Class Meeting: Lectures 11 & 12: BGP and Measurement

Kyle Jamieson

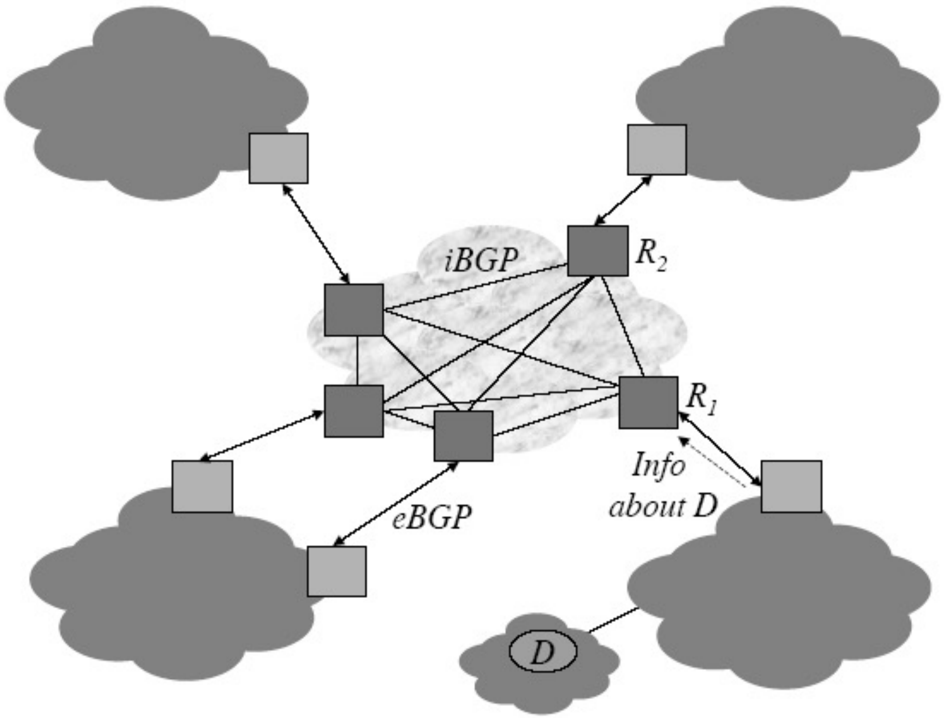
COS 461: Computer Networks

[Portions adapted from H. Balakrishnan, N. Feamster, B. Karp, and J. Rexford]

Context: Autonomous Systems

- A routing domain is called Autonomous System (AS)
  - Each AS known by unique 16-bit number
  - AS owns one or handful of address prefixes; allocates addresses under those prefixes
  - AS typically a commercial entity or other organization
  - ASes often competitors (e.g., different ISPs)
- Interior Gateway Protocols (IGPs) (e.g., DV, LS) route within individual ASes
- Exterior Gateway Protocols (EGPs) (e.g., BGP) route among ASes

