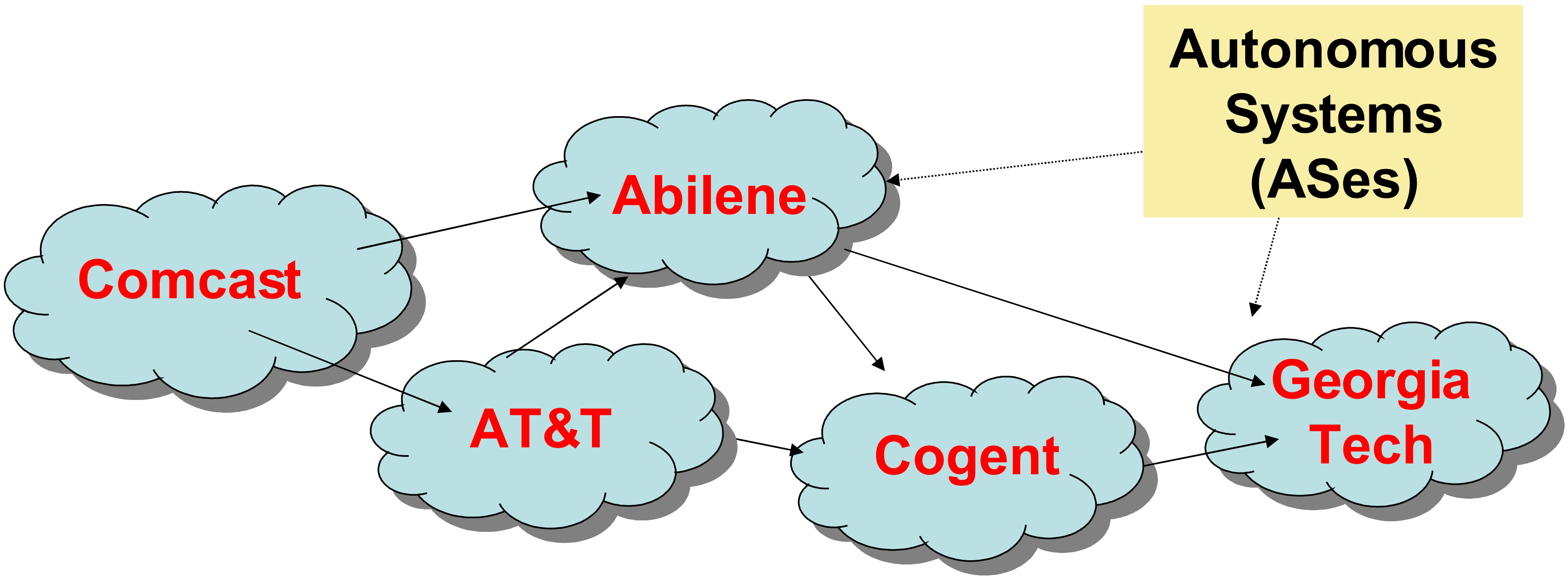
Intradomain Topology and Routing

Nick Feamster
CS 7260
January 17, 2007

Administrivia

- Problem Set 1: Slight delay
- Project groups: Next week
- Project ideas will go up over the weekend

Internet Routing Overview



- **Today:** Intradomain (*i.e.*, "intra-AS") routing
- **Monday:** Interdomain routing

Today: Routing Inside an AS

- Intra-AS **topology**
  - Nodes and edges
  - *Example:* Abilene

- Intradomain **routing** protocols
  - Distance Vector
    - Split-horizon/Poison-reverse
    - *Example:* RIP
  - Link State
    - *Example:* OSPF

Key Questions

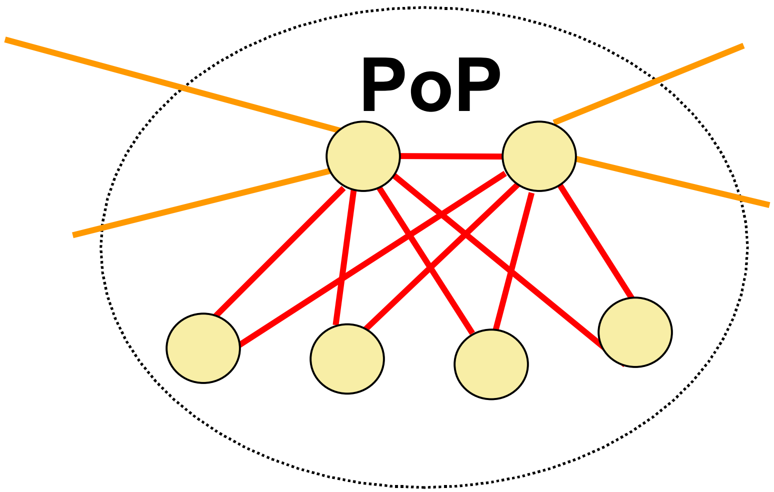
- **Where to place "nodes"?**
  - Typically in dense population centers
    - Close to other providers (easier interconnection)
    - Close to other customers (cheaper backhaul)
  - *Note:* A "node" may in fact be a group of routers, located in a single city. Called a "Point-of-Presence" (PoP)
- **Where to place "edges"?**
  - Often constrained by location of fiber

