Inter-Domain Routing: BGP II

Brad Karp

UCL Computer Science

(drawn mostly from lecture notes

by Hari Balakrishnan and Nick Feamster, MIT)

CS 3035/GZ01

4th December 2014

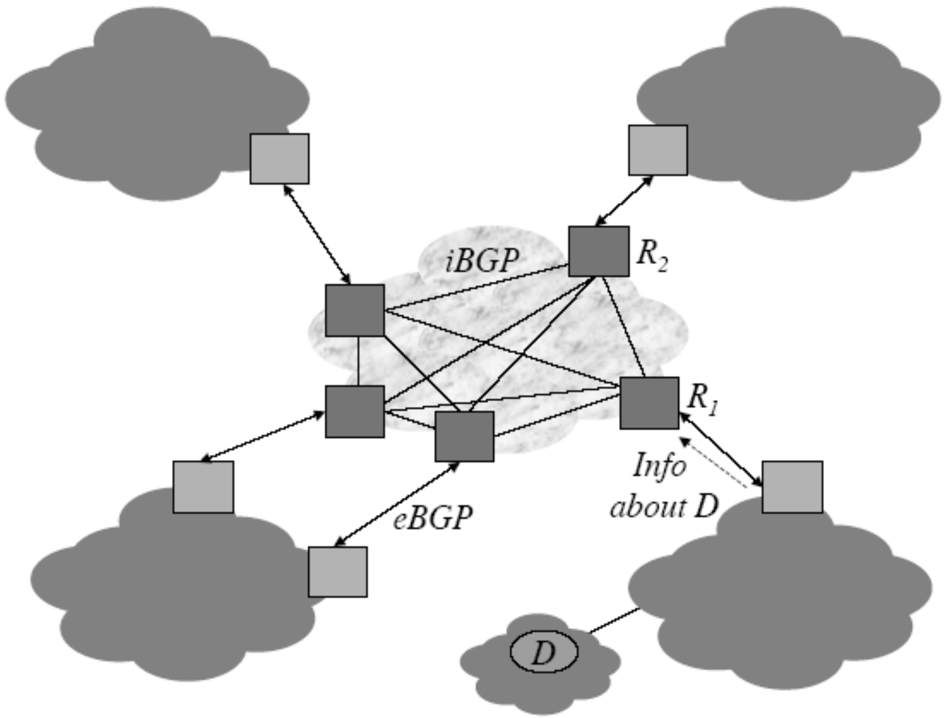
BGP Protocol (cont'd)

- BGP doesn't chiefly aim to compute shortest paths (or minimize other metric, as do DV, LS)
- Chief purpose of BGP is to announce reachability, and enable policy-based routing
- BGP announcement:
  - IP prefix: [Attribute 0] [Attribute1] […]

Outline

- Context: Inter-Domain Routing
- Relationships between ASes
- Enforcing Policy, not Optimality
- BGP Design Goals
- BGP Protocol
- eBGP and iBGP
- BGP Route Attributes
- Synthesis: Policy through Route Attributes
- War Story: Depeering

eBGP and iBGP



- eBGP: external BGP advertises routes between ASes
- iBGP: internal BGP propagates external routes throughout receiving AS

eBGP and iBGP (cont'd)

- Each eBGP participant hears different advertisements from neighboring ASes
- Must propagate routes learned via eBGP throughout AS
- Design goals:
  - Loop-free forwarding: forwarding paths over routes learned via eBGP should not loop
  - Complete visibility: all routers within AS must choose same, best route to destination learned via eBGP

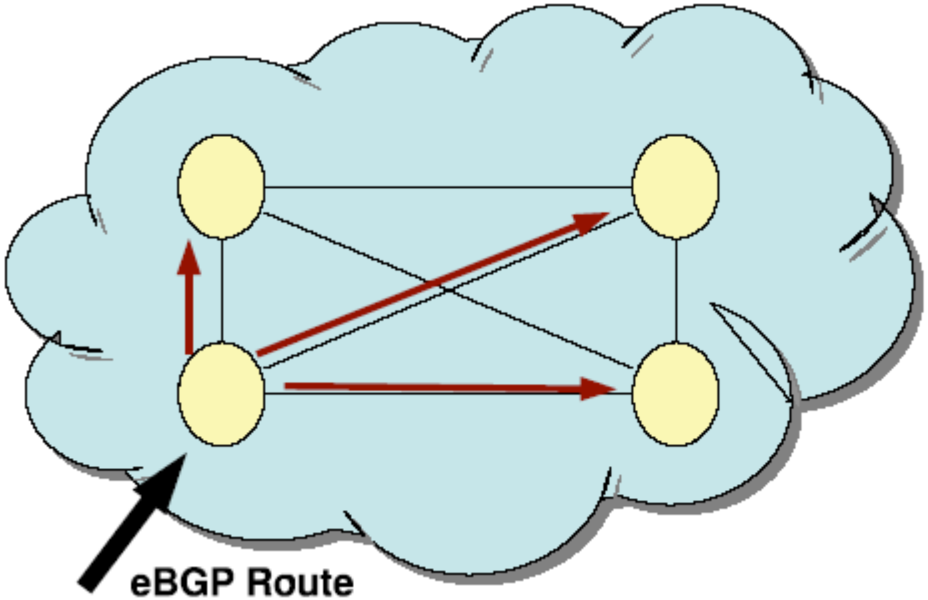
eBGP and iBGP (cont'd)

- Each eBGP participant hears different advertisements from neighboring ASes

Within AS1, choosing external route to destination in AS2 amounts to **choosing egress router within AS1**

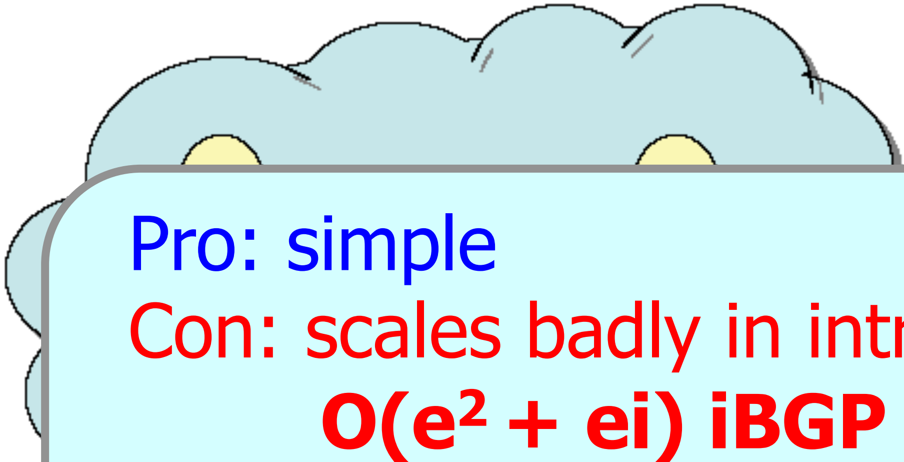
  - Loop-free forwarding: forwarding paths over routes learned via eBGP should not loop
  - Complete visibility: all routers within AS must choose same, best route to destination learned via eBGP

Simple iBGP: Full Mesh



- How to achieve complete visibility?
  - Push all routes learned via eBGP to all internal routers using iBGP
- Full Mesh: each eBGP router floods routes it learns to all other routers in AS
- Flooding done over TCP, using intra-AS routing provided by IGP (e.g., link state routing)

Simple iBGP: Full Mesh



- How to achieve complete visibility?
  - Push all routes learned via eBGP to all internal

Pro: simple

Con: scales badly in intra-AS router count:

$O(e^2 + ei)$ iBGP sessions

(where e eBGP routers, i iBGP routers)

More scalable iBGP uses route reflectors or confederations; details in lecture notes

IGP (e.g., link state routing)

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

Outline

- Context: Inter-Domain Routing
- Relationships between ASes
- Enforcing Policy, not Optimality
- BGP Design Goals
- BGP Protocol
- eBGP and iBGP
- BGP Route Attributes
- Synthesis: Policy through Route Attributes
- War Story: Depeering

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

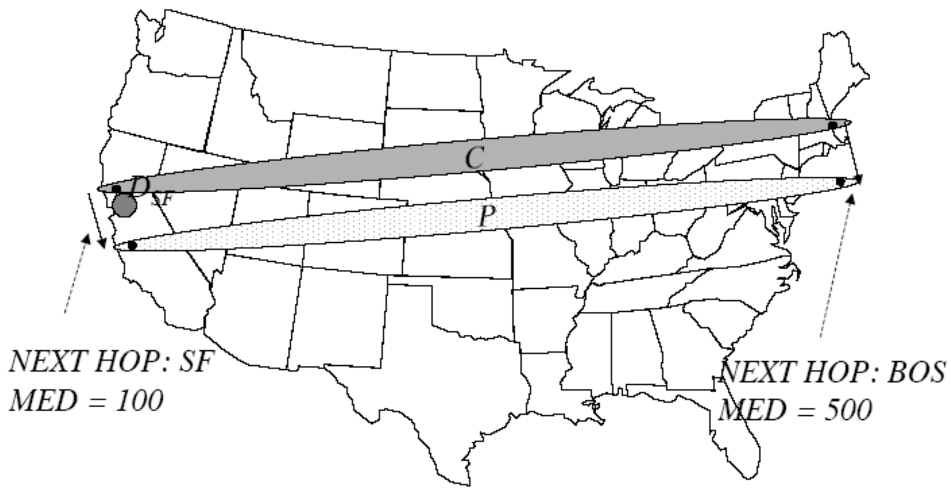
ASPATH Attribute: Path Vector Routing

- Contains full list of AS numbers along 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

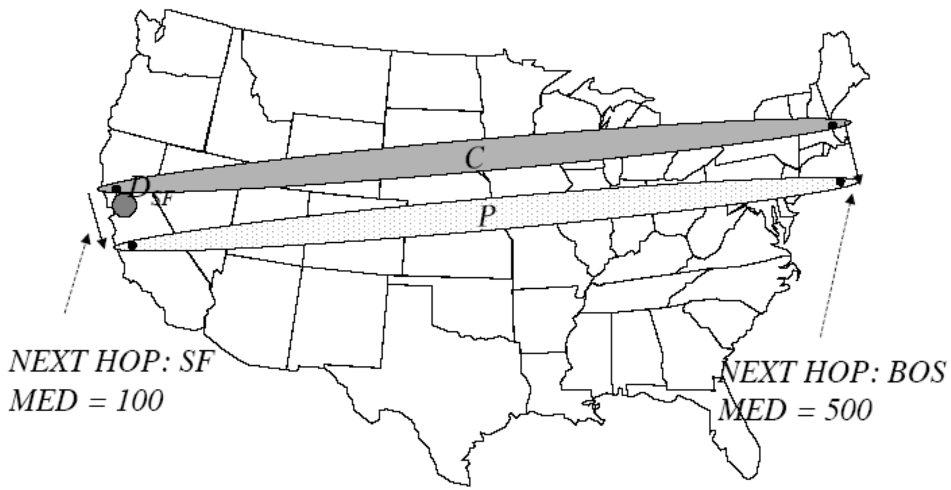
- ASes often connect at multiple points (e.g., global backbones)
- ASPATHs will be same length
- But AS' administrator may prefer a particular transit point
  - …often the one that saves him money!
- MED Attribute: Multi-Exit Discriminator, allows choosing transit point between two ASes

MED Attribute: Example



- Provider P, customer C
- Source: Boston on P, Destination: San Francisco on C
- Whose backbone for cross-country trip?
- C wants traffic to cross country on P

MED Attribute: Example (cont'd)



- 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

MED Attribute: Example (cont'd)

- C adds MED attribute to advertisements of routes to D_{SF}

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?)

