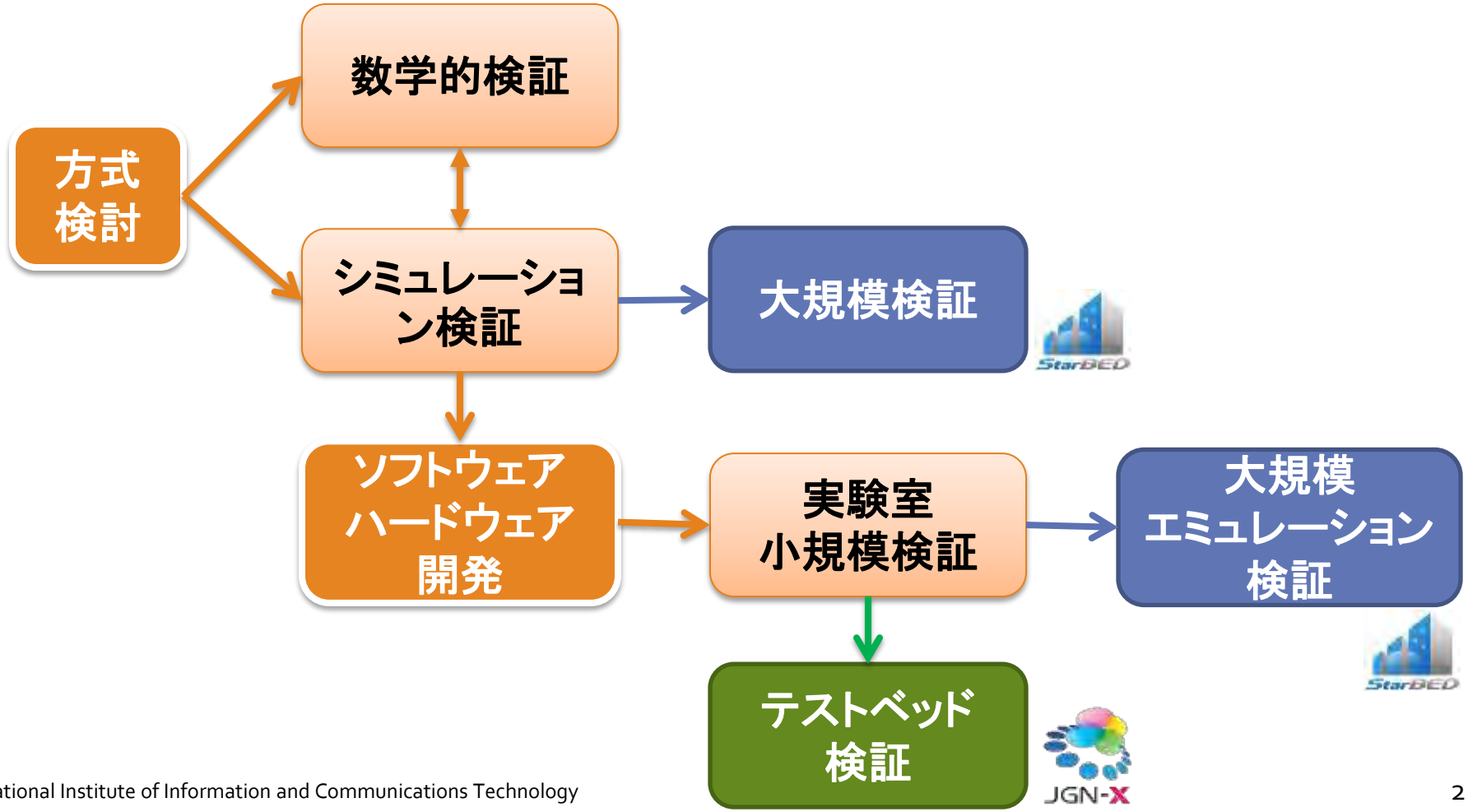
StarBEDを用いた大規模ネットワークにおける分散システム検証

原井 洋明 (harai@nict.go.jp)
情報通信研究機構
光ネットワーク研究所 ネットワークアーキテクチャ研究室長
実験実施者：藤川賢治 、戸室知二、田崎創、大西真晶、森岡和行、福島裕介

2015年10月20日
新世代ネットワーク推進フォーラム
第22回テストベッドネットワーク推進WG

StarBEDといえば、、、。

- プロジェクトでは纏まった数のサーバ調達が困難
- 1000台を超える実機を用いて検証ができる
- 発表者の研究室スタッフの充実に伴ない、使わせて戴くことにした



利用したStarBEDサーバ

グループ名	PCサーバの型式	設置数	導入年	利用数		
F	NEC Express5800 110Rg-1	168	2006	HA	R	HI
G	Proside Amazelast neo920	150	2007			
H	HP ProLiant DL320 G5p	240	2009	80		
I	Cisco UCS C200 M2	192	2011		60	
J	Cisco UCS C200 M2	82*	2011	10		
K	Cisco UCS C200 M2	144	2011	10		20
L	Cisco UCS C200 M2	120	2011	100	60	100
M	Cisco UCS C200 M2	14	2011			
N	DELL PowerEdge C6220	224	2013	10		
O	SeaMicro SM15000-XE	64	2013			

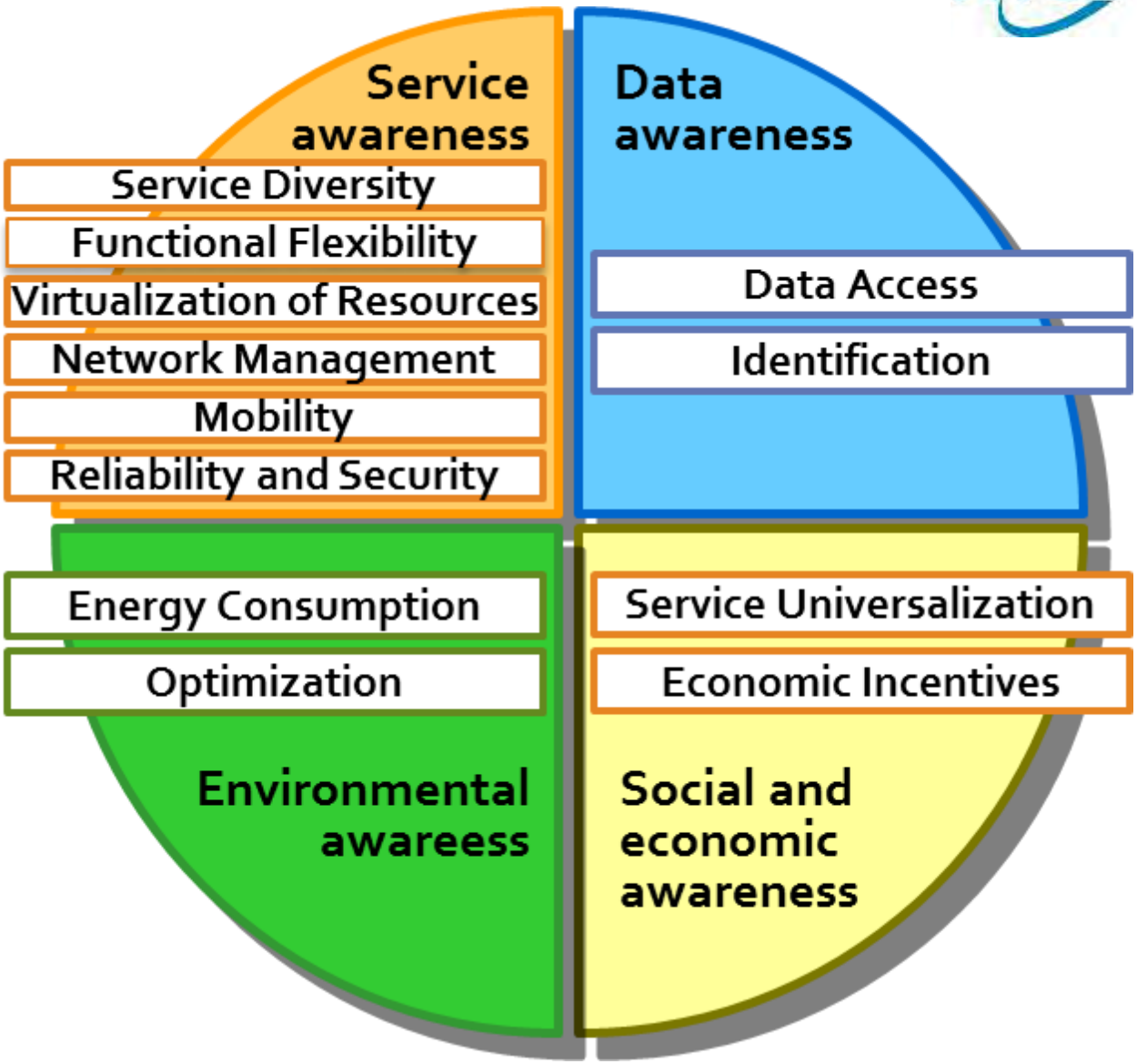
講演の内容

- 研究の方向性と内容
- StarBED^{3}をさまざまな分散システムの検証に
  - 階層型自動番号割当 HANA
    - インターネット規模でサーバ協調によりアドレス空間を配布
  - 局所情報を用いた分散経路制御
    - リンク・ノードのjoin/leaveが時々刻々と変動する環境で全域情報の把握に頼らず到達性を求める
  - ID・ロケータ分離機構 HIMALIS
    - 故障時の経路切替検証

