Experiences in Emulating 10K AS Topology with Massive VM Multiplexing

Shinsuke MIWA@NICT

danna@nict.go.jp

(presentator: Hiroaki Hazeyama)

Cooperate with:

Mio Suzuki@NICT, Hiroaki Hazeyama@NAIST
Satoshi Uda@JAIST, Toshiyuki Miyachi@NICT
Youki Kadobayashi@NAIST, Yoichi Shinoda@NICT

Background

- **We need a large-scale internet-like test environment for testing the actual running codes of internet scale applications**
  - Controllable, manageable, tractable
  - Observing everything
  - Reasonable cost
- **Requirements for Large-scale Network Experiments**
  - Scale, Topology, Link quality, Routing policy, Background traffic, ...**High-fidelity Internet emulation on a testbed**

Project Goal

- `# make internet` or `# make world`
  - Make the Internet for any experiment on a testbed
    - Get snapshot of the Internet
      → Pick up required sub-graph
      → Create configurations
      → Allocate/Construct an emulated Internet
      on a testbed

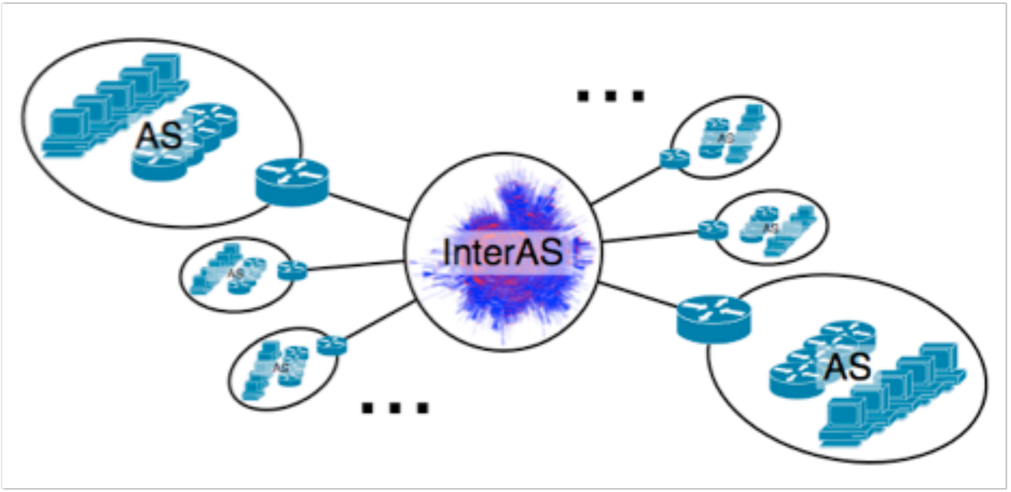
Issue

- **How can we get “the Internet like” environment on a testbed?**
  - Scale of the Internet
    - over 600,000,000 hosts (ISC, January 2009)
    - over 30,000 advertised AS (potaroo.net, May 2009)
  - Other characteristics...
    - bandwidth, host behavior, traffic, …..

As the first step, we tackle on the scale issue

First Step is "Emulating AS Topology"

- **Reason 1:**
  Inter-AS Network is a core part of the Internet
  - AS: Autonomous System is a management/policy domain of the Internet



- **Reason 2: Reasonable size**
  - About 30,000 ASs advertised on the Internet
  - Over 1,000 PC servers on StarBED
  - If we can construct 30 ASs per 1 PC server..., **Yes, we can!**

Contribution

- **Develop tools for constructing emulated inter-AS network**

- **Successfully provide inter-AS network up to 10K size**
  - We can easily set up around 500 size topology in one hour
  - We can construct 10k size topology, but …..

