純な線形補間で補うことは誤差の増大を招く恐れがあると考えられる．

図 4 は外れ値の割合の分布を表す．この場合の外れ値とは，75 % 信頼区間に含まれない計測値を指す．外れ値が全く計測されなかった経路が全体の 4.4 % 存在した．また，外れ値が 10 % 未満であった経路は全体の 84.7 % を占めた．外れ値の割合が大きい経路は図 1 (b) のような，標準偏差は小さいがしばしば異常に大きな通信遅延が計測されるような経路だと判定できる．

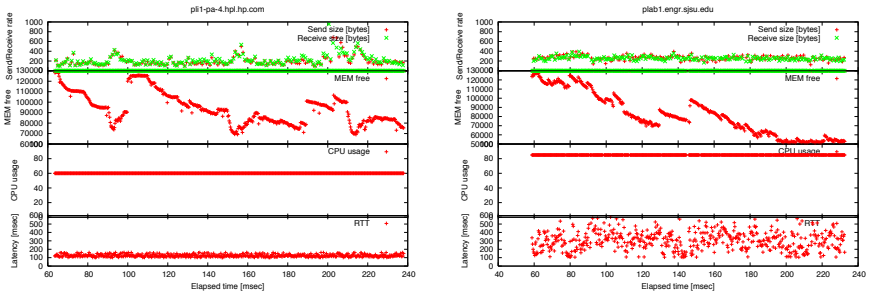
注意すべきなのは，外れ値が少ない経路の計測値が必ずしも安定しているとは限らない点である．図 5 は外れ値が全くなかった経路の $2\sigma$ 値の分布を示したものである．このうち $2\sigma$ 値が ±1 ミリ秒未満であった経路は全体の 6.2 % であった．また，±10 ミリ秒未満のものは全体の 29.4 % であった．一方で ±100 ミリ秒以上のものは全体の 61.6 % を占めた．これらの経路では図 1 (d) のように，通信遅延値が比較的大きく揺らいでいるといえる．このことから，通信遅延の分布の異常を検出するには，外れ値と $2\sigma$ 値の両方を監視する必要があるといえる．

図 6 は計測値の中央値と最低値の差の分布を表す．通信遅延はパレート分布を示すことが知られている[23]．よって，中央値と最低値の差が小さいほど，多くの通信で理想値に近い通信遅延が観測されたことを示す．また，差が大きいものは通信遅延の変動が大きいと読みとれる．今回の計測では，エンドホスト間経路の 50.7 % で中央値と最低値の差が 1 ミリ秒未満であった．また，65.8 % が 10 ミリ秒未満，77.3 % が 100 ミリ秒未満であった．200 ミリ秒以上であったのは全体の 5.6 % であった．このことから，図 1 の (d) に相当するパターンの全体にしめる比率は低いと考えられる．

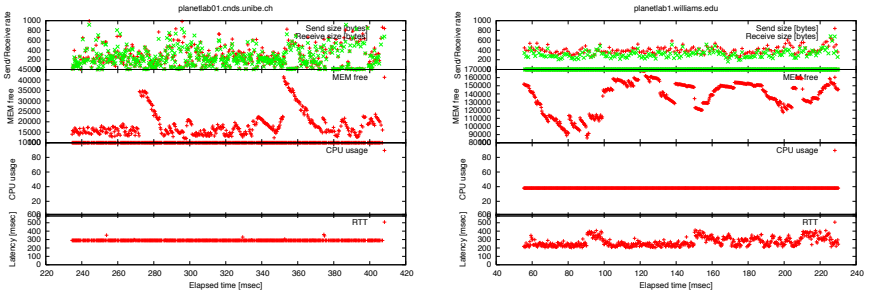
### 4.2 通信遅延の変動とエンドホストの負荷情報との関係

仮に観測元ホスト内の負荷によって通信遅延が大きく変わるのであれば，エンドホストの負荷を観測することで通信遅延の変動が予測できることになる．そこで，エンドホストの負荷の変動と通信遅延の変動との関係を調査し，ホストの負荷から通信遅延の変動が予測できるかどうか確かめた．ここではありふれた負荷の指標である CPU 使用率，空きメモリ量お