Outline

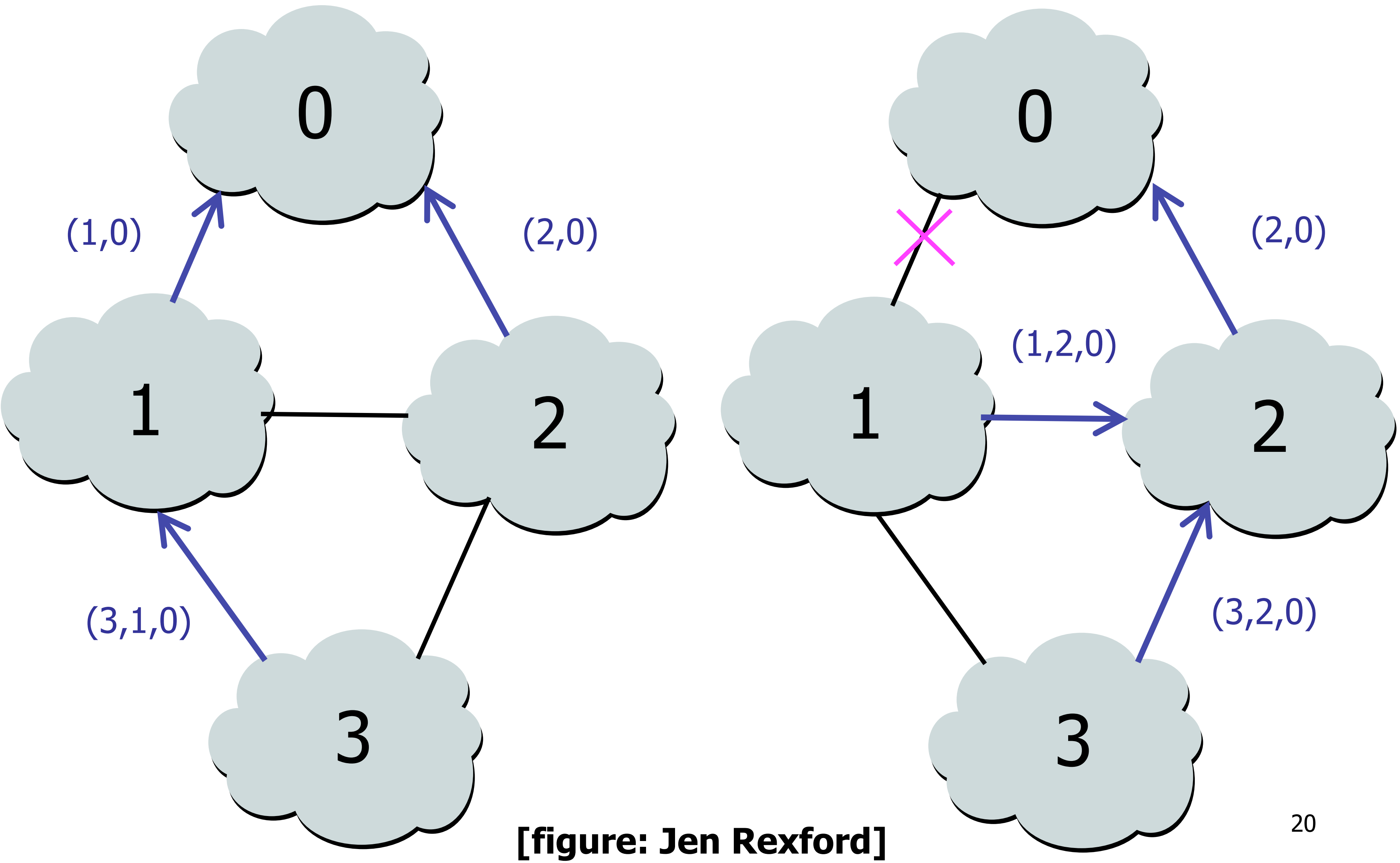
- Context: Inter-Domain Routing
- Relationships between ASes
- Enforcing Policy, not Optimality
- BGP Design Goals
- BGP Protocol
- eBGP and iBGP
- BGP Route Attributes
- Synthesis: Policy through Route Attributes
- War Story: Depeering

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**

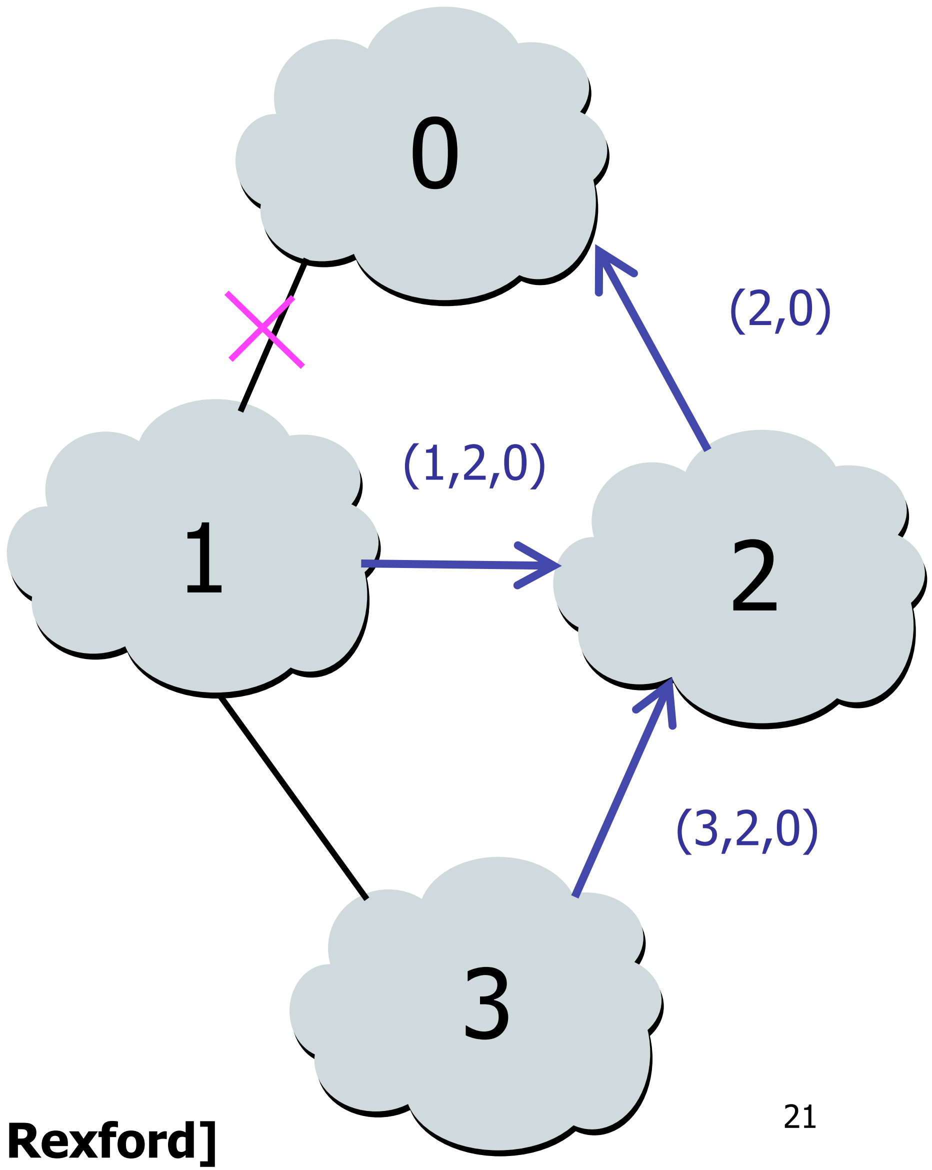
BGP Dynamics



[figure: Jen Rexford]

BGP Dynamics: Path Exploration

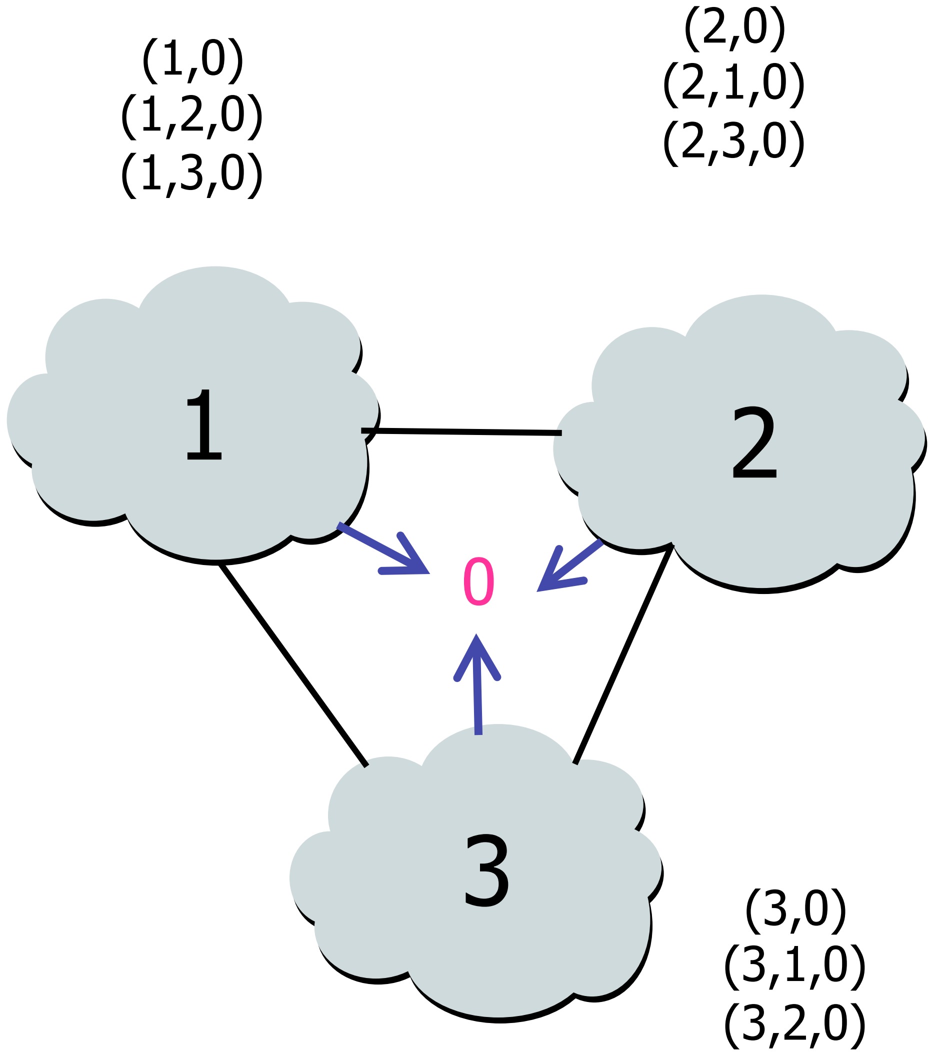
- AS 1
  - Delete the route (1,0)
  - Switch to next route (1,2,0)
  - Announce route (1,2,0) to AS 3
- AS 3
  - Sees (1,2,0) replace (1,0)
  - Compares to route (2,0)
  - Switches to using AS 2



[slide: Jen Rexford]

Path Exploration: Slower Example

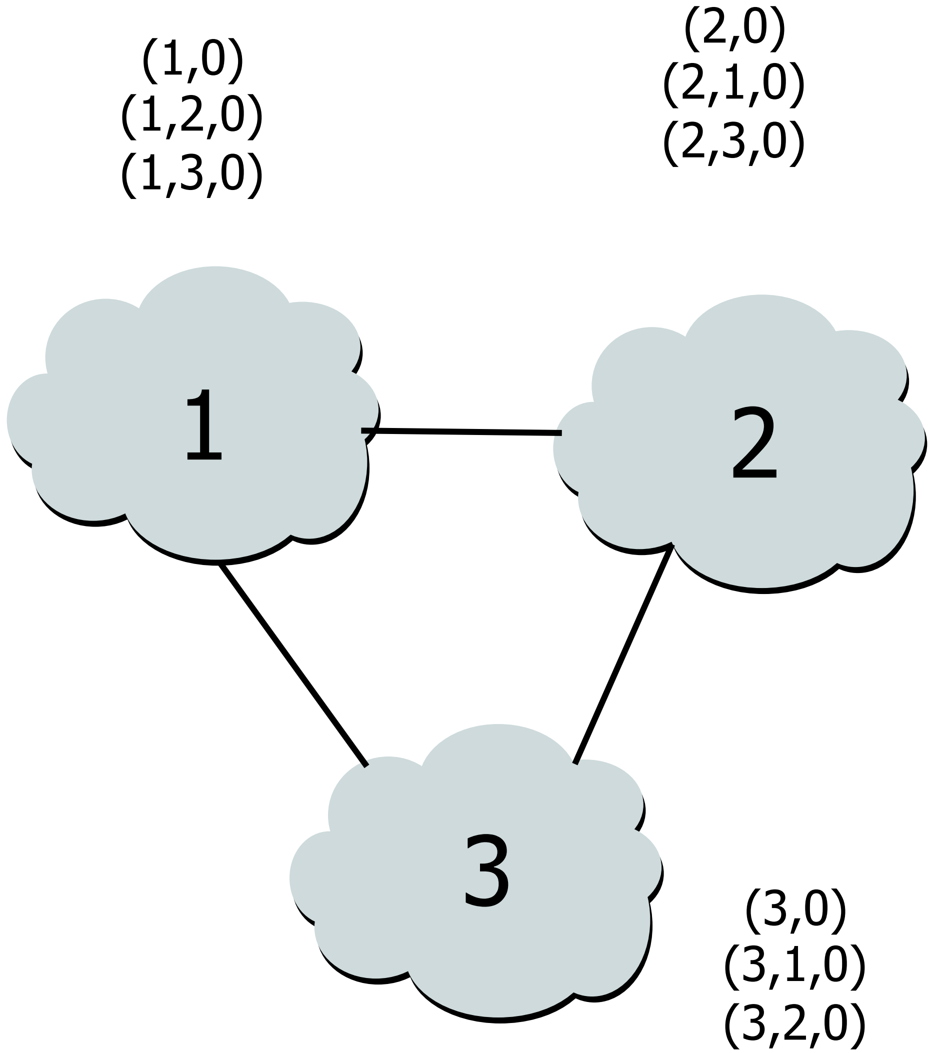
- Initial situation
  - Destination 0 is alive
  - All ASes use direct path
- When destination dies
  - All ASes lose direct path
  - All repeatedly switch to longer paths
  - Eventually withdrawn
- e.g., AS 2
  - (2,0) → (2,1,0)
  - (2,1,0) → (2,3,0)
  - (2,3,0) → (2,1,3,0)
  - (2,1,3,0) → null



