Inter-domain Routing

Ken Keefe
Jon Osting
Nathaniel Cook

CS 538 - February 9, 2016

Historical context

- Original ARPAnet (late 1960s - late 1980s)
  - Global policy, single routing protocol
- CSNET (1981), NSFNET (1985), ESnet (1986), etc
  - Different network policies, different routing protocols
  - Development of backbone services
    - 56K network
    - T1 Network
    - T3 Network
- Privatization of the internet (1995+)
  - Backbone developed by NSF moved to commercial providers
  - Establishment of Network Access Points (NAPs)
- Modern Internet is a huge, complex ecosystem
  - Large international ISPs
  - Regional/Local ISPs
  - Content Providers/hosting

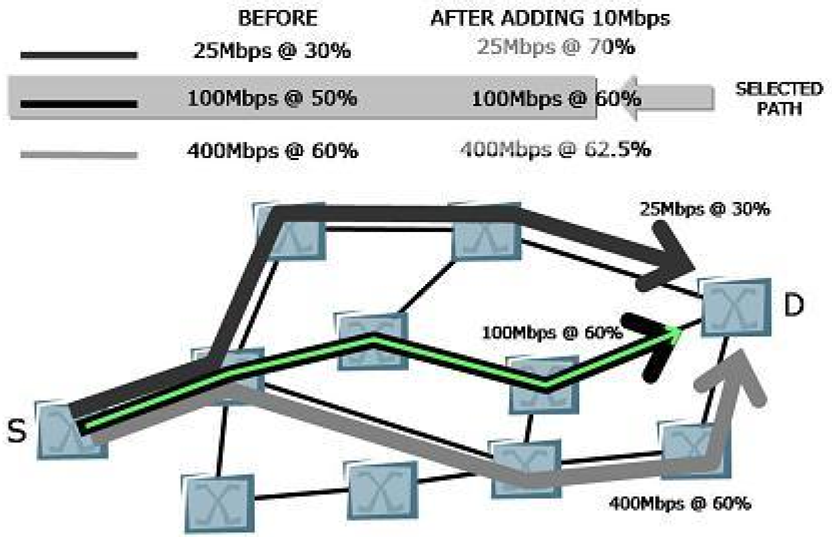
Key Questions / Issues

How to move data across the internet?

- **Physical Infrastructure**
  - Routers
  - Lines
- Routing Protocols/Addressing
- **Network Topology**
- Business relationships & economic costs
- Efficiency
- Reliability/Robustness
- **Traffic engineering**

Traffic Engineering

- Minimize the maximum utilization of a network
- Given multiple parallel paths, devise a strategy to balance traffic
- Objective is to provide
  - Reliability
  - Performance
- Current methods fall into two categories:
  - Offline
  - Online
- Shortest path is not always the best policy.

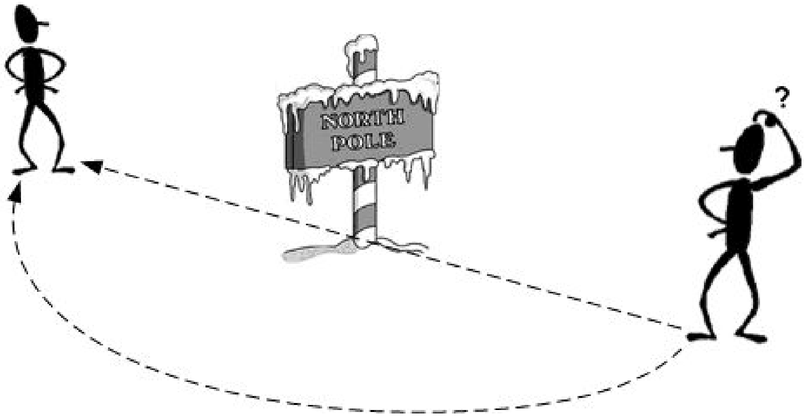


Offline Methods

- Traffic policies do not adapt to current conditions or factors such as
  - Time of day
  - Traffic type
  - Attacks
  - Network path changes
- Examples include
  - OSPF
  - MPLS Multi-commodity Flow Optimizer
- Inefficient because
  - Uses long-term behavior of traffic to set long-term policies
  - Tries to handle network failure by precomputing plans for some failure scenarios