何度もStarBED を利用させていただき、
改めて、この場をお借りして御礼申し上げます

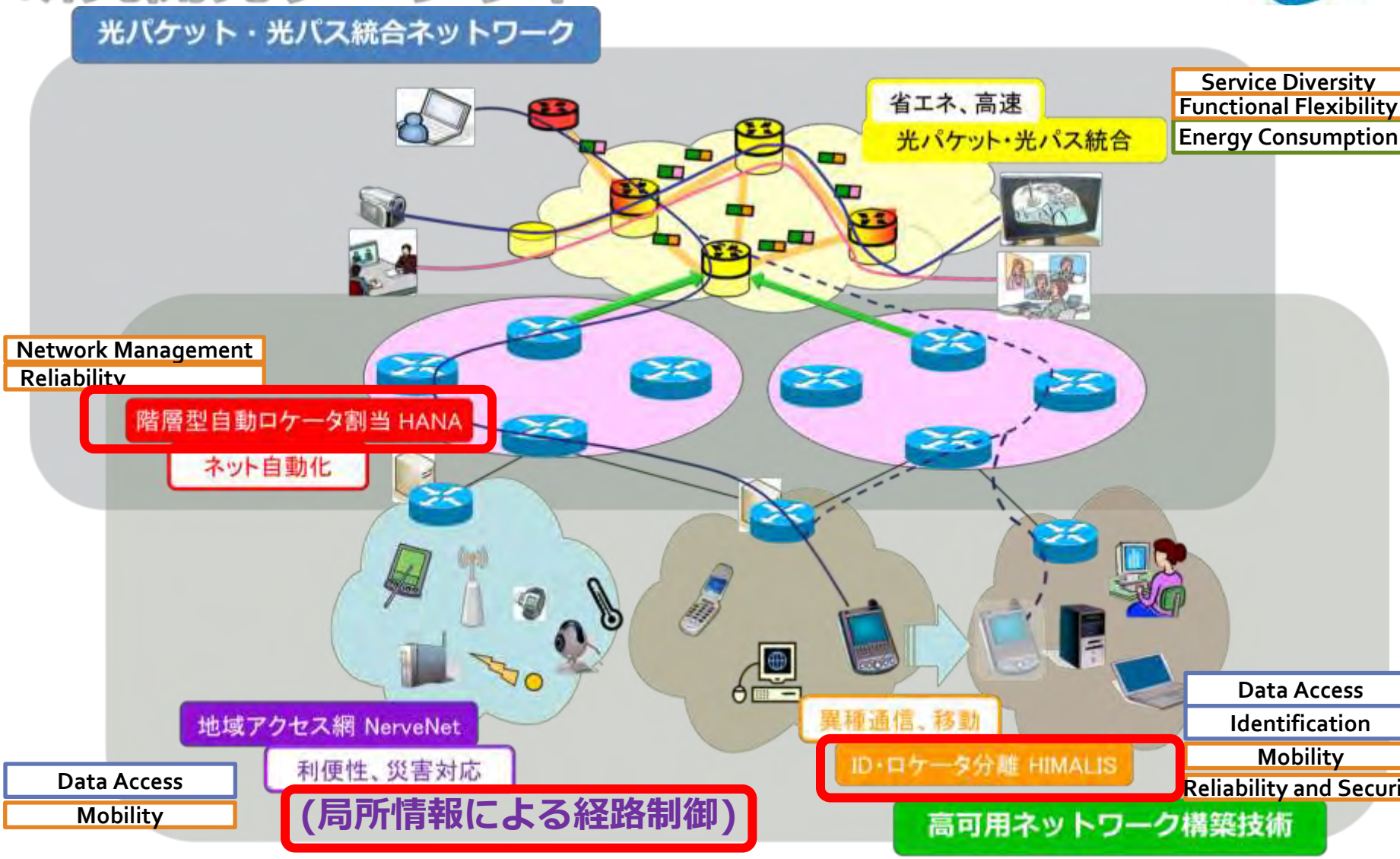
将来ネットワークITU-T勧告Y.3001

- 4つの目的
- 12個の設計目標



Rec ITU-T Y.3001, Future Networks: Objectives and Design Goals

ネットワークアーキテクチャ研究室
研究開発ターゲット



2020年の未来社会を支えるネットワークをつくります

ネットワーク設定が煩わしい．．．

ネットワーク構築がメンドーでしょ？

- 保全対策
- 安全対策
- 冗長化
- 接続ポートタグ付け
- IP アドレス割当
- VLAN割当
- ネットワーク更改
- ネットワーク増設

大規模ネットワーク構築で設定ミスしたらたいへんですね･･･

確認、確認、指差し確認

収容SW	ポート番号	サーバホスト名
HRKc302pf400100	0/1	HRKc302kssv0100
	0/2	HRKc302kssv0200
	0/3	HRKc302kssv0300
	0/4	HRKc302kssv0400
	0/5	HRKc302kssv0500
	0/6	HRKc302kssv0600
	0/7	HRKc302kssv0700

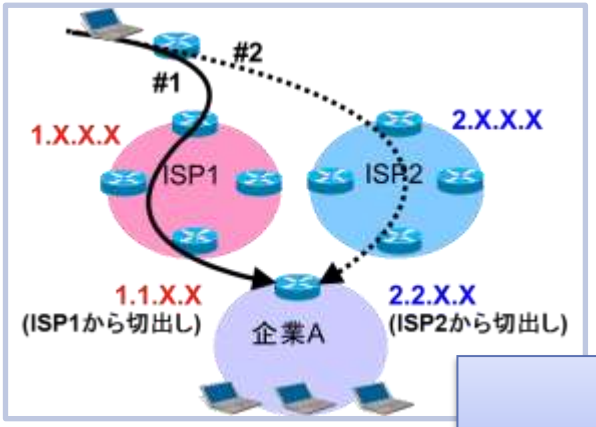
ホスト名	IP	GW IP
HRKc104krsv0100	10.192.127.20	10.192.
HRKc104krsv0101	10.192.127.1	10.192.
HRKc104krsv0102	10.192.127.2	10.192.
HRKc104krsv0103	10.192.127.3	10.192.
HRKc104krsv0104	10.192.127.4	10.192.
HRKc104krsv0105	10.192.127.5	10.192.
HRKc104krsv0106	10.192.127.6	10.192.
HRKc104krsv0107	10.192.127.7	10.192.
HRKc104krsv0108	10.192.127.8	10.192.
HRKc104krsv0109	10.192.127.9	10.192.
HRKc104krsv01ma	10.192.127.17	10.192.
KHNcd02krsv0100	10.194.127.20	10.194.
KHNcd02krsv0101	10.194.127.1	10.194.
KHNcd02krsv0102	10.194.127.2	10.194.
KHNcd02krsv0103	10.194.127.3	10.194.
KHNcd02krsv0104	10.194.127.4	10.194.

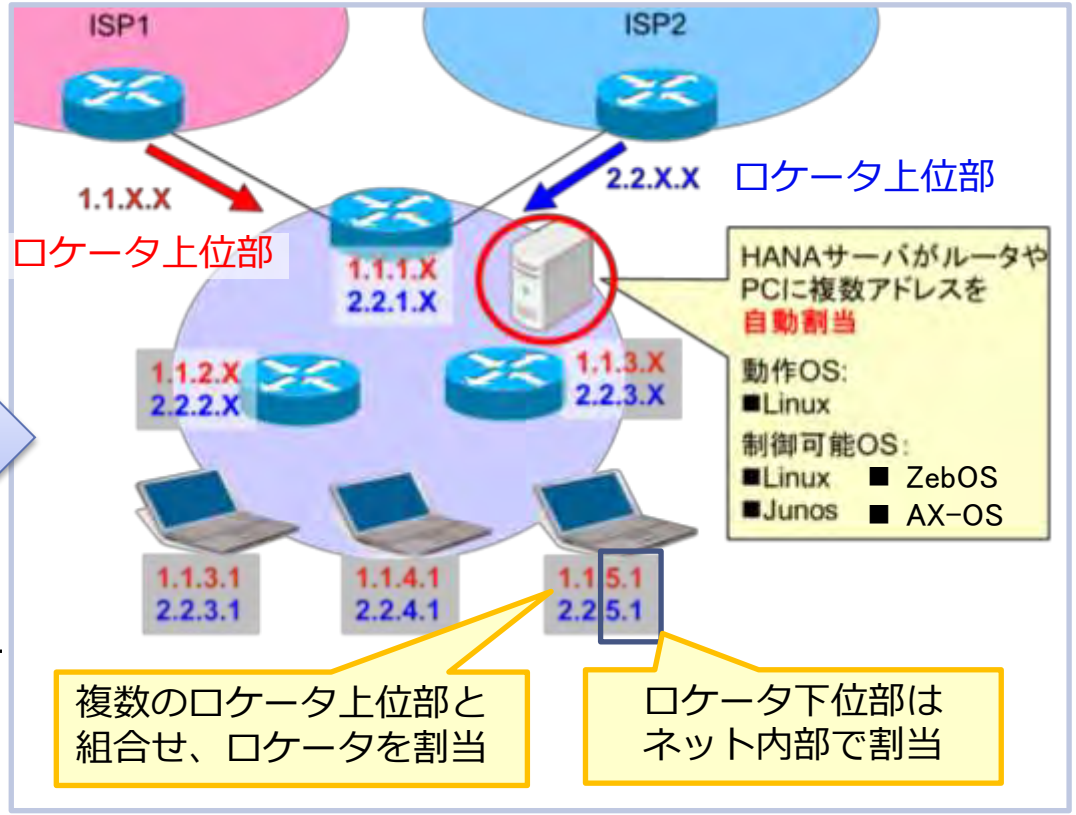
0/1		HRKc303kssv0100
0/2		HRKc303kssv0200
0/3		HRKc303kssv0300
0/4		HRKc303kssv0400
0/5		HRKc303kssv0500
0/6		HRKc303kssv0600
0/7		HRKc303kssv0700
0/8		HRKc303kssv0800
0/9		HRKc303kssv0900

消費者向けサービス事業で、公表される規模の事故で人為要因は2件（全体の2%）

階層型自動番号割当 [HANA]

人手をかけず稼働率の高いネットワークを構築する技術





特長

- 同時複数経路で耐障害性向上
- 自動割当で管理者負担軽減

  PC千台の網で番号設定負担 1/100
- アドレス更新も簡単
- IPv4でも、もちろんIPv6でも
- 小さな経路表で網安定と省エネ

ネットワーク検証実績

- JGN-Xにネットワーク構築、動作検証
- **46,000ネットワークでの動作検証**
- SDNとの連携
- 耐災害の観点からの検証

JGN-X

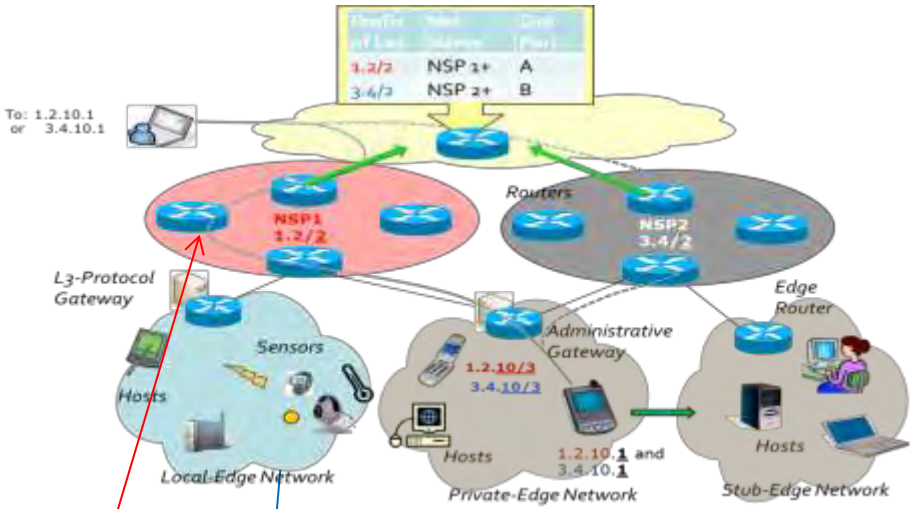
StarBED

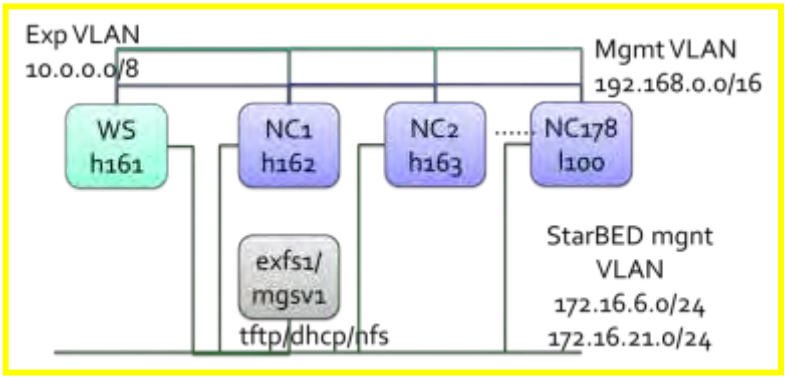
仙台

StarBED³におけるHANA検証 (2011年)

