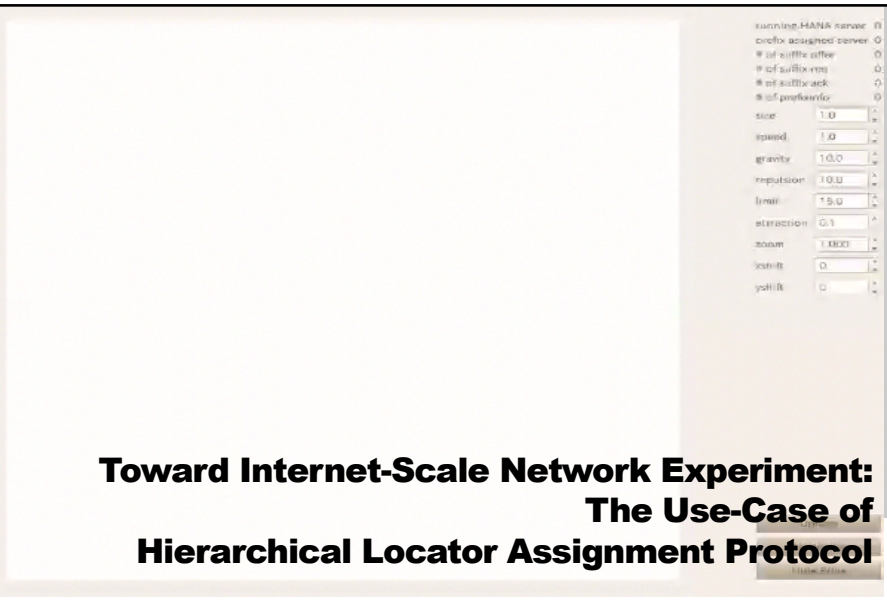


Toward Internet-Scale Network Experiment: The Use-Case of Hierarchical Locator Assignment Protocol

Hajime Tazaki, Kenji Fujikawa, Tomoji Tomuro, Hiroaki Harai
(NICT)

IEICE TC-NV July, 2012

Copyright©2012 National Institute of Information and Communications Technology. All Rights Reserved.

Summary

Future Internet を実現するアーキテクチャの大規模評価

- 現存インターネット上で新アーキテクチャが動作するか?

StarBED³をつかった大規模ネットワークエミュレーション

- 例題:HANA でインターネットスケール実験

This is not NV arch proposal nor NV use case, but testbed usage as an NV application

2

Copyright©2012 National Institute of Information and Communications Technology. All Rights Reserved.

ロケータとIDに関する研究

2つの新世代ネットワークアーキテクチャ研究

- 階層的ロケータ (IPアドレス)割当機構
- ロケータ・ID分離機構

現存インターネットの問題解決を指向

- IP アドレス = 位置識別子と通信端点識別子がオーバーロード
- 様々な問題
  - 移動透過性なし
  - FIBサイズを抑えたマルチホーム手法の欠如

2つの提案

- Hierarchical Automatic Number Allocation (HANA) [1]
- Heterogeneity Inclusion and Mobility Adaptation through Locator ID Separation (HIMALIS) [2]

[1] Fujikawa at al. "The Basic Procedures of Hierarchical Automatic Locator Number Allocation Protocol HANA,", AWFIT 2011
[2] Kafle et al., "HIMALIS: Heterogeneity Inclusion and Mobility Adaptation through Locator ID Separation in New Generation Network", IEICE .E93-B , 2010
Copyright©2012 National Institute of Information and Communications Technology. All Rights Reserved.

3

Motivation

本当にそんな仕組みが動くのか?

- Network Emulation Testbed (NET) の活用したProof Of Concept
- 仮説:
  - 現存インターネットの構造で，階層化アドレス割当は効果的である
  - 規模:現存AS(Autonomous System)でアクティブな約36,000ASのAS間トポロジを利用
- 実証:
  - メッセージ交換が完了するか?
  - トポロジ状態(リンク増減など)の変動に，耐えうるか?