Open Shortest Path First

- Builds topological graph from available routers
- Uses Dijkstra's algorithm to find shortest path tree for each route
- Handles failures by looking at next shortest path in the tree

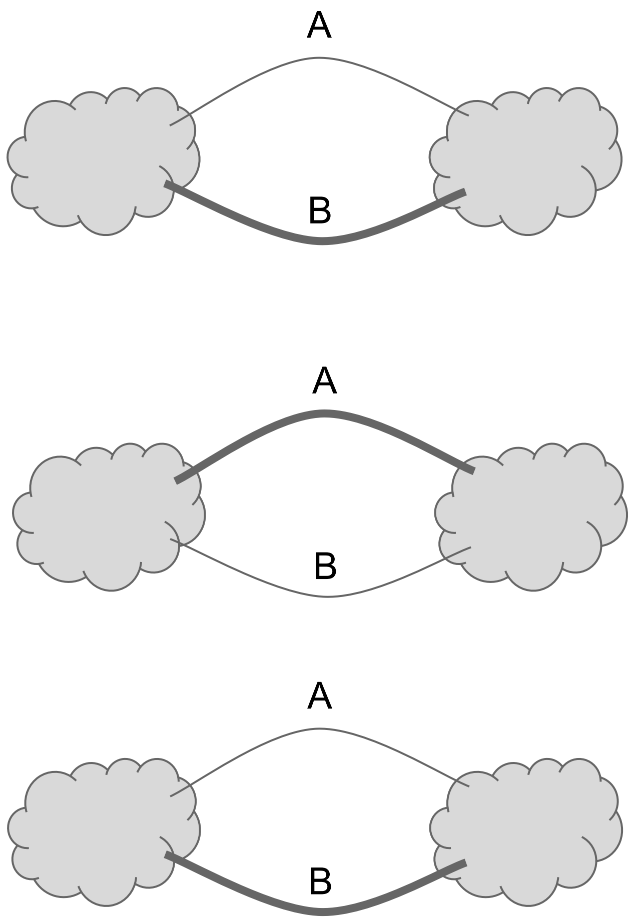


Multi-Protocol Label Switching

- Protocol independent
- Data packets are assigned labels that correspond with paths through the network instead of end points
- Routers follow the path prescribed in the label
- Traffic can be given priorities
- Paths are determined offline to build a pool of labels

Online Methods

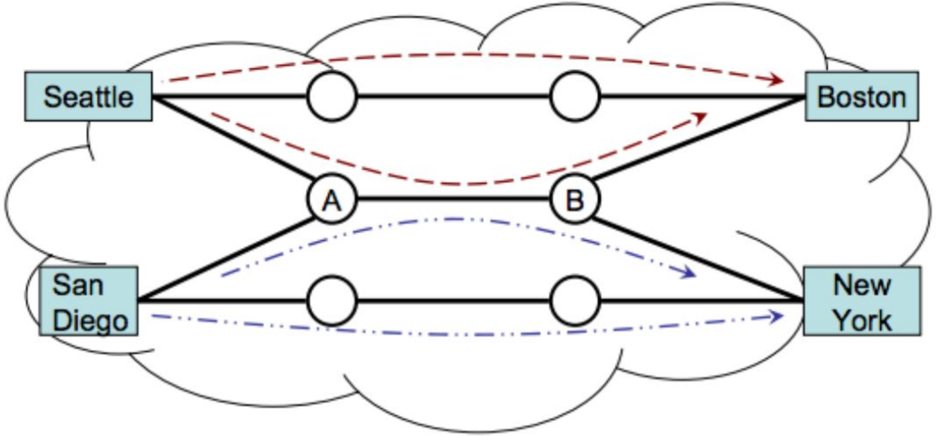
- Respond to traffic data by altering policy
- Must avoid
  - oscillations
  - network instability caused by online traffic engineering
- Approaches:
  - Central Authority
  - Central Oracle / Distributed Authority
  - Distributed
    - TeXCP



Problem Formalization

- Optimization problem:

$$\min_{x_{sp}} \max_{l \in L} u_l,$$



- $\{x_{sp}\}$ - traffic split ratios
- L - all links for IE pair s on path p
- u_l - utilization of link l

  - Discussion: Is this the right optimization? What about latency? Security?