⋆1 通信遅延は非正規分布を示すのでチェビシフの不等式に従って算出している．




(a) pli1-pa-4.hpl.hp.com　(b) plab1.engr.sjsu.edu




(c) planetlab01.cnds.unibe.ch　(d) planetlab1.williams.edu

**図 7** 通信遅延の変動と CPU 使用率，メモリ使用率およびデータの送受信レートとの関係（送信先ホストは planet-lab2.sfc.wide.ad.jp）．(a) および (b) は CPU 使用率，空きメモリ量，送受信レートが似たような傾向を示しているが，(b) は通信遅延が大きく変動している．一方で，(c) および (d) では (c) の方が CPU 使用率やメモリ使用率が高く送受信レートも大きいが，(d) の方が通信遅延の揺らぎが大きい．

よびネットワークの送受信レートを ICMP パケット送信時に計測した．それぞれの指標と観測した通信遅延値との相関を検定し，ホストの負荷の変動と通信遅延の変動に関係があるか調査した．

図 7 は計測結果の一部である．それぞれのグラフは上から送受信レート，空きメモリ量，CPU 使用率，通信遅延を表している．図 7 (a) および (b) は 送受信レートがほぼ一定かつ空きメモリ量が減少している．しかし，(b) の方が明らかに通信遅延の変動が激しい．(a) と (b) では (b) の方が CPU 使用率が総じて高いので，CPU 使用率が高いと通信遅延が増大するとも考えられる．ところが (b) と (c) を比較すると，(c) は CPU 使用率がほぼ 100 % であるにも関わらず通信遅延はほぼ一定であるのに対して，(b) は (c) よりも CPU

**表 1** 通信遅延の変動と CPU 使用率，空きメモリ量，データの受信量および送信量の変動との相関係数の分布（総サンプル数 164,910）．

| | 有意な相関が認められた割合 | 相関係数 $r$ の範囲 | $\mid r \mid > 0.2$ である経路の割合 |
|---|---|---|---|
| CPU 使用率 | 0.04 % | $[-0.29, 0.32]$ | 0.0061 % |
| 空きメモリ量 | 11 % | $[-0.73, 0.79]$ | 1.3 % |
| ネットワーク受信量 | 4.0 % | $[-0.82, 0.50]$ | 0.034 % |
| ネットワーク送信量 | 4.1 % | $[-0.53, 0.50]$ | 0.033 % |

使用率が低いにもかかわらず通信遅延の変動が大きい．また (c) と (d) を比較したところ，(c) は CPU 使用率とメモリ使用量が大きく送受信レートの変動が激しいにも関わらず，(d) よりも通信遅延の変動が小さい．これらの結果から，通信遅延の変動と CPU 使用率や空きメモリ量，ネットワークの送受信レートとの間に相関が見られないことが予想できる．

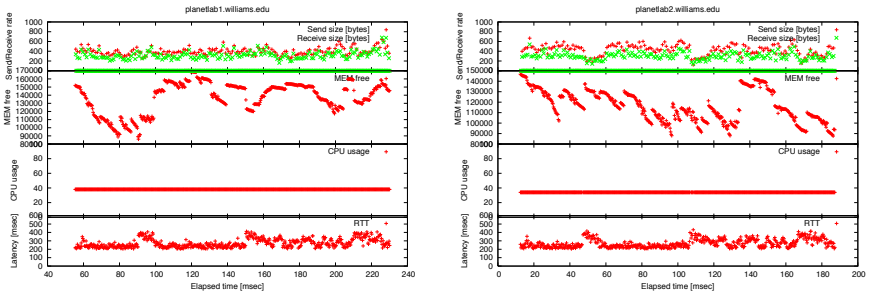
通信遅延の計測値と CPU 使用率や空きメモリ量，ネットワークの送受信レートとの関係を定量的に確認するために，統計手法の一つである無相関検定を行った．ここでは通信遅延の計測値の分布が非正規分布であることから，スピアマンの順位相関係数に基づく無相関検定を行った．検定は CPU 使用率，空きメモリ量，送受信レートそれぞれに対して行った．帰無仮説 $H_0$ は「通信遅延の計測値との相関は 0 である」，対立仮説 $H_1$ は「通信遅延の計測値との相関は 0 ではない」である．