4

Copyright©2012 National Institute of Information and Communications Technology. All Rights Reserved.

Network Experiments

シミュレーション/エミュレーション/実環境

- 得意分野の違い

シミュレーション

- アルゴリズム検証などは得意
- 結果の信憑性についての議論あり[4]

エミュレーションベーステストベッド (Planetlab/CoreLab)

- Proof of Concept 型
- 規模が限定される (PlanetLab: 1K order)

実環境

- 信憑性は問題ないが，様々なパラメタ変更は困難

	シミュレーション	エミュレーション	実環境
実時間性	NO	YES	YES
実験制御	HIGH	HIGH	LOW
状態範囲	LARGE	LARGE	SMALL
現実性	LOW	MEDIUM	HIGH
実験コスト	LOW	MEDIUM	HIGH
容易性	HIGH	MEDIUM	LOW

ネットワーク実験の手法 [3]

[3] R. Beuran, et al. QOMB: A wireless network emulation testbed. IEEE GlobeCom 2009
[4] S. Kurkowski, et al., "MANET simulation studies: the incredibles. ", ACM MC2R 2005

Copyright©2012 National Institute of Information and Communications Technology. All Rights Reserved.

5

Network Experiments (cont'd)

	シミュレーション	エミュレーション	実環境
実時間性	NO	YES	YES
実験制御	HIGH	HIGH	LOW
状態範囲	LARGE	LARGE	SMALL
現実性	LOW	MEDIUM	HIGH
実験コスト	LOW	MEDIUM	HIGH
容易性	HIGH	MEDIUM	LOW

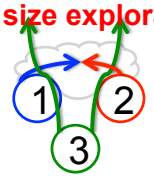
StarBED[3]を利用した大規模エミュレーションによる検証

Copyright©2012 National Institute of Information and Communications Technology. All Rights Reserved.

6

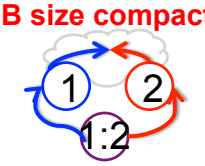
What is HANA?

FIB size exploration



Conventional Style of Multi-homing (PI address advertisement)

FIB size compaction



HANA Style of Multi-homing (PA address advertisement)

HANA (Hierarchical Automatic locator Number Allocation)

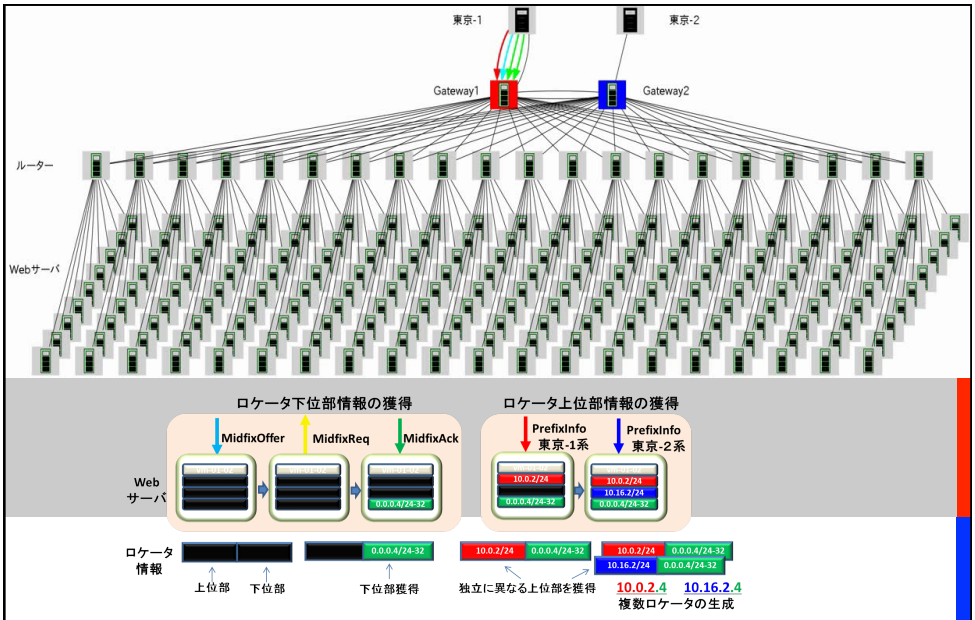
- グローバルネットワークのアドレス (Locator) 割当プロトコル
- 階層化シグナリングにて
  - PA (Provider Aggregated)アドレスの利用を促進
  - FIB Size Exploration を解決
  - マルチプレフィクス割当