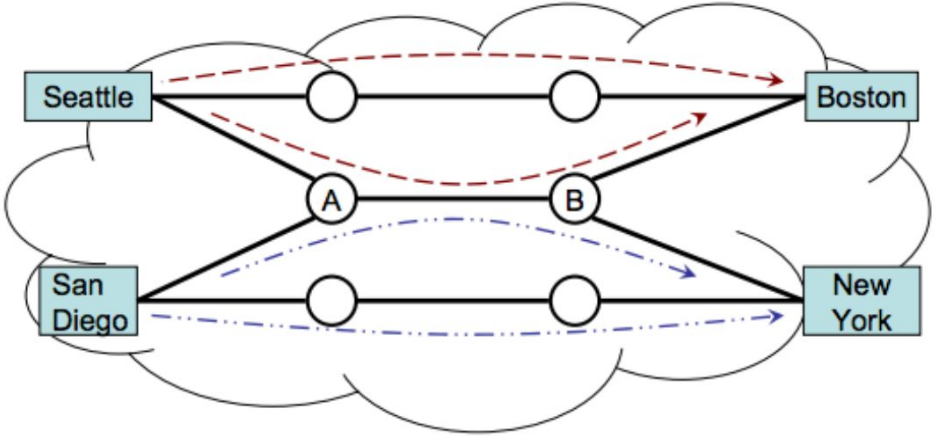
TeXCP

- Two pieces
  - Load Balancer
  - Feedback controller
- Load balancer
  - Uses knowledge of network state
  - Balances traffic among paths to minimize utilization
- Feedback controller
  - Each path has a controller
  - Tracks traffic to ensure
    - Performance
    - Stability
  - Works at a faster time scale than the Load Balancer
- These two components feed input to each other

Images: http://storyglitz.com/wp-content/uploads/2015/01/balancing-and-carrying-bricks.jpg
https://www.cxcompany.com/content/original/Blogposts/feedback2.jpg

TeXCP Operation

- Ingress router in each IE pair runs a TeXCP agent
- Agent is supplied with a set of paths
- Agent probes each path to discover utilization and failure state
- Agent splits traffic across paths to minimize the max-utilization
- Agents update in real time
- Agents do not communicate with each other

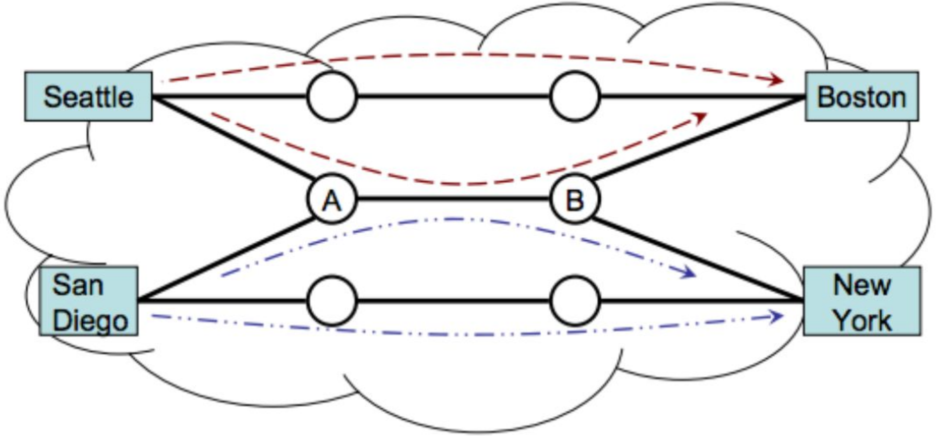


- BGP manages IP prefix to egress point map

- Discussion: What could be gained from a little data sharing? What would agents share?