- 10,000 AS エミュレーション
- Debian GNU/Linux 6.0の LXC に実装

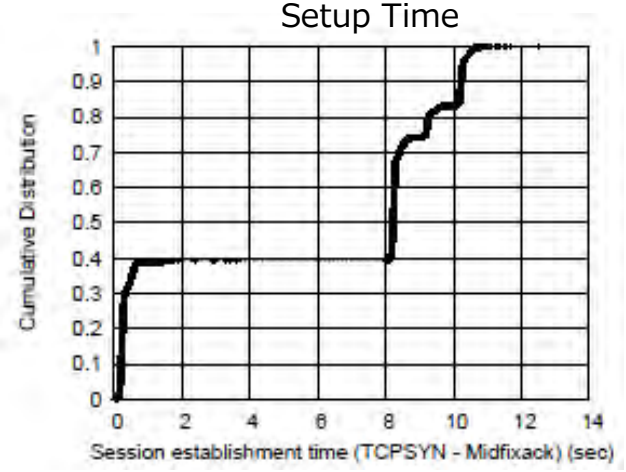
Emulation on StarBED³
Group H 80 nodes (HP ProLiant DL320)
Group L 100 nodes (Cisco UCS C200 M2)



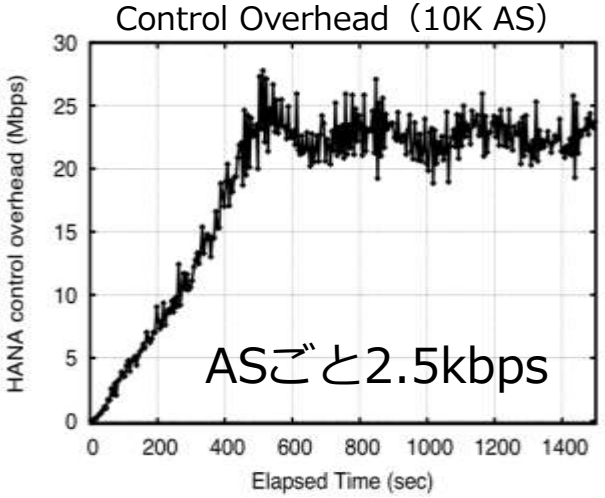


Locator Assign Visualization

Setup Time



Control Overhead（10K AS）



K. Fujikawa, H. Tazaki, H. Harai, SAINT2012

HANAを、北陸StarBED技術センターの皆様に宣伝いただきました。御礼申し上げます。

「これが世界最大規模のテストベッドの全貌だ：
潜入！ 北陸StarBED技術センター」
IT Media Inc 記事 @IT Master of IP Network (2013.3.26)

記事によると「すでに1万ASでの検証は完了し、今は**3万6000ASでの大規模検証が行われている**」とのこと。。。

発表文献をくださいと聞かれる．．．

やるにはやったんですが、発表はStarBEDのこの会までとっておきました

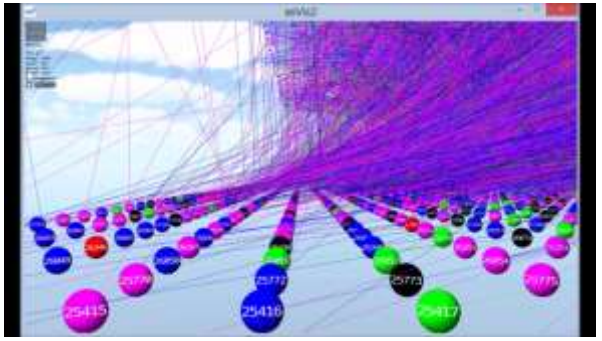
そうこうしてる間に見えるAS が**4万6000**に増えてましたが。。。

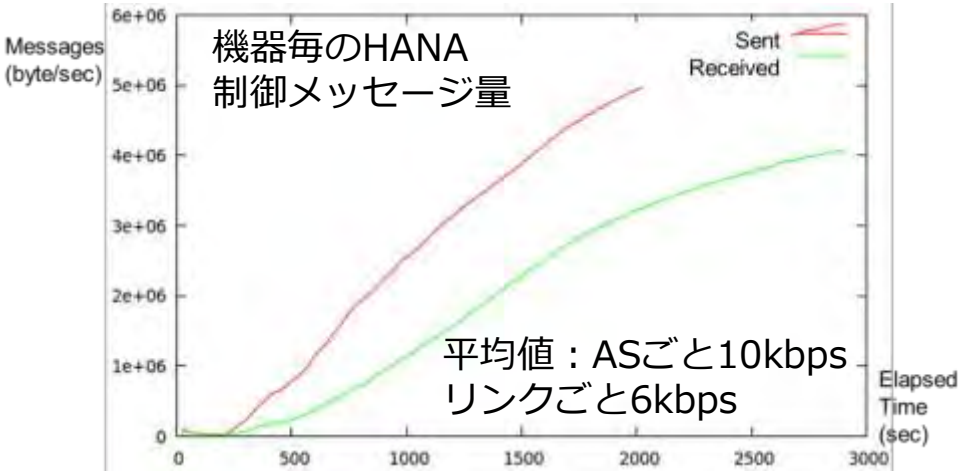
HANAのインターネット規模性検証（2012～）

- 実インターネットAS規模の46,000ノードでHANAを稼動
  - ノード間でのアドレス空間割当動作を検証
  - IPv4およびIPv6アドレス体系ともに検証
  - 制御量はAS平均10kbps

HANA 実験の構造 (CAIDAの実測データベースより作成)	
最上位 AS数	821
中間 AS数	6,234
最下位 AS数	39,122
合計 AS 数	46,177
親を沢山もつASの親数	44
子を沢山もつASの子数	4,275
BGPで利用するリンク数	177,397
HANAで利用するリンク数	81,661