検証

- シミュレーション有
- CoreLab (2011 IEEE ICC demo)
- 大規模環境での検証なし

7

Copyright©2012 National Institute of Information and Communications Technology. All Rights Reserved.



Visualization at INTEROP 2012@幕張

8

Copyright©2012 National Institute of Information and Communications Technology. All Rights Reserved.

関連研究

Miniature Internet

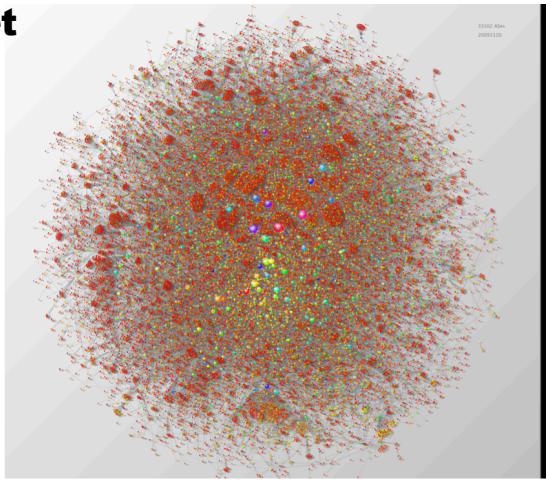
SIGCOM VISA 2009 [5]

10K ASes の BGP ネットワークをStarBED上に構築

XENによるVM多重

CAIDAによるAS Relationship dataset の再現

専用Viewerによる模擬状態の表示



[5] Miwa et al., "Experiences in Emulating 10K AS Topology with Massive VM Multiplexing", SIGCOMM VISA 2009

9

Copyright©2012 National Institute of Information and Communications Technology. All Rights Reserved.

インターネットクラスの大規模ネットワーク実験

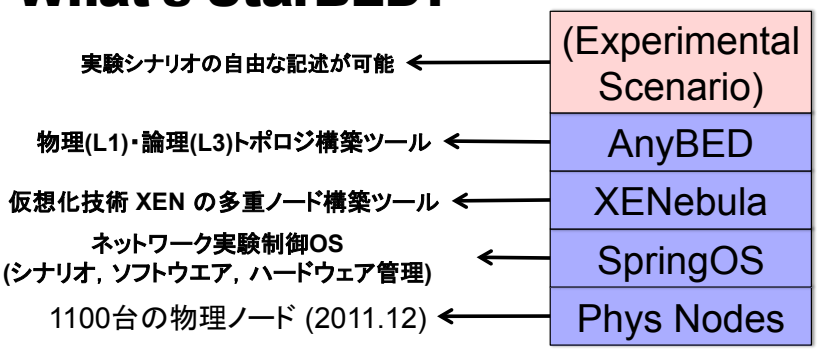
-- HANA

More discussion is available:
Fujikawa, et al., "**Inter-AS Locator Allocation of Hierarchical Automatic Number Allocation in a 10,000-AS Network**, " SAINT, July 2012

10

Copyright©2012 National Institute of Information and Communications Technology. All Rights Reserved.

What's StarBED?



既存ツールの利用

- SpringOS, AnyBED,CAIDA topology dataset

HANA実験用にカスタマイズ

- LXCNebula, AnyBED (HANA configuration generator)

参考：http://starbed.nict.go.jp/

Copyright©2012 National Institute of Information and Communications Technology. All Rights Reserved.

11

実験概要

現存Internet Inter-AS BGPネットワーク(トポロジ)において，階層化ロケータ割当がどういう挙動をみせるか?