[slide: Jen Rexford]

Path Exploration: Slower Example

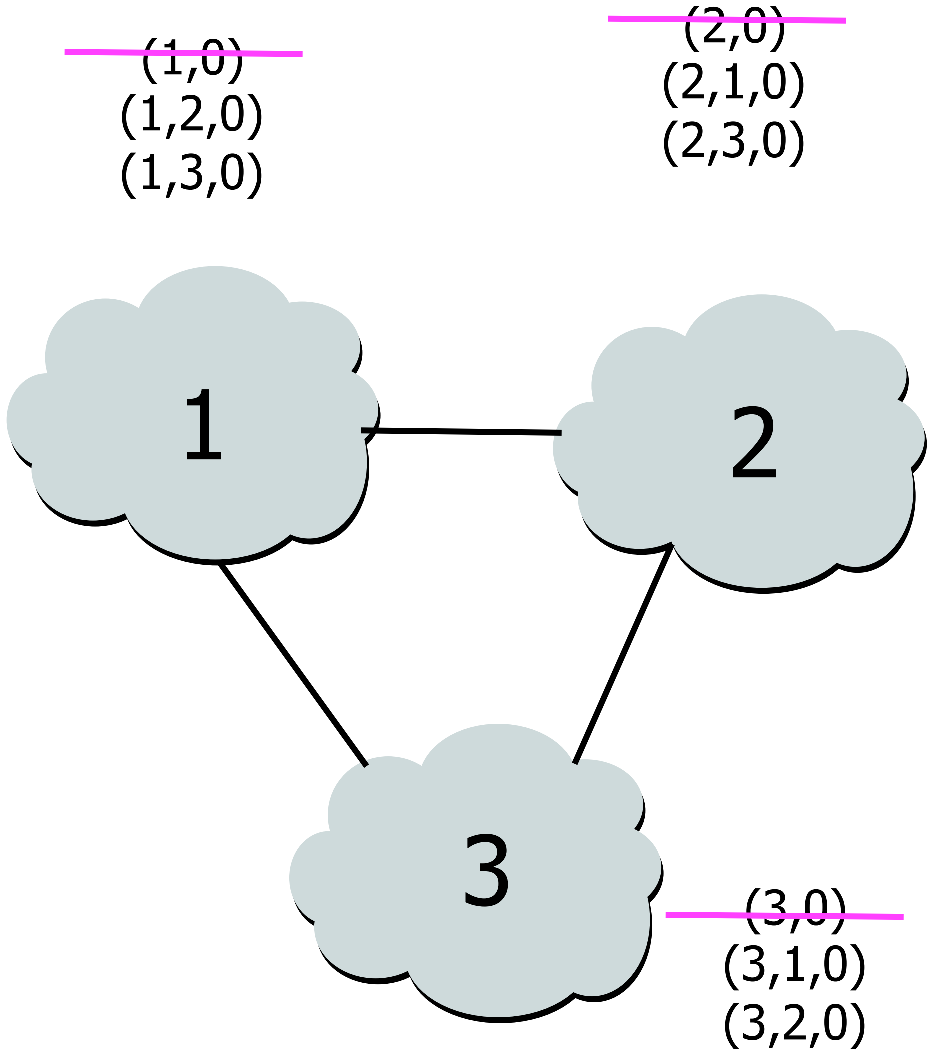
- Initial situation
  - Destination 0 is alive
  - All ASes use direct path
- When destination dies
  - All ASes lose direct path
  - All repeatedly switch to longer paths
  - Eventually withdrawn
- e.g., AS 2
  - (2,0) → (2,1,0)
  - (2,1,0) → (2,3,0)
  - (2,3,0) → (2,1,3,0)
  - (2,1,3,0) → null



[slide: Jen Rexford]

Path Exploration: Slower Example

- Initial situation
  - Destination 0 is alive
  - All ASes use direct path
- When destination dies
  - All ASes lose direct path
  - All repeatedly switch to longer paths
  - Eventually withdrawn
- e.g., AS 2
  - (2,0) → (2,1,0)
  - (2,1,0) → (2,3,0)
  - (2,3,0) → (2,1,3,0)
  - (2,1,3,0) → null



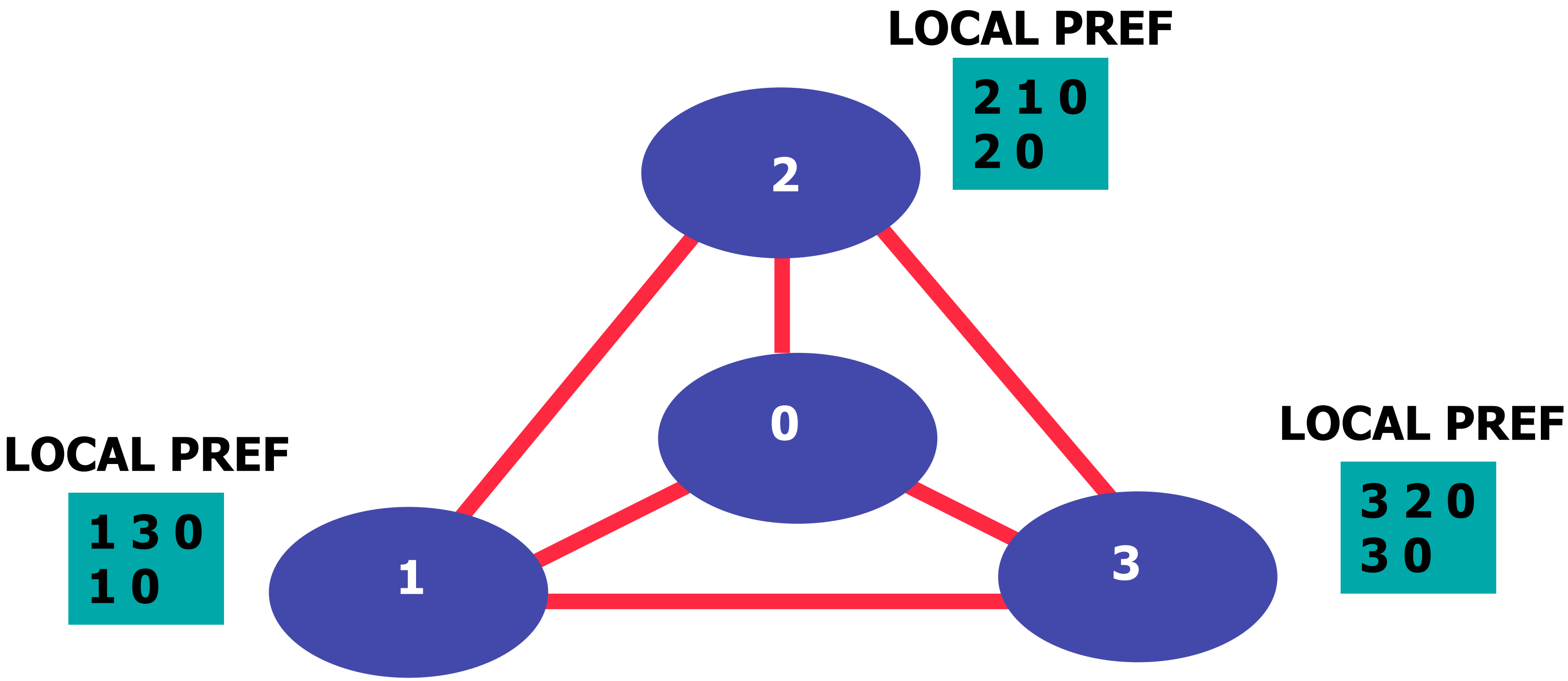
[slide: Jen Rexford]

Limiting Update Traffic

- Minimum route advertisement interval (MRAI)
  - Minimum spacing between announcements
  - For a particular (prefix, peer) pair
- Advantages
  - Provides a rate limit on BGP updates
  - Allows grouping of updates within interval
- Disadvantages
  - Adds delay to convergence process
  - e.g., 30 seconds for each step

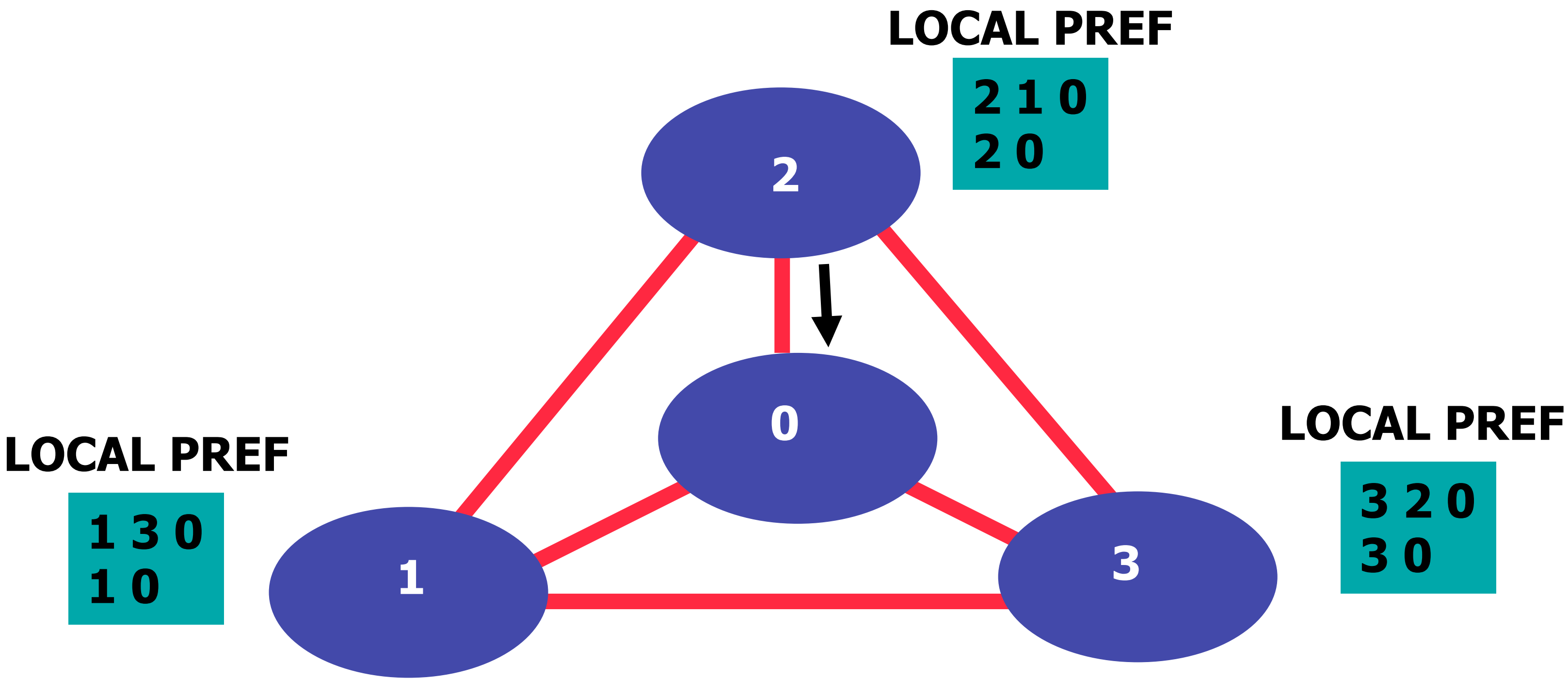
[slide: Jen Rexford]

Policies May Cause Persistent Oscillations ("Dispute Wheels")