eBGP and iBGP



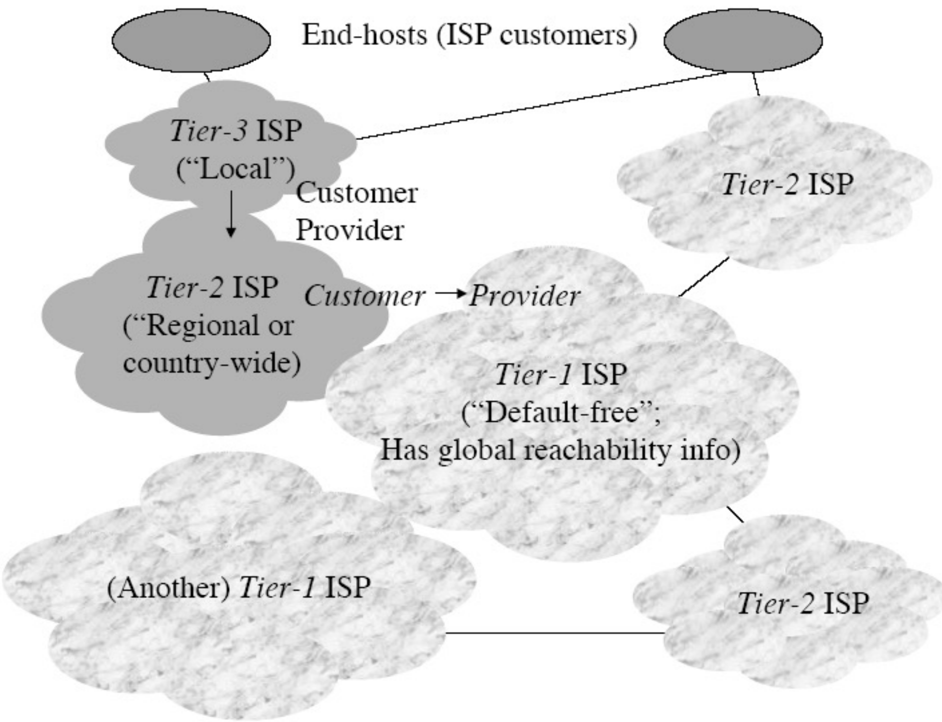
- eBGP: external BGP advertises routes between ASes

- iBGP: internal BGP propagates external routes throughout receiving AS

Synthesis:
Routing with IGP + iBGP

- Every router in AS now learns two routing tables
  - IGP (e.g., link state) table: routes to every router within AS, via interface
  - EGP (e.g., iBGP) table: routes to every prefix in global Internet, via egress router IP
- Produce one integrated forwarding table
  - All IGP entries kept as-is
  - For each EGP entry
    - find next-hop interface i for egress router IP in IGP table
    - add entry: <foreign prefix, i>
  - End result: O(prefixes) entries in all routers' tables

Global Internet Routing



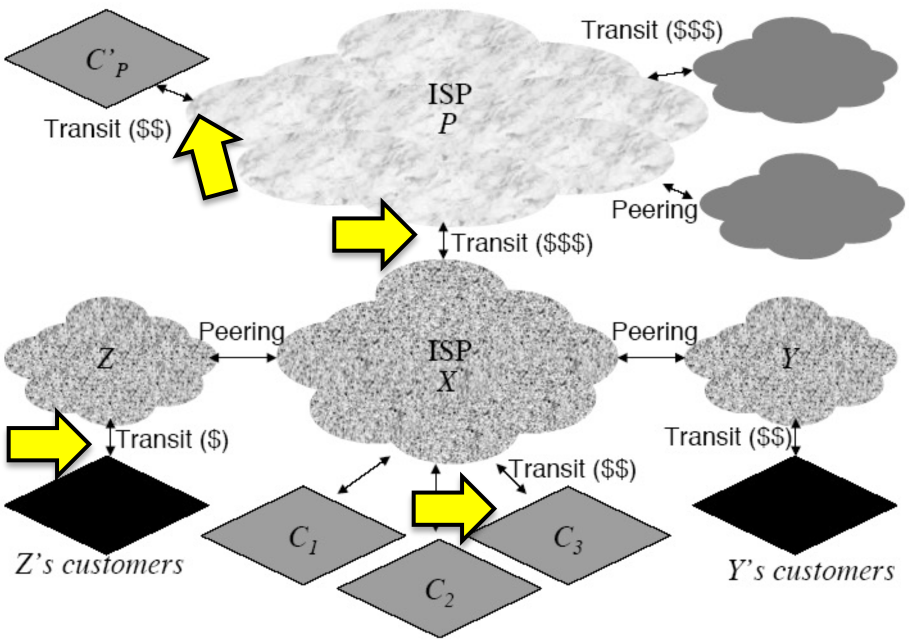
- Tiers of ISPs:
  - Tier 1: geographically global, ISP customers, no default routes
  - Tier 2: regional geographically
  - Tier 3: local geographically, end customers