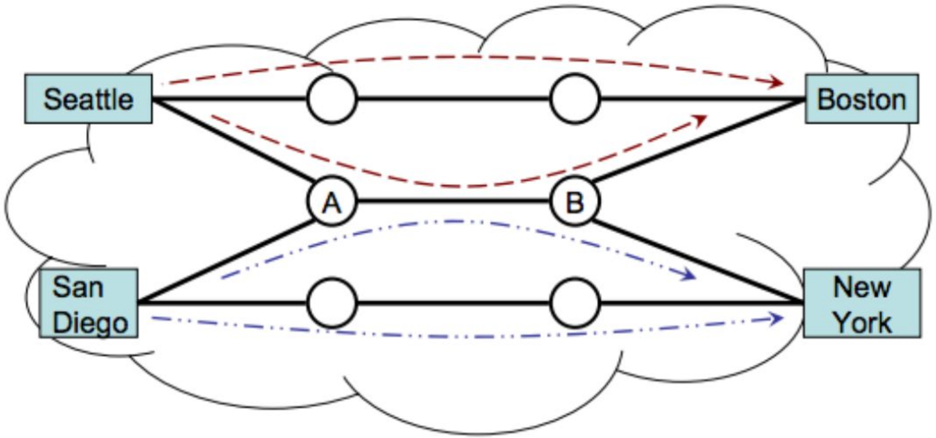
Path Selection

- Paths are selected based solely on topology
    - This is done offline
    - Rarely recomputed
- TeXCP chooses K-shortest paths (length is propagation delay)
- TeXCP may actually use less than K paths depending on failure and utilization state.



Probing Network State

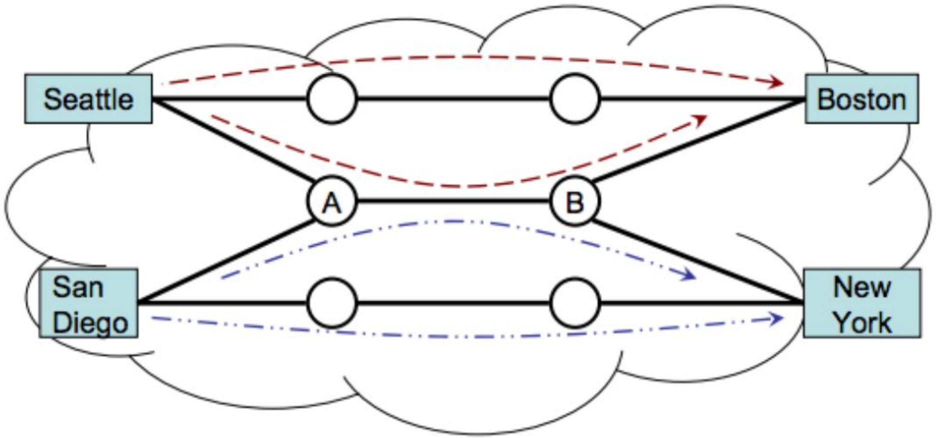
- Agent sends a probe on each path at some rate T_p
- Intermediate routers update utilization in probe if their output link has a greater utilization
- Probe is returned by E to I and I gives it to correct agent
- Probe loss indicates failure or congestion



  - Discussion: Will intermediate routers require additional hardware/software to handle these probes? Can a sufficient probe be achieved with existing hardware/software?

