模倣インターネット環境の構築

— AS 間ネットワーク構築の試行 —

三輪 信介* 鈴木 未央† 櫨山 寛章† 宇多 仁‡
宮地 利幸* 門林 雄基† * 篠田 陽一* ‡

インターネットに新しい技術を導入するためには、その技術が妥当であるのか、悪影響を及ぼさないのかなどを、実践的に検証する必要がある。特に広くインターネット上に展開する予定の技術について検証するためには、実際のインターネットに近い現実的な環境が求められる。そこで、このような技術の実験を支援するために、我々はテストベッド上に模倣インターネットを構築する試みに取り組んでいる。本稿では、テストベッド上に模倣インターネットを構築するための手法について述べるとともに、我々が行った StarBED 上での XEN を用いた仮想ノードによる AS 間ネットワーク構築の試みについて述べ、その評価と考察を行う。試行では 5000AS からなる AS 間ネットワークを 100 物理ノードで実現できることを確認した。

Building up the Internet on a Testbed

— Trial: Building the Inter AS Networks —

Shinsuke MIWA* Mio SUZUKI† Hiroaki HAZEYAMA† Satoshi UDA‡
Toshiyuki MIYACHI* Youki KADOBAYASHI† * Yoichi SHINODA* ‡

In order to evaluate new technologies which will be introduced to the Internet, these technologies should be practically experimented to confirm their adequacy and whether or not there are side-effects. To experiment some of these technologies, which will be widely deployed on the Internet, realistic environment which seems similar to the Internet are demanded. In order to support to experiment on realistic Internet like environment, we are now trying to make the Internet on a testbed. In this paper, we describe our method to construct the Internet like environment on the testbed, and we also report our trials of inter AS networks on StarBED with XEN and our tools. We achived 5000 ASes on 100 testbed nodes, and also estimated its performance and fidelity.

1 はじめに

インターネットのような大規模な分散環境に新しい技術を導入する場合、その技術がインターネット上で適切に動作するのか、既存のインターネット環境に悪影響を及ぼさないのかなどを、バイナリ実装レベルで実践的に検証しておく必要がある。このような検証を行うためには、現実のインターネットに則した実験環境が必要である。

現在のインターネットは、5 億台前後という非常に多くのホストやルータとそれらの間のネットワークで構成されている（ISC のインターネットホスト数統計 [1] による）。ネットワークは、運用の単位である AS（Autonomous System; 自律システム）に分割されており、AS 内ネットワークと AS 間ネットワークに大別することができる。インターネット上

*情報通信研究機構, NICT: National Institute of Information Communications Technology
†奈良先端科学技術大学院大学, NAIST: NAra Institute of Science and Technology
‡北陸先端科学技術大学院大学, JAIST: Japan Advanced Institute of Science and Technology

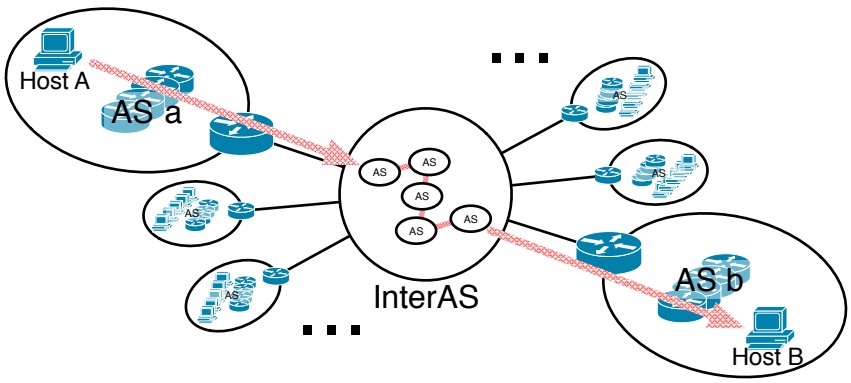

図 1: インターネットの構成

には、およそ 25,000 の活動中の AS がある（CAIDA の AS Ranking[2] による）。

ある AS（AS a）上のホスト（ホスト A）から別の AS（AS b）上のホスト（ホスト B）への通信は、ホスト A から AS a の AS 内ネットワークを通り、AS 間ネットワークを介して、AS b に到達し、AS b の AS 内ネットワークを経てホスト B に到達する。よって、インターネットは、図 1 のように模式化できる。

