

# CIS 700/005

## Networking Meets Databases

Boon Thau Loo  
Spring 2007  
Lecture 15

(Presentation courtesy of slides from Matt Caesar (SIGCOMM) and Jelger/Tschudin (Dagstuhl seminar))

### Announcement

- 2 minute informal spiel on class project on Thursday
- Will miss class on:
  - Apr 5 : DARPA meeting
  - Apr 10: NetDB/NSDI 2007.

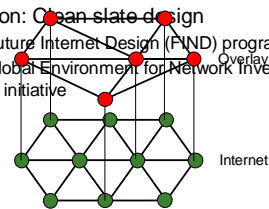
### Era of change for the Internet

*"in the thirty-odd years since its invention, new uses and abuses, ..., are pushing the Internet into realms that its original design neither anticipated nor easily accommodates...."*

Overcoming Barriers to Disruptive Innovation in Networking,  
NSF Workshop Report '05

### Efforts at Internet Innovation

- Evolution: Overlay Networks
  - Commercial (Akamai, VPN, MS Exchange servers)
  - P2P (filesharing, telephony)
  - Research prototypes on testbed (PlanetLab)
- Revolution: Clean-slate design
  - NSF Future Internet Design (FIND) program
  - NSF Global Environment for Network Investigations (GENI) initiative



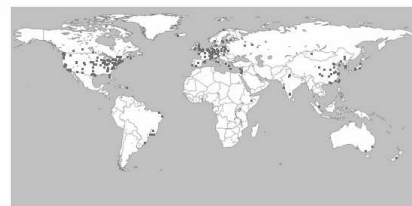
### Context of ROFL paper

- Up to this point in class:
  - Distributed Hash tables
    - Network Data Independence -> Decouple location from identity
  - Overlay networks
    - Internet Indirection Infrastructure (i3)
  - Active networks
    - Extreme solution
  - Declarative networks / PLAN / ANTs
    - "Safe languages" of active networks
  - Clean-slate radical architectural changes
    - TRAIID, IPNL, LFN/DOA, HIP, SNF, ROFL.....

#### PlanetLab

PlanetLab is a global research network that supports the development of new network services. Since the beginning of 2003, more than 1,000 researchers at top academic institutions and industrial research labs have used PlanetLab to develop new technologies for distributed storage, network mapping, peer-to-peer systems, distributed hash tables, and query processing.

PlanetLab currently consists of 754 nodes at 363 sites.



<http://www.planet-lab.org/>

## GENI Global Environment for Network Innovations

**Home**

Overview

FAQ

**Research Opportunities**

Challenges

Questions

Success

**Key Concept**

Goals

Requirements

Snapshot

**Community**

Discussion List

Meetings

Working Groups

Broad Participation

**References**

Design Documents

**What is GENI?**

Global Environment for Network Innovations (GENI) is a facility concept being explored by the US computing community. The goal of GENI is to increase the quality and quantity of experimental research outcomes in networking and distributed systems, and to accelerate the transition of these outcomes into products and services that will enhance economic competitiveness and secure the Nation's future. Ultimately, research performed on GENI is expected to lead to a transition of the Internet as we know it today.

With the support of the NSF Directorate for Computer and Information Science and Engineering (CISE), the computing community is currently developing the conceptual design of the facility. A number of workshops have already taken place, and under the leadership of a planning group an initial strawman design has been completed. This design is being shared broadly herein for comment. CISE will support a number of town hall meetings in the early spring of 2006 to gather community input that further informs and refines the design.

Using a competitive merit review process, CISE is planning to support the establishment of a GENI Coordinating Consortium (GCC) in 2006. The GCC will represent the computing community's broad research interests in the GENI facility.

This web site is currently maintained by the GENI planning group as a forum for discussing GENI, which will in

<http://www.geni.net> (\$300 million plan)

## STANFORD UNIVERSITY

### Clean Slate Design for the Internet

An Interdisciplinary Research Program at Stanford University

**Home**

**Weekly Seminars**

**Past Seminars**

**Submit A Proposal**

**Projects**

**Faculty**

**Companies**

**Links**

**Contact**

**Retreats**

**Overview**

We believe that the current Internet has significant deficiencies that need to be solved before it can become a unified global communication infrastructure. Further, we believe the Internet's shortcomings will not be resolved by the conventional incremental and backward-compatible style of academic and industrial networking research. The proposed program will focus on unconventional, bold, and long-term research that tries to break the network's ossification. To this end, the research program can be characterized by two research questions: "What do we know today, if we were to start again with a clean slate, how would we design a global communications infrastructure?" and "How should the Internet look in 15 years?" We will measure our success in the long-term. We intend to look back in 15 years time and see significant impact from our program.