- メッセージ交換がどのくらいの時間で完了するか?
- トポロジ状態(リンク増減など)の変動に，耐えうるか?

実験トポロジ

- CAIDA AS-Relationship dataset，2011/1/16 データ [6]
- AS間のトポロジを，BGP RouteView より抽出

HANA実装

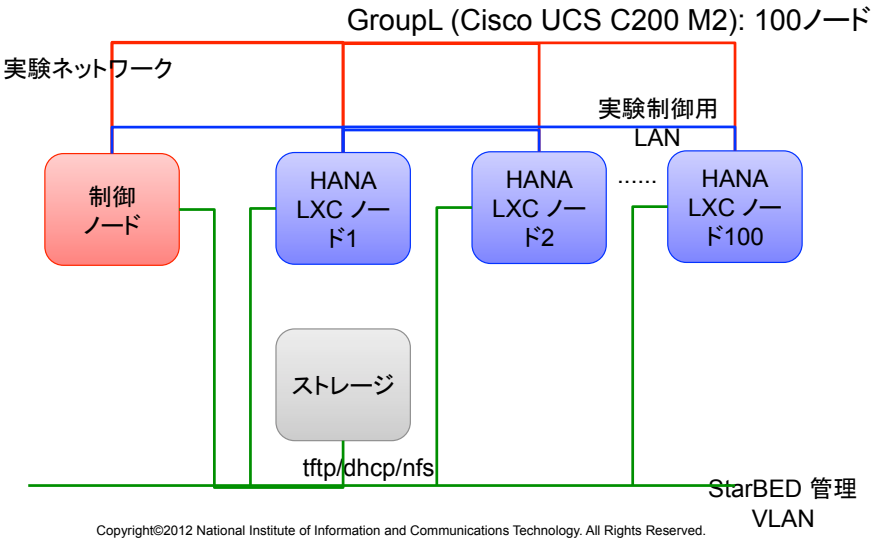
- Socket ベースユーザランドアプリケーション
- 各LXCコンテナ上に1HANAプロセス動作

[6] The CAIDA AS Relationships Dataset, <2011-01-16>. http://www.caida.org/data/active/as-relationships/

Copyright©2012 National Institute of Information and Communications Technology. All Rights Reserved.

12

Layer-1/Layer-2 Configuration



Copyright©2012 National Institute of Information and Communications Technology. All Rights Reserved.

仮想ノードConfiguration

Linux Container (LXC)によるノード仮想化

- Group Lノード100台 => 10000ノード

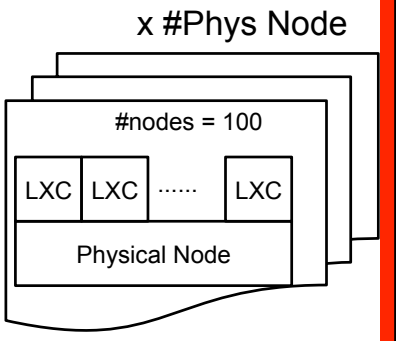
pickup/wipeoutによるノード複製 (SpringOS)

AnyBED によるIPアドレス，仮想インターフェース設定

aufs (another union file system)による各仮想ノード差分管理

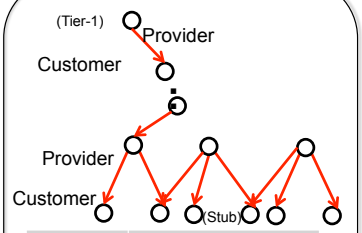
WSノードによる一括起動停止スクリプト (make-world-lxc.sh)

- Physノードにて，pcap保存



Copyright©2012 National Institute of Information and Communications Technology. All Rights Reserved.

