

An Overlay Control Plane for Wide Area Routing

Sharad Agarwal

Chen-Nee Chuah, Randy H. Katz

{sagarwal,randy}@eecs.berkeley.edu, chuah@ece.ucdavis.edu.

Outline

- Introduction
 - Problem statement
 - Related work : inadequate solutions
- OPP architecture
 - Overview
 - Completed work : AS relationship and topology map
 - Evaluation

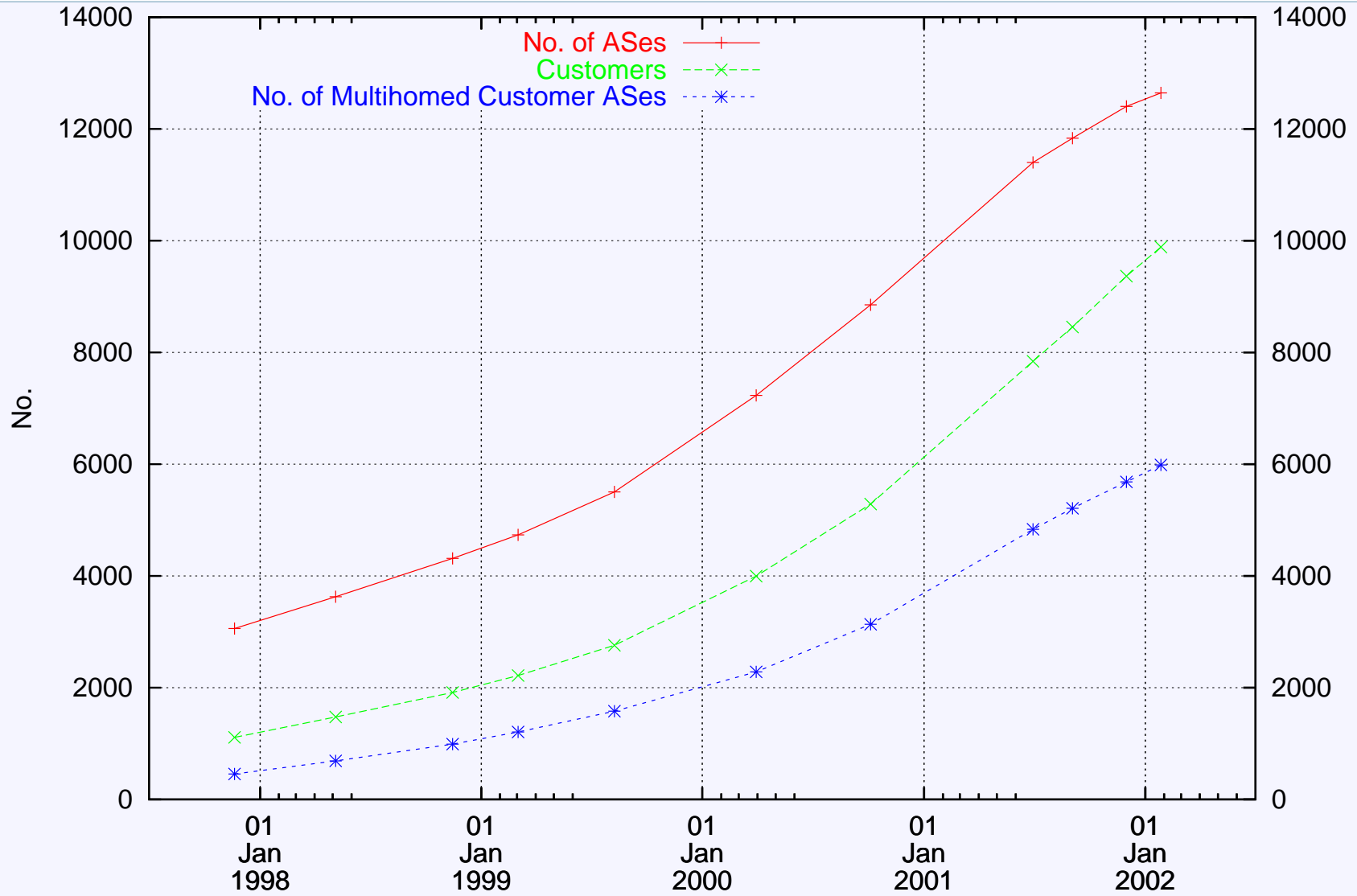
Introduction

- Internet composed of 13,000 domains (AS's)
 - Large ISPs like Sprint, AT&T
 - Small orgs like UC Berkeley
- Connected via inter-domain routing (BGP)

Introduction

- Internet composed of 13,000 domains (AS's)
 - Large ISPs like Sprint, AT&T
 - Small orgs like UC Berkeley
- Connected via inter-domain routing (BGP)
- Recent growth in BGP participation
 - Internet connectivity more important
 - Specifically
 - Adaptive to congestion & failures

Multihoming of Customer ASes



■ Large AS growth due to multihomed leaf AS's

■ Multihoming is \$\$

[<-->]

Multihoming

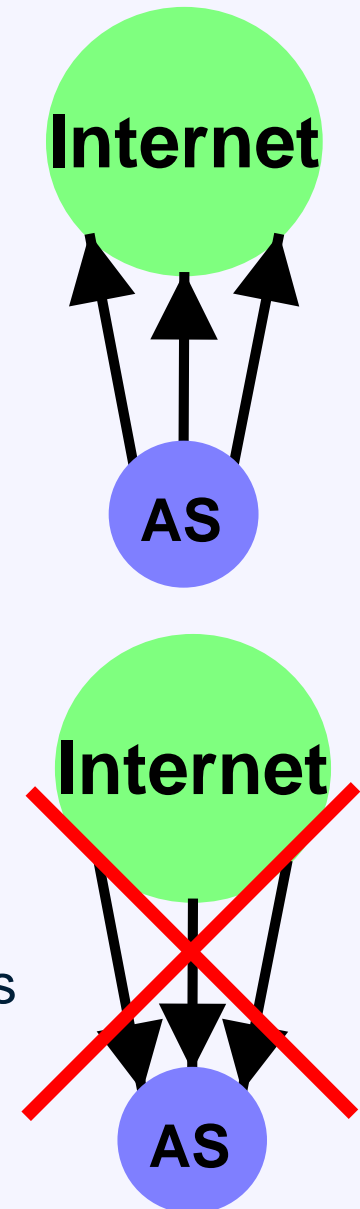
- Failover
 - Primary provider + redundant access links
 - However, limited by BGP : ~15 minutes

Multihoming

- Failover
 - Primary provider + redundant access links
 - However, limited by BGP : ~15 minutes
- Traffic load balancing
 - Outgoing traffic
 - Use smart BGP route selection
 - See papers by J. Rexford
 - See Routsience et al

Multihoming

- Failover
 - Primary provider + redundant access links
 - However, limited by BGP : ~15 minutes
- Traffic load balancing
 - Outgoing traffic
 - Use smart BGP route selection
 - See papers by J. Rexford
 - See Routsience et al
 - Incoming traffic
 - Not possible today ... (sort of)
 - Can pollute BGP with weird routes
 - Local optimizations have global ramifications
 - Can't scale, not effective enough



Problem Statement

■ Goals

- Improve fail over time from ~15 minutes
- Improve time to shift incoming traffic between paths
 - Current BGP techniques offer no control
- Secondary goals
 - May reduce routing table growth
 - May reduce route flapping

Problem Statement

■ Goals

- Improve fail over time from ~15 minutes
- Improve time to shift incoming traffic between paths
 - Current BGP techniques offer no control
- Secondary goals
 - May reduce routing table growth
 - May reduce route flapping

■ Constraints

- Coexist with deployed IGP/EGP
- Allow incremental deployment
 - Incremental replacement of BGP
- Detect & avoid oscillations, divergence due to conflicts
- Be scalable