このように、対象となるホスト群とそれを含む AS 内ネットワークと、AS 間ネットワークを模倣することができれば、実験に必要なインターネットの主要な部分を模倣できると考えられる。

我々は、この模式化に従い、バイナリ実装の実証実験などに際し、インターネットの代わりとして利用できる模倣インターネット環境をテストベッド上に構築することを目指して、いくつかの取り組みを始めた。本稿では、模倣インターネット環境のうち、AS 間ネットワークの構築手法について述べ、テストベッド上への AS 間ネットワーク構築の試行とその評価、考察を行う。

2 関連研究

実験で BGP を用いるために、Zebra[3] bgpd を ns-2[4] や GTNetS[5] などのシミュレータで利用できるようにした BGP++[6] などの試みに見られるように、シミュレータで大規模な AS 間ネットワークを模倣する試みは既に行われている。

また、ns-2 をベースとしたエミュレーション環境 DETER などで、大規模な AS 間ネットワークを構築し、BGP に関連する技術 [7] の実験も行われている。

しかし、これらのシミュレータやエミュレータでは、本研究が目的とするようなバイナリ実装の検証や実パケットの観測を必要とするような実験を行うことはできない。

3 AS 間ネットワークの構築手法

ここでは、インターネットの主要な中間要素である AS 間ネットワークに着目し、テストベッド上での構築手法について述べる。

なお、インターネットの模倣を行うには、対象となるホストの模倣や AS 内ネットワークの模倣などについても検討を行う必要があるが、本稿では扱わない。

3.1 要求仕様

本研究では、バイナリ実装の実証実験などを行う際に、インターネットの代替として利用可能な環境の構築を目指している。そのため、シミュレーションによる実現ではなく、バイナリ実装が動作する環境を用意し、通信では実パケットがやりとりされるような環境を構築することを目標とする。また、実際のインターネットに則した環境とするために、模倣 AS 間ネットワークのトポロジは実際のインターネットを元に構成することとする。

インターネット上には、前述の通りおよそ 25,000 の活動中の AS があり、複雑な AS 間ネットワークが形成されている。AS 間の到達性情報は、EGP（Exterior Gateway Protocol）で交換され、すべての AS に必要な到達性情報が伝搬することで AS 間の通信

を可能とする。いくつかのEGPがあるが、ここでは単純化のため、代表的なEGPであるBGP4（Border Gateway Protocol Version 4）のみを取り扱うこととし、AS間ネットワークはBGP4によって到達性情報を交換し、それに従って経路制御される網とする。

よって、BGPスピーカとBGP4の到達性情報に基づき経路制御するルータ、ASと他のASをつなぐリンクが各ASの最低限の構成要素であり、これらを実際のインターネットのAS間ネットワークのトポロジに則して構成することで、AS間ネットワークを模倣することとする。また、各リンクは帯域や遅延、パケットロス率などの性能が異なる。これらについても、実際のインターネットの各リンクの性能に則して模倣することとする。

規模に関しては、対応できる規模が大きいほど良いと考えられるが、まずは現状のインターネットの規模を最大目標値に設定し、AS数では25,000程度、経路数では250,000程度[8]を目標とする。

3.2 前提環境

テストベッド上へのAS間ネットワークの構築手法を検討するにあたり、前提とするテストベッドについて確認しておく。

多くのホストやルータから成るインターネットを模倣するために、多数のノードから成るクラスタ環境であると想定する。ノード間は、何らかのネットワークリンクで接続され、必要とされるトポロジに合わせて柔軟にトポロジを変更できるものとする。

また、バイナリ実装の実証実験などを目的とするため、各ノードは、OSやアプリケーションソフトウェアのバイナリ実装が動作するPCのノードを動作環境と想定するが、バイナリ実装が動作するなら、物理ノードであるか仮想ノードであるかは問わないこととする。