- Each ISP is an AS
  - AS operator sets policies for how to route to others, how to let others route to them

AS-AS Relationships: Customers and Providers

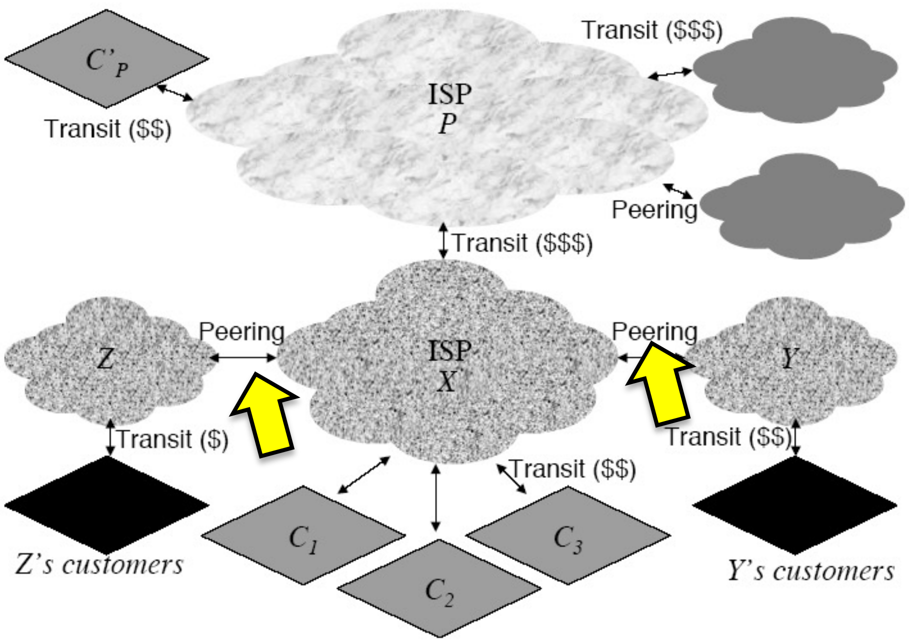
- Smaller ASes (corporations, universities) typically purchase connectivity from ISPs
- Regional ISPs typically purchase connectivity from global ISPs
- Each such connection has two roles:
  - Customer: smaller AS paying for connectivity
  - Provider: larger AS being paid for connectivity
- Other possibility: ISP-to-ISP connection

AS-AS Relationship: Transit



- Provider-Customer AS-AS connections are called **transit**
- Provider allows customer to route to (nearly) all destinations in its routing tables
- Transit nearly always involves payment from customer to provider

AS-AS Relationship: Peering



- **Peering:** two ASes (usually ISPs) mutually allow one another to route to some of the destinations in their routing tables

- By contract, but usually no money changes hands, so long as traffic ratio is narrower than, e.g., 4:1

Financial Motives: Peering and Transit

- Peering relationship often between competing ISPs
- Incentives to peer:
  - Typically, two ISPs notice their own direct customers originate a lot of traffic for the other
  - Each can avoid paying transit costs to others for this traffic; shunt it directly to one another
  - Often better performance (shorter latency, lower loss rate) as avoid transit via another provider
  - Easier than stealing one another's customers
- Tier 1s must typically peer with one another to build complete, global routing tables

The Meaning of Advertising Routes

- AS **A** advertises a route for destination **D** to AS **B:** effectively an offer to forward all traffic from AS **B** to **D**
- Forwarding traffic costs bandwidth
- AS' incentive to control which routes they advertise:
  - no one wants to forward packets without being compensated to do so
  - e.g., when peering, only let neighboring AS send to specific own customer destinations enumerated peering contract