表 1 は検定の結果を示す．検定の結果，帰無仮説が棄却できた経路は高々 11 % であった．またその中で，相関係数の絶対値が 0.2 を超えたのは高々 1.3 % であった．これは，一般に弱い相関があるといえる基準を超えた経路は高々 1.3 % であったと読み取れる．この結果から，ホスト内の負荷の変動と通信遅延の変動との相関は定量的に極めて弱いといえる．

### 4.3 同一ネットワーク内ホストでの通信遅延変動の類似性

同一ネットワーク内にあるホストは，送信先のホストまでの経路が同じである可能性が高い．このことから，同一ネットワーク内のホストの通信遅延変動は類似性が高いと予想できる．そこで，同一ネットワーク内のホスト間における通信遅延変動の類似性を調査した．

図 8 は williams.edu ネットワーク内のホストにおける計測値を示したものである．グラフが示すとおり，これら 4 台のホストの CPU 使用率，メモリ空き容量およびデータの送受信レートは異なっている．しかし，通信遅延の変動の様子は似通ったものになっている．この類似性は測定不能期間の発生においても見られる．図 9 は同一ネットワーク内において測定不能な期間があったホストの例である．(a) および (b) で発生した測定不能期間の時刻がおおよそ似通ったものとなっていることがわかる．



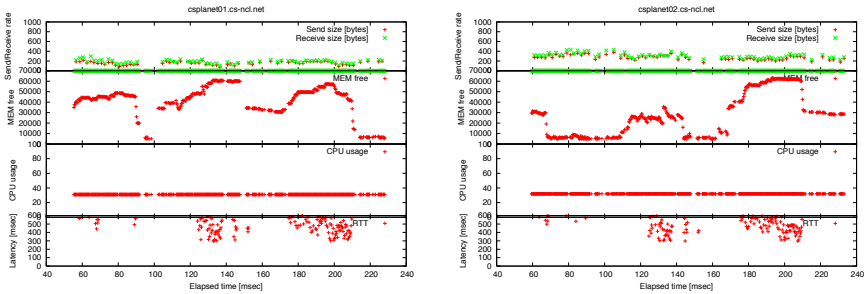

(a) planetlab1.williams.edu (b) planetlab2.williams.edu

(c) planetlab3.williams.edu (d) planetlab4.williams.edu

**図 8** 同一ネットワーク内のホストにおける，通信遅延の変動と CPU 使用率，メモリ使用率およびデータの送受信レート（送信先ホストは planetlab2.sfc.wide.ad.jp）．4 台で似たような通信遅延の変動が起きている．一方で，CPU およびメモリ使用率や送受信レートの変動の様子は異なっている．

同一ネットワーク内のホストとその他のホストとにおける通信遅延の変動の違いを確認するために，ホストごとの通信遅延の分布を箱ひげ図で表したものが図 10 である．横軸が通信遅延，縦軸がホストのドメイン名をサブドメイン区切りで反転して表記したものである．このグラフから，ドメイン内で似通った通信遅延の分布を示す傾向があることが読み取れる．

一方で，同じドメインであっても異なる通信遅延の分布を示すものがあることも読み取れる．たとえば pli1-br-2.hpl.hp.com や pli1-pa-4.hpl.hp.com は hpl.hp.com ドメイン内のホストであるが，他の 2 台と明らかに異なる通信遅延の分布を示している．IP アドレスを調べてみたところ，pli1-br-2.hpl.hp.com の IP アドレスは 192.6.10.50，pli1-pa-4.hpl.hp.com の IP アドレスは 204.123.28.55 であったのに対して，他の 2 台は 192.6.26.31 と 192.6.26.33 であった．よって，pli1-br-2.hpl.hp.com と pli1-pa-4.hpl.hp.com はドメインが同じではあるが，設置されたネットワークは他と異なると考えられる．一方/24 が一致し同じネットワークのホストであると考えられる pli2-pa-1.hpl.hp.com と pli2-pa-3.hpl.hp.com はよく似た通信遅延の分布を示していた．traceroute コマンドでルーティング経路を調べてみたところ，この 2 台は同じ経路を辿っていた．このことからこの 2 台は同じネットワークに設置されているといえる．