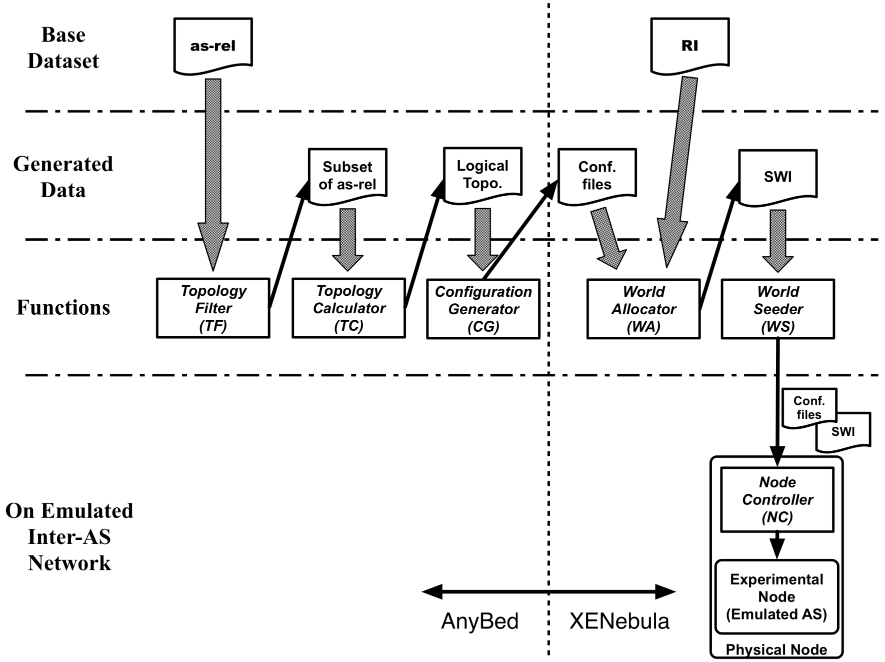
Overview of Emulation

- **BGP inter-AS topology according to the real Internet**
  - CAIDA AS Relationship and AS Ranking dataset
- **Emulate 1 AS = 1 BGP router**
  - Quagga **`bgpd`** `and` **`zebra`** run on each AS node
  - No other BGP speakers and No other hosts on one AS
- **Scale boost = multiplexing using Xen**
  - Dom0: VMKnoppix
  - DomU: ttylinux
- **ASs can be controlled from management LAN**
  - TFTP, PXEboot, NFSRoot
  - ssh / telnet,syslog

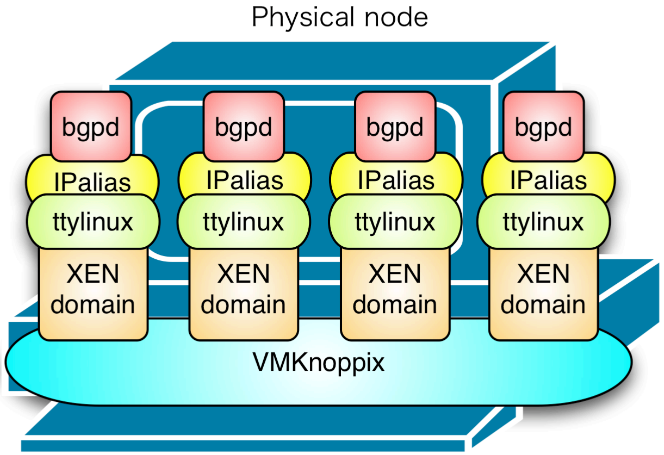
Overview of Our Tools

- **AnyBed**
  - Generate topology and configurations of AS nodes from base dataset
- **XENebula**
  - Create/manage Xen VMs according to **AnyBed** results

TF: Topology Filter
TC: Topology Calculator
CG: Configuration Generator
WA: World Allocator
WS: World Seeder
NC: Node Controller



Multiplexing and Memory Allocation



- **Default settings**
  - Dom0
    - 1024 MB
  - domU
    - Base memory : 24MB
    - Additional memory : 24MB per 25 neighbors
  - Allocation Policy
    - Allocate domU to dom0 from top rank AS node
    - ∑(domU on dom0(i)) ≒ ∑(domU on dom0(i+1))

Experiments

- Feasibility Study
- Merging into the real Internet
- The current Maximum Scale

Feasibility Study

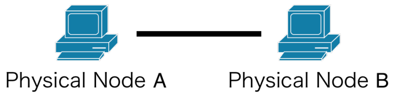
- 250 ASs on 5 physical nodes
- scale down to 1/10

Spec. of Physical Nodes

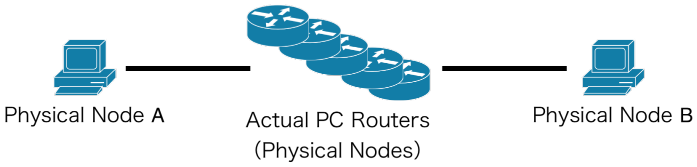
Item	Spec.
CPU	Intel Pentium4 3.2GHz
Memory	2GB
NIC	1000Base-T x 4
HDD	SATA 80GB x 2