Related Work

- Limit prefix length, NOPEER, flap limiting
 - Don't solve underlying issue

Related Work

- Limit prefix length, NOPEER, flap limiting
 - Don't solve underlying issue
- MPLS / DiffServ based Intra-domain TE solutions
 - Would follow BGP routes
 - We don't expect open MPLS clouds everywhere

Related Work

- Limit prefix length, NOPEER, flap limiting
 - Don't solve underlying issue
- MPLS / DiffServ based Intra-domain TE solutions
 - Would follow BGP routes
 - We don't expect open MPLS clouds everywhere
- RON, Routing Arbiter, Nimrod
 - Unscalable in our scenario

Outline

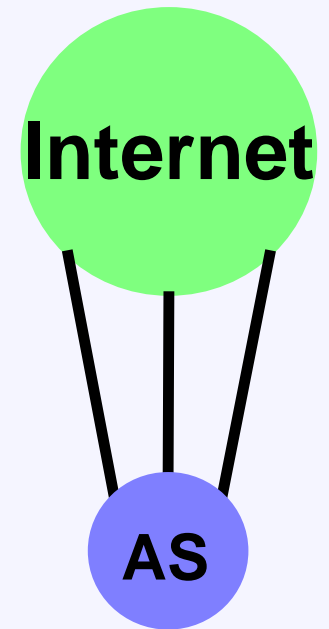
- Introduction
 - ~~Problem statement~~
 - ~~Related work : inadequate solutions~~
- OPP architecture
 - Overview
 - Completed work : AS relationship and topology map
 - Evaluation

Challenges

- How to design routing control structure?

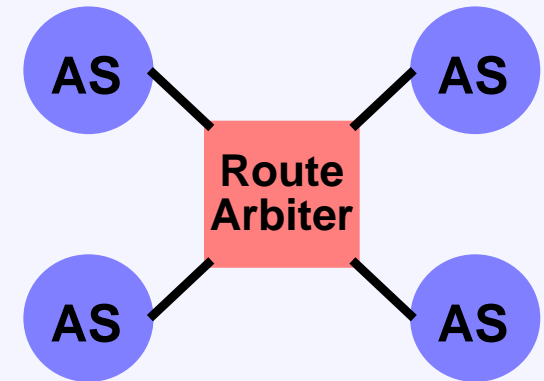
Challenges

- How to design routing control structure?
 - Local optimization isn't enough
 - Locus of control is remote



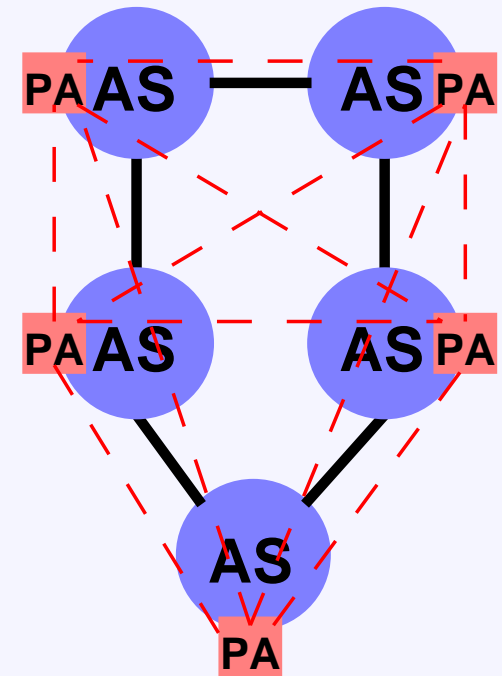
Challenges

- How to design routing control structure?
 - Local optimization isn't enough
 - Locus of control is remote
 - Global optimization unattainable
 - Computationally complex
 - Link state
 - Scalability is an issue
 - Full disclosure of policies bad