Point-of-Presence (PoP)

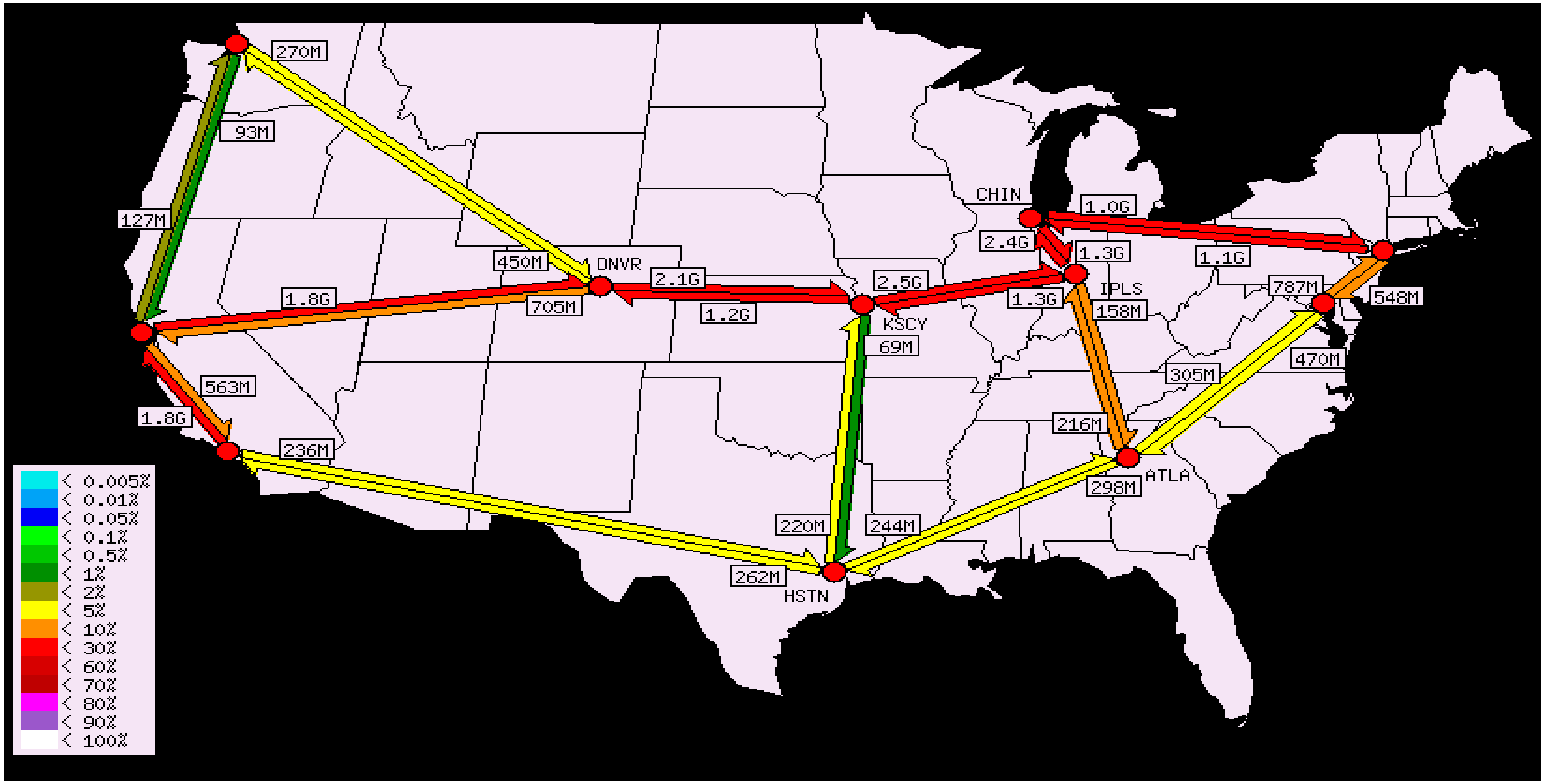
- A "cluster" of routers in a single physical location