Advertising Routes for Transit Customers

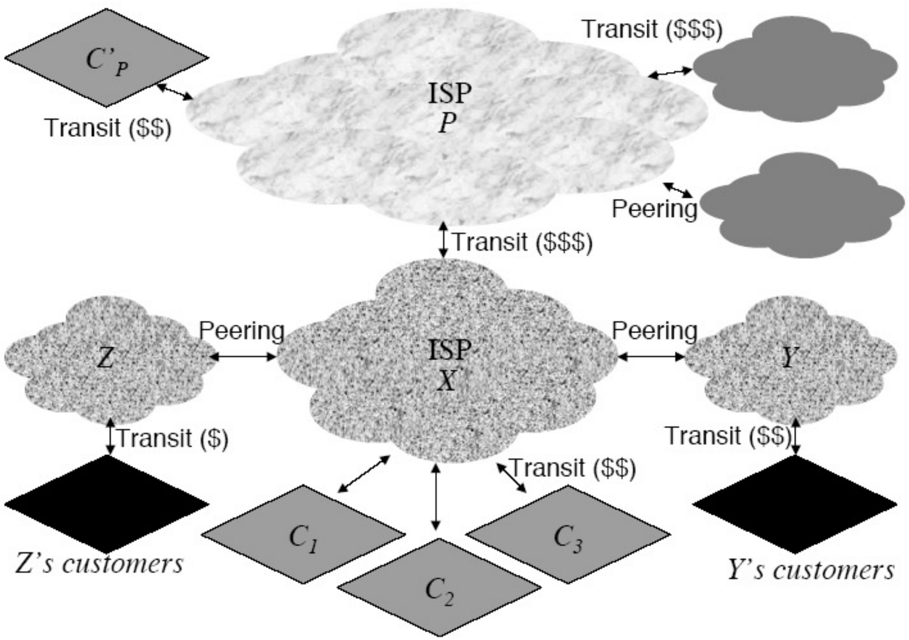
- ISP motivated to advertise routes to its own customers to its transit providers
  - Customers paying to be reachable from global Internet
  - More traffic to customer, faster link customer must buy
- If ISP hears route for its own customer from multiple neighbors, should favor advertisement from own customer

Routes Heard from Providers

- If ISP hears routes from its provider (via a transit relationship), to whom does it advertise them?
  - Not to ISPs with peering relationships; they don't pay, so no motivation to provide transit service for them!
  - To own customers, who pay to be able to reach global Internet

Example: Routes Heard from Providers

- Provider ISP P announces route to C'_P (its own customer) to X



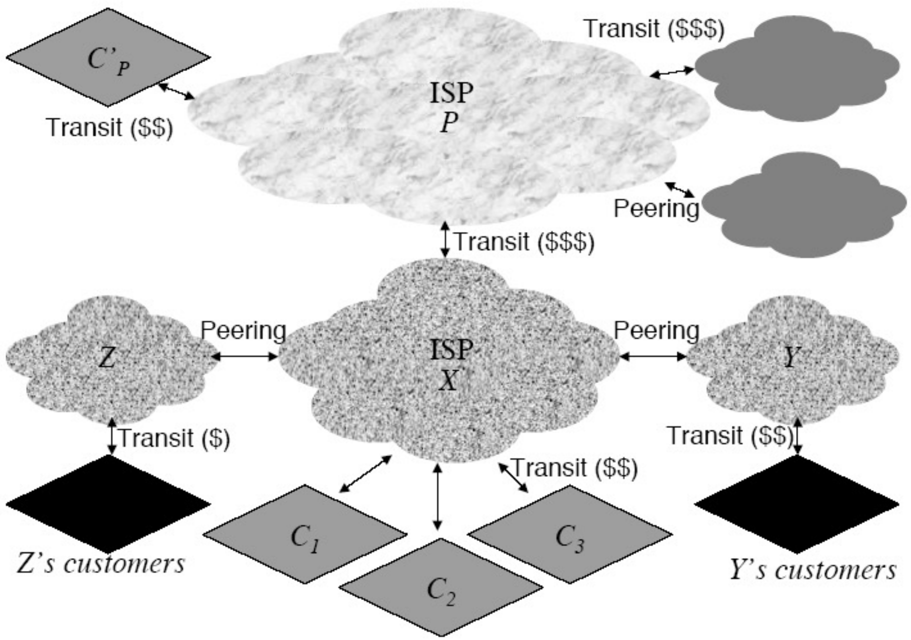
- X doesn't announce C'_P to Y or Z; (no revenue from peering)
- X announces C'_P to C_i; (they're paying to be able to reach everywhere)