我々が研究開発してきたStarBED[9]は、830台のPCから成るネットワーク実証実験用クラスタで、すべてのノードはEthernetもしくはATMによって接続されており、VLANやVC/VPによってネットワークトポロジを柔軟に変更可能であり、前提に則している。よって、本研究では、StarBEDを用いて模倣インターネットをテストベッドに構築する試みを行うこととした。

3.3 構築手法

要求を元に前提とする環境上に、模倣AS間ネットワークを自動的に構成する手法について述べる。

3.3.1 提案手法の概要

実際のインターネットのAS間ネットワークのトポロジに則して構成するために、CAIDAのAS Relationshipデータを元にして模倣AS間ネットワークは構成する。CAIDAのAS Relationshipデータは、AS番号の組とAS間の関係（customer, peer, provider, sibling）を示す数値からなるAS間の関係性のリストである。具体的には、CAIDAのAS Relationshipデータを我々が研究開発してきたAnybed[10]のトポロジ記述に変換し、そのトポロジ記述に基づいてAS間ネットワークの物理・論理トポロジを構成した。

本来ASは、AS内ネットワークと複数のAS境界ルータから成っているが、ここでは単純化のために、各ASを単一のノードで実現することとし、ノード上ではBGPスピーカとBGPによるルーティングデーモンを動作させる。隣接AS数に応じてノードには仮想ネットワークインタフェイスを用意し、すべてのASは隣接ASと直接接続される。BGPスピーカとルーティングデーモンとしては、Quagga[11]の`bgpd`, `ospfd`, `zebra`を用いた。[1]

各BGPスピーカの設定は、CAIDAのAS Relationshipデータを元にしたAnybedのトポロジ記述から生成し、AS間の関係に基づいて設定するリンクの方向に従って経路フィルタを行うようにした。

ネットワークリンクは、当初各AS間リンクにEthernetのIEEE802.1Qタグ付きVLANを一つ割り当てることを想定した。しかし、タグの数が最大でも4096で不足だったため、すべてのAS間リンクを一つのVLAN上で構成し、IPアドレスレベルでネットワークの分割を行うこととした。各ネットワークリンクの性能は、ネットワークエミュレータ`netem`[12]でエミュレートする。しかし、現状では各ネットワークリンクの性能を導出する手法が定まっていないため、今回の試行では、ネットワークリンクの性能については変更していない。