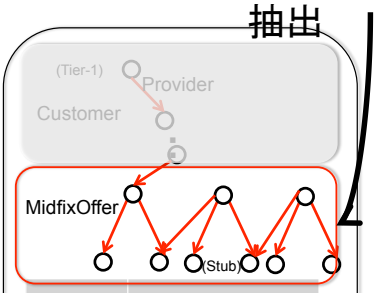
実験トポロジ



	#AS
#tier-1	33
#stub	30458
Others	5637
(total)	36128

CAIDA AS-relationship 2011.1.16

2階層トポロジのみ抽出



#links	
#stub	10000 (max)

HANA実験トポロジ

15

Copyright©2012 National Institute of Information and Communications Technology. All Rights Reserved.

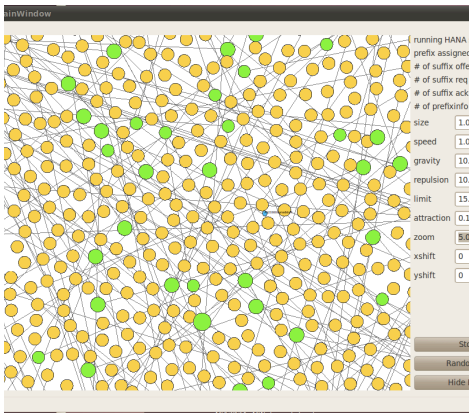
可視化(HANA Viewer)

Internet Viewer[7]をカスタマイズ

ロケータ割当状態の可視化

論理HANAリンク単位

トポロジ全体の接続関係の可視化



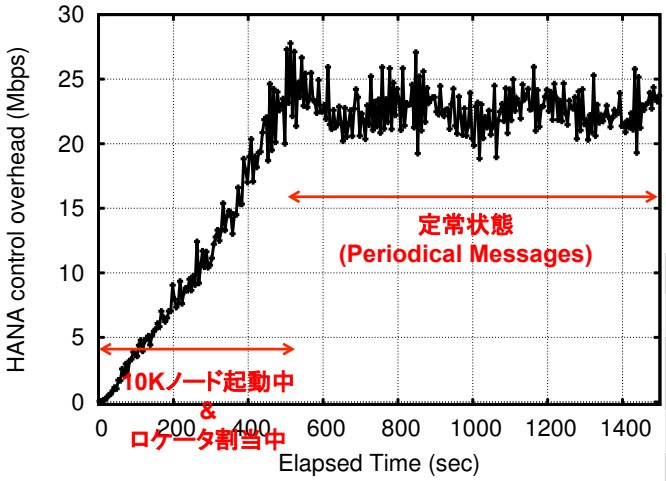
ロケータ割り当て状況の可視化

[7] NAIST IPLab, Internet Viewer, http://intertrack.naist.jp/internet-viewer-0.0.1.tar.gz

16

Copyright©2012 National Institute of Information and Communications Technology. All Rights Reserved.

Result (制御メッセージ量)



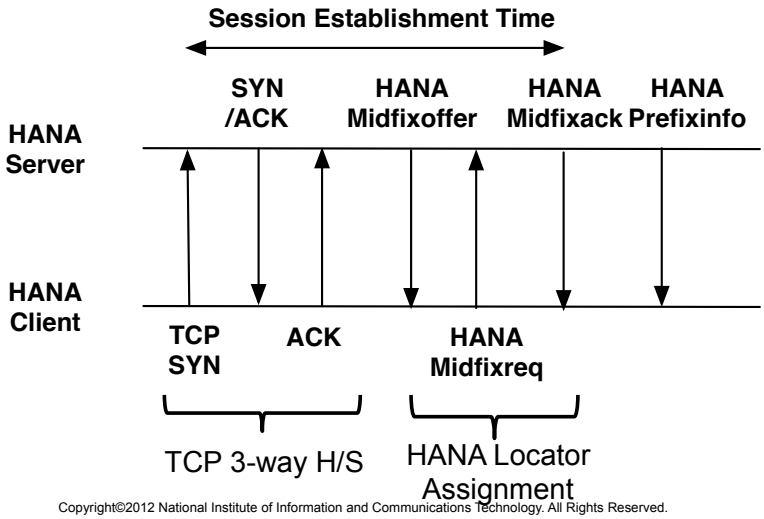
内部メッセージ数検証

#Links	17359
PrefixInfo	17299
MidfixOffer	17299
MidfixReq	17299
MidfixAck	17299

Copyright©2012 National Institute of Information and Communications Technology. All Rights Reserved.