- Inter-PoP links
  - Long distances
  - High bandwidth

- Intra-PoP links
  - Cables between racks or floors
  - Aggregated bandwidth

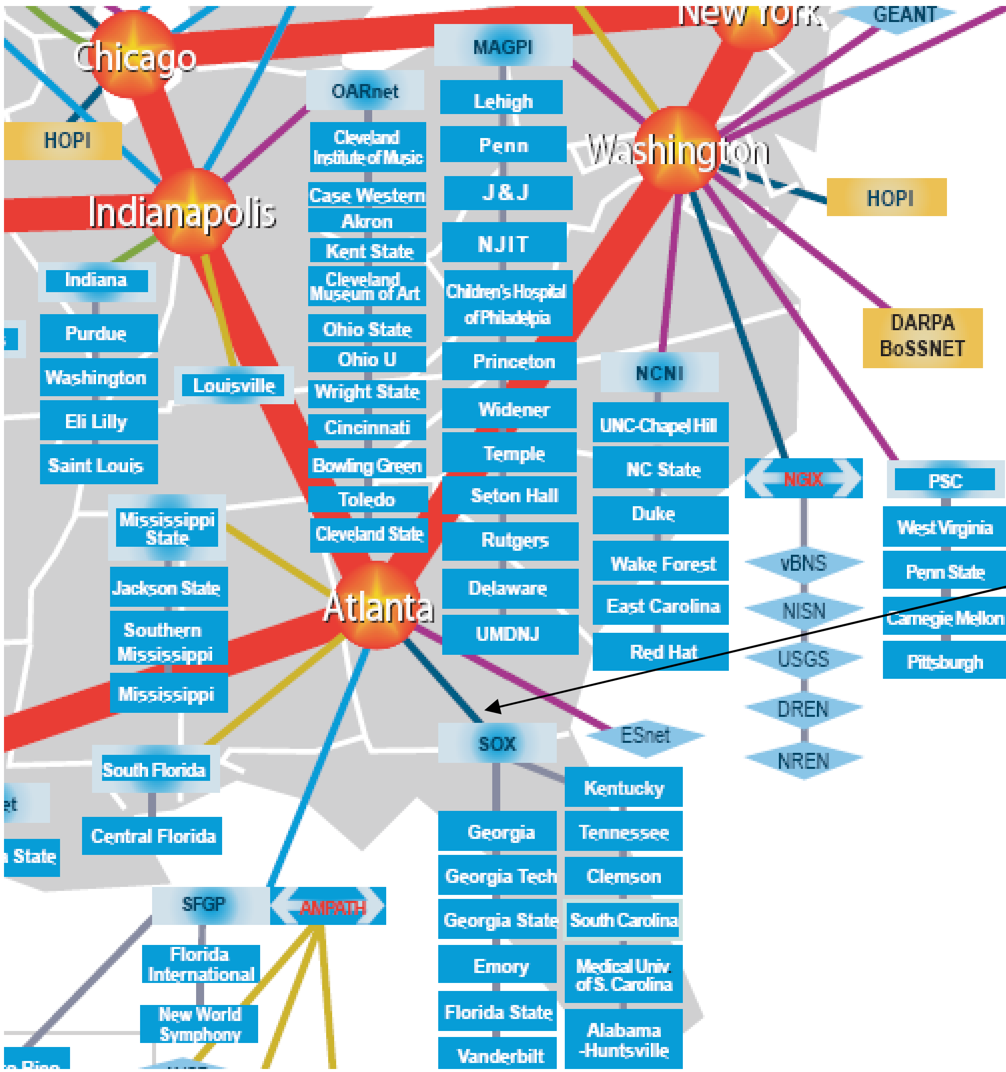


Example: Abilene Network Topology



- Problem Set 1 will have a problem dealing with Abilene router configurations/topology.