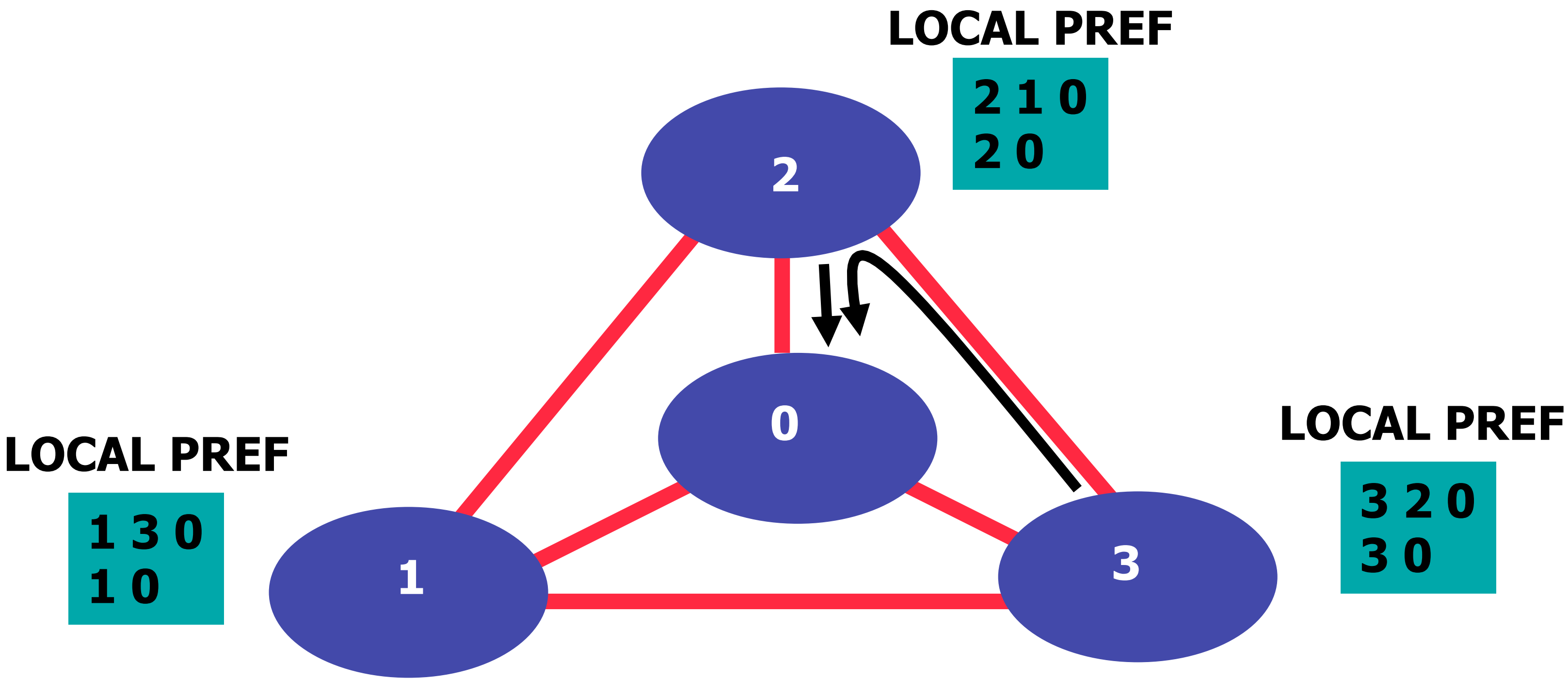
[figure: Jen Rexford]

Policies May Cause Persistent Oscillations ("Dispute Wheels")



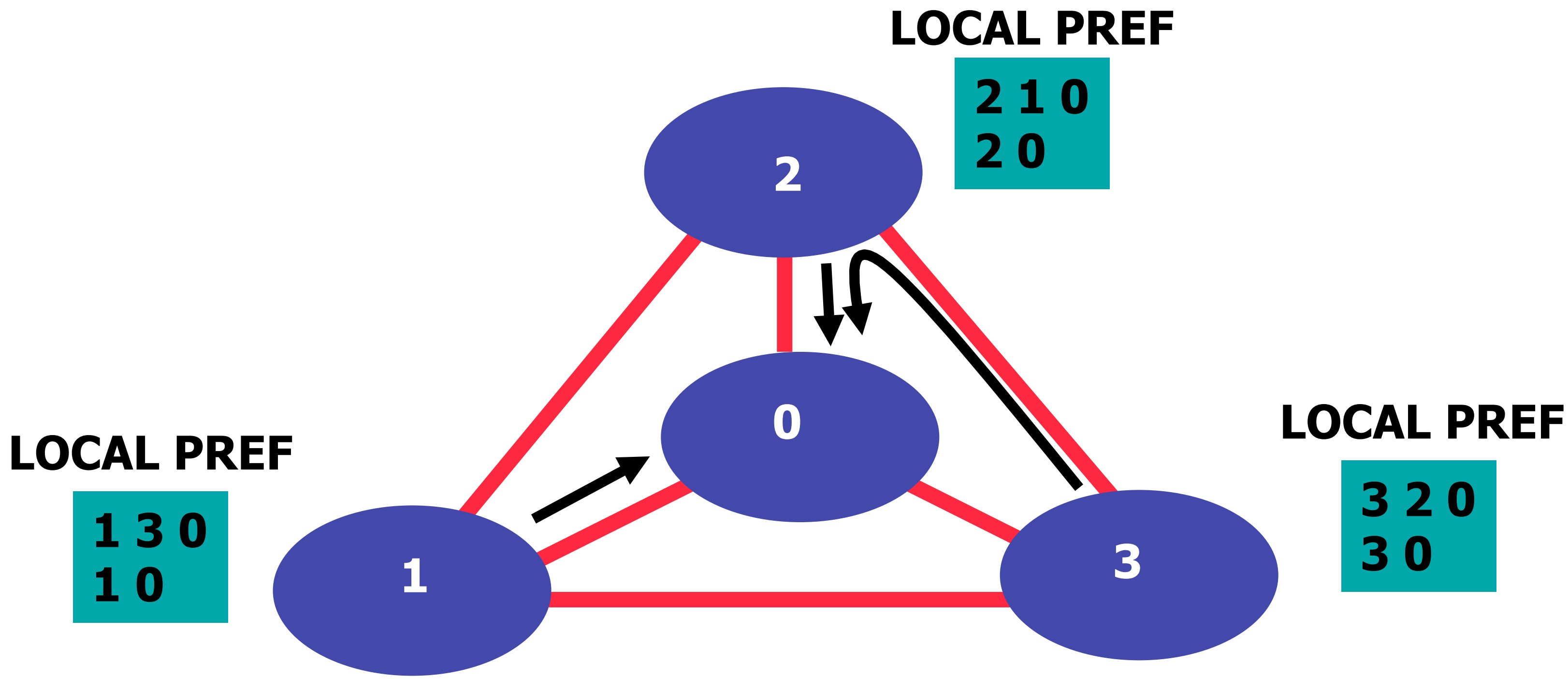
[figure: Jen Rexford]

Policies May Cause Persistent Oscillations ("Dispute Wheels")



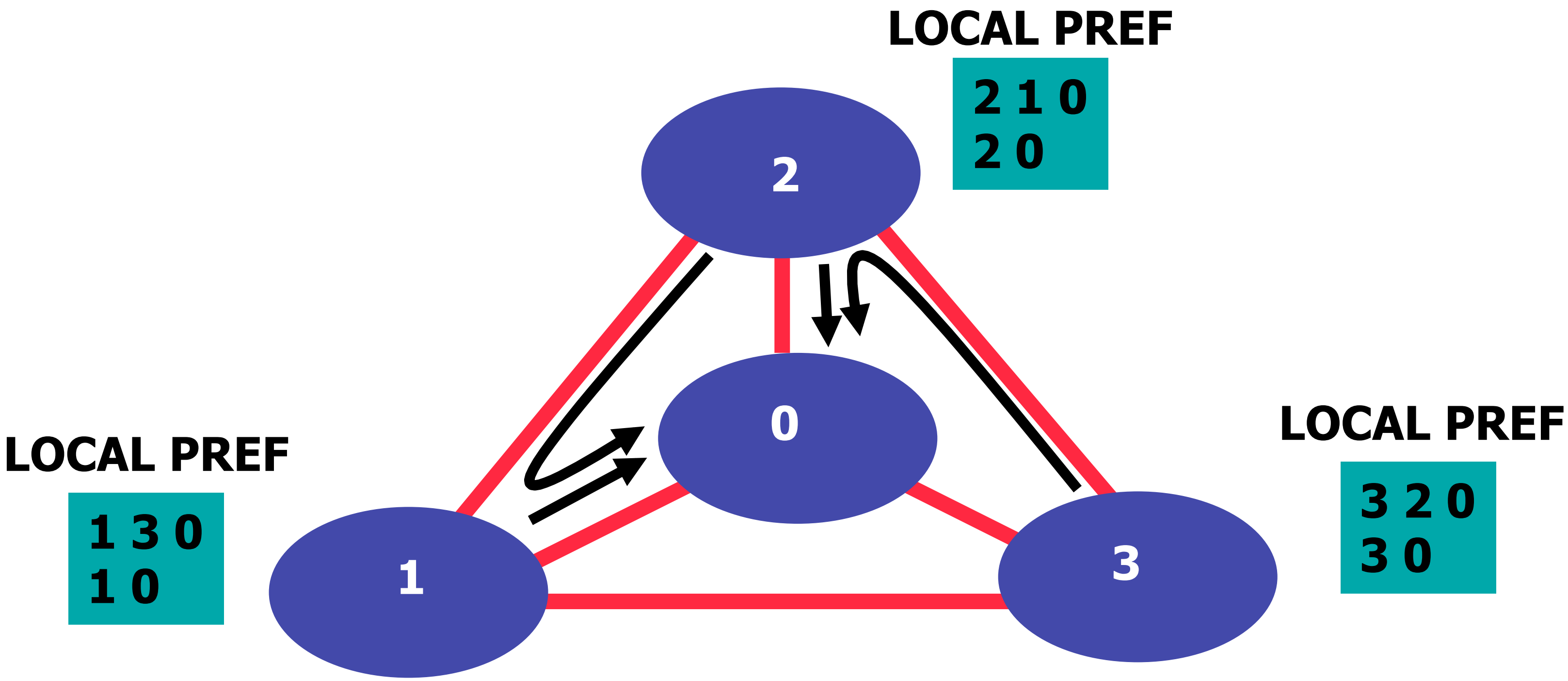
[figure: Jen Rexford]

Policies May Cause Persistent Oscillations ("Dispute Wheels")



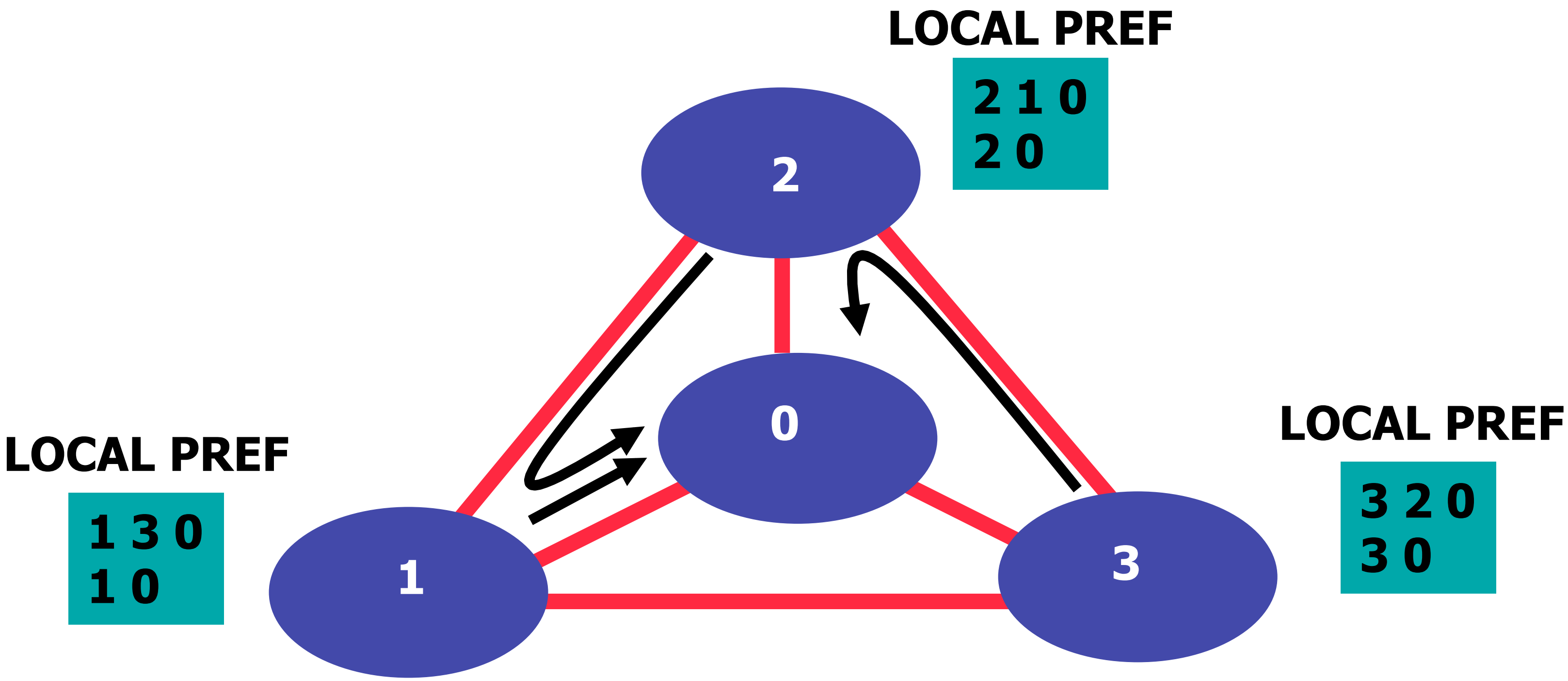
[figure: Jen Rexford]