Comparison of Delay and Throughput on each environment

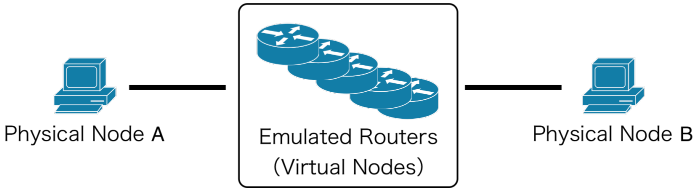
	ping (Avg., ms)	iperf (Mbps)
1)	0.122	961
2)	0.936	961
3)	1.489	160



1) No Routing



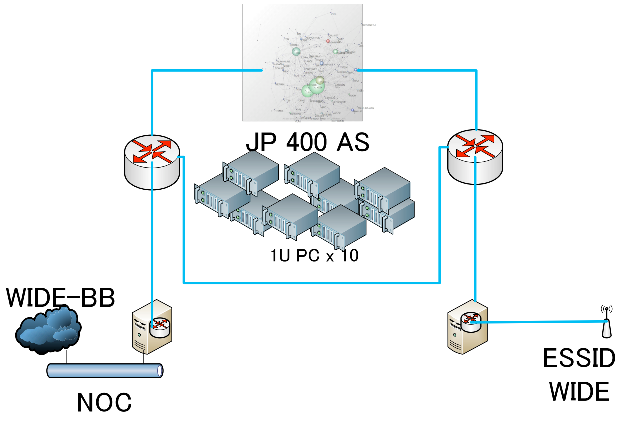
2) via Five Actual PC Routers



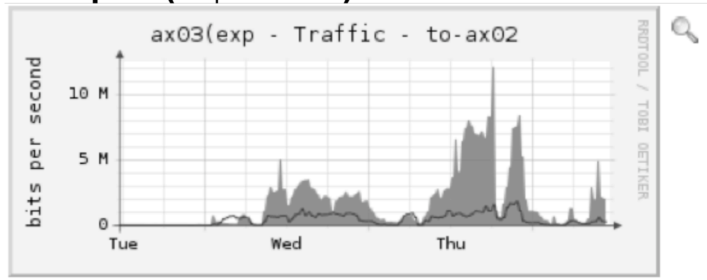
3) via Five Emulated Routers

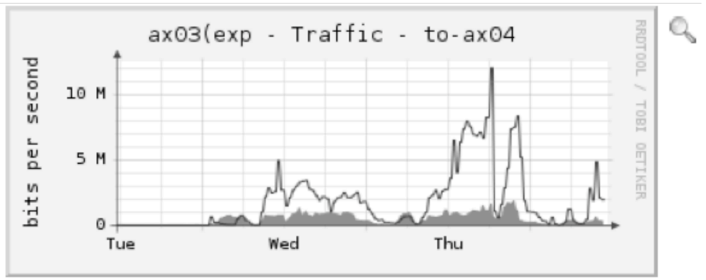
Merging into the real Internet

- **Stably operated**
  - Emulated Japanese (445ASs) inter-AS topology on 10 physical nodes
  - Put in between the Internet and an academic conference network by static route setting
  - Spent 30 min. for setup
  - No trouble during 3 days



- **Throughput**
  - 30-60Mbps (`netperf`)
  - 98Mbps (`iperf`)