In the spirit of past successful interdisciplinary research programs at Stanford, the program will be driven by research projects from the ground up. Rather than build a grand infrastructure and tightly coordinated research agenda, we will create a loosely-coupled breeding ground for new ideas. Some projects will be very small, while others will involve multiple researchers, our goal is to be flexible, creating the structure and identifying and focusing funds to support the best research in clean-slate

<http://cleanslate.stanford.edu>

## 100x100 Clean Slate Project

**The Project**

- ↳ Mission
- ↳ FAQ
- ↳ Background

**Participants**

- ↳ Researchers
- ↳ Institutions
- ↳ Internal site

**Communications**

- ↳ News Room
- ↳ Papers
- ↳ Project Retreat
- ↳ Presentations

**100 Mbps for 100 million American homes!!**

**100x100 Clean Slate Project**

The Internet is one of the most successful technology achievements. In less than 30 years, the Internet has grown from a small experimental network that served as a playground for researchers to a global infrastructure that connects hundreds of millions of people. IP, the technical foundation of Internet, is widely regarded, by both the general and technical communities, to be the convergence technology layer for all communication infrastructures and services. To date, network researchers have focused on solutions that incrementally improve the Internet with the implicit assumption that radical new solutions are not needed or have no chance of ever being deployed.

This project takes the opposite approach. We ask: **Given the benefit of hindsight and our current understanding of network requirements and technologies, if we were not bound by existing design decisions and would be able to design the network from first principles (a clean slate design), how should we do it?**

The scope of the 100x100 Clean Slate Design Project is necessarily broad, as the design of

<http://www.100x100network.org/>

## One of the Key Goals of "Clean-slate": Evolution of Networks

- How much of an architecture can evolve over time?
  - hypothesis: no more "one size fits all" network
  - what must remain stable? what is the least common denominator, i.e.
  - the meta-arch?

- Common question: must addresses be explicitly defined/supported by a network??
  - architecture? (should it be address-centric?)

Jelger/Tschudin – Dagstuhl Seminar – Oct/Nov. 2006

## When Addresses rule the Net

- In the ARPANET, addresses were fixed and could actually be used as "long-lived names"
  - Address 1 is UCLA-NMC [RFC-597, 1973]
  - HOST.TXT file is mapping a static symbol into another static symbol
- Same with the early Internet
  - 1.0.0.0/8 is BBN-PR [RFC-820, 1982]
- After the introduction of DNS in 1984 addresses are still very long-lived:
  - RFCs still contain the static list of assigned addresses
  - estimated number of hosts = 1,000
- RFC-990 (1986) is the last one to contain a list of assigned addresses
  - estimated number of hosts <= 10,000

Jelger/Tschudin – Dagstuhl Seminar – Oct/Nov. 2006

## Internet Addressing Scheme (Review)

- Class-based addressing schemes:
  - 32 bits divided into 2 parts:
 

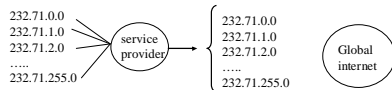
0	8	network	host	{	126 nets
					~16M hosts
  - Class A
 

0	16	network	host	{	~16K nets
					~65K hosts
  - Class B
 

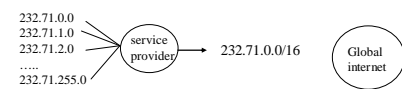
0	24	network	host	{	~2M nets
					254 hosts
- Classless Inter-domain Routing (CIDR)
  - Variable prefix – prevents wastage
  - Prefix aggregation – lower overhead of Inter-domain routing
  - Routers match to longest prefix

## Scalability: Routing Table Size

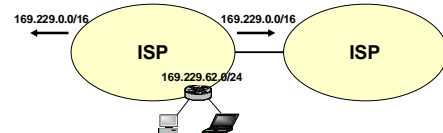
### Without CIDR:



### With CIDR:



## What's wrong with Internet addressing today?

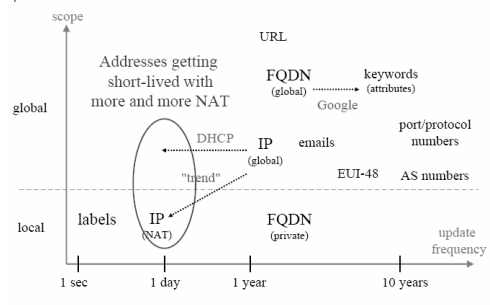


- Hierarchical addressing allows excellent scaling properties
- But, forces addressing to conform to network topology
- Since topology is not static, addresses can't persistently identify hosts
  - Host can have multiple locations (mobility)
  - Host can have multiple identities (multi-homing)