Policies May Cause Persistent Oscillations ("Dispute Wheels")



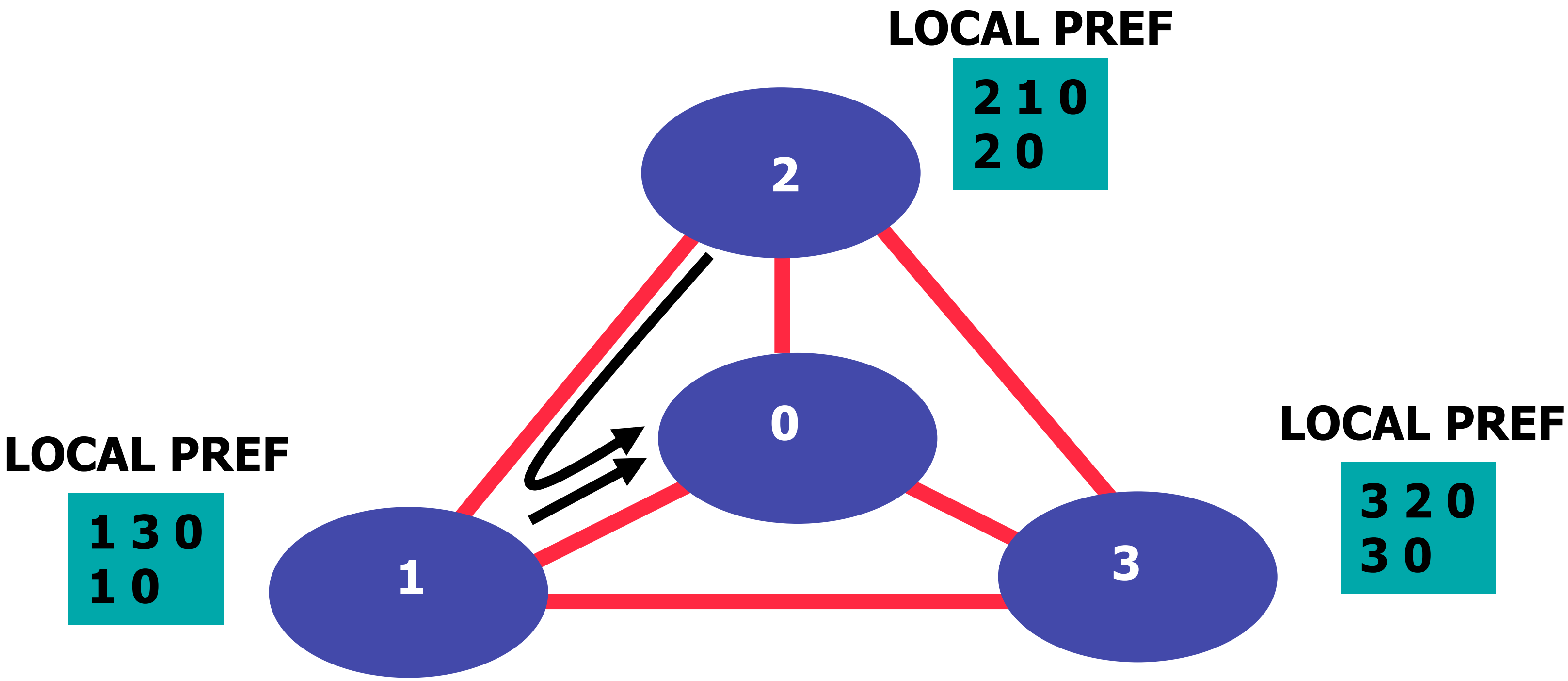
[figure: Jen Rexford]

Policies May Cause Persistent Oscillations ("Dispute Wheels")



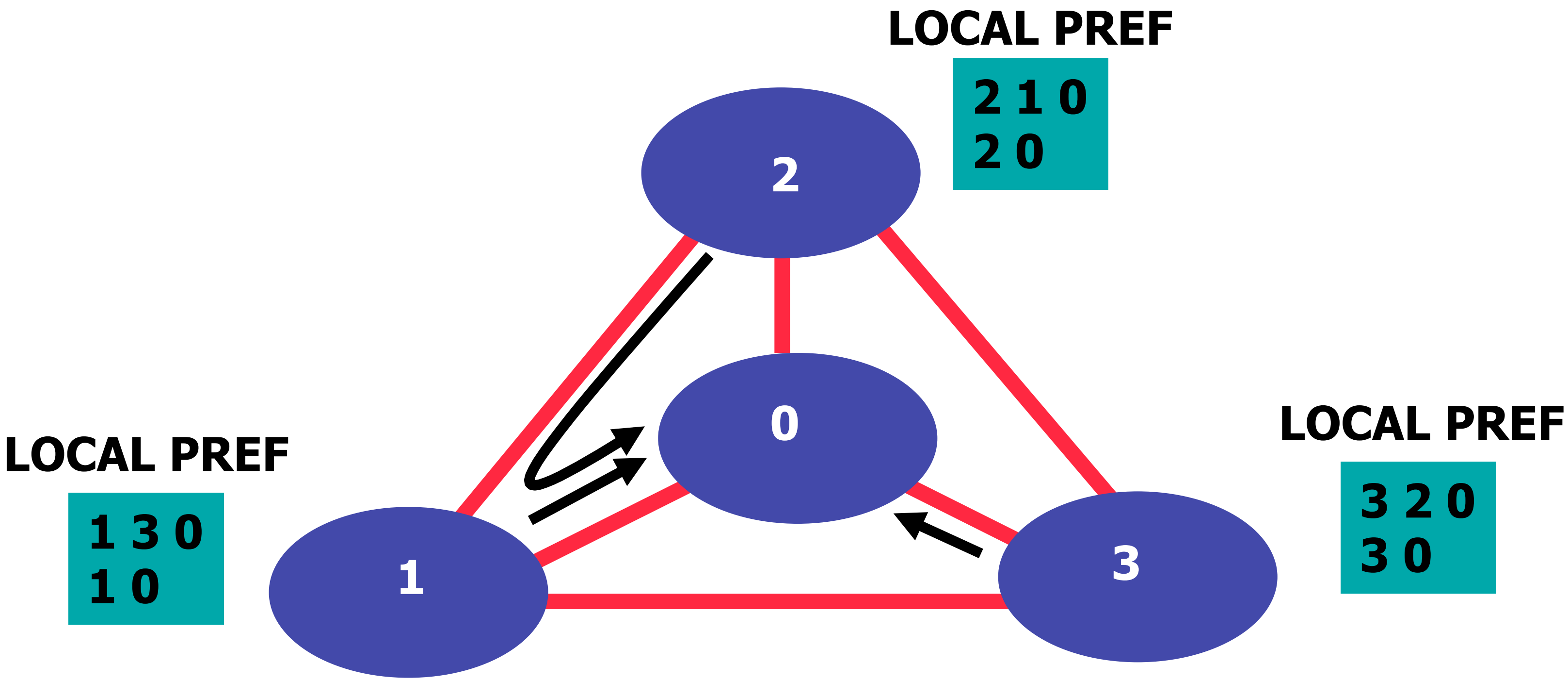
[figure: Jen Rexford]

Policies May Cause Persistent Oscillations ("Dispute Wheels")



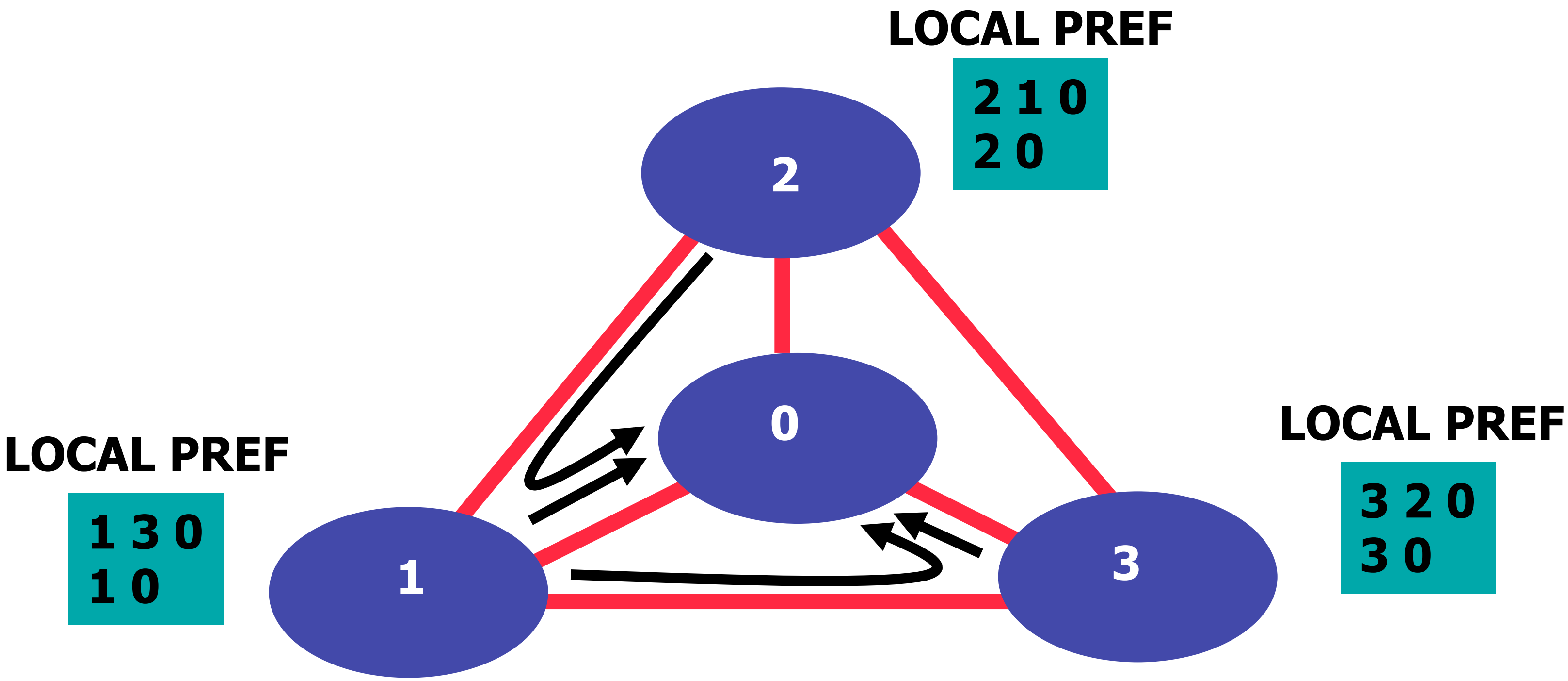
[figure: Jen Rexford]

Policies May Cause Persistent Oscillations ("Dispute Wheels")