最下位ASへのアドレス空間割当

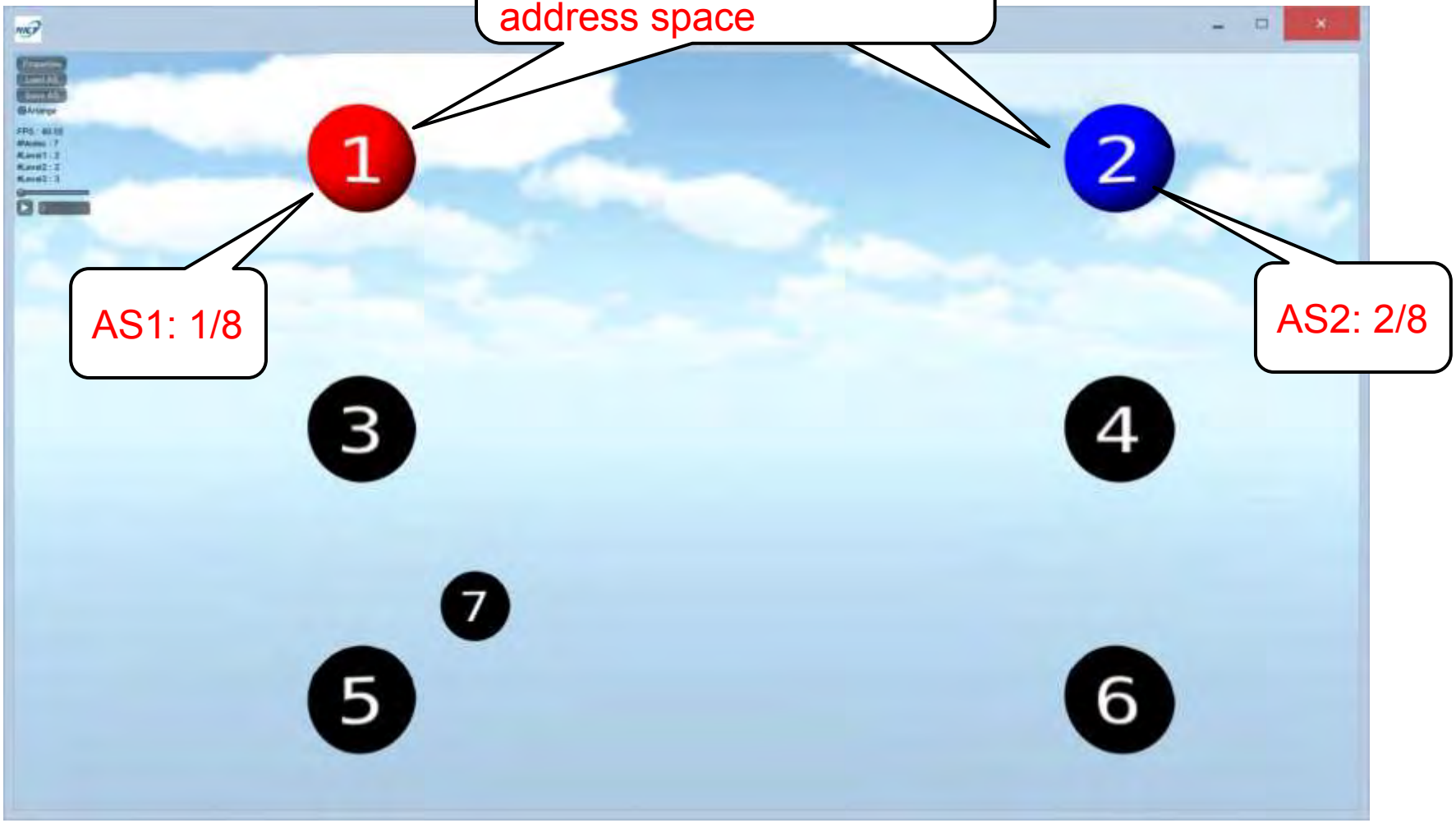




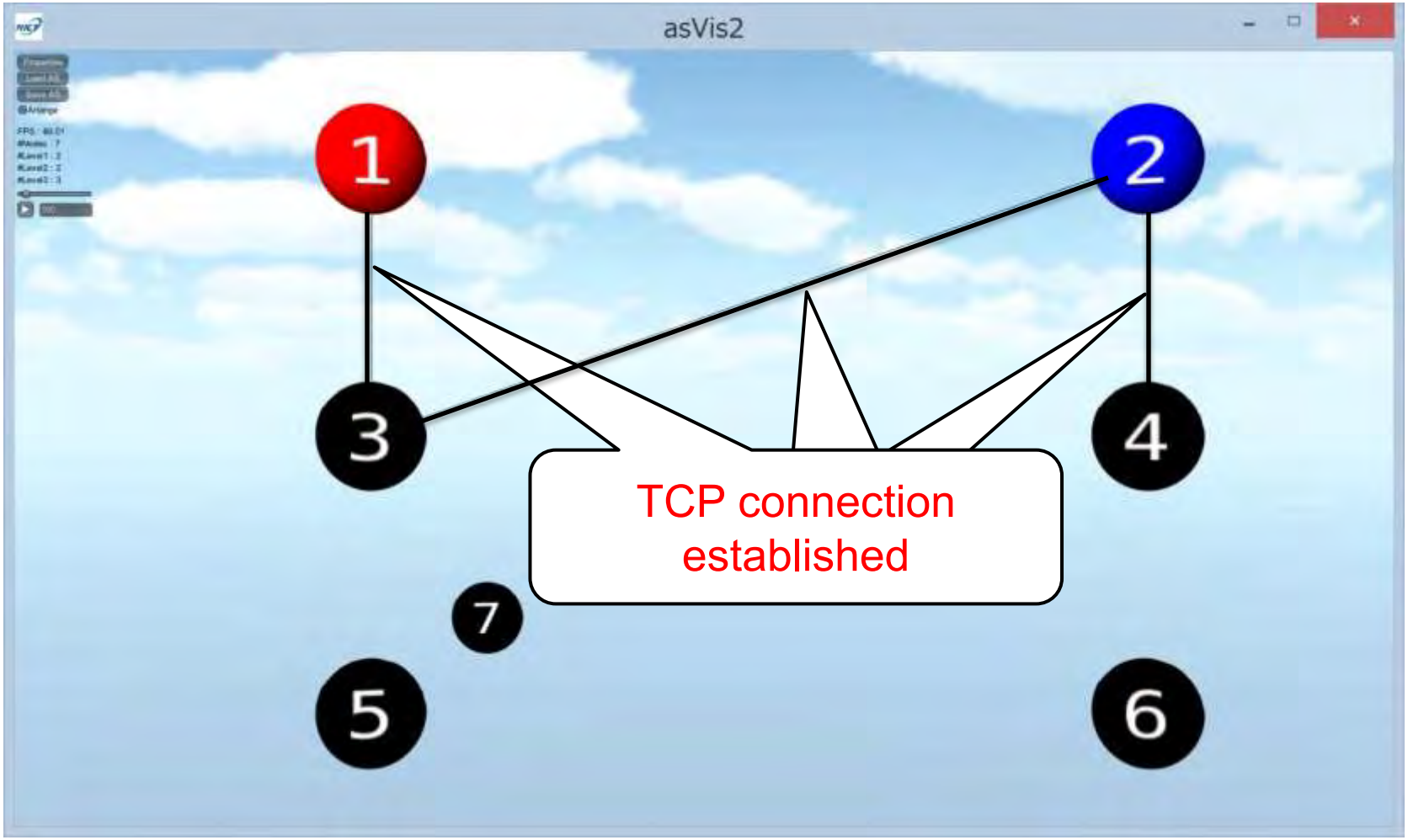
StarBED3 で利用した10 機の物理サーバの諸元

- CPU Intel Xeon E5 2650 (2.00GHz 8 core) x 2
- Memory 128 GB (DDR3-1333)