17

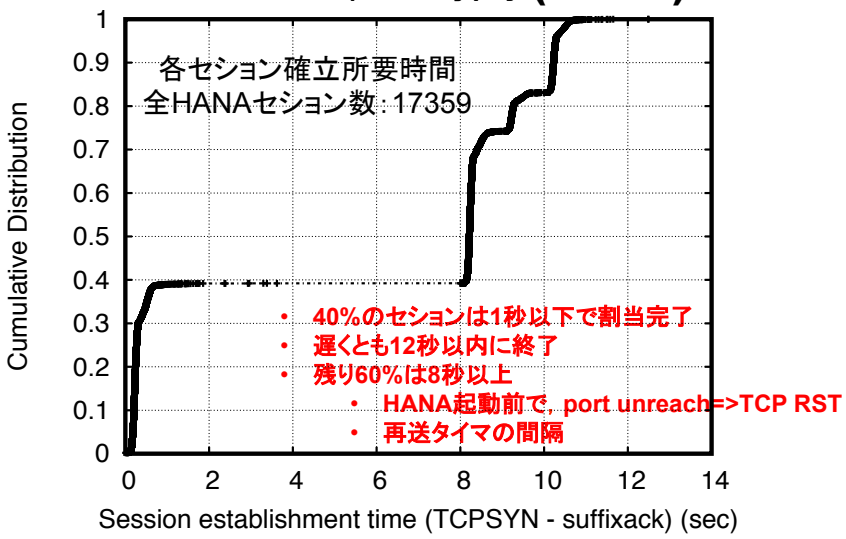
HANAセション確立時間 (Session Establishment Time)



Copyright©2012 National Institute of Information and Communications Technology. All Rights Reserved.

18

HANAセション確立時間 (cont'd)



- 40%のセションは1秒以下で割当完了
- 遅くとも12秒以内に終了
- 残り60%は8秒以上
  - HANA起動前で，port unreach=>TCP RST
  - 再送タイマの間隔

Copyright©2012 National Institute of Information and Communications Technology. All Rights Reserved.

19

Lessons learned

この実験固有の細かな実験テクニック

- 起動タイミングの制御
  - プロセス単位･コンテナ単位で一部のみ欠ける事がある

大規模(インターネット規模)実験のテクニック

- 解析を単純化するための計測情報
  - 36Kノード分の情報の，集約の仕方が実験解析の肝
  - 必要最小限の情報のみの出力
- 実験フローの定型化
  - 安定した実験･再現性のある実験環境の構築
  - 試行錯誤をするためのフローが長い
    - e.g., スクリプト修正=>検証=>失敗･チューニング などが半日

Try & Catch のサイクルを円滑化する事が重要

Copyright©2012 National Institute of Information and Communications Technology. All Rights Reserved.

20

SUMMARY

実験よりの観測

- 10000 AS にて，限定されたトポロジ(2階層)において
  - プロトコル初動時間の検証
  - 安定状態遷移までの時間・インパクトを観測実験よりの観測
- 10000ノード起動・TCP 3w HS完了まで，600秒
- ロケータ割当完了までの時間：最大12秒

これまでの研究室内検証では抽象化されたシミュレーションでしか実現不可能であった検証が，変更不要な実装物バイナリを用いて実施できた

今後の予定

- 現インターネットに観測されるアクティブAS数(36861)にて検証
- トポロジの複雑度を現存のものと同一に(階層限定なし)

21

Copyright©2012 National Institute of Information and Communications Technology. All Rights Reserved.

Thank you

Acknowledgement
NAIST櫨山 寛章氏
NICT 鈴木未央氏，宮地利幸氏，三輪信介氏

22

Copyright©2012 National Institute of Information and Communications Technology. All Rights Reserved.