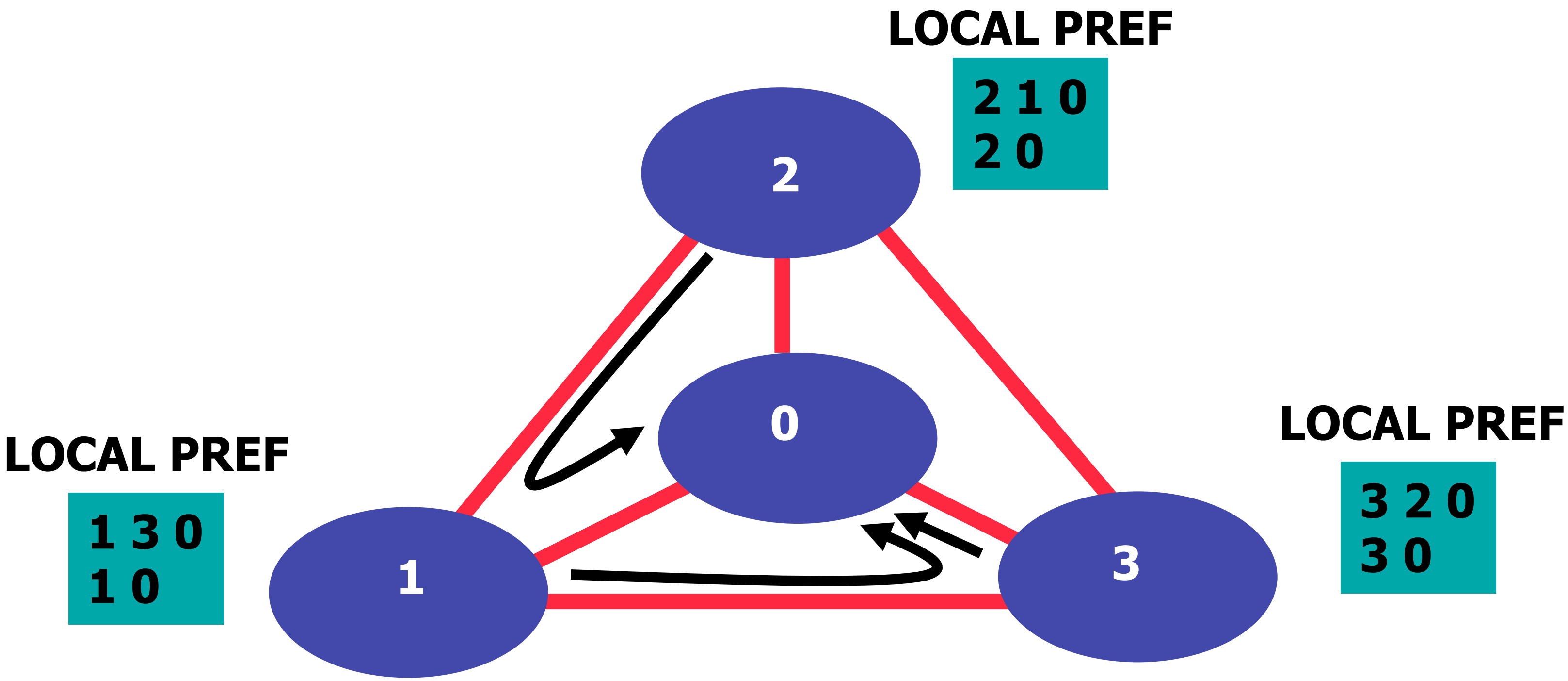
[figure: Jen Rexford]

Policies May Cause Persistent Oscillations ("Dispute Wheels")



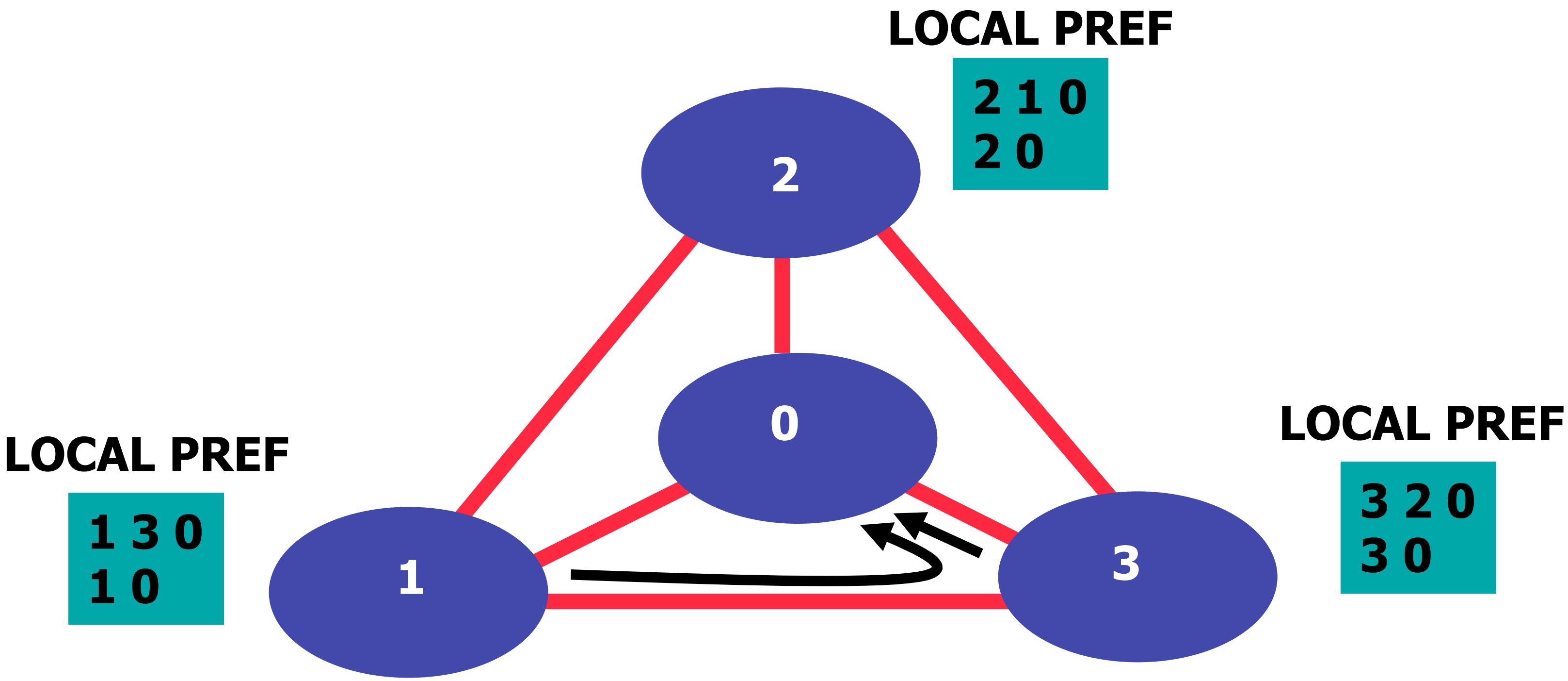
[figure: Jen Rexford]

Policies May Cause Persistent Oscillations ("Dispute Wheels")



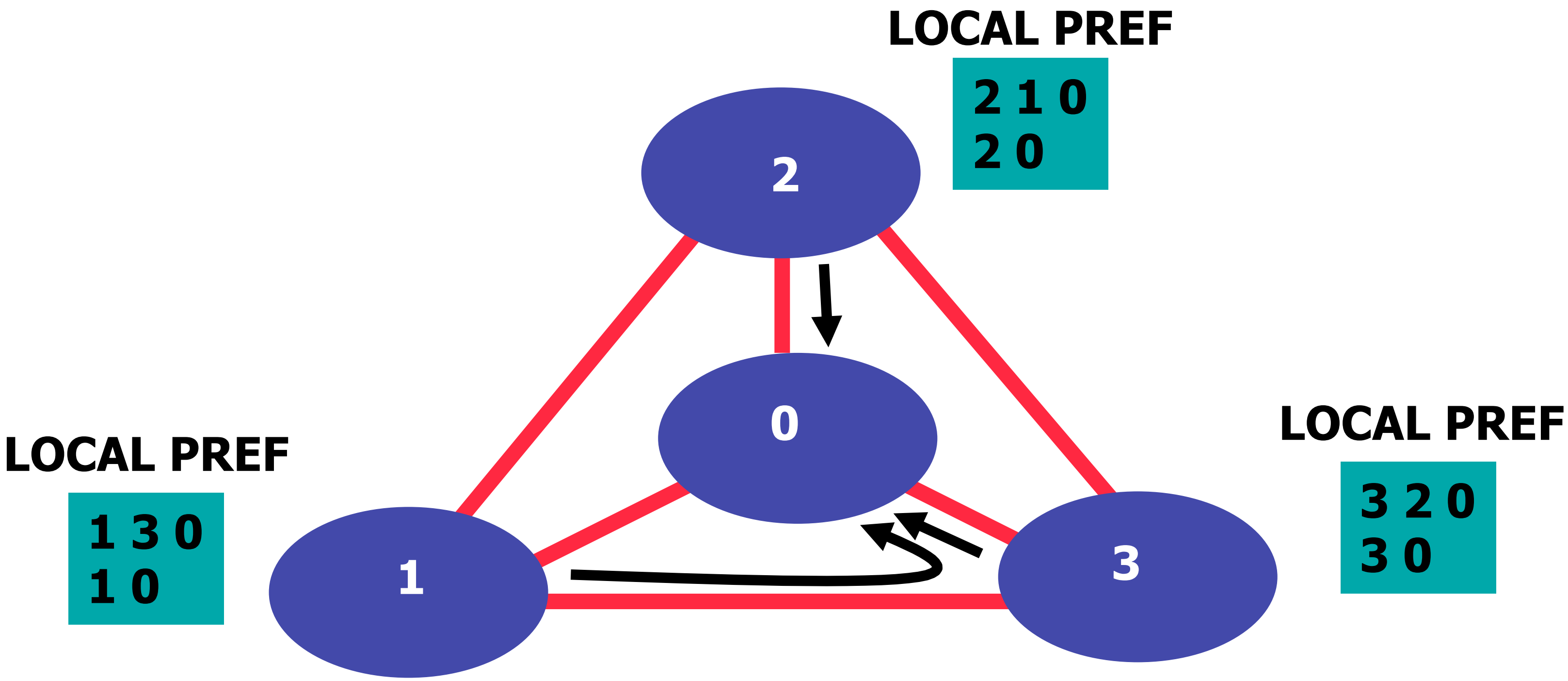
[figure: Jen Rexford]

Policies May Cause Persistent Oscillations ("Dispute Wheels")



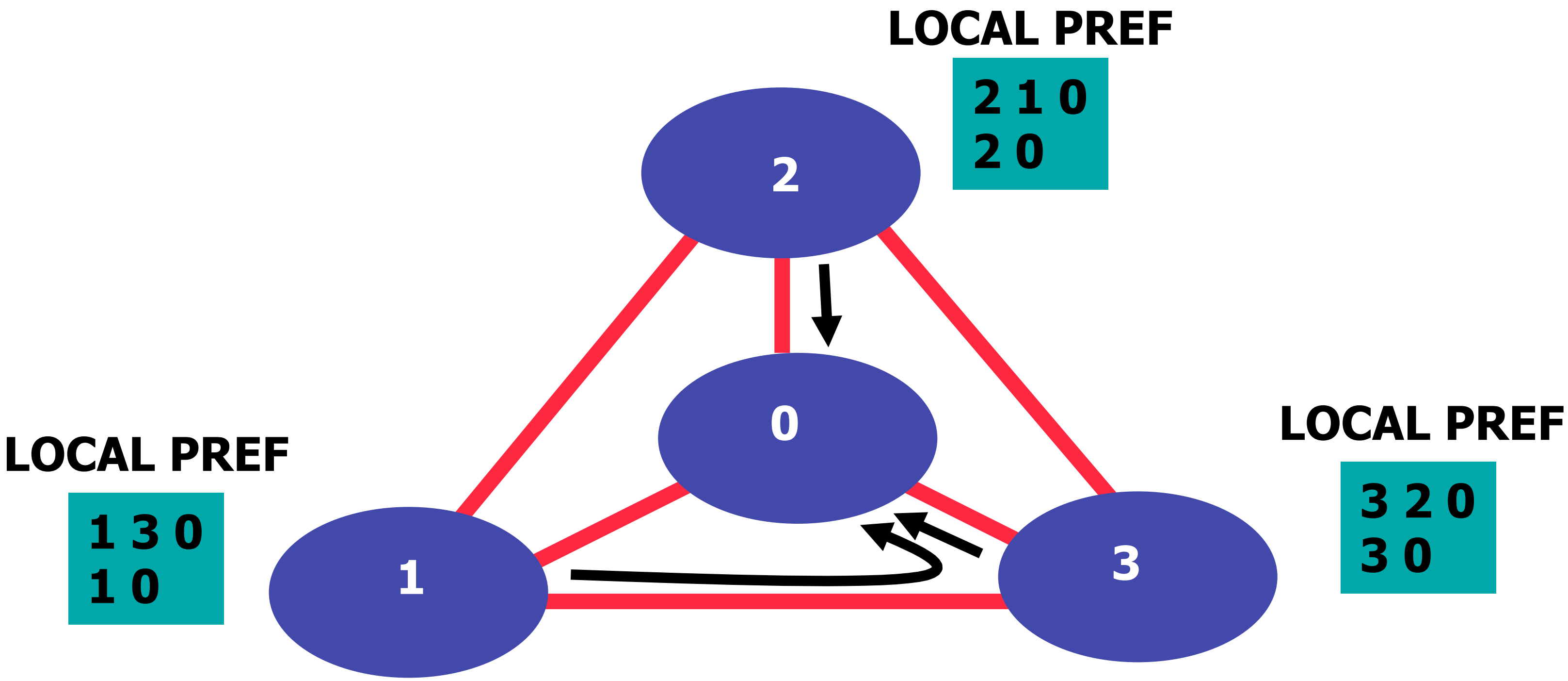
[figure: Jen Rexford]

Policies May Cause Persistent Oscillations ("Dispute Wheels")



[figure: Jen Rexford]

Policies May Cause Persistent Oscillations ("Dispute Wheels")



- Suppose each AS prefers two-hop path to direct one
- Repeats forever!