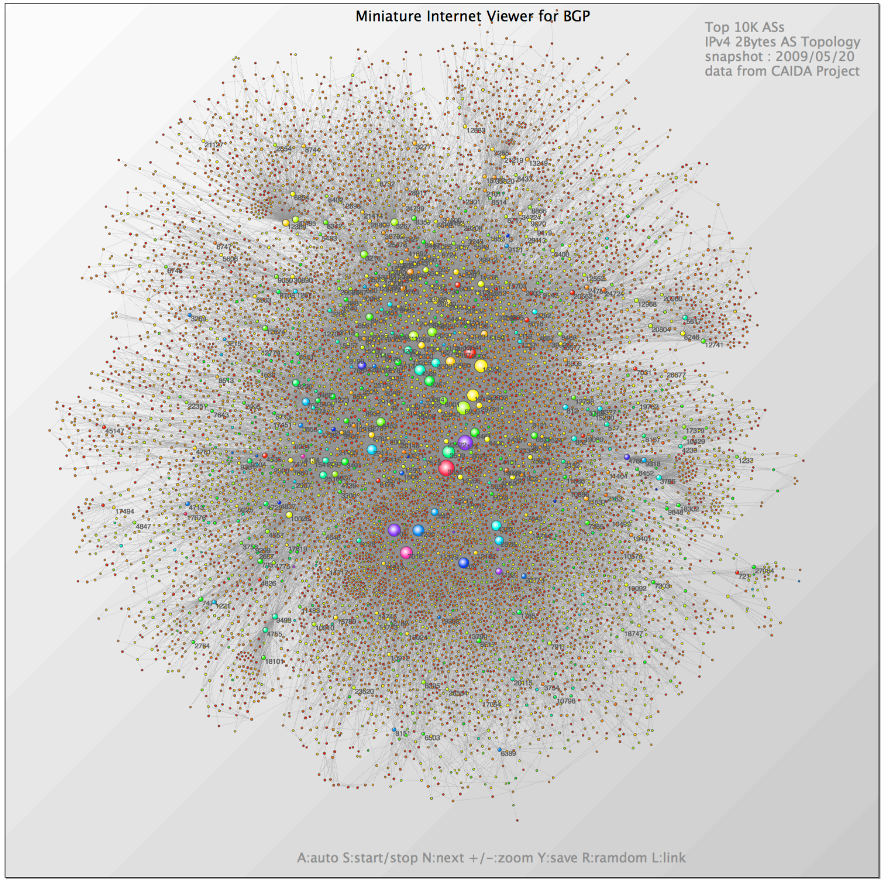




The current Maximum scale

- **Basic spec.**
  - 10K inter-AS topology on 150 physical nodes
    - Top 10K from CAIDA AS Ranking
  - About 75 ASs per 1 physical node

- **Tune-up !**
  - domU base memory size is up to 72MB per 25 neighbor Ass
    - Estimated by experience
  - No swap memory
    - to get good response
  - Linux Kernel tuning for dom0
    - SOMAXCONN, FD_SETSIZE, INR_OPEN, NR_OPEN
  - Linux Kernerl tuning for dom0
    - vIRQ size, etc.
  - Locating Core AS nodes onto physical nodes


Miniature Internet Viewer for BGP
Top 10K ASs
IPv4 2Bytes AS Topology
snapshot : 2009/05/20
data from CAIDA Project
A:auto S:start/stop N:next +/-:zoom Y:save R:ramdom L:link

The current Maximum Scale (cont.)

- **We did it, but it is not stable yet**
  - Message storms
    - ARP Storm
    - BGP OPEN and RESET Storm on initial booting
    - Broadcasting due to the overflow of FDB on L2 switches
    - Small / middle class transit AS's bgpd daemon is easily killed by linux oom-killer
  - The allocated memory is enough to store full routes of the 10k topology,
  - but it is not enough to handle BGP message storms, especially at the initial booting stage

The current Maximum Scale (cont.)

- **To be stable**
    - According to our experience, we need a few days
        - Boot domU slowly, insert enough interval between each domU for waiting calculation of route changes and for avoiding route flapping
    - Will more memory give stability ???
        - Allocate additional memory to tranist AS domU nodes for message storm
        - We don’t know the details, yet :-p
            - Under construction for tomorrow demonstration

Future Work

- Solve Some Technical Issues to scale up
- Emulating Other Parts of the Internet
- Stability / Fidelity / Optimization study

Conclusion

- **Our tools for constructing emulated inter-AS topology on a testbed could successfully emulate a part of the Internet**
- **We are now working on:**
  - Health check/management of virtual node
  - Study of resource allocation algorithm
  - Emulating intra-AS OSPF network
- **Demo. "Emulating over 10K AS topology with massive VM multiplexing" will be made**
  - at 12:30 on Tuesday 2009-08-18
  - now we are constructing on the remote testbed (StarBED)
  - You can see emulated 10K AS topology via our viewer

FYI.

- AnyBed
  - `http://sourceforge.net/projects/anybed/`
- XENebula
  - `http://tbn.starbed.org/XENebula/`
- Contact
  - Please send an e-mail to `nerdbox-freaks@wide.ad.jp`
  - `hiroa-ha@is.naist.jp, danna@nict.go.jp`

After demo session …..

- **101 bgpd nodes haven’t booted up yet !!!**
  - 7 nodes doesn’t any response from network
  - In other 94 nodes, linux kernel remounted /dev/sda1 before starting bgpd.
    - linux kernel remounted /dev/sda1 due to Buffer I/O error caused by the netfront overflow of domU
    - So, bgpd could not create /var/run/bgpd.pid
  - Other booted bgpds still throw reset messages each other ….
    - What is wrong ????

Conclusion (updated)

- **We are still struggling to boot up all 10k bgpd nodes**
  - Long long road to get a controllable miniature of the Internet