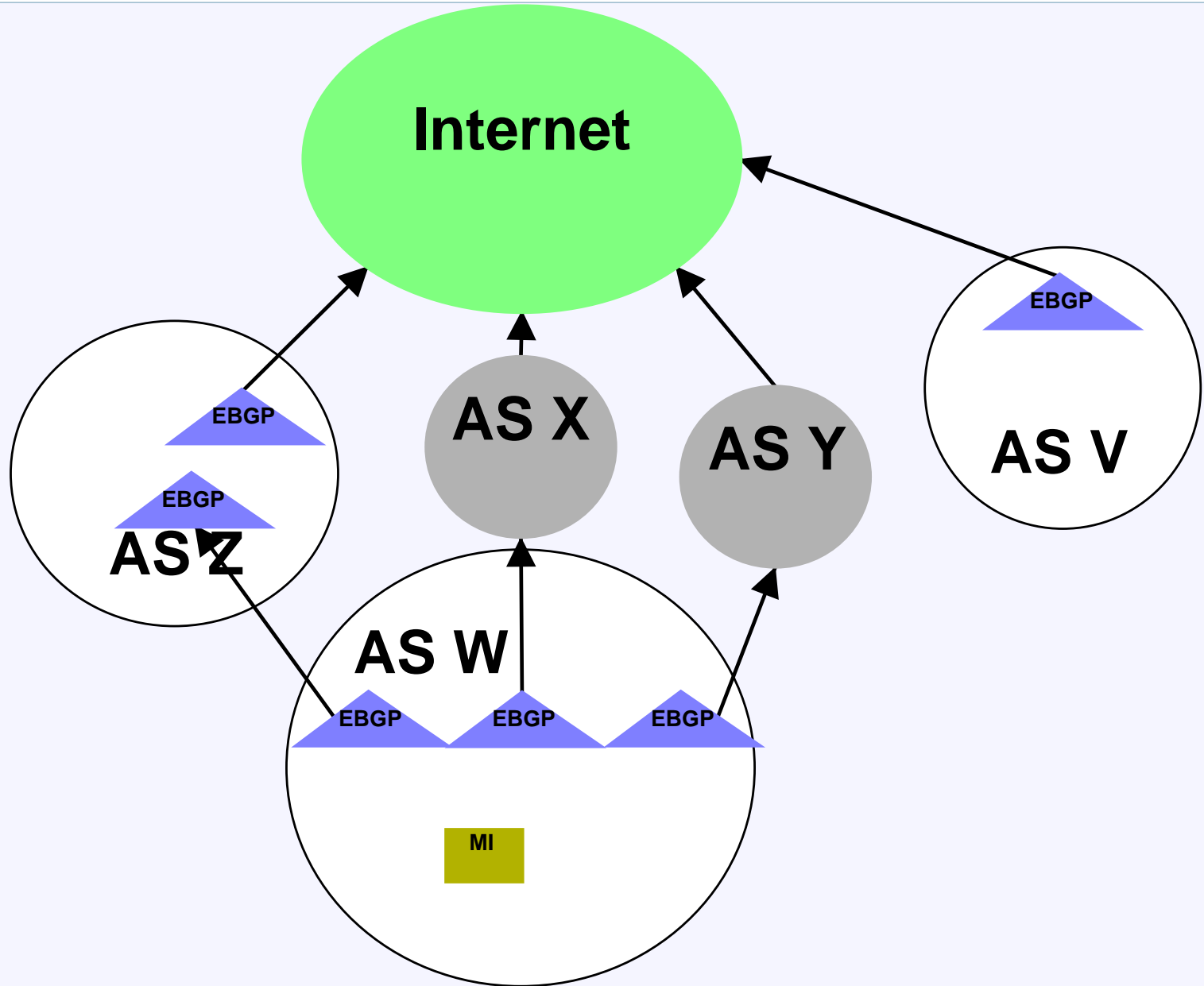


Challenges

- How to design routing control structure?
 - Local optimization isn't enough
 - Locus of control is remote
 - Global optimization unattainable
 - Computationally complex
 - Link state
 - Scalability is an issue
 - Full disclosure of policies bad
 - Middle ground
 - Logically separate routing control plane
 - Find loci of control
 - Negotiate policy control
 - Adapt to non-responsiveness, network change