Routes Advertised to Peers

- Which routes should an ISP advertise to ASes with whom it has peering relationships?
  - Routes for all own downstream transit customers
  - Routes to ISP's own addresses
  - Not routes heard from upstream transit provider of ISP (peer might route via ISP for those destinations, but doesn't pay)
  - Not routes heard from other peering relationships (same reason!)

Example: Routes Advertised to Peers

- ISP X announces C_i to Y and Z



- ISP X doesn't announce routes heard from ISP P to Y or Z
- ISP X doesn't announce routes heard from ISP Y to ISP Z, or vice-versa

Route Export: Summary

- ISPs typically provide selective transit
  - Full transit (export of all routes) for own transit customers in both directions
  - Some transit (export of routes between mutual customers) across peering relationship
  - Transit only for transit customers (export of routes to customers) to providers

- These decisions about what routes to advertise motivated by policy (money), not by optimality (e.g., shortest paths)

Route Import

- Router may hear many routes to same destination
  - Identity of advertiser very important

- Suppose router hears advertisement to own transit customer from other AS
  - Shouldn't route via other AS; longer path!
  - Customer routes higher priority than routes to same destination advertised by providers or peers

- Routes heard over peering higher priority than provider routes
  - Peering is free; you pay provider to forward via them
- customer > peer > provider

Using Route Attributes

- Recall: BGP route advertisement is simply:
  - IP Prefix: [Attribute 0] [Attribute 1] […]

- Administrators enforce policy routing using attributes:
  - filter and rank routes based on attributes
  - modify "next hop" IP address attribute
  - tag a route with attribute to influence ranking and filtering of route at other routers

NEXT HOP Attribute

- Indicates IP address of next-hop router
- Modified as routes are announced
  - eBGP: when border router announces outside of AS, changes to own IP address
  - iBGP: when border router disseminates within AS, changes to own IP address
  - iBGP: any iBGP router that repeats route to other iBGP router leaves unchanged

ASPATH Attribute: Path Vector Routing

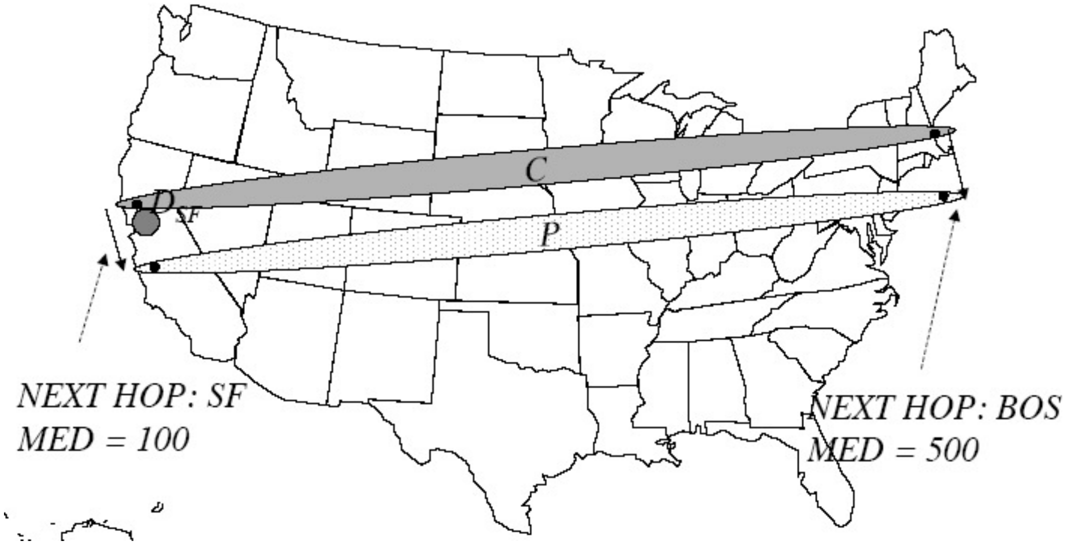
- Contains full list of AS numbers on path to destination prefix
- Ingress router prepends own AS number to ASPATH of routes heard over eBGP
- Functions like distance vector routing, but with explicit enumeration of AS "hops"
  - Barring local policy settings, shorter ASPATHs preferred to longer ones
  - If reject routes that contain own AS number, cannot choose route that loops among ASes!