各ASを単一のノードで実現したとしても、すべてのASを物理ノードで構築した場合、25,000程度の物理ノードが必要となり、現実的ではない。しか

[1] 今回の実験では、1AS1ノード構成とし、`ospfd`は利用していないが、1AS多ノード構成も可能となっており、その場合には`ospfd`も利用する。

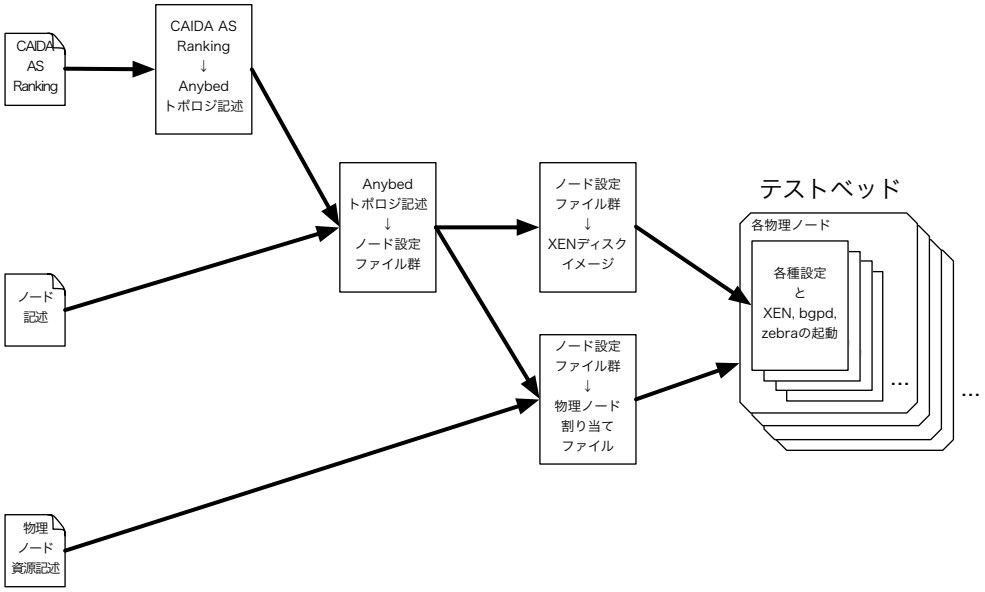

図 2: 提案方式の概要

し、シミュレーションではバイナリ実装に関する検証環境とはなりえない。そこで、仮想化技術を利用し、ハードウェア仮想化や OS 仮想化による仮想ノードを用いることとした。具体的には、XEN[13] によって Debian Linux を準仮想化し用いた。

物理ノードの OS としては、OS パッケージとして他種類の仮想化技術が導入済みで、Debian Linux をベースとした平易なパッケージである VMKnoppix[14] を採用した。VMKnoppix をネットワーク起動に対応するよう改良し、物理ノードはディスクレスでネットワーク起動する。

各物理ノードと各仮想ノードには、AS 間リンクとは別に、実験管理用のネットワークインタフェイスを用意し、それを介して各種の操作を行う。

3.3.2 構築の手順

前述の概要に従い、以下に提案手法による構築の手順について述べる。

0. 入力として、仮想ノードのノード設定の記述（以降、ノード記述）と物理ノードの資源記述、CAIDA AS Relationship のデータを与える
1. CAIDA の AS Relationship データから指定した AS 数の分を Anybed のトポロジ記述に変換
2. トポロジ記述とノード記述から各ノードの設定ファイル群（`bgpd.conf`, `zebra.conf` など）を生成
3. 設定ファイルを各仮想ノード用の XEN ディスクイメージに挿入
4. 各仮想ノードをノードの設定と物理ノードに資源状況に合わせて割り当て、割り当てファイルを作成
5. 割り当てファイルに従い、物理ノードを起動し、XEN の設定・起動、物理ネットワークの設定を行う
6. 各物理ノード上で割り当てられた仮想ノードを起動する
7. 各仮想ノード起動時に仮想ネットワークインタフェイスの設定と BGP スピーカおよびルーティングデーモンを起動する

上記の手順で、入力されたノード記述と CAIDA AS Relationship のデータから、自動的に模倣 AS 間ネットワークが構成される。提案方式の概略を図 2 に示す。

表 1: 物理ノード グループ F の構成

項目	構成
CPU	Pentium4
Memory	2GB
NIC	1000Base-T×6

表 2: 計測結果

環境	ping 平均 (ms)	iperf (Mbps)
物理ノード直結	0.122	961
物理ノード経由	0.936	961
模倣 AS 間ネットワーク経由	1.489	160

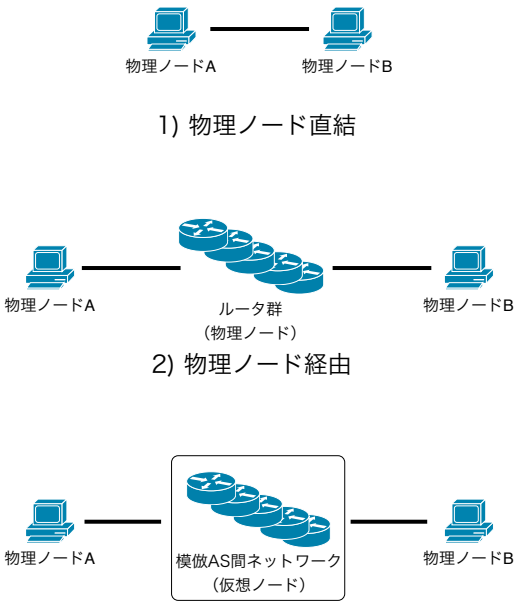

図 3: 評価環境

4 試行と評価

提案方式を用いて、実際に StarBED 上への模倣 AS 間ネットワークの構築について、いくつかの試行を行った。各試行には下記のルールで名称を付けてある。以降では、これに従って試行を区別する。

模倣対象. 試行範囲. 元にしたトポロジ

ここでは、

- 簡単な性能評価を行った “InterAS.Top250.CAIDA070430”
- 日本のインターネットの模倣を目指した “InterAS.JPNIC0707.CAIDA070430”
- 現時点での最大 AS 模倣数を目指した “InterAS.Top5000.CAIDA070430”