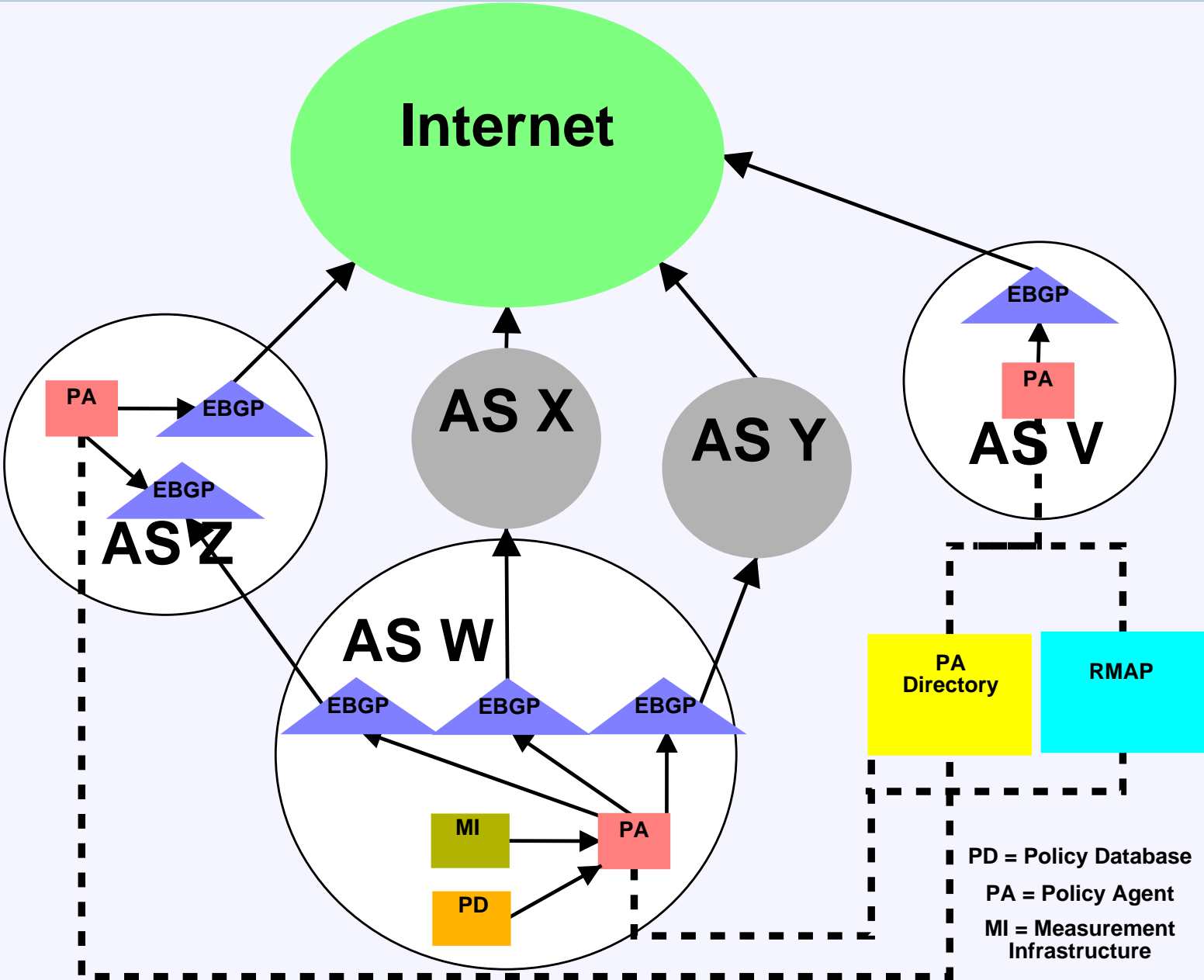


OPP Architecture



[<-->]