The Load Balancer

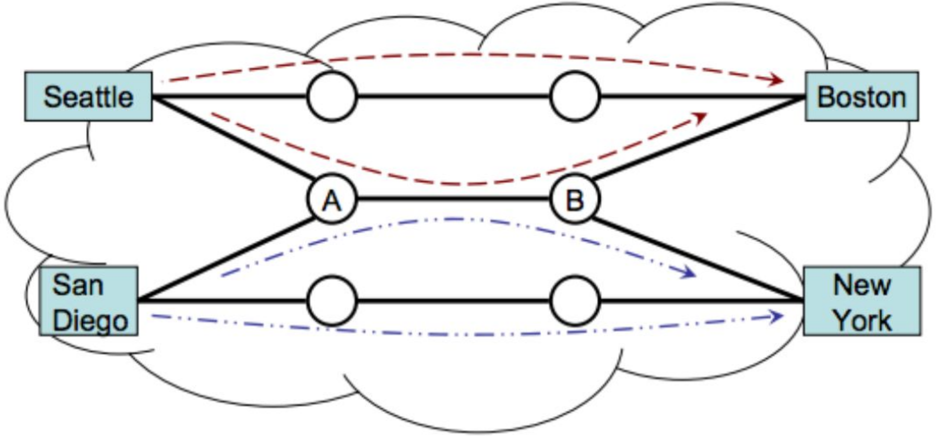
- Incrementally changes fraction of traffic on a path
- If path utilization is above average, reduce fraction
- If path utilization is below average, increase fraction
- Make changes proportional to current fraction
- Average utilization is normalized by fraction of traffic on path



- For paths with zero traffic being sent by the agent, if utilization is below average, a small constant bootstraps the ramping up of traffic.

Oscillations and Congestion

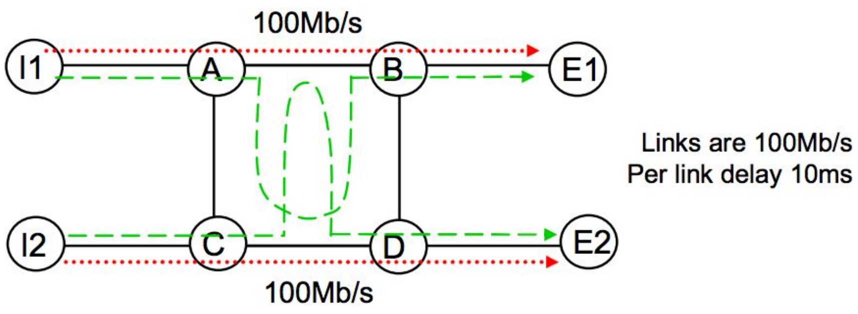
- See diagram
- Now imagine links with dozens of TeXCP agents sending traffic
- What can we do?
- Alter the feedback the router sends to the TeXCP agents so that combined agents cannot overshoot the capacity of the link



- Discussion: Other options for dealing with this?

Shorter Paths First

- See diagram
- If both agents choose green paths, the utilization is the same as if both use red.
- Additional weight is given to shorter paths in cases where utilization is equal



IXP: Internet Exchange Points

Basic Terminology

AS - autonomous system - collection of network infrastructure (routers, etc) under common control

NAP - network access point, ancestor of modern IXPs, funded by NSF when NSFNET was privatized

IXP - internet exchange point - location that connects together multiple ASes

ISP - internet service provider

CDN - content delivery network

IXP Members: ASes

Relationships between ASes

Transit (provider-customer) - Provider paid by customer to deliver packets

Peer (provider-provider) - Provider will carry packets from another provider to its customers for free