## What's wrong with Internet addressing today?

- When the Internet was first invented:
  - Nodes are mostly static, so location = identity.
  - Easy to impose hierarchy
- ...But most network applications today require persistent identity in the presence of changing/multiple locations.
- It's hard to provide persistent identity in presence of hierarchical addressing
  - Need to decouple identity from addressing
  - Drastically complicates network configuration, mobility, address assignment
- Seems difficult to achieve...

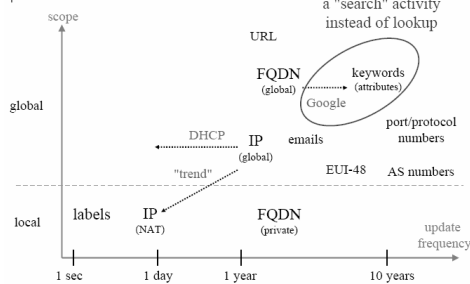
### Two trends (1)



Jeiger/Tschudin - Dagstuhl Seminar - Oct/Nov. 2006

Jeiger/Tschudin - Dagstuhl Seminar - Oct/Nov. 2006

### Two trends (2)



Jeiger/Tschudin - Dagstuhl Seminar - Oct/Nov. 2006

Jeiger/Tschudin - Dagstuhl Seminar - Oct/Nov. 2006

## Solutions we have studied...

- Level of indirection
  - Location-identity split
  - Name resolution
    - Route using a name
    - Name turns into location IP address
    - Location is ephemeral, name is long-term identifier
  - E.g. Mobile IP, Internet Indirection Infrastructure

## ROFL: Is there an alternative?

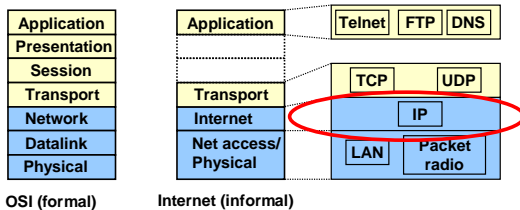
- Why not route on “flat” host identifiers?
  - Assign addresses independently of network topology
  - Get rid of location and hierarchy altogether!!
  - *Be a Purist*: Architectural uniformity to the extreme.
- Benefits:
  - No separate name resolution system required
  - Simpler network config/allocation/mobility
    - Allocation of identifiers need only ensure uniqueness
    - Need not worry about adherence to network topology
  - Simpler (more meaningful) network-layer access controls
    - Identifier-based instead of IP-based

## Is it possible to route on flat identifiers?

- Challenge: flat identifiers break aggregation
- Is it possible to scalably route without aggregation?
  - Paper is more about addressing the feasibility of flat naming without hierarchy, and generating discussions.
  - “Postman delivers using SSN, not address”
- Not a straightforward application of DHTs
  - Assumes point-to-point routing. Cannot address the problem of building “from scratch” a network
  - Doesn’t support routing policies

## OSI vs. Internet

- OSI: conceptually define services, interfaces, protocols
- Internet: provide a successful implementation



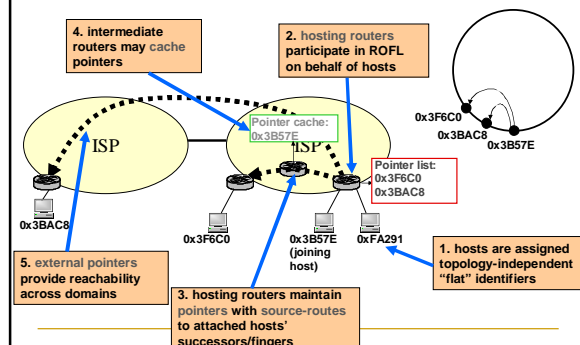
## Basic idea behind ROFL

- Scalable routing on flat identifiers
- Goal #1: Scale to Internet topologies
- Goal #2: Support for BGP policies
- Highly challenging problem, but solution would arguably give a number of benefits

## Basic mechanisms behind ROFL

- Goal #1: Scale to Internet topologies
  - Mechanism: DHT-style routing, maintain source-routes to successors/fingers
  - Provides: Scalable network routing without aggregation
- Goal #2: Support for BGP policies
  - Mechanism: Intelligently choose successors/fingers to conform to ISP relationships
  - Provides: Support for policies, operational model of BGP