OPP Architecture



[<-->]

Components of OPP

- Policy database
 - Important AS's (e.g. \$\$ customers)
 - Local application servers & mirrors
 - SLAs & pricing constraints
 - Objective function

Components of OPP

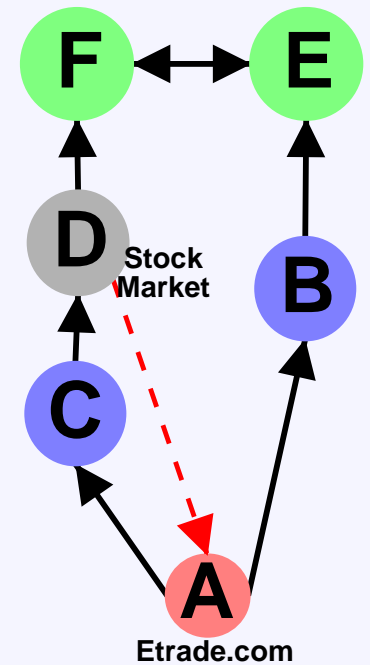
- Policy database
 - Important AS's (e.g. \$\$ customers)
 - Local application servers & mirrors
 - SLAs & pricing constraints
 - Objective function
- Measurement infrastructure
 - Already exists in most AS's
 - E-BGP link info, SNMP traffic

Components of OPP

- Policy database
 - Important AS's (e.g. \$\$ customers)
 - Local application servers & mirrors
 - SLAs & pricing constraints
 - Objective function
- Measurement infrastructure
 - Already exists in most AS's
 - E-BGP link info, SNMP traffic
- Distant PAs
 - 1 PA directory or many (e.g. DNS)
 - Some PAs uncooperative; check MI
 - Some requests may conflict

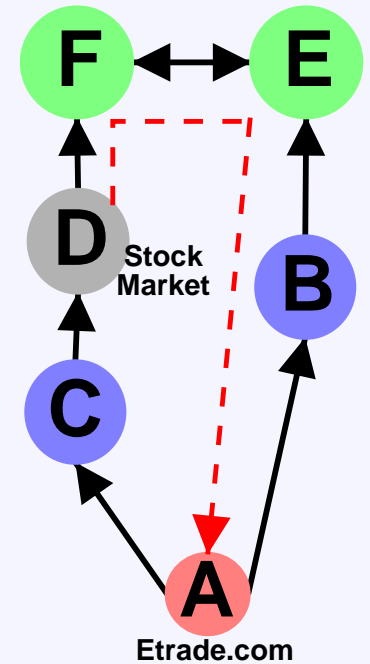
Components of OPP

- Relationship & Topology Map
 - 1 RMAP or many
 - Find likely route, transit / peering relationships



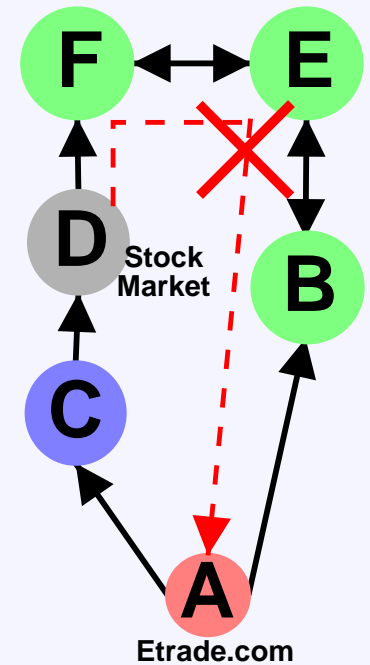
Components of OPP

- Relationship & Topology Map
 - 1 RMAP or many
 - Find likely route, transit / peering relationships
 - A wants D to use path through F
 - A may need to contact F
 - F sends route to D