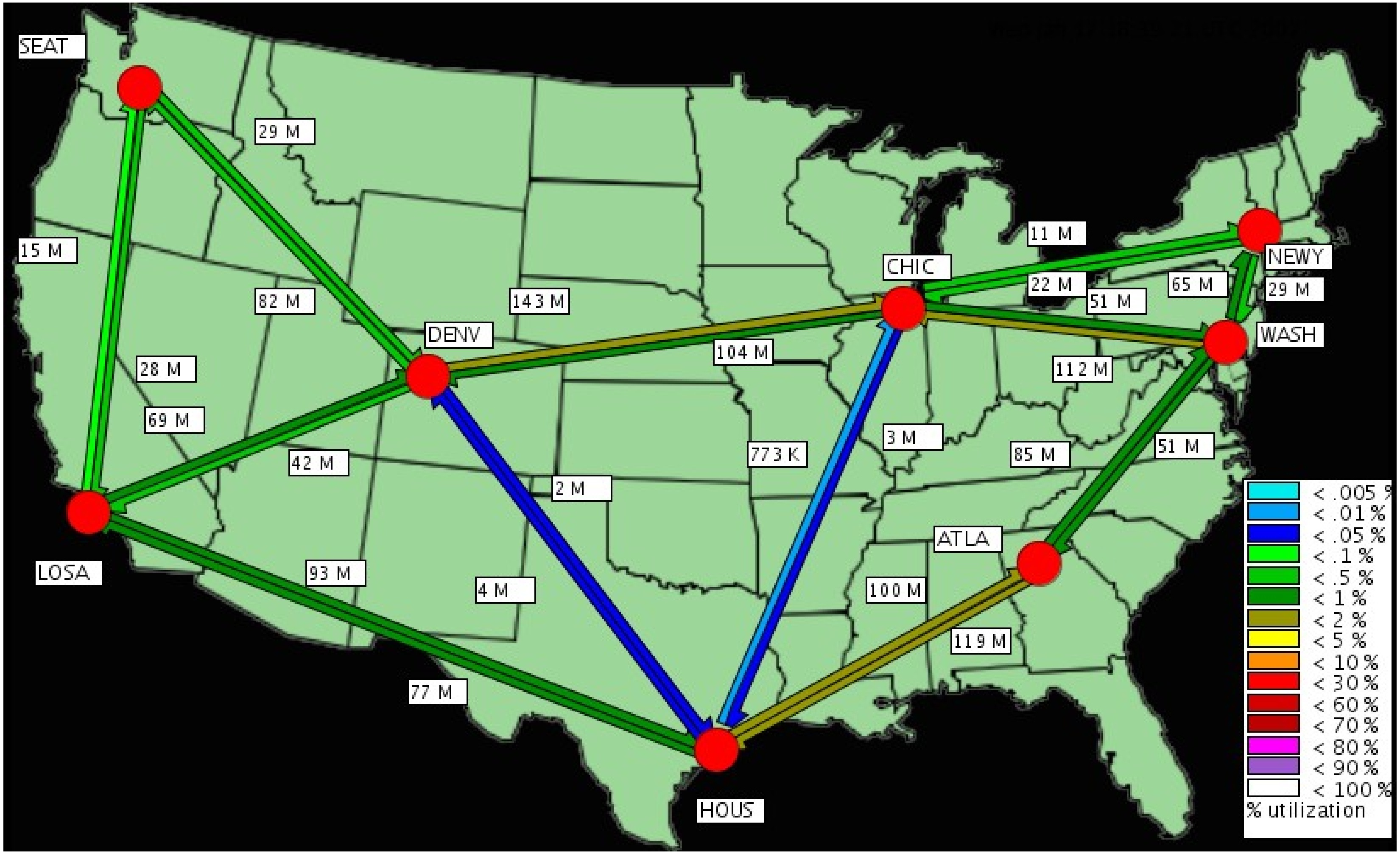
Where's Georgia Tech?