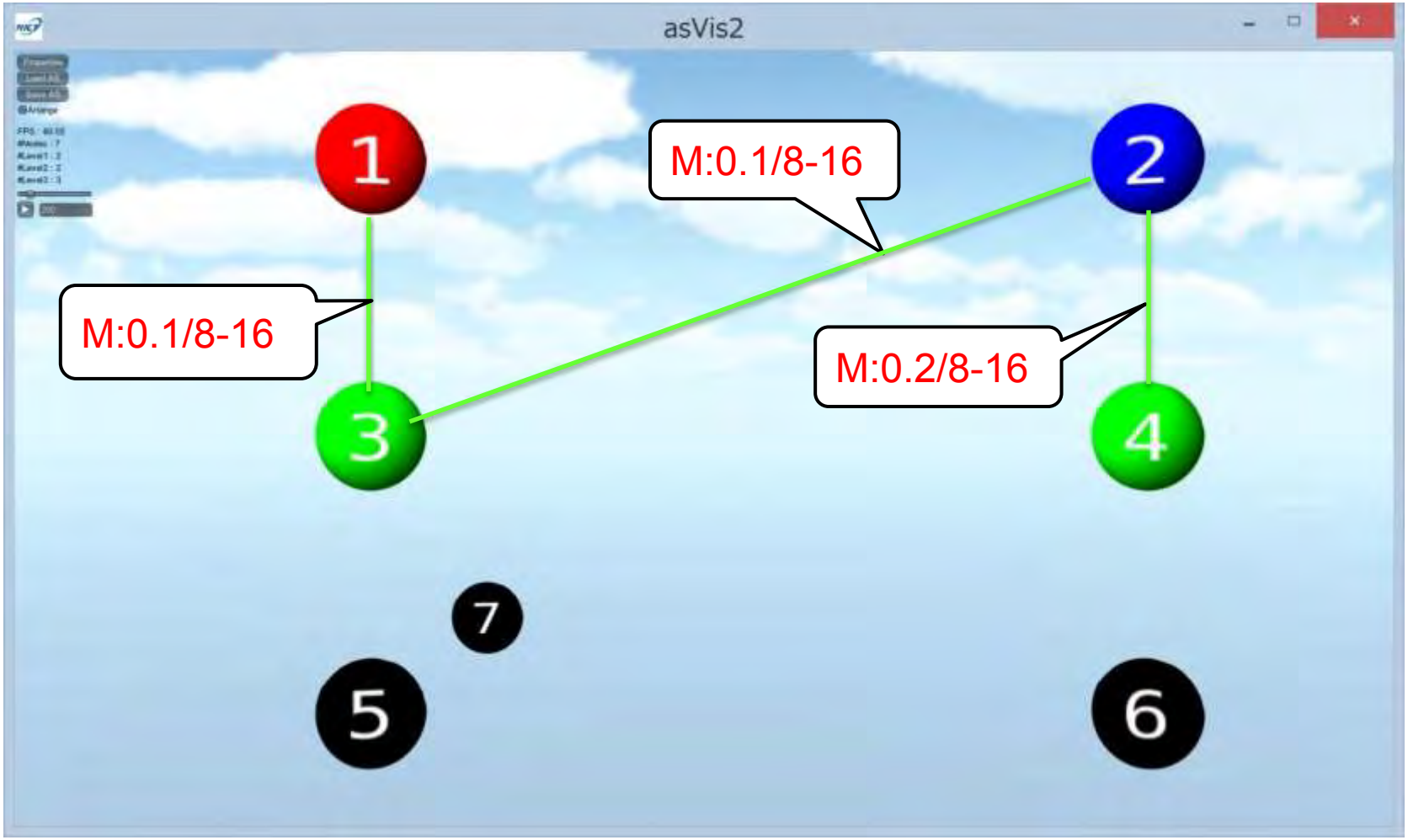
Simulation Example (1) Initial Status



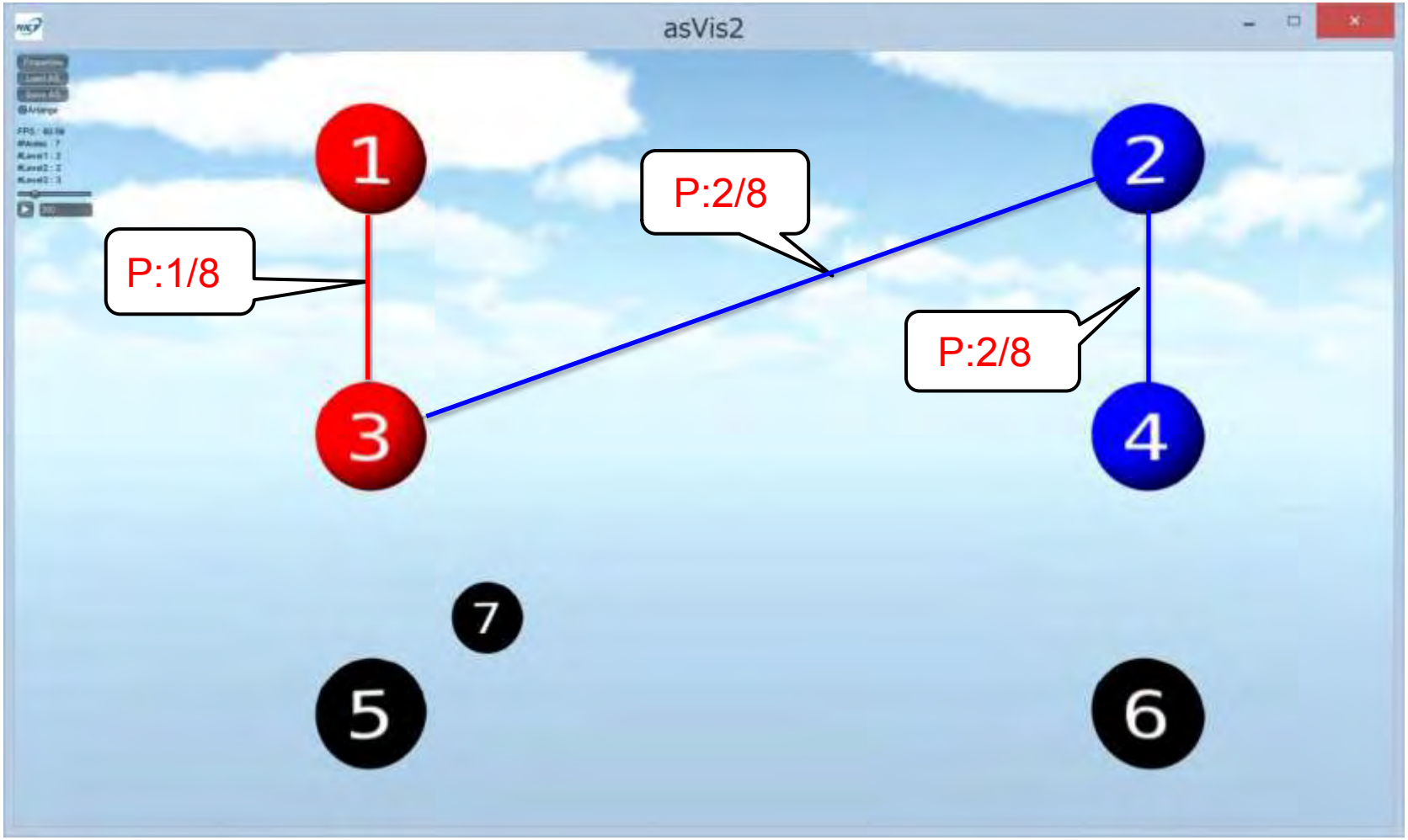
Simulation Example (2) TCP connection



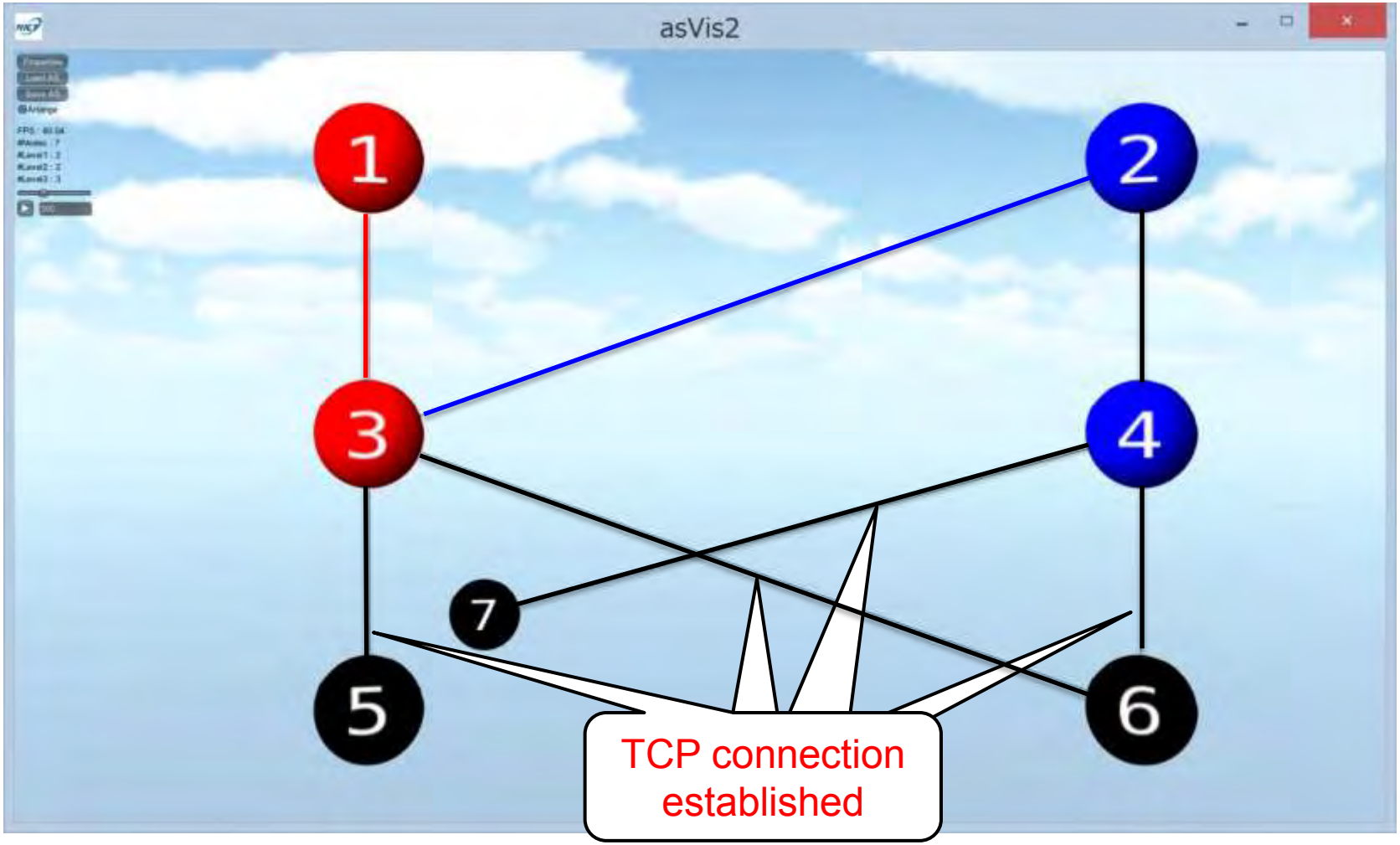
Simulation Example (3) Midfix allocation



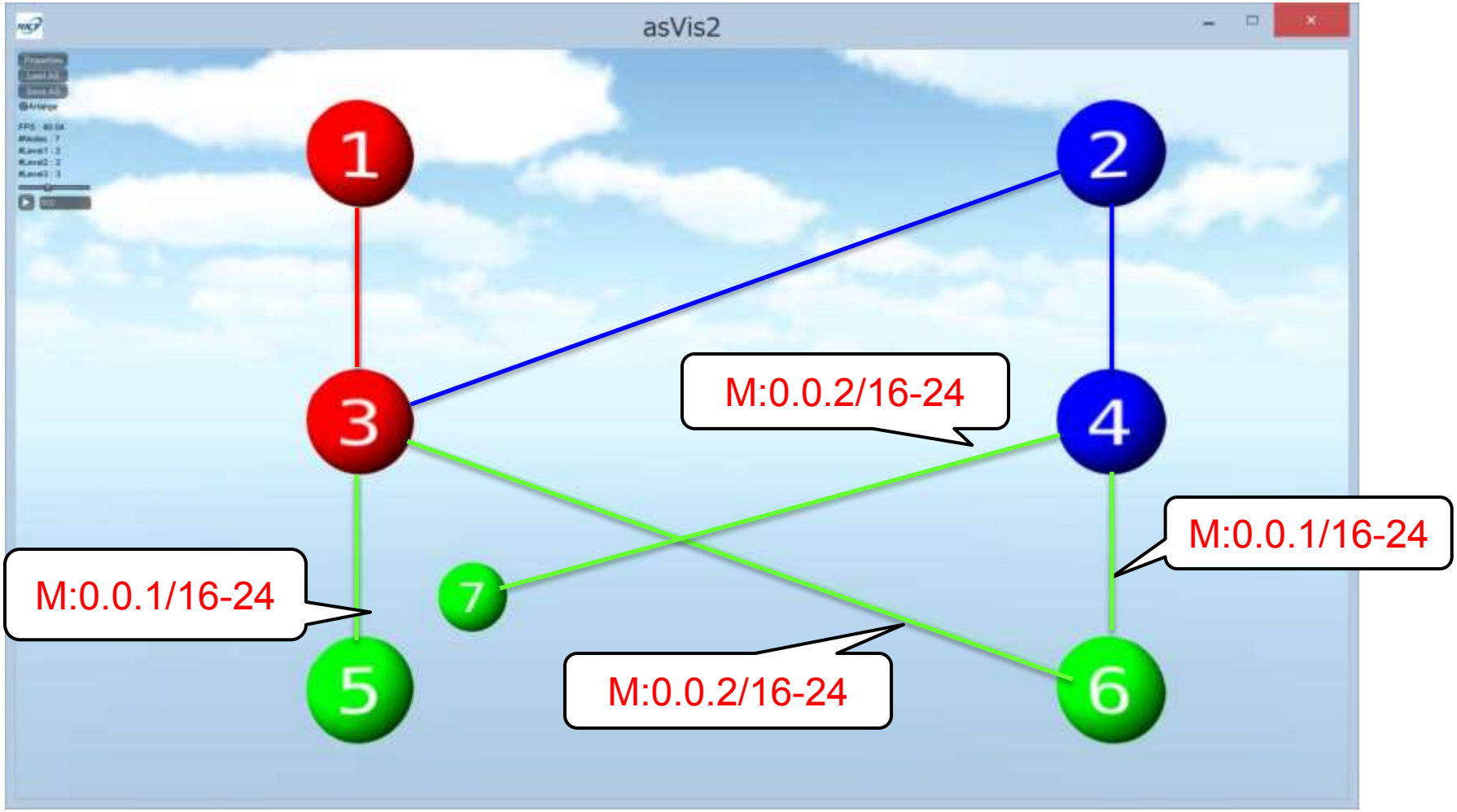
Simulation Example (4) Prefix distribution