について述べる。今回のすべての試行は 2007 年 4 月 30 日時点の CAIDA AS Relationship のデータに基づいてトポロジを作成している。

また、すべての試行は、StarBED のグループ F を用いて行った。グループ F のノード構成については表 1 に示す。グループ F は 168 台あるが、今回はそのうち必要な台数のみを用いた。

4.1 上位 250AS — InterAS.Top250.CAIDA070430

CAIDA の AS Relationship の上位 250AS を単純に抜き出し、AS 間ネットワークの模倣を試みた。5 物理ノードにそれぞれ 50AS ずつを割り当て、模倣した。この模倣 AS 間ネットワーク環境を用いて、簡単な性能評価を行った。評価としては、遅延（RTT）と帯域について、物理ノードの性能との差異を計測した。

評価は、物理ノードと物理ノードを直結した場合（「物理ノード直結」）、物理ノード上で直接ルーティングデーモンを動作させて 5 ホップの経路を経由した場合（「物理ノード経由」）、模倣 AS 間ネットワークで 5 ホップの経路を経由した場合（「模倣 AS 間ネットワーク経由」）の 3 種について（図 3 参照）、`ping` コマンドと `iperf`[15] で計測した。計測結果を表 2 に示す。

物理ノード直結の場合の RTT は 0.1 程度（100 回試行平均値）で、物理ノード経由では 1.0 程度、模倣 AS 間ネットワークを経由した場合は 1.5 程度と、差異は認められるが大きな差異はない。帯域は、物理ノード直結の場合は 960Mbps 程度、物理ノード経由の場合も同様となったが、模倣 AS 間ネットワークを経由した場合は 160Mbps 程度であり、6 倍程度の差はあるが、100Mbps 程度のネットワークを模倣する実験には実用上十分といえる。

4.2 JPNIC — InterAS.JPNIC0707.CAIDA070430

CAIDA AS Relationship のデータから 2007 年 7 月時点の日本ネットワークインフォメーションセンター（JPNIC）の AS 番号登録情報に記載のある 394 の AS を抜き出し、日本の AS 間ネットワークの模倣を試みた。仮想ノードへのメモリ割り当ては前述の通りで、実行した結果、5 物理ノードで 394AS を模倣することができた。

しかし、経路を評価した結果、複数の AS で経路の断絶が観測された。これは、単純に 394AS を抜き出したため、394AS に属さない AS を経由してしか BGP 経路情報を受け取れない AS が多くあったためと思われる。

4.3 上位 5000AS — InterAS.Top5000.CAIDA070430

CAIDA の AS Relationship の上位 5000AS を単純に抜き出し、AS 間ネットワークの模倣を試みた。各仮想ノードのメモリは、隣接 AS 数 25 個あたり 24MB を割り当てることとし、実行した結果、100 物理ノードで 5000AS を模倣することができた。

ただし、50-100 仮想ノードが正常に起動せず、再起動を必要とした。この現象は、1000AS 以上の模倣から散見されるようになり、仮想ノード数の増加に従って増えている。

また、模倣 AS 間ネットワークでは、ほぼ同時期にすべての BGP スピーカが起動するが、BGP では多数の AS が同時に立ち上がる状況を想定していないためか、BGP 情報が大量に増減を繰り返す現象がしばらく続いた。

仮想ノードへのメモリ割当量は経験値で、`bgpd` のメモリ利用量が隣接 AS 数に応じて増加することと、およそ 25 隣接 AS あたり 24MB 程度を割り当てた場合には、メモリ不足のエラーを生じなくなることから、経験的に割り出した。この値は、上位 5000AS 程度までの実験では普遍的に利用できるが、これより大きな AS 数を模倣する場合にも適用できるのかについては、検証が必要であり、今後より多数の物理ノードを用いた実験を計画している。