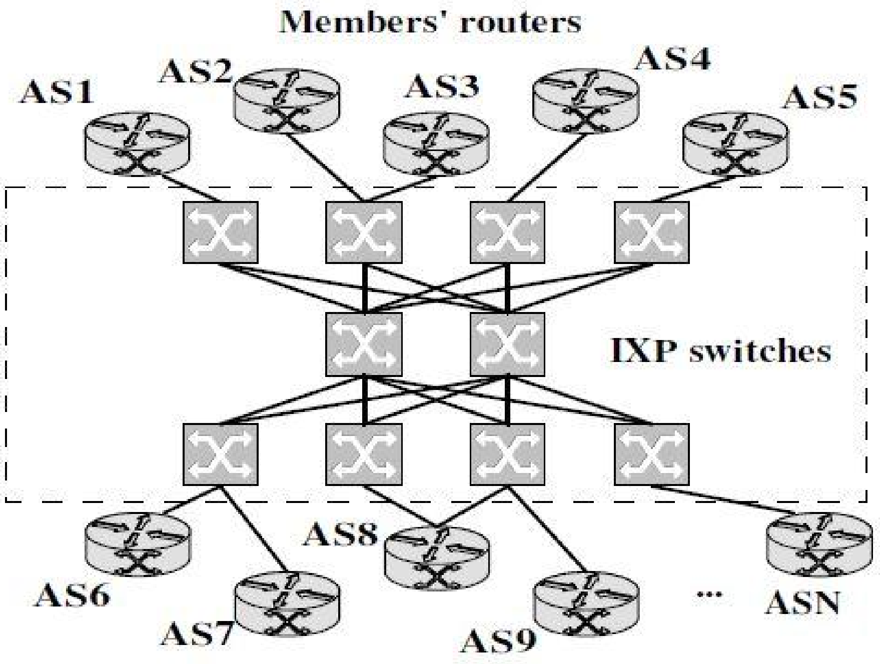
Basic Tiering Structure of the internet

Tier-1:   can connect/reach other networks via its peer relationships

Tier-2: has a mixture of peering/transit relationships

Tier-3/Leaf: has only transit relationships

Structure of an IXP



Source: Anatomy of a Large European IXP

- Each member of the IXP has an access point at the exchange
- Members agree to share data via mutual agreement
- Costs are shared among members of the exchange
- Why do members choose to belong to an exchange?
  - Cost efficiency
    - Shared Infrastructure
    - Peering relationships
    - Increased competition
  - Redundancy

Analysis of a large European IXP

Key findings and questions

- View of the internet as a network of networks or ASes
  - Who are the various members of an IXP?
  - What is the role of an IXP in internet topology?
- Internet peering ecosystem
  - What drives ASes to interconnect?
  - How do various members of an IXP interact?
- Quantity and quality of internet inter-domain traffic
  - What type of traffic is being exchanged?
  - What patterns does it exhibit?

ASes Peering at IXPs

Prior Research

- Traceroute data or BGP routing data revealed a small number of peer links (literature from 2009 suggested 35-45K P-P links for entire internet)

This IXP research paper

- Used sFlow data from 2011
  - Recorded IP and TCP headers - understanding of destination and application
  - Represents actual traffic between networks (not just exchange of routing information)
- Results from IXP public switching network shows a huge number of P-P links
  - 67% of possible bidirectional pairings between 400 members = 50K+ in a single IXP

Replication of traceroute/BGP research

Table 2: Overview of routing and looking glass datasets for November. The numbers show P-P links.

Dataset	Unique LGs / ASN	Visible links	only in this dataset
RV	78	5,336	1,084
RIPE	319	10,913	5,460
NP	723	3,419	684
RV+RIPE+NP	997	13,051	10,472
LG	821 / 148	4,892	2,313
RV+RIPE+NP+LG	1,070	15,364	15,364

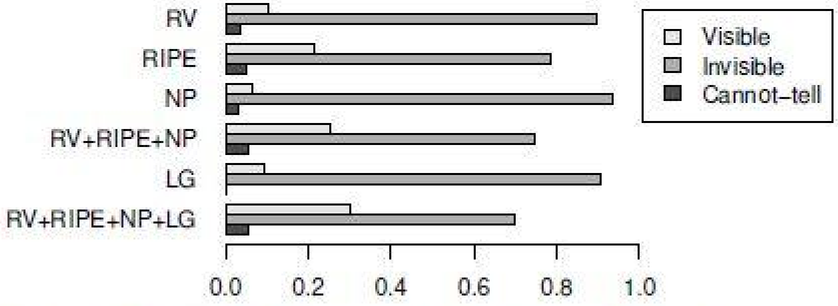


Figure 2: Peering links and visibility in control/data plane (normalized by number of detected P-P links).