(a) csplanet01.cs-ncl.net (b) csplanet02.cs-ncl.net

**図 9** 不定期に応答がない状態における通信遅延の変動と CPU 使用率，メモリ使用率およびデータの送受信レート（送信先ホストは planetlab2.sfc.wide.ad.jp）．通信障害の発生パターンも，同一ネットワーク内で似たようなものになる．一方で，CPU およびメモリ使用率や送受信レートの変動の様子は異なっている．

これらの検証から，通信遅延の分布は送信先ホストまでのネットワーク経路に大きく依存し，同一ホストでの通信遅延の変動は類似性が高いと考えられる．ただし，cnds.unibe.ch ドメイン内のように外れ値の分布が異なることもある．このような同一ネットワーク内における通信遅延変動の相違は，TCP の asymmetric routing のような，下位レイヤのプロトコルによる影響であると考えられる．

同一ネットワーク内における通信遅延変動の類似性の有無を定量的に評価するために検定を行った．ここでは通信遅延が非正規分布であることから，有意水準 5 % でフリグナー・キリーン検定を行い評価した．帰無仮説 $H_0$ は「同一ネットワーク内のホストで計測された通信遅延の中央値は等しい」，対立仮説 $H_1$ は「同一ネットワーク内のホストで計測された通信遅延の中央値に差がある」である．検定の結果，計測値が十分にあり検定が可能であったのべ 31,939 のネットワーク中 64.3 % で帰無仮説が採択された．また，計測失敗のなかったネットワークに絞ったところ，のべ 18,508 のネットワーク中 58.9 % で帰無仮説が採択された．なお，フリグナー・キリーン検定はひとつでも分布が他と異なるものがあっ

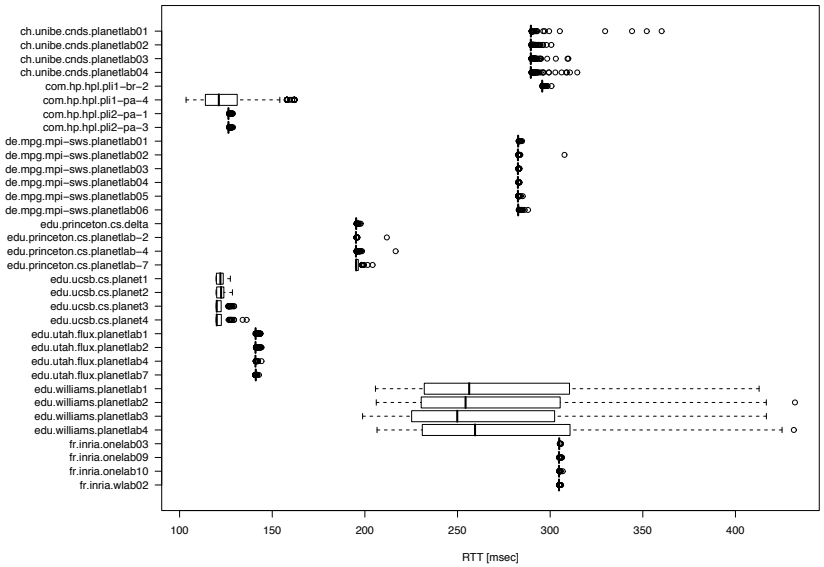



**図 10** planetlab2.sfc.wide.ad.jp に対する通信遅延の分布.

た場合帰無仮説は棄却される．したがって実際はこの値よりも多くのネットワークで類似性が認められると考えられる．

## 5. 導出した設計上のアイディア

通信遅延の振る舞いに関する調査から，我々は本記録システムの設計に関して 2 点のアイディアを導き出した．1 点目は同一ネットワーク内で協調動作し，パケット送信数を抑えながら異常状態を検出するというアイディアである．4 章の検証の通り，ホストの負荷情報からでは通信遅延の変動を予見することは難しい．そのためそれぞれのホストで独立して通信遅延の変動を記録しようとすると，非決定的に発生する通信遅延の異常を記録するために細かい間隔で計測用のパケットを送信しなければならなくなる．そこで，同一ネットワーク内のホストで通信遅延の分布の類似性が高いことを利用して計測する方法が考えられる．計測用パケットの送信間隔を広げることで送信パケット数を減らす代わりに，同一ネットワーク内のホストで送信タイミングをずらすことで計測間隔を見かけ上短くする．あるホストが通信異常を検出すると，計測の間隔を狭めると同時に異常を検出した旨を同一ネットワーク内の他のホストに通知する．これにより，計測用のパケットを減らしつつ計測の精度を高めることが期待できる．