5 課題と考察

AS 間ネットワークの模倣について試行した結果、得られた知見にもとづき、提案方式の課題とインターネットを模倣する試みに関して考察を加える。

5.1 提案方式の課題

現在の提案方式は、すべての AS 間リンクを一つの VLAN 上に展開し、IP エリアスで L3 レベルでの分離が図られているのみである。そのため、いくつかの問題を生じる。一つは、BGP だけではなく、OSPF などネットワークセグメント上でのマルチキャストやブロードキャストを用いるルーティングデーモンを導入し、境界ルータ間連携や AS 内ネットワークを模倣する場合には、IP エリアスでは混信し対応できない。また、`netem` は IP エリアスの場合、各リンクの性能を分離することができないため、AS 間リンクごとにリンク性能をエミュレーションすることができない。トポロジ中の必要な部分を抜き出し、一部分のみ VLAN で分離するなどの対策が必要となる。

物理ノードへの仮想ノードの割り当ては、経験値に基づいており、どのような場合でも確実かつ効率良く物理ノードに仮想ノードを配置できるとはいえない。また、現状では単純に空いている物理ノードに順に仮想ノードを配置しているが、資源の効率的な分散などを考えた場合、偏りがあるといえる。さらに、現状ではメモリのみを資源として扱っているが、プロセッサの負荷なども考慮に入れる必要があると考えられる。よって、割り当てようとする仮想ノードの引き起こしうる各資源の消費量の見積を正確にすることと、それに合わせて、資源の残量の大きさをもとに割り当て対象を決めるなど、より高度な割り当ての仕組みが必要と考えられる。

5.2 模倣の厳密さ

PC 上のソフトウェア実装がそのまま動作するといっても、仮想化された環境では時間遷移や環境の違いから動作が本来と異なることがないとはいえない。より厳密な模倣が求められる場合には、仮想ノードではなく物理ノードを利用するようにするなどの工夫が必要である。

また、ルータに関しては、現在の実装では仮想ノード上のルーティングソフトウェアで実現されており、実際のルータ機器との間では動作に違いがあると思

われる。CISCO 7200 Simulator[16] などのようにより厳密なルータ機器のエミュレーションを試みる方法や Schooner[17] のように実際に利用されているルータ機器そのものを用いる方法などが必要となる。

さらには、各ネットワークリンクの性能の模倣についてはネットワークエミュレーションソフトウェア（netem）によるエミュレーションであり、その模倣能力に大きく依存しているといえる。ネットワークエミュレータの模倣能力に関しては検証が必要であるが、模倣能力が不十分な場合には、実際のネットワークを環境中に挿入するなど、何らかの手法で模倣能力を補完する必要がある。

なお、模倣インターネットの模倣の厳密さがどの程度であるのかは、何らかの手法で検証する必要がある。物理ノードで構成した環境と仮想ノードによる環境との差異から推定するなどが考えられるが、環境の同一性の担保や計測結果の正確さの保証など、正確な推定には困難が予想される。

5.3 模倣の規模

模倣インターネットの究極の目標はインターネットのすべてのホスト、すべてのルータなどの接続機器、すべての回線を模倣することである。よって、必然的に模倣の規模は巨大になる。しかし、模倣の規模の拡大に従っていくつかの問題が顕在化すると考えられる。

インターネット上のホスト数と同じホストを用意するのは現実的ではないため、仮想化技術など何らかの多重化手段を用いて、少ない物理ノードで多くのホストを模倣する必要がある。この場合、同じ物理ノード数であれば、規模を大きくすればするほど、多重化に用いる抽象度を上げねばならない。よって、大きな規模に対応しようとすればするほど、模倣の厳密さを下げねばならないというトレードオフが生まれることになる。同様に、多重化すればするほど、性能は分割されるため、規模に合わせて性能は低下せざるを得ない。

さらに、今回の試行では、1000AS を超えた規模から正常に起動しない仮想ノードが散見されており、仮想化技術の耐規模性や安定性などの問題が、規模の拡大によって顕在化するというリスクも孕んでいると考えられる。

5.4 規模と管理・監視

実験においては、実験環境を必要な条件に合わせて管理し、実験を行う必要がある。しかし、模倣インターネットの規模が巨大になり、インターネットと同じになった場合には、インターネットと同じ規模の分散システムを管理することとなる。また、実際に構成した模倣インターネットがどうなっているのかは、規模が大きくなった場合にはインターネットと同じ種類の計測手法で、観測的に確認せねばならない。