## How ROFL works



### Internet policies today

- **Economic relationships:** peer, provider/customer
- **Isolation:** routing contained within hierarchy

### Canon DHT (background)

- Basis for Inter-domain routing in ROFL

- Merge rings in a bottom-up fashion with two conditions.
- For an identifier  $id_a$  ring 1 with external pointer to  $id_b$  in ring 2:
  - No identifiers between  $[id_a, id_b]$
  - $id_b$  will be  $id_a$ 's successor
  - $O(\log n)$  number of pointers

### Isolation in ROFL (Canon)

→ Traffic between two hosts traverses no higher than their lowest common provider in the AS hierarchy

### Policy support in ROFL

Goal: prefer peer route over provider route

Mechanism: Convert peering relationships to Virtual ASes

- Peering
- Provider-customer
- Backup

→ Traffic respects peering, backup, and provider-customer relationships

### Scalability in ROFL

- Two extensions to improve locality:
  - Maintain proximity-based fingers in a policy-safe fashion
  - Pointer caching strategies: prefer nearby, popular pointers

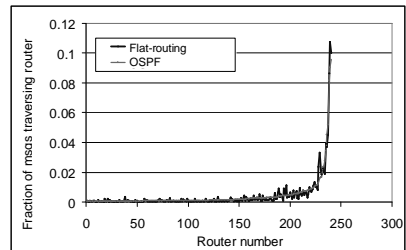
### More Routing Concerns

- Routing control
  - Inter-domain: endpoint-based negotiation
  - Intra-domain: leverage the inter-domain design (Isolation property)
- Enhanced delivery services
  - Anycast
  - Multicast
- Security
  - Default off (Hotnets '05) – Hosts reachable only from fingers
  - Use capabilities (a capability is a cryptographic token designating that a particular source is allowed to contact the destination)

## Evaluation

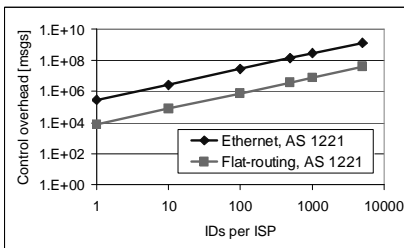
- Distributed packet-level simulations
  - Deployed on cluster across 62 machines, scaled to 300 million hosts
  - Inferred Internet topology from Routeviews, Rocketfuel, CAIDA skitter traces
- Implementation
  - Ran on Planetlab as overlay, covering 82 ASes
  - Configured inter-ISP policies from Routeviews traces
- Metrics: stretch, control overhead, router memory usage

## Enterprise/ISP simulations: Load balance



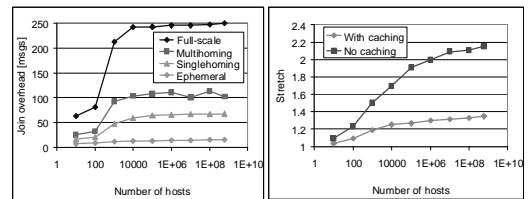
- No "hot spots" introduced by protocol

## Enterprise/ISP simulations: Reduction in churn



- 37 to 181x less control overhead
- 34-1200x less memory requirements

## Internet-scale simulations



- Join overhead <300 msgs, stretch < 1.4
- Root-server lookups inflate latency from 54ms to 134ms, Flat IDs has no penalty

Metric	ROFL	BGP+DNS
Join overhead	450 per-host "lightweight joins": 14 per-host	typically 0 per host 40,000 per prefix
Latency	Baseline: 135ms With pointer caches: 70ms	With lookup: 137ms No lookup: 54ms
State	2 million pointer cache entries in core, ~100 entries at edge	150 thousand entries at core and edge 77 million DNS entries at root servers
Failure	Cached pointers equivalent of backup paths, failures only affect successors and fingers	Undergoes convergence process, failures have global impact

## Conclusion

- Routing On Flat Labels should not leave you Rolling On the Floor Laughing
  - Because performance is tolerable
  - Because it provides several benefits
- ROFL is one point in the design space
  - ROFL as lookup
  - ROFL for content routing
    - Better support for ONYX, INS, PIER, P2, ... ?
    - "The results are close enough to tempt, but not enough to satisfy"
- Later in the semester:
  - Virtual Ring Routing – DHT on wireless networks
    - Similar to ROFL's Inter-domain design