MED Attribute: Choosing Among Multiple Exit Points

- ASes often connect at multiple points (e.g., global backbones)
- ASPATHs will be same length
- But AS' administrator may prefer a particular transit point
  - …often the one that saves them money!
- MED Attribute: Multi-Exit Discriminator, allows choosing transit point between two ASes

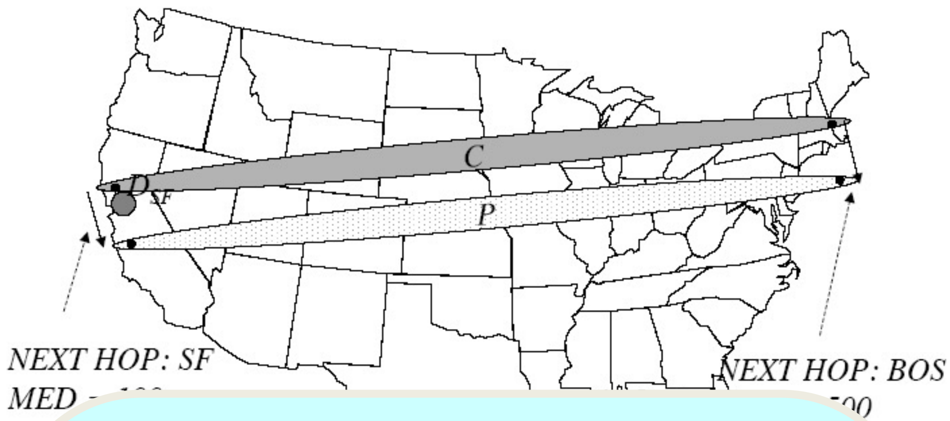
MED Attribute: Example (1/2)

- Provider P, customer C
- Source: Boston on P, Destination: AS D_{SF} (San Francisco) on C



- Whose backbone for cross-country trip?
- C wants traffic to cross country on P

MED Attribute: Example (2/2)



AS need not honor MEDs from neighbor

AS only motivated to honor MEDs from other AS with whom financial settlement in place; i.e., not done in peering arrangements

Most ISPs prefer shortest-exit routing: get packet onto someone else's backbone as quickly as possible

Result: highly asymmetric routes! (why?)

- C adds MED attribute to advertisements of routes to D_{SF}
  - Integer cost
- C's router in SF advertises MED 100; in BOS advertises 500
- P should choose MED with least cost for destination D_{SF}
- Result: traffic crosses country on P

Synthesis:
Multiple Attributes into Policy Routing

- How do attributes interact? Priority order:

Priority	Rule	Details
1	LOCAL PREF	**Highest LOCAL PREF (e.g., prefer transit customer routes over peer and provider routes)**
2	ASPATH	**Shortest ASPATH length**
3	MED	**Lowest MED**
4	eBGP > iBGP	**Prefer routes learned over eBGP vs. over iBGP**
5	IGP path	**"Nearest" egress router**
6	Router ID	**Smallest router IP address**