RV - Route-views route table data

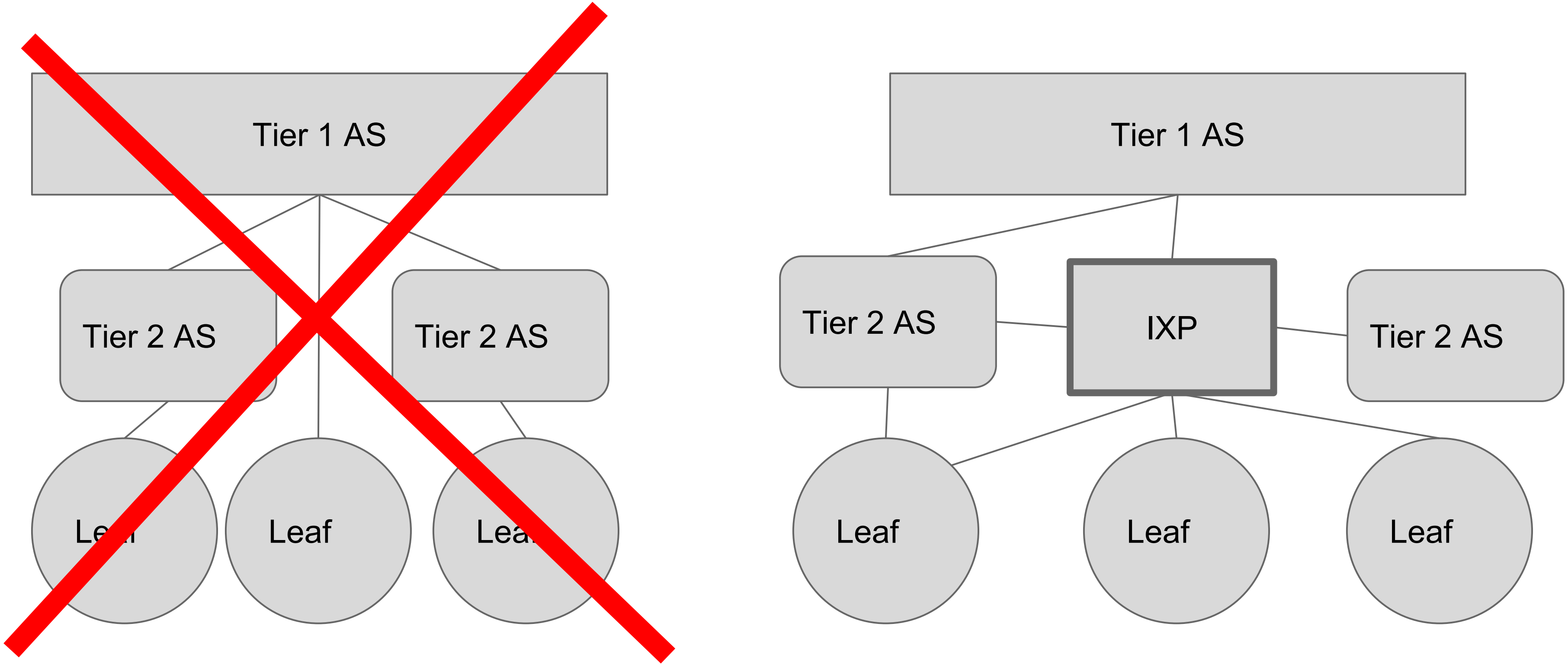
RIPE - RIPE NCC route table data

NP - Non-public route table data

LG - traceroutes using public looking glass servers

Source: Source: Anatomy of a Large European IXP

Implications for Internet topology



Who is peering?

Relationships

- Tier 1 ISPs connect peer with relatively few other ASes (less than 25%)
  - Consistent with their business model as transit providers
- Tier 2/Tier 3 ISPs peer with many other ASes (70% of avaliable connections)
- CDNs often have open peering policies

Traffic Structure

- Locality
- Temporal
  - Diurnal
- (Non)-Sparsity
  - ISPs typically have sparse traffic due to careful planning
  - IXPs have non-sparse traffic
- Low Rank
  - Highly structured and predictable traffic patterns
  - Possible to design better measurement techniques (require less data)

Discussion Questions

- Will online route changes hurt TCP congestion control due to packet reordering? How can we deal with this?
- Because TeXCP focuses on utilization, could this lead to significant use of long paths?
- How can we increase speed of failure detection of probes?
- What threats could a network using TeXCP be vulnerable to?
- How could we secure TeXCP?
- The paper analyzed a large European IXP, are IXPs in the US likely to be similar? Why or why not?
- Why are so many peer links missing from the traceroute and BGP routing tables?

Email: josting2@illinois.edu