Components of OPP

- Relationship & Topology Map
 - 1 RMAP or many
 - Find likely route, transit / peering relationships
 - A wants D to use path through F
 - A may need to contact F
 - F sends route to D
 - However, if E & B are peers instead
 - Wouldn't work



RMAP : Results

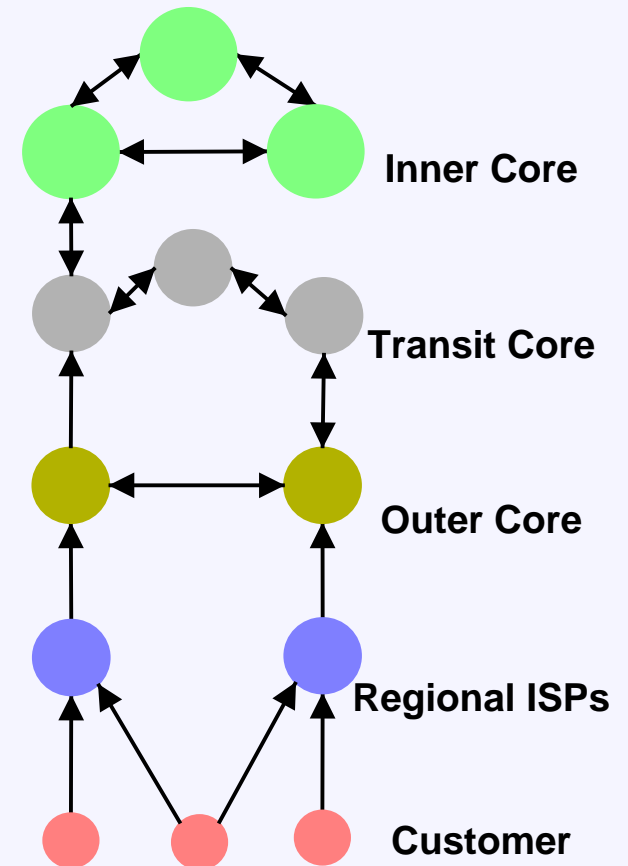
Inferred Relationships for 23,935 AS Pairs

Relationship	# AS pairs	Percentage
Provider-customer	22,621	94.51%
Peer-peer	1,136	4.75%
Unknown	178	0.74%

Distribution of AS's in Hierarchy

Level	# of AS's
Inner core (0)	20
Transit core (1)	129
Outer core (2)	897
Regional ISPs (3)	971
Customers (4)	8898

- INFOCOM 2002
- *“Characterizing the Internet Hierarchy from Multiple Vantage Points”*
- L. Subramanian, S. Agarwal, J. Rexford, R. H. Katz



Research Issues

- Goal
 - Reduce fail over time, finer grained load balancing

Research Issues

- Goal
 - Reduce fail over time, finer grained load balancing
- Measure side effects
 - Table growth, flapping, traffic, scalability

Research Issues

- Goal
 - Reduce fail over time, finer grained load balancing
- Measure side effects
 - Table growth, flapping, traffic, scalability
- Deployment
 - Cooperative architecture, like BGP
 - Keep history of uncooperating PAs
 - Distribution of PAs
 - Benefits leaf AS's
 - But need PA's in core (at aggregation points)
 - Leaf AS's are customers of core
 - Large benefits will create pressure
 - More participants, better RMAP

Evaluation Methodology

- Analytical
 - Use real topologies, real BGP tables
 - Potential improvement
- Implement and emulate
 - Code complete PA, PD
 - Millennium cluster, run SW BGP speakers (GNU Zebra)

Summary

- Hypothesis
 - Available, congestion adaptive connectivity is lacking
 - An overlay control plane can achieve this
- Many interesting research issues
 - How to balance local optimization and global optimization
 - Fail over time, load balancing, traffic impact, scalable, deployment, ...
- Measureable success
 - Real BGP tables and traffic patterns
 - Real BGP implementations in emulation testbed