2 点目はエンドホスト間経路の異常を統計処理によって認識し，動的に記録頻度を高めるというアイディアである．過去の中央値および $2\sigma$ 値と現在の計測値を比較したり計測不能数の変動を監視することで，現在のエンドホスト間経路に異常が発生しているかどうか判定することができると考えられる．例えば，$2\sigma$ 値よりも大きい通信遅延が観測された場合はエンドホスト間経路に異常が起きている可能性が高い．また，$2\sigma$ 値が以前より増加傾向にある場合や，一定値より大きい場合は，異常が発生して通信遅延が大きく変動している可能性が高い．他に，最低値よりも極端に大きい通信遅延が観測された場合も，異常が発生している可能性が高いと判断できる．このタイミングで計測の頻度を上げれば異常状態を記録できると考えられる．この手法は図 1 の (b) 以外の状態を判定することが可能であると考えられる．(b) の種類の通信遅延変動を記録するには，本質的な解決手法は計測頻度を上げることである．これは 1 点目に挙げたアイディアにより，同一ネットワークからの計測値で補完することによって記録できると考えている．

## 6. おわりに

我々は現在インターネット上で発生した通信遅延の変動を記録するシステムを構築中である．本論文では本記録システムの設計をする際に行った予備調査の結果を報告した．調査では 169,804 のエンドホスト間経路においてのべ 97,500,532 回の ICMP パケットを送信し通信遅延の変動を観測した．分析の結果次の 3 点を確認した．1 点目は，エンドホスト間のリンク特性の変動は大きく 4 種類に分けられるということである．2 点目は，通信遅延の変動とホストの負荷の変動との相関は弱いということである．3 点目は，同一ネットワーク内にあるホストは同一の傾向を示す場合が多いということである．この分析結果から設計上のアイディアを 2 点導出した．1 点目は，同一ネットワーク内のホスト間で協調し，非決定的に発生する通信遅延の異常状態を記録するというものである．2 点目は，統計手法を用いて通信遅延の異常状態を検出し，動的に記録の頻度を高めるというものである．

今後は記録基盤の実装を行い，実際に非決定的に発生する通信遅延の異常が記録できるかどうか確かめる予定である．まずはエミュレーション環境である Emulab 上で実験を行い，人工的に発生させた通信遅延の揺らぎが精度よく記録できるか確かめる．その後，PlanetLab