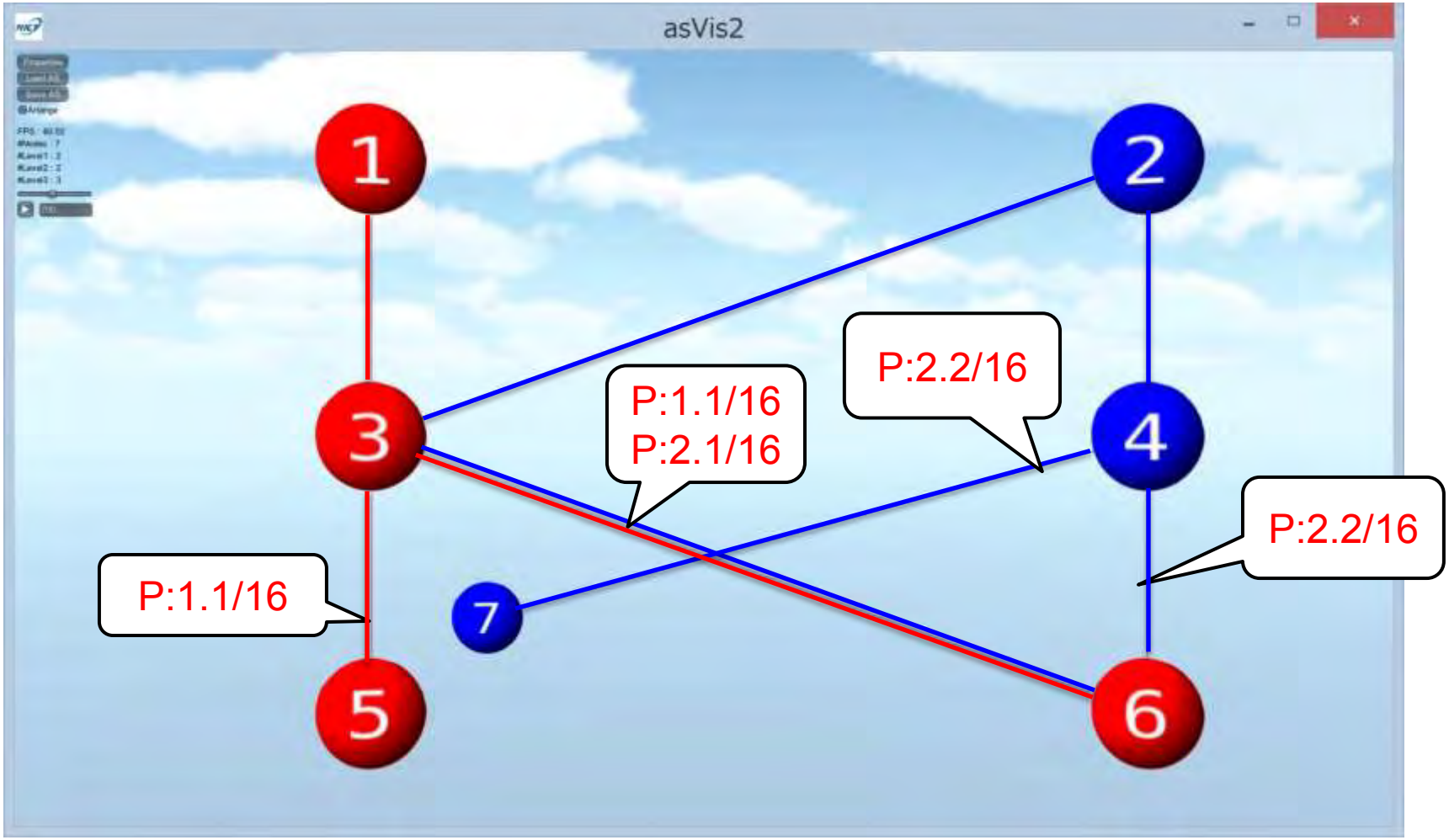
Simulation Example (5) TCP connection



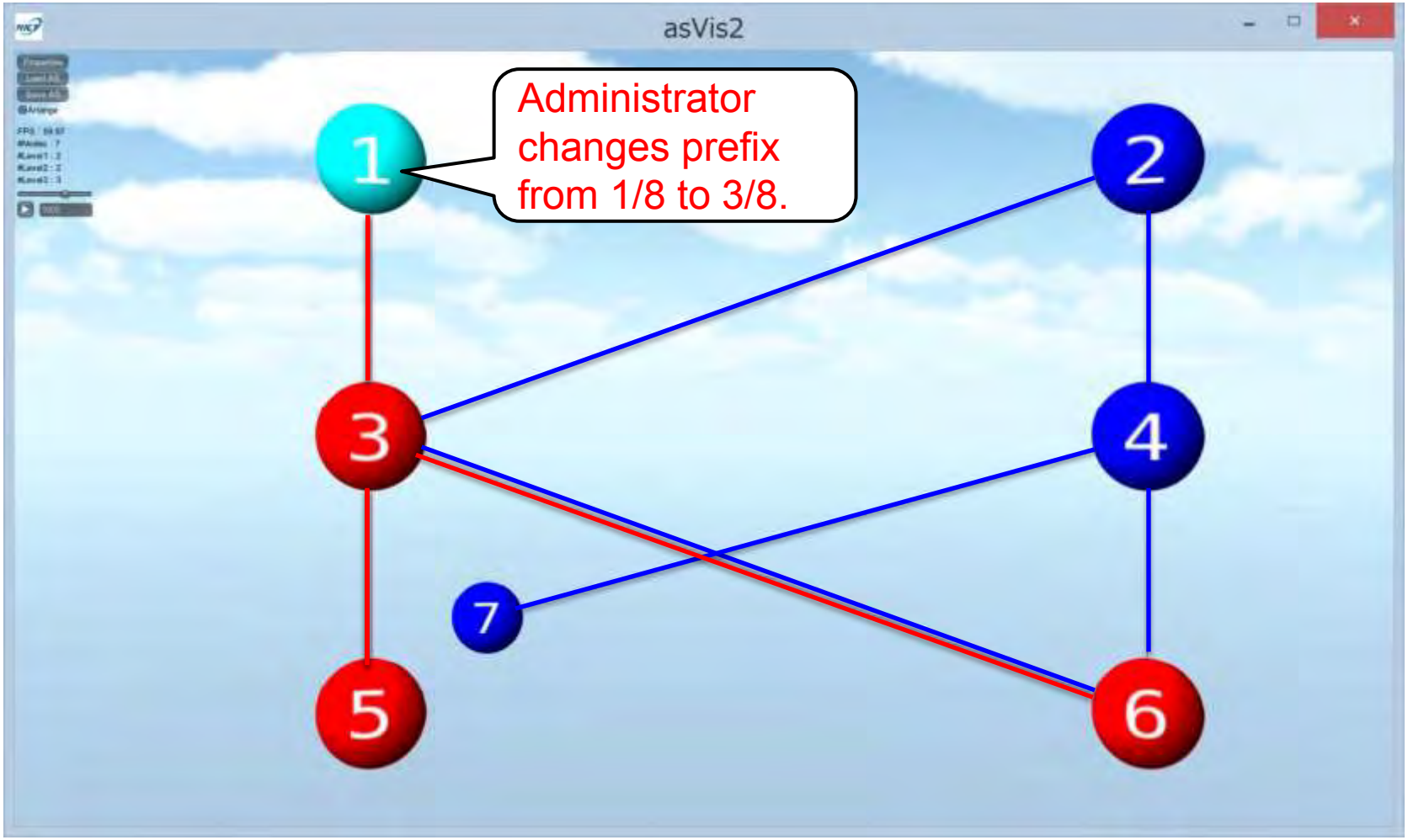
Simulation Example (6) Midfix allocation



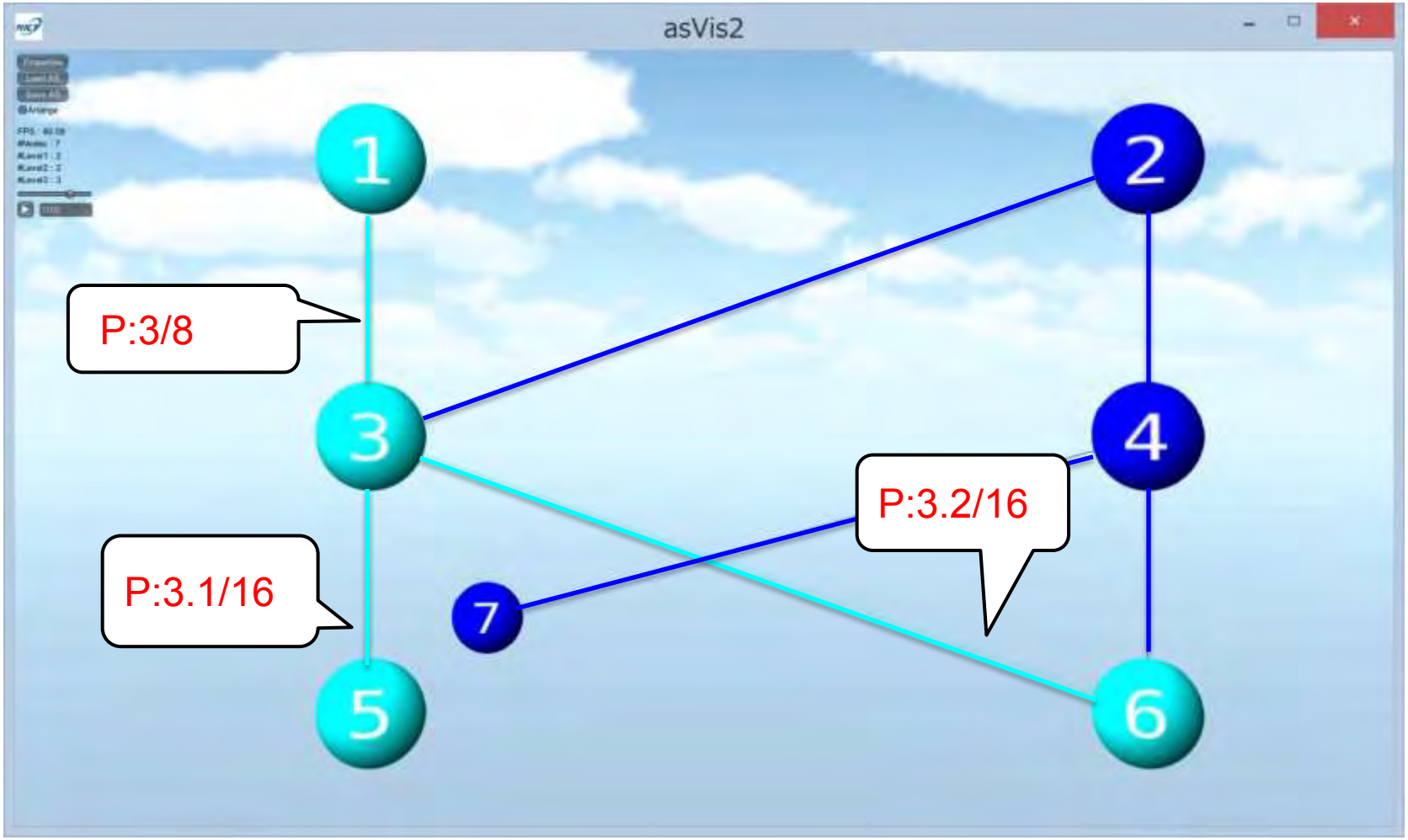
Simulation Example (7) Prefix distribution



Simulation Example (8) Prefix is changed

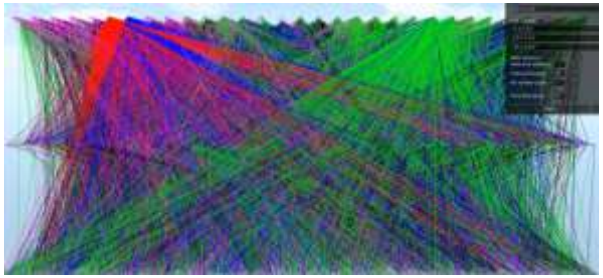


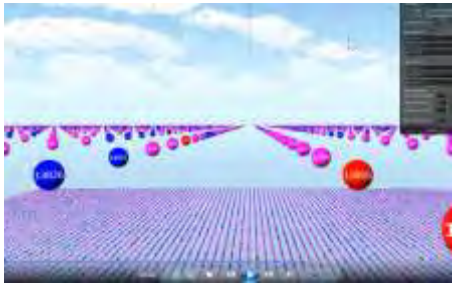
Simulation Example (9) Changed prefix distribution