[figure: Jen Rexford]

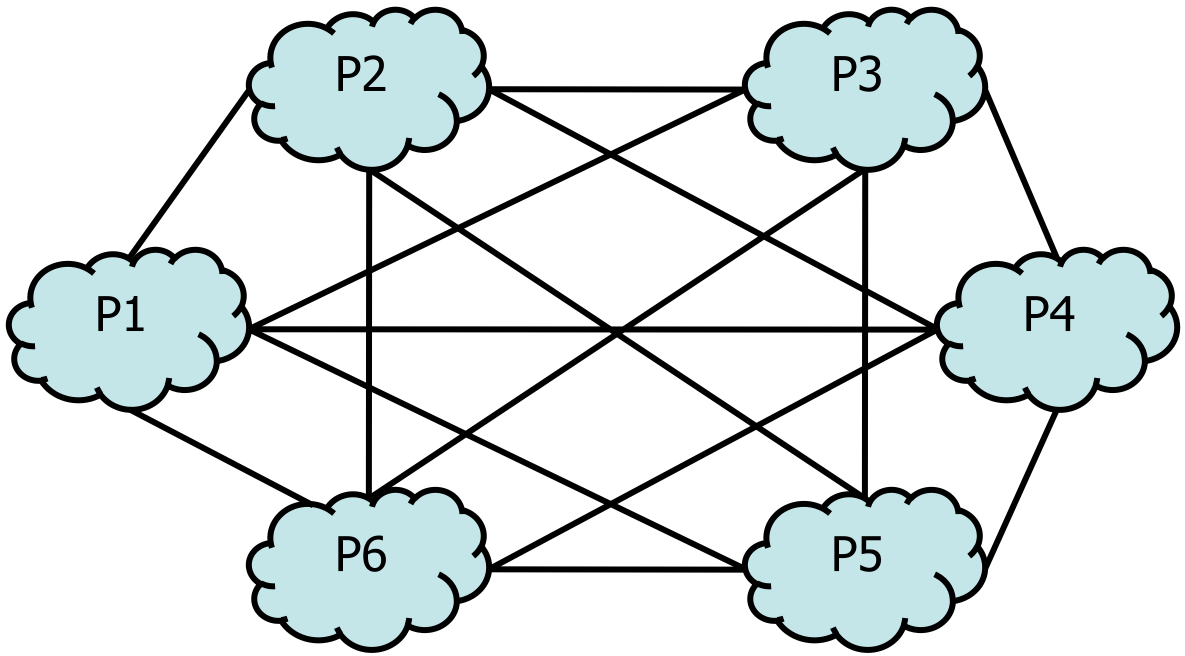
War Story: Depeering

- 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

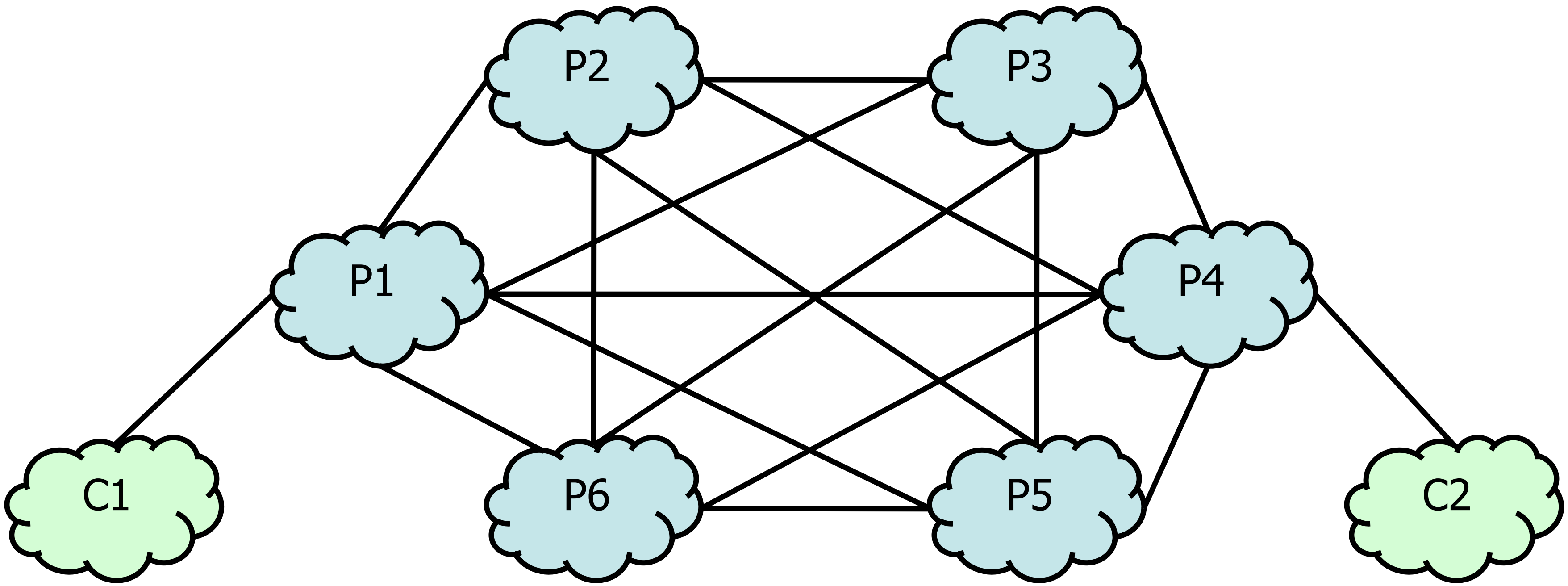
ISPs with no Transit Provider (as of January 2009)

- Qwest (AS209)
- Verizon (AS701)
- Sprint (AS1239)
- Telia (AS1299)
- XO (AS2828)
- NTT (AS2914)
- Level 3 (AS3356)
- Global Crossing (AS3549)
- Savvis (AS3561)
- Teleglobe (AS6453)
- Abovenet (AS6461)
- AT&T (AS7018)