提案方式では、これらの問題を避けるために実験管理用のネットワークですべての物理・仮想ノードは管理・監視している。よって、管理用のネットワークからノードの各種情報の取得や操作が可能となっている。しかし、実際に AS 間ネットワークが正しく構成されているかなどを確認する際には、この種の問題が発生している。模倣インターネットを考える上で、規模と管理・監視の問題は一つの大きな課題である。

6 おわりに

本稿では、バイナリ実装の実践的検証に利用できる、テストベッド上への模倣インターネットの構築について、我々が StarBED 上で行った AS 間ネットワーク構築の試行とその結果を述べた。提案手法では、100 物理ノードで 5000AS から成る模倣 AS 間ネットワークを構築でき、160Mbps 程度の帯域を有することを確認した。

提案手法は、まだ試行途上であり、今後もさらに大規模な実験によって、改良を行う予定である。また、今後は AS 内ネットワークの模倣も行い、それらを組み合わせて最終的にインターネットを模倣しきることを目指していくつもりである。

参考文献

[1] Internet Systems Consortium, Inc., “ISC Domain Survey: Number of Internet Hosts”, http://www.isc.org/index.pl?/ops/ds/host-count-history.php.

[2] Cooperative Association for Internet Data Analysis (CAIDA), “AS ranking”, http://as-rank.caida.org/.

[3] IP Infusion Inc., “GNU Zebra – routing software”, http://www.zebra.org/.

[4] “The Network Simulator - ns-2”, http://www.isi.edu/nsnam/ns/index.html.

[5] G. F. Riley, “Using the Georgia Tech Network Simulator”, http://www.ece.gatech.edu/research/labs/MANIACS/GTNetS/.

[6] X. A. Dimitropoulos, G. Riley, “Efficient Large-Scale BGP Simulations”, Elsevier Science Publishers, Elsevier Computer Networks, Special Issue on Network Modeling and Simulation, vol. 50, num. 12, 2006.

[7] G. Carl, G. Kesidis, S. Phoha, B. Madan, “Preliminary BGP Multiple-Origin Autonomous Systems (MOAS) Experiments on the DETER Testbed”, *In proceedings of DETER community workshop*, Jun. 2006.

[8] G. Huston, “BGP Routing Table Analysis Reports”, http://bgp.potaroo.net/.

[9] T. Miyachi, K. Chinen, Y. Shinoda, “StarBED and SpringOS: Large-scale General Purpose Network Testbed and Supporting Software”, *In proceedings of International Conference on Performance Evaluation Methodlogies and Tools (Valuetools) 2006*, ACM Press, ISBN 1-59593-504-5, Oct. 2006.

[10] M. Suzuki, H. Hazeyama, Y. Kadobayashi, “Expediting experiments across testbeds with AnyBed: a testbed-independent topology configuration tool”, *In proceedings of Second International Conference on Testbeds and Research Infrastructures for the Development of Networks & Communities* (TridentCom2006), Mar. 2006.

[11] “Quagga Software Routing Suite”, http://www.quagga.net/.

[12] S. Hemminger, “Network Emulation with NetEm”, linux.conf.au 2005, Apr. 2005.

[13] S. Crosby, D. E. Williams, J. Garcia, “Virtualization With Xen: Including XenEnterprise, XenServer, and XenExpress”, Syngress Media Inc.,ISBN 1-597-49167-5, 2007.

[14] 産業技術総合研究所, “VMKNOPPIX: Collection of Virtual Machine”, http://unit.aist.go.jp/itri/knoppix/vmknoppix/.

[15] A. Tirumala, F. Qin, J. Dugan, J. Ferguson, K. Gibbs, “NLANR/DAST : Iperf - The TCP/UDP Bandwidth Measurement Tool”, http://dast.nlanr.net/Projects/Iperf/.

[16] C. Fillot, “Cisco 7200 Simulator”, http://www.ipflow.utc.fr/index.php/Cisco_7200_Simulator.

[17] P. Barford, L. Landweber, “Bench-style Network Research in an Internet Instance Laboratory”, *In Proceedings of SPIE ITCom*, Aug. 2002.