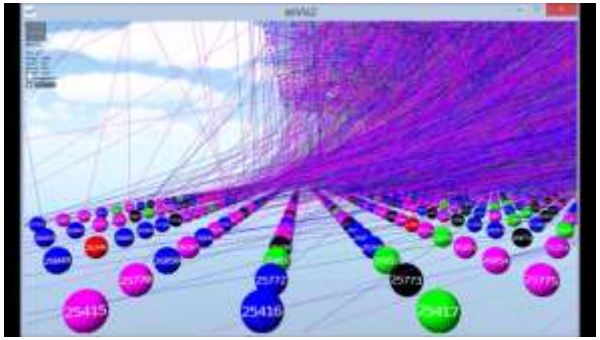
HANAのインターネット規模性検証（可視化）

全体へのドレス空間割当



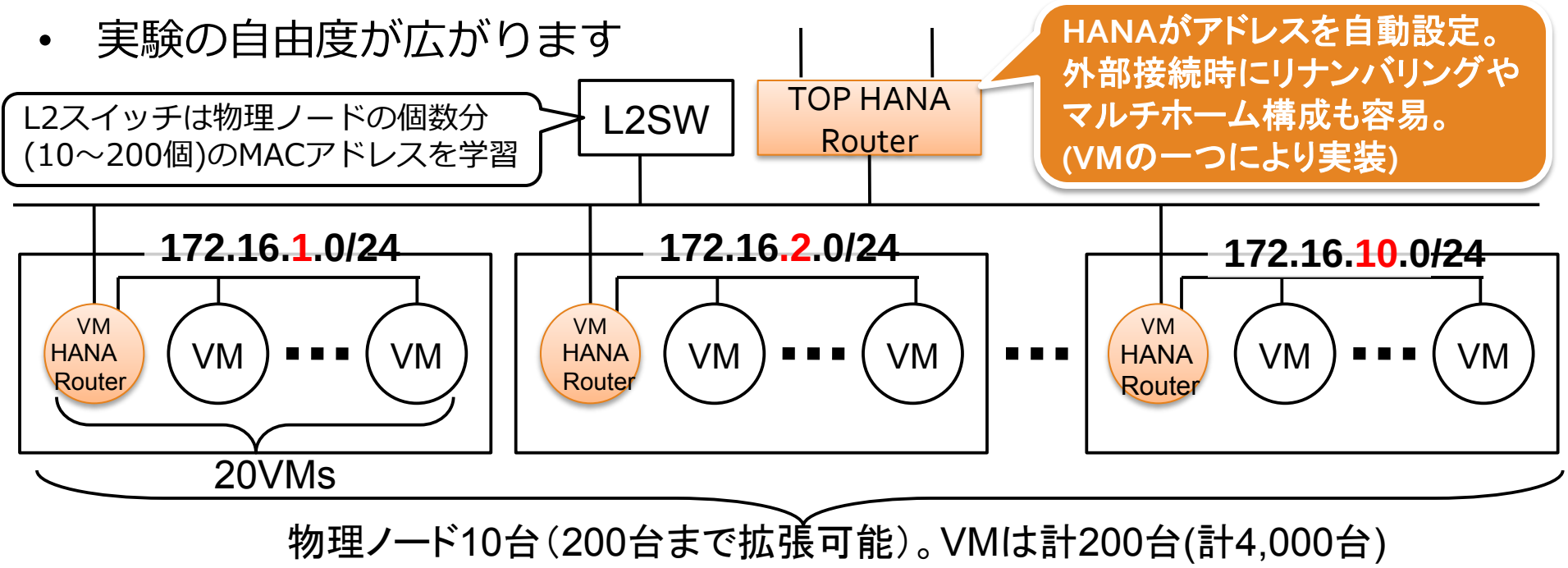


最下位ASへのアドレス空間割当

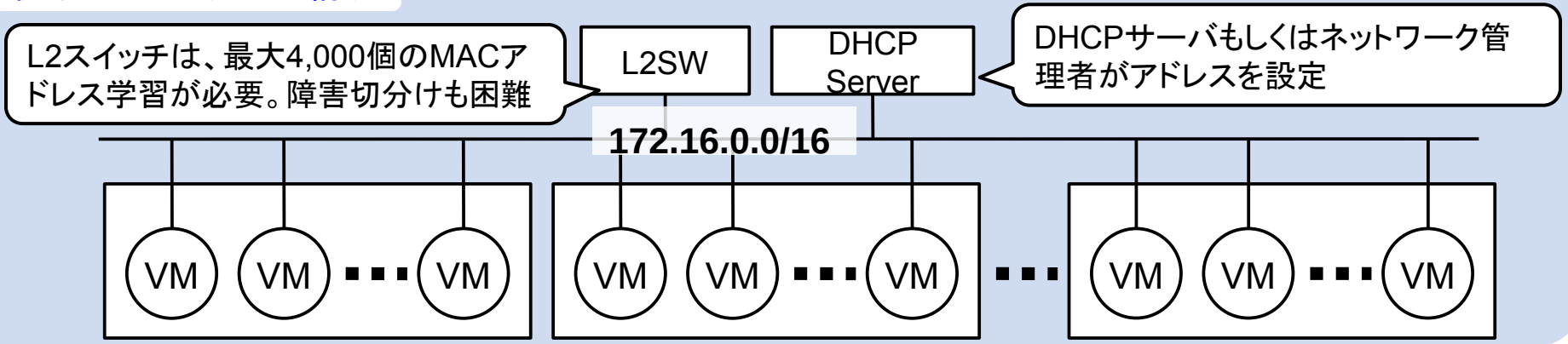


StarBED利用者向けネットワークをHANAで構築

- 階層的なIPネットワークをStarBEDで簡単に構築できるようになりました
- 実験の自由度が広がります



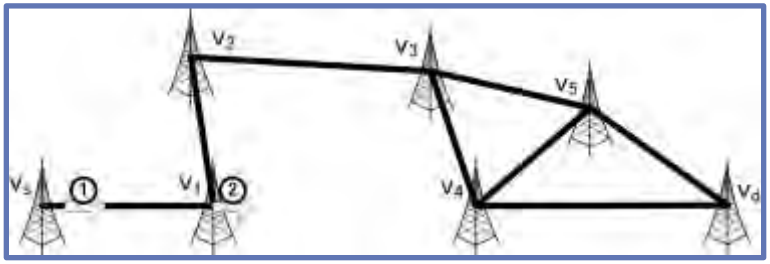
従来までのフラットな構成

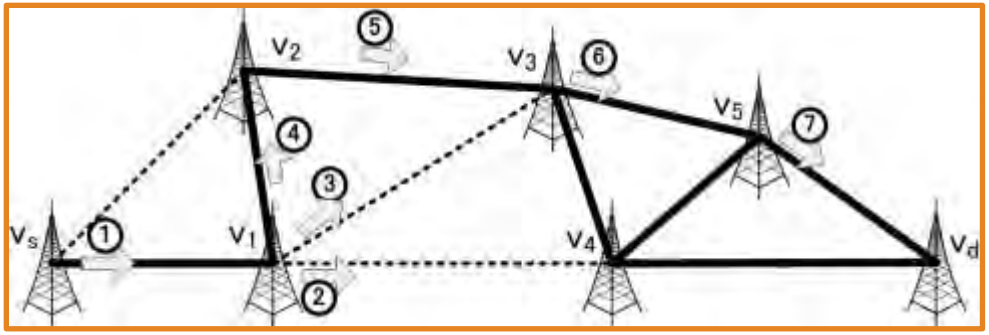


局所情報を用いた分散経路制御
ドロネーオーバレイ迂回ルーティング

Cf) M. Ohnishi, H. Harai, APSITT 2015.

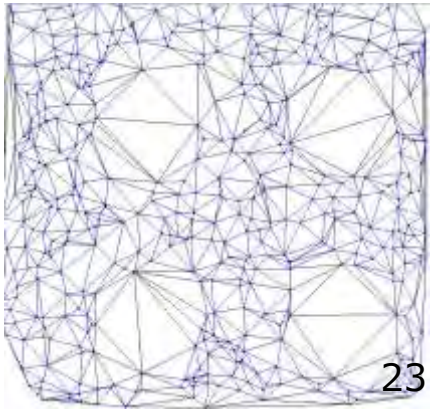
- 災害時等トポロジーが変動する
- ネットワークの全体トポロジを把握せずに経路表を構築
  - 全方位の近隣ノードに対してのみ経路表を作成
- 位置座標に基づいてデータを**目的地に近い方向**へ転送
  - **近隣だが転送の仕組み上、到達不能の場所がある**
  - **近隣だが物理制約上、到達不能の場所がある**





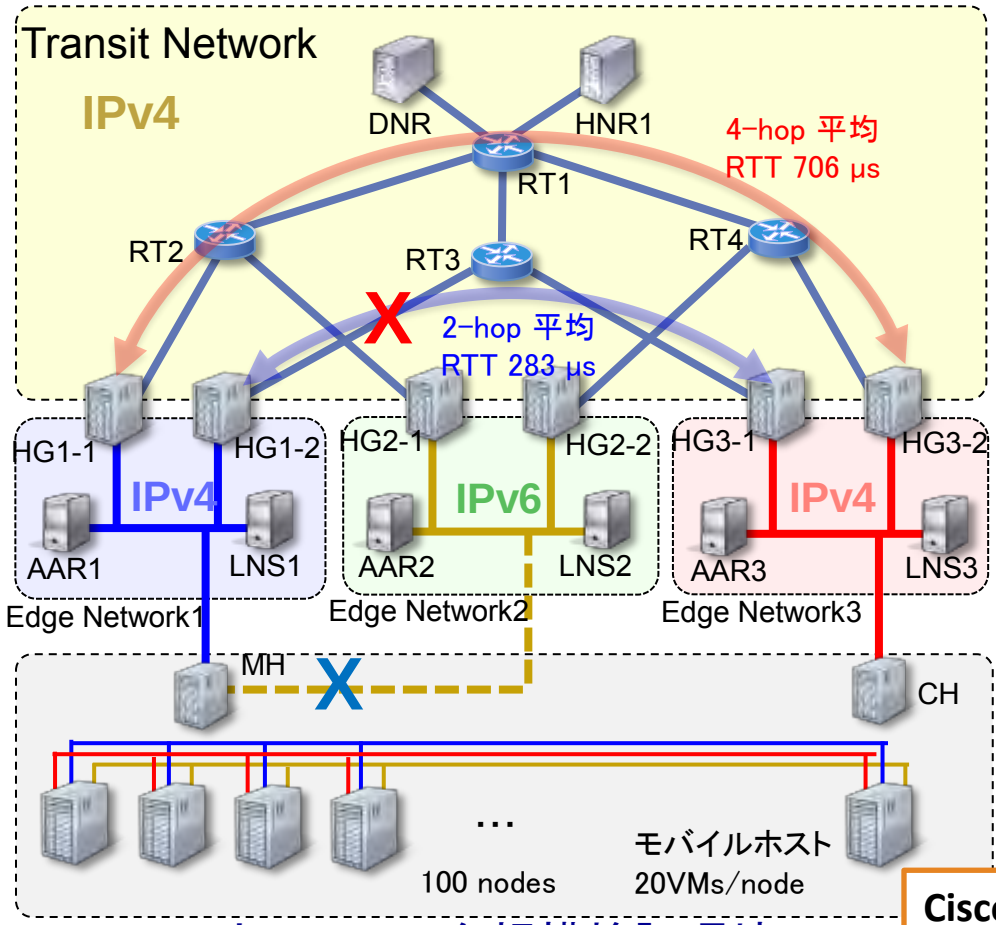
- 迂回経路を作る => 局所の情報で作る
- JAVA VM を用いた実装の2,500ノード検証

Cisco UCS C200 M2 を最大125機
- **CPU Intel Xeon X5670 (2.93GHz 6 core) x 2**
- **Memory 48 GB**