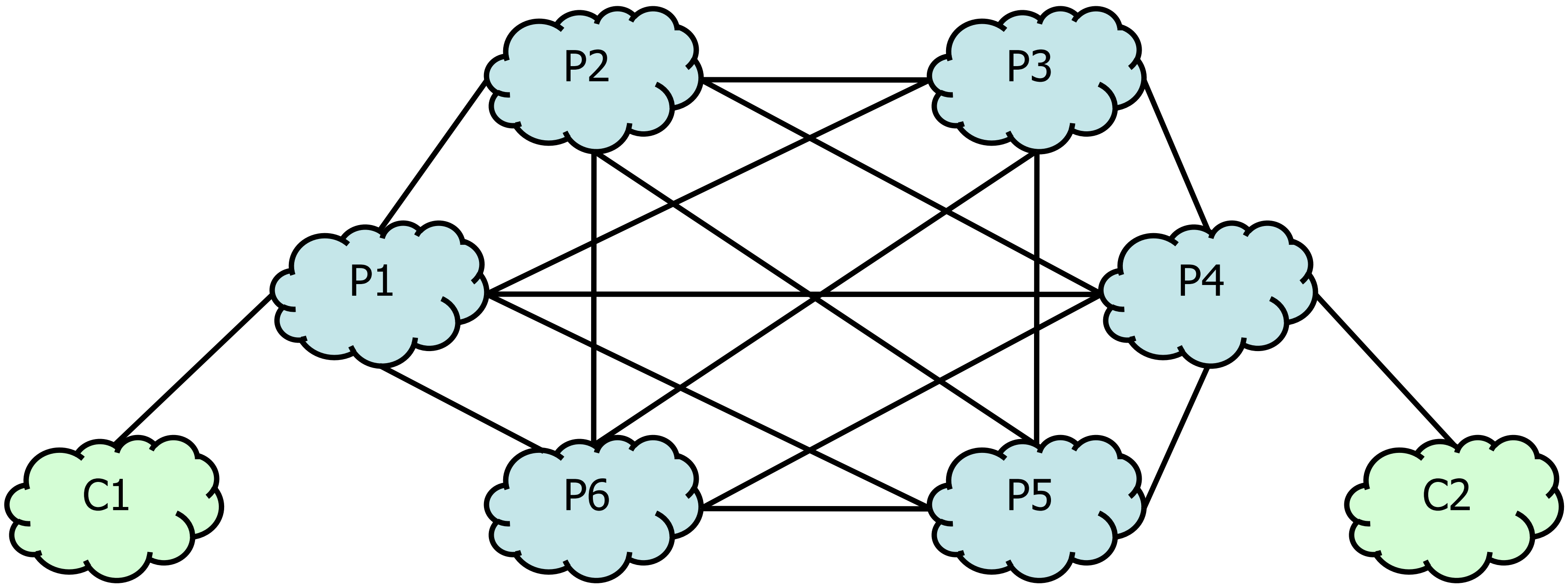
Full-Mesh Peering



Full-Mesh Peering

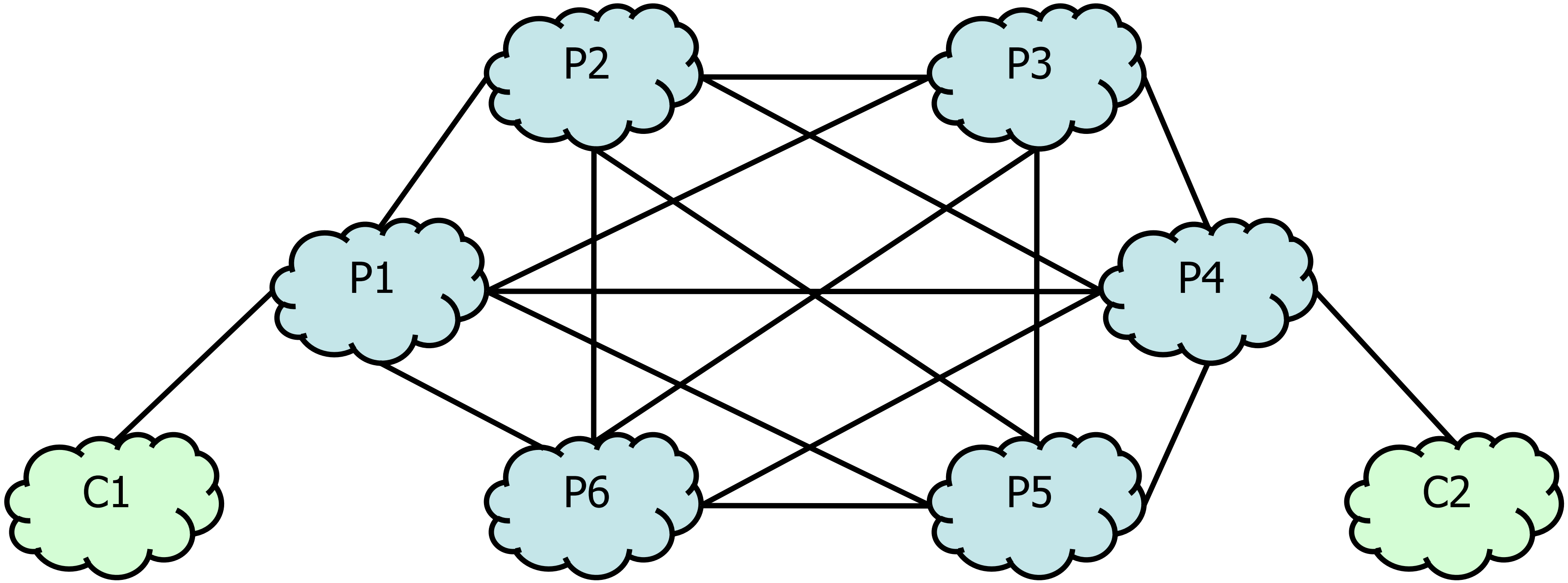


Full-Mesh Peering

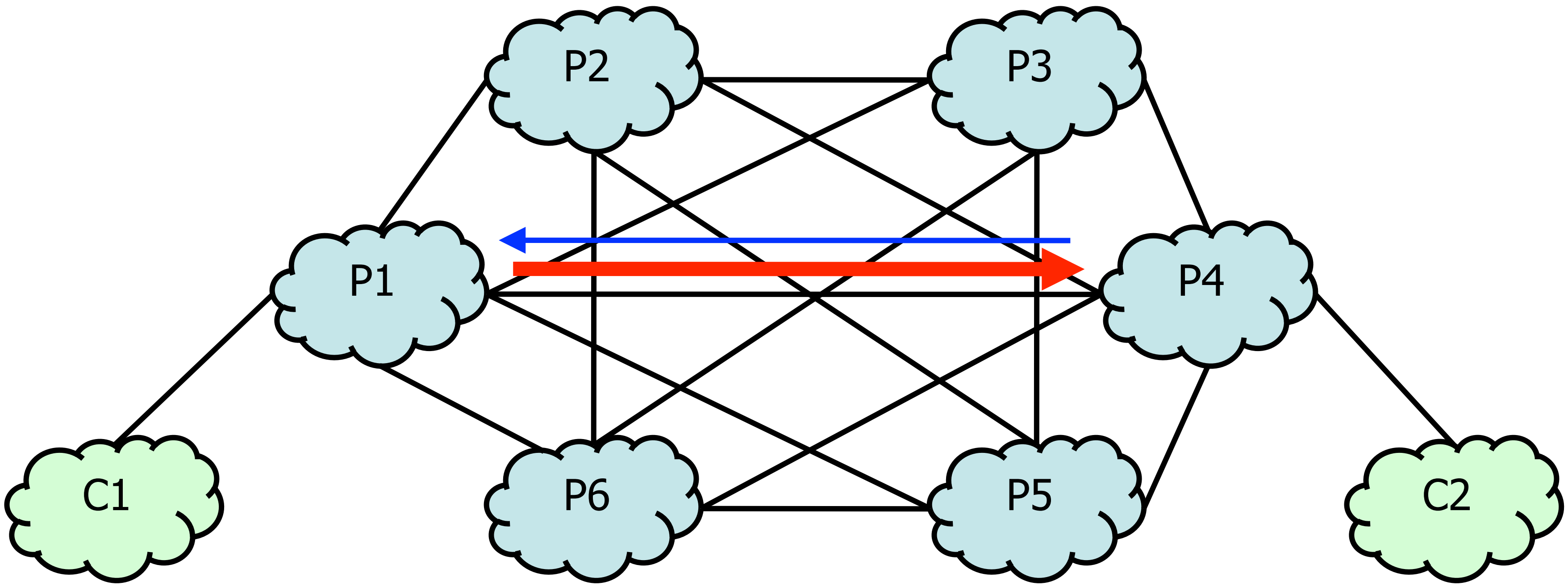


For Internet to be connected, **all** ISPs who do not buy transit service **must** be connected in full mesh!

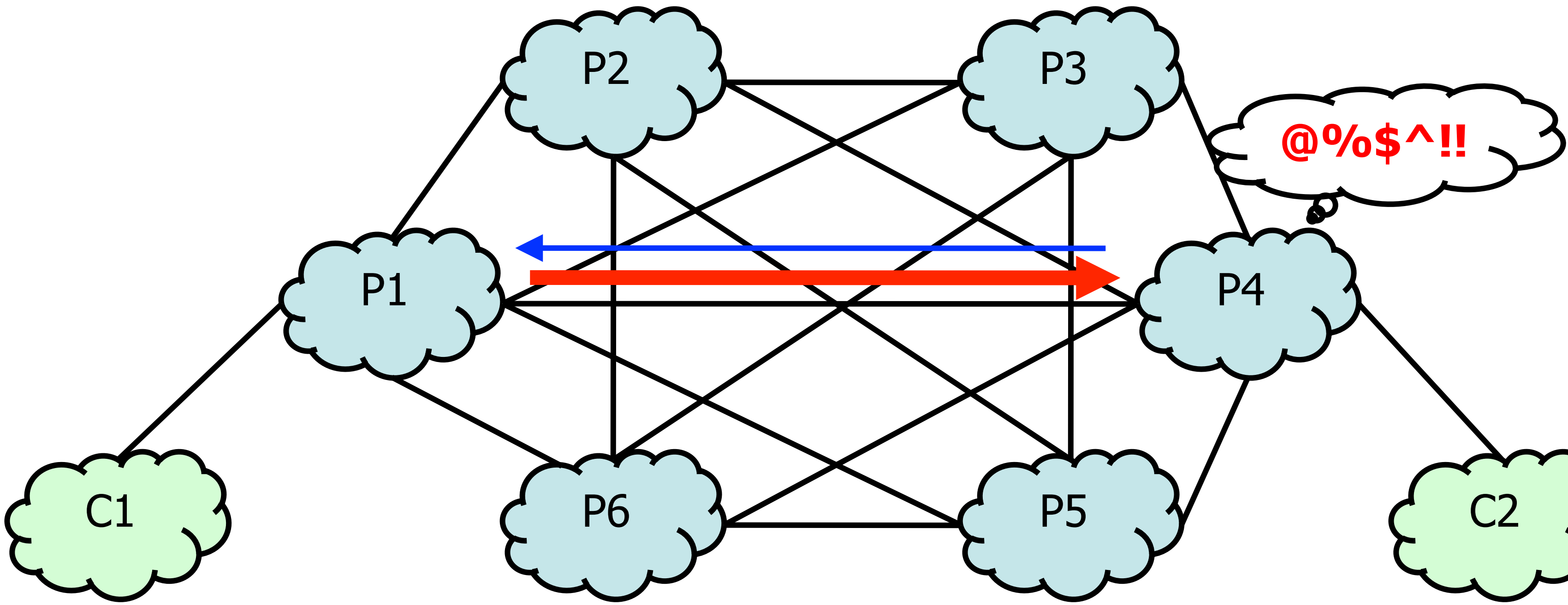
A Peers' Quarrel: Depeering



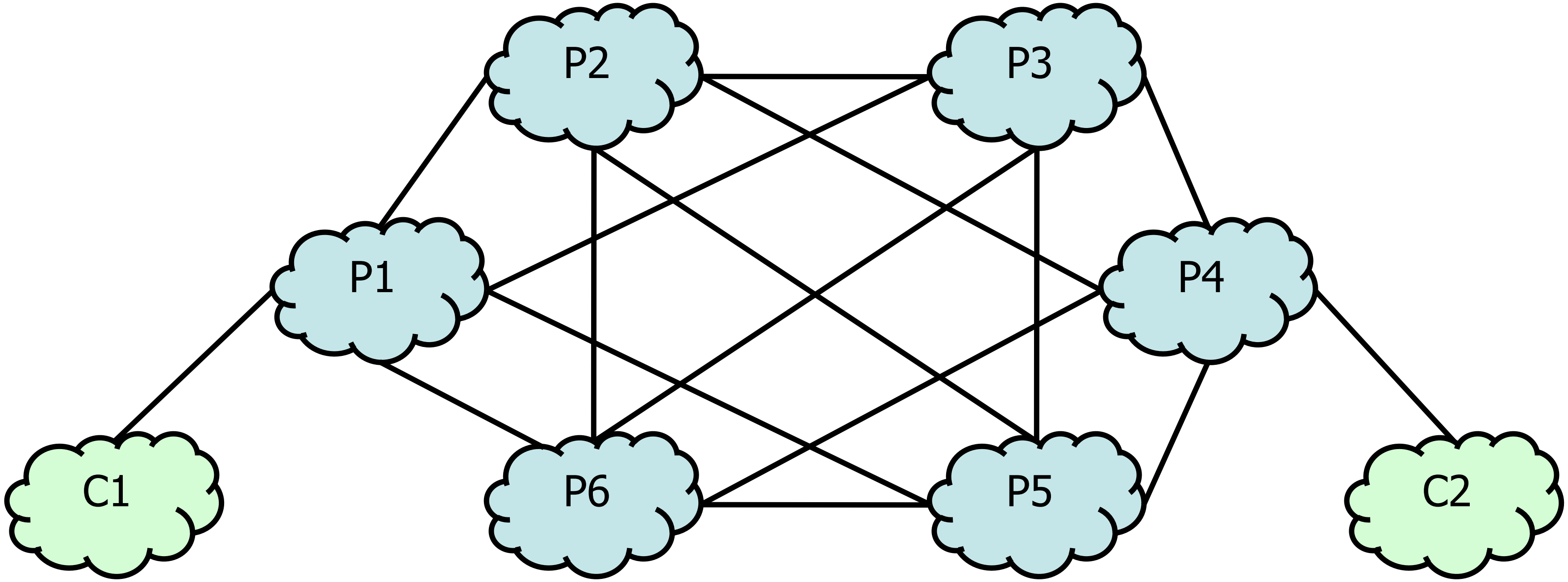
A Peers' Quarrel: Depeering



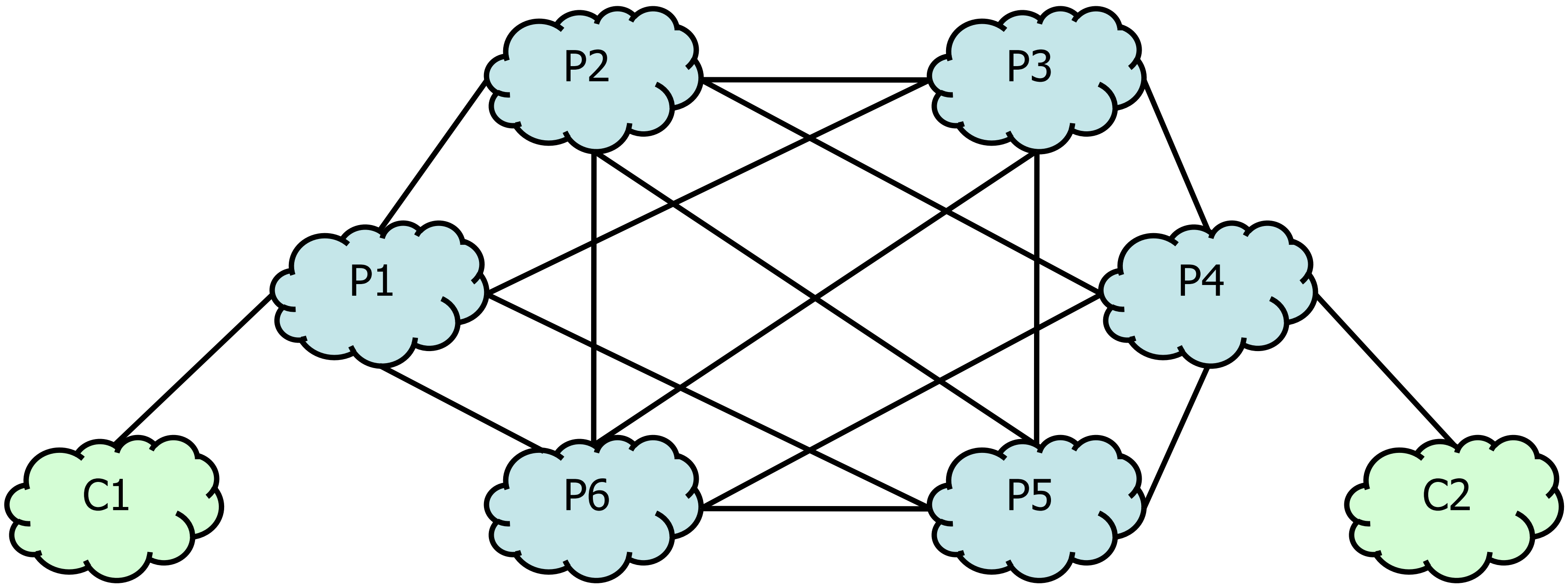
A Peers' Quarrel: Depeering



A Peers' Quarrel: Depeering



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., New York Court System, 128 educational institutions, Pfizer, Merck, Northup Grumman, ...

Summary: Inter-Domain Routing with BGP

- Inter-domain routing chiefly concerned with policy, not optimality
  - Economic motivation: cost of carrying traffic
  - Different relationships demand different routing: customer-provider vs. peering
- BGP: Path-Vector inter-domain routing protocol
  - Scalable in number of ASes
  - Route attributes support policy routing
  - Loop-free at AS granularity
  - Shortest ASPATHs achieved, after policy enforced
- Behavior and configuration of BGP very complex and poorly understood; open research problem!