等のインターネット環境で記録を行い，インターネット上で発生した通信遅延の変動が記録できるかどうか検証する予定である．

**謝辞** 本研究の一部は，文部科学省科学研究費補助金特定領域研究「情報爆発時代に向けた新しい IT 基盤技術の研究」による支援を受けている．

## 参 考 文 献

1) Dilley, J., Maggs, B.M., Parikh, J., Prokop, H., Sitaraman, R.K. and Weihl, W.E.: Globally Distributed Content Delivery., *IEEE Internet Computing*, Vol.6, No.5, pp. 50–58 (2002).
2) Cohen, B.: Incentives Build Robustness in BitTorrent., *Proc. of Workshop on Economics of Peer-to-Peer Systems*, 5 pages (2003).
3) Eugene Ng, T. S. and Zhang, H.: Predicting Internet Network Distance with Coordinates-Based Approaches., *Proc. of IEEE INFOCOM*, pp.170–179 (2002).
4) Dabek, F., Cox, R., Kaashoek, F. and Morris, R.: Vivaldi: A Decentralized Network Coordinate System., *Proc. of ACM SIGCOMM*, pp.15–26 (2004).
5) Ledlie, J., Gardner, P. and Seltzer, M.: Network Coordinates in the Wild, *Proc. of USENIX Symp. on Networked Systems Design and Implementation*, pp.299–311 (2007).
6) Bavier, A.C., Bowman, M., Chun, B.N., Culler, D.E., Karlin, S., Muir, S., Peterson, L.L., Roscoe, T., Spalink, T. and Wawrzoniak, M.: Operating Systems Support for Planetary-Scale Network Services., *Proc. of USENIX Symp. on Networked Systems Design and Implementation*, pp.253–266 (2004).
7) Vahdat, A., Yocum, K., Walsh, K., Mahadevan, P., Kostic, D., Chase, J. and Becker, D.: Scalability and Accuracy in a Large-Scale Network Emulator, *Proc. of USENIX Symp. on Operating Systems Design and Implementation*, pp.271–284 (2002).
8) White, B., Lepreau, J., Stoller, L., Ricci, R., Guruprasad, S., Newbold, M., Hibler, M., Barb, C. and Joglekar, A.: An Integrated Experimental Environment for Distributed Systems and Networks, *Proc. of USENIX Symp. on Operating Systems Design and Implementation*, pp.255–270 (2002).
9) Madhyastha, H.V., Isdal, T., Piatek, M., Dixon, C., Anderson, T.E., Krishnamurthy, A. and Venkataramani, A.: iPlane: An Information Plane for Distributed Services., *Proc. of USENIX Symp. on Operating Systems Design and Implementation*, pp.367–380 (2006).
10) VINT project: The Network Simulator ns-2, http://www.isi.edu/nsnam/ns/.
11) Ricci, R., Duerig, J., Sanaga, P., Gebhardt, D., Hibler, M., Atkinson, K., Zhang, J., Kasera, S.K. and Lepreau, J.: The Flexlab Approach to Realistic Evaluation of Networked Systems, *Proc. of USENIX Symp. on Networked Systems Design and Implementation*, pp.201–214 (2007).
12) Vishwanath, K. and Vahdat, A.: Swing: Realistic and Responsive Network Traffic Generation., *IEEE/ACM Trans. on Networking*, Vol.17, No.3, pp.712–725 (2009).
13) Madhyastha, H. V., Katz-Bassett, E., Anderson, T., Krishnamurthy, A. and Venkataramani, A.: iPlane Nano: Path Prediction for Peer-to-Peer Applications., *Proc. of USENIX Symp. on Networked Systems Design and Implementation*, pp. 137–152 (2009).
14) Juniper Networks, Inc.: Netscreen Series. http://www.juniper.net/.
15) Vuze.com: Azureus BitTorrent Client. http://azureus.sourceforge.net/.
16) AOL Nullsoft: Gnutella (2003).
17) Freedman, M.J., Freudenthal, E. and Mazières, D.: Democratizing Content Publication with Coral., *Proc. of USENIX Symp. on Networked Systems Design and Implementation*, pp.239–252 (2004).
18) Freedman, M.J., Lakshminarayanan, K. and Mazières, D.: OASIS: Anycast for Any Service., *Proc. of USENIX Symp. on Networked Systems Design and Implementations*, pp.129–142 (2006).
19) Stoica, I., Morris, R., Karger, D.R., Kaashoek, M.F. and Balakrishnan, H.: Chord: A scalable peer-to-peer lookup service for internet applications, *Proc. of ACM SIGCOMM*, pp.149–160 (2001).
20) Ratnasamy, S., Francis, P., Handley, M., Karp, R.M. and Shenker, S.: A Scalable Content-Addressable Network., *Proc. of ACM SIGCOMM*, pp.161–172 (2001).
21) Rowstron, A. I.T. and Druschel, P.: Pastry: Scalable, Decentralized Object Location, and Routing for Large-Scale Peer-to-Peer Systems, *Proc. of IFIP/ACM Int'l Conf. on Distributed Systems Platforms*, pp.329–350 (2001).
22) Zhao, B.Y., Huang, L., jeremy Stribling, Rhea, S.C., Joseph, A.D. and Kubiatowicz, J.: Tapestry: a resilient global-scale overlay for service deployment, *IEEE Journal on Selected Areas in Communications*, Vol.22, No.1, pp.41–53 (2004).
23) Allman, M. and Paxson, V.: On Estimating End-to-end Network Path Properties, *ACM SIGCOMM Computer Communication Review*, Vol.31, No.2 supplement, pp. 124–151 (2001).