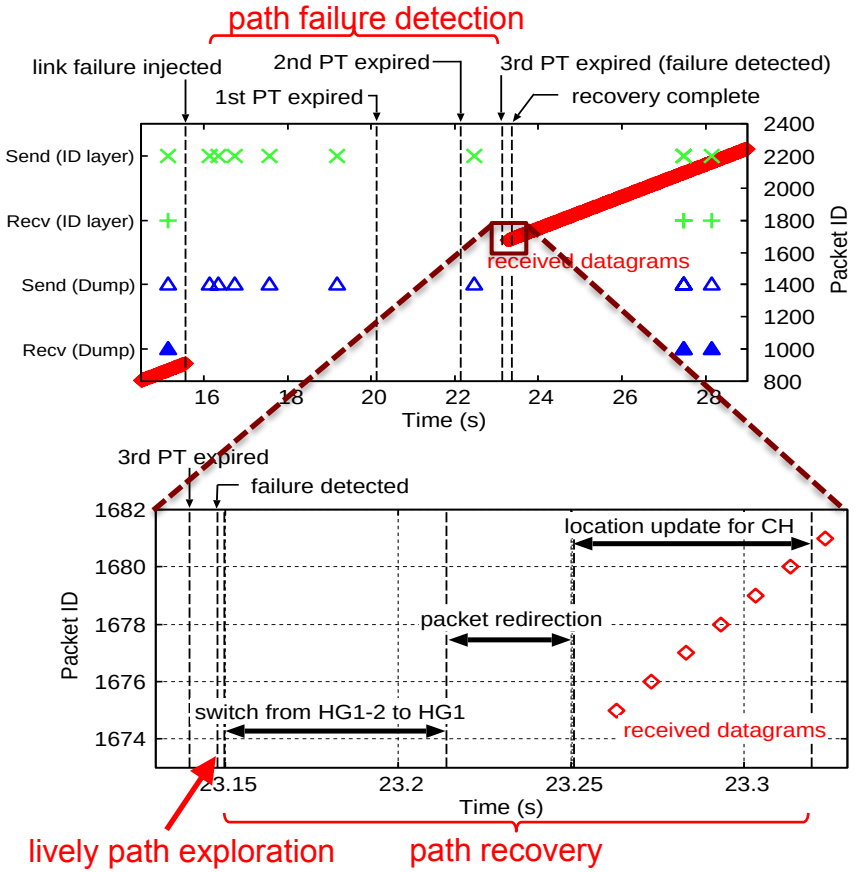


ID通信機構 HIMALIS のホスト2000台規模ネットワーク構築

- IPアドレスに依存しない通信、モビリティ、マルチホーム、認証、**故障回復**
- 通信障害検知・回復等のゲートウェイ間連携、レジストリ間連携など分散システムの検証・評価



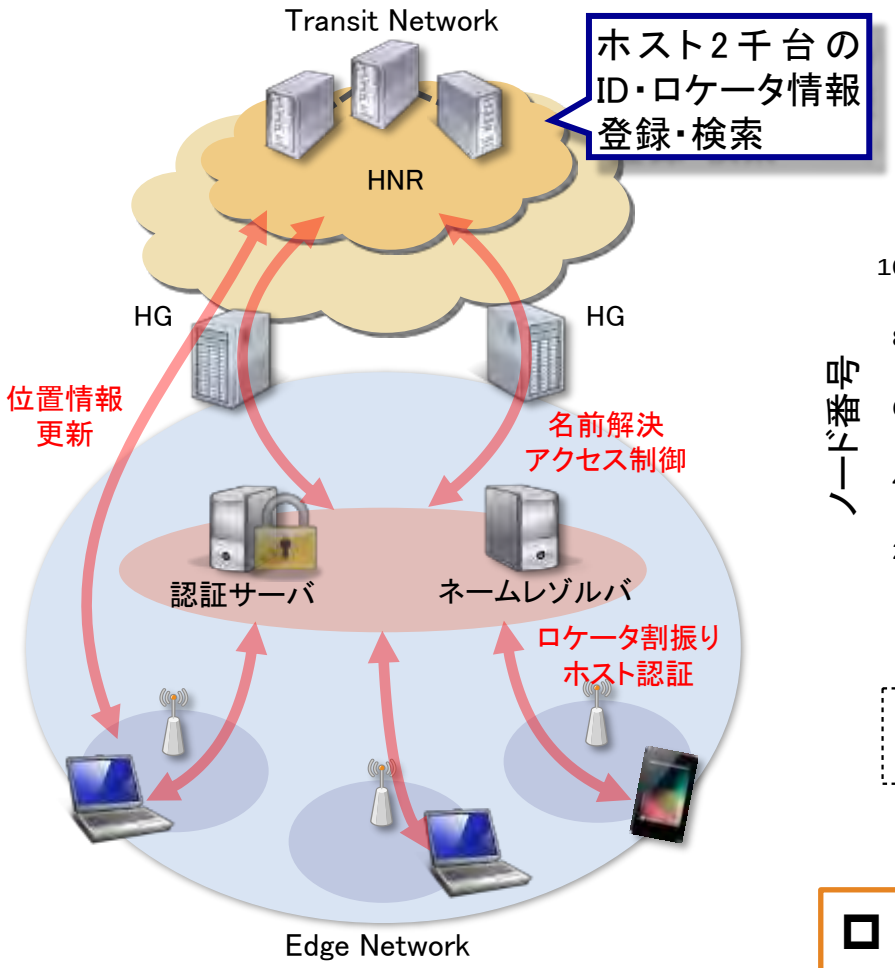
ホスト2,000台規模検証環境



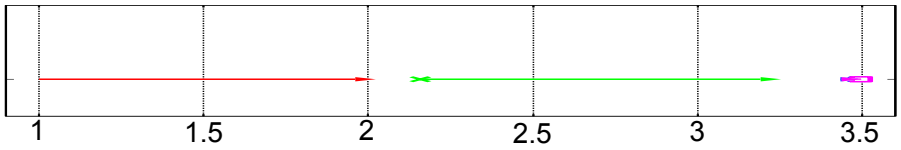
Cf) Y. Fukushima et al., ICUFN 2014.

Cisco UCS C200 M2 (グループ K, L)
- **CPU Intel Xeon X5670 (2.93GHz 6 core) x 2**
- **Memory 48 GB, 6 GbE NICs**

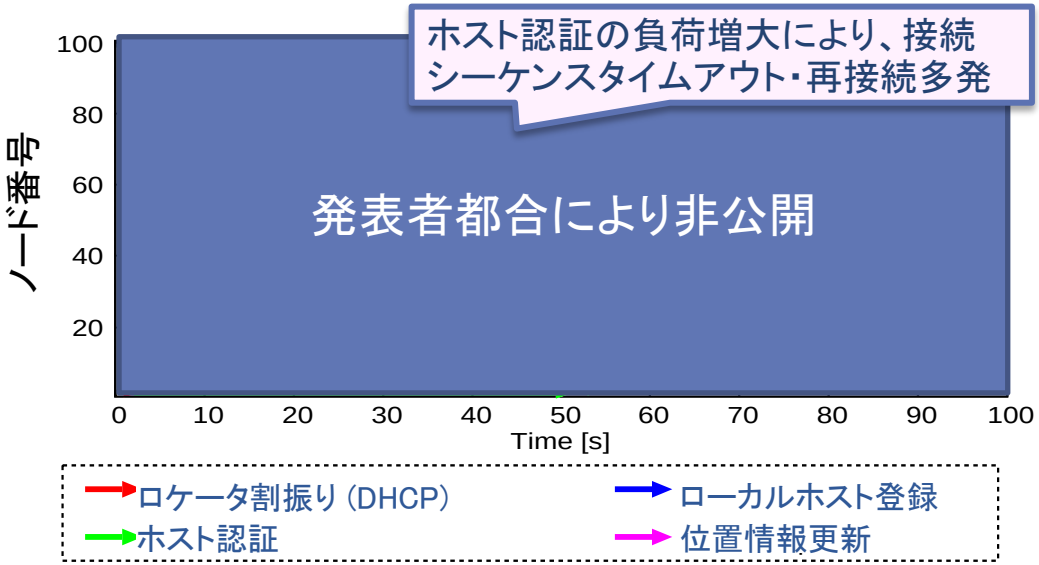
HIMALISシグナリングストームの検出・改善



HIMALISの認証付き制御プレーン



ホスト1台のときのネットワーク接続時間（秒）



ホスト100台のネットワーク接続（2014年4月時点）

- 認証サーバ、ネームレゾルバの負荷増大による接続シーケンスの爆発（ストーム）を検出
  => ソフトウェアの改善策導出
- ホスト1,000台の接続を確認（2015年現在）

StarBED³ 利用 Tips (2013年12月)

- ノード操作がより簡単に、ディスク読み書込み速度が改善
  - 今年度操作スクリプト一新、350GBのディスクコピーが2時間程度
  - 以前は、信用のおける定量データがないので感覚で、割当てマシンスペックに大きく依存と思いますが、2011年ごろのマシンでは終夜ディスクコピーしてた（250GB以上と思います）。
- 100台使用すると数台は同じ作業を行っても結構な割合で不具合あると感じる
  - 理由は色々：ディスク障害、メモリ障害、NIC、BIOS設定不備など）。
- 利用要望書は早め提出、時間が足りない時は延長申請可
  - 来月の利用申請は毎月15日、2013年度現在の利用要望は300%超
- KVMコンソールが応答しない、昨日できたことができないとき
  - StarBED側の問題である可能性大、すぐに連絡、あまり頑張らない

一部非公開

まとめ

- 分散システムの大規模検証
  - 階層型自動番号割当 HANA
  - 局所情報経路制御
  - ID・ロケータ分離 HIMALIS
- レイヤ3のネットワークを組む際のアドレッシングが煩わしい方、ご相談ください
- 新世代ネットワークの研究開発に大規模なサーバ群を用いた検証は不可欠
  - 動作実績
  - 性能面でのバグ検出
  - 5G、IoT など数が必要な場面は登場している

関連文献

- K. Fujikawa, H. Harai, and M. Ohta, “The Basic Procedures of Hierarchical Automatic Locator Number Allocation Protocol HANA,” Proc. Asia Workshop on Future Internet Technologies (AWFIT 2011), pp. 124--131, October 2011.
- **K. Fujikawa, H. Tazaki, H. Harai, “Inter-AS Locator Allocation of Hierarchical Automatic Number Allocation in a 10,000-AS Network, ”Proc. SAINT 2012, July 2012.**
- M. Ohnishi, M. Inoue, and H. Harai, “Incremental distributed construction method of Delaunay overlay network on detour overlay paths,” Journal of Information Processing (JIP), Vol. 21, No. 2, February 2013.
- **M. Ohnishi and H. Harai, "Delaunay Overlay Network Construction Method for Super-Wide Area Disaster Situations," APSITT 2015 (10th Asia-Pacific Symposium on Information and Telecommunication Technologies), pp. 88—90, August 2015.**
- V. P. Kafle, R. Li, D. Inoue, H. Harai, "Design and Implementation of Security for HIMALIS Architecture of Future Networks," IEICE Transactions on Information and System 2013, Vol. E96-D, No. 2, pp. 226--237, February 2013.
- **Y. Fukushima, V. P. Kafle, T. Tomuro, and H. Harai, “Implementation of Communication Path Recovery Mechanism in a Multihomed ID/Locator-split Network,” The Sixth International Conference on Ubiquitous and Future Networks (ICUFN 2014), pp.322—327, July 2014.**