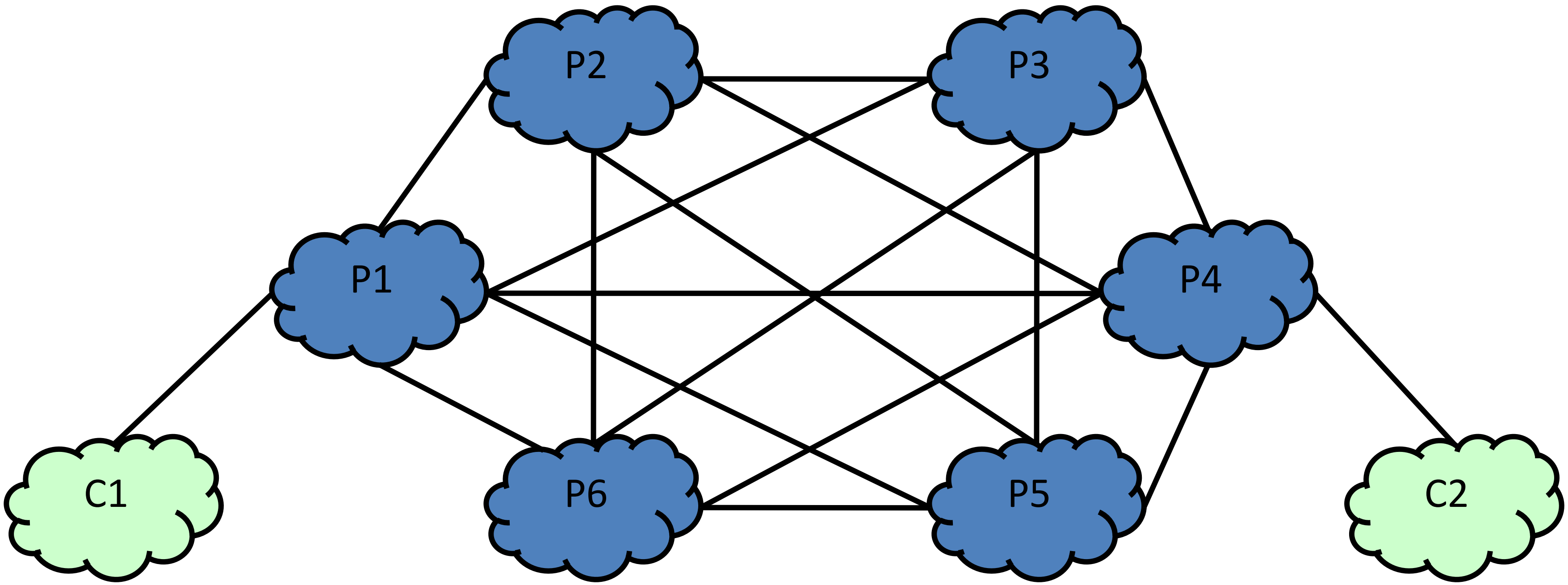
War Story: Depeering

- All tier-1 ISPs peer directly with one another in a full mesh

- True tier-1 ISPs do not pay for peering and buy transit from no one

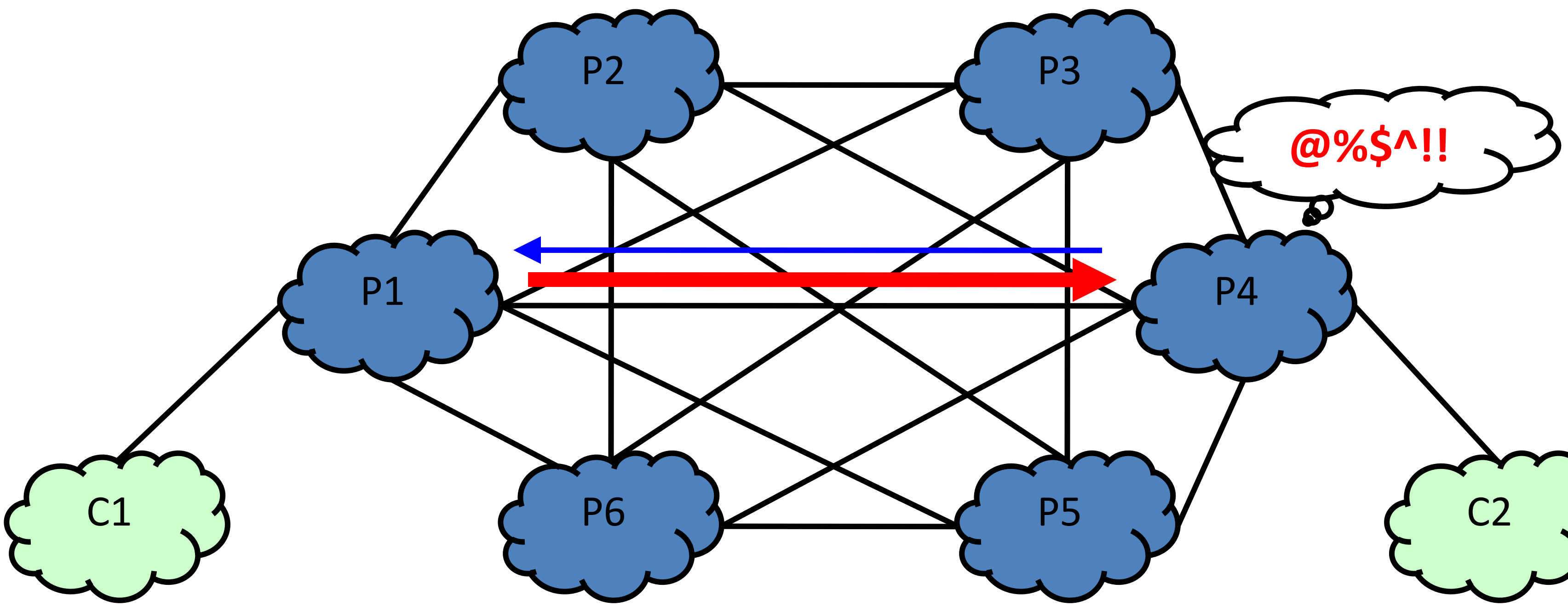
- A few *other* large ISPs pay no transit provider:
  - they peer with all tier-1 ISPs…
  - …but pay settlements to one or more of them

Full-Mesh Peering



For Internet to be connected, all ISPs who do not buy transit service must be connected in full mesh!

A Peers' Quarrel: Depeering



When P4 terminates BGP peering with P1, C1 and C2 can no longer reach one another, if they have no other transit path!

P4 has partitioned the Internet!

Depeering Happens

- 10/2005: Level 3 depeered Cogent
- 3/2008: Telia depeered Cogent
- 10/2008: Sprint depeered Cogent
  - lasted from 30th October – 2nd November, 2008
  - 3.3% of IP prefixes in global Internet behind one ISP partitioned from other, including NASA, Maryland Dept. of Trans., NY Court System, 128 educational institutions, Pfizer, Merck, Northup Grumman, ...

Measurement: BGP Monitoring

Motivation for BGP Monitoring

- Visibility into external destinations
  - What neighboring ASes are telling you
  - How you are reaching external destinations
- Detecting anomalies
  - Increases in number of destination prefixes
  - Lost reachability or instability of some destinations
- Input to traffic-engineering tools
  - Knowing the current routes in the network
- Workload for testing routers
  - Realistic message traces to play back to routers

BGP Monitoring: A Wish List

- Ideally: know what the router knows
  - All externally-learned routes
  - Before applying policy and selecting best route
- How to achieve this
  - Special monitoring session on routers that tells everything they have learned
  - Packet monitoring on all links with BGP sessions
- If you can't do that, you could always do…
  - Periodic dumps of routing tables
  - BGP session to learn best route from router

Conclusions

- Inter-domain routing chiefly concerned with policy, not optimality
- Behavior and configuration of BGP complex and not fully understood
- Measurement is crucial to network operations
  - Measure, model, control
  - Detect, diagnose, fix