10GigE (10GbpS uplink) Southeast Exchange

(SOX) is at 56 Marietta Street

Recent Development: NLR Packet Net

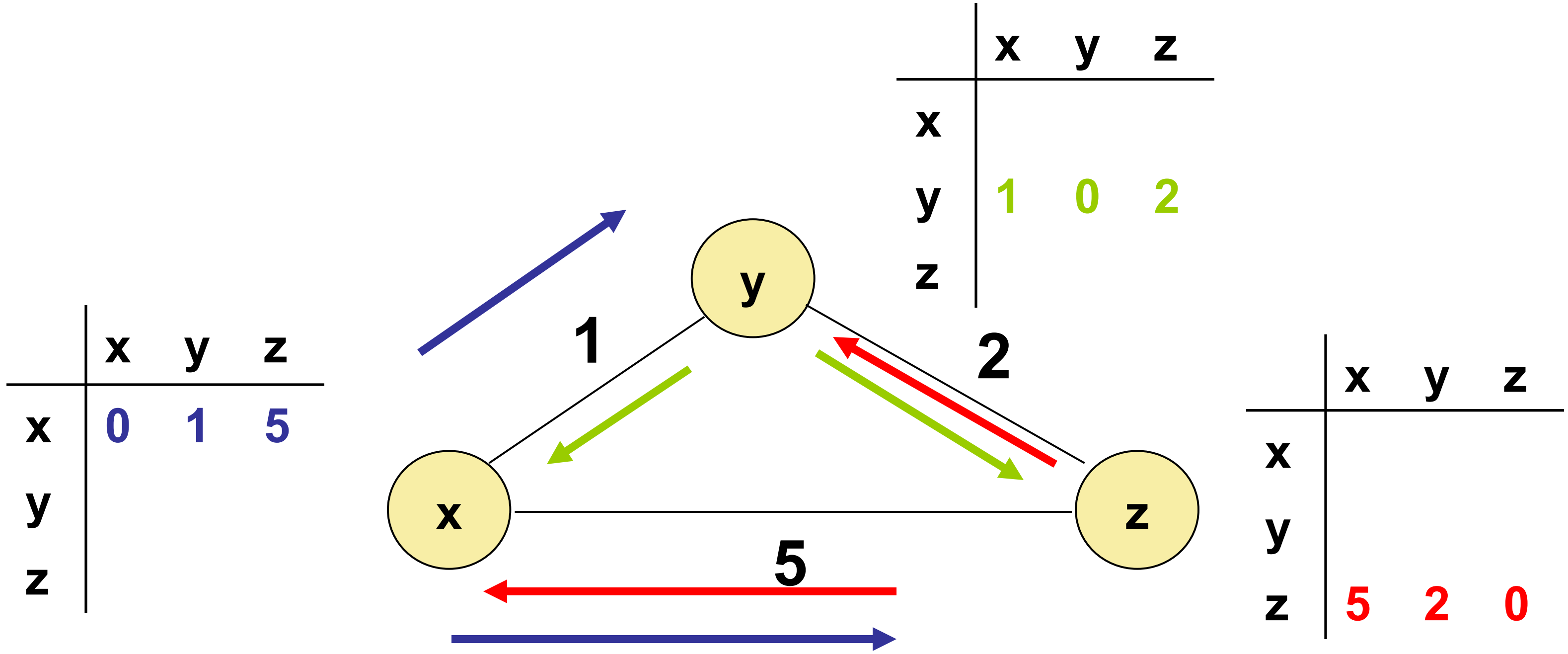


Problem: Routing

- **Routing:** the process by which nodes discover where to forward traffic so that it reaches a certain node

- **Within an AS:** there are two “styles”
  - Distance vector
  - Link State

Distance-Vector Routing

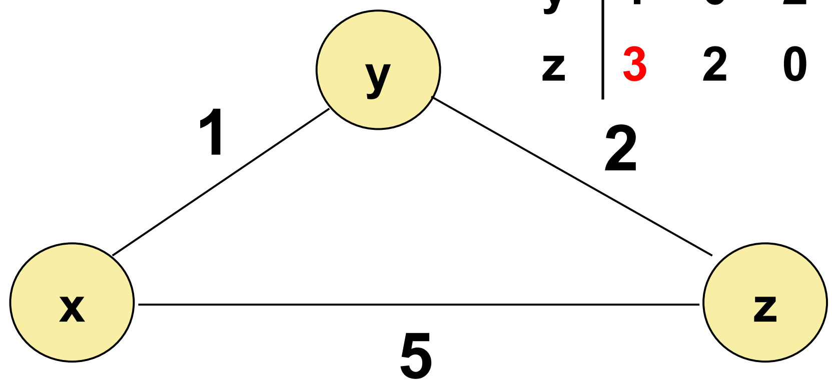


- Routers send routing table copies to neighbors
- Routers compute costs to destination based on shortest available path
- Based on Bellman-Ford Algorithm
  - $d_x(y) = \min_v\{ c(x,v) + d_v(y) \}$
  - Solution to this equation *is* x's forwarding table

Good News Travels Quickly

	x	y	z
x	0	1	3
y	1	0	2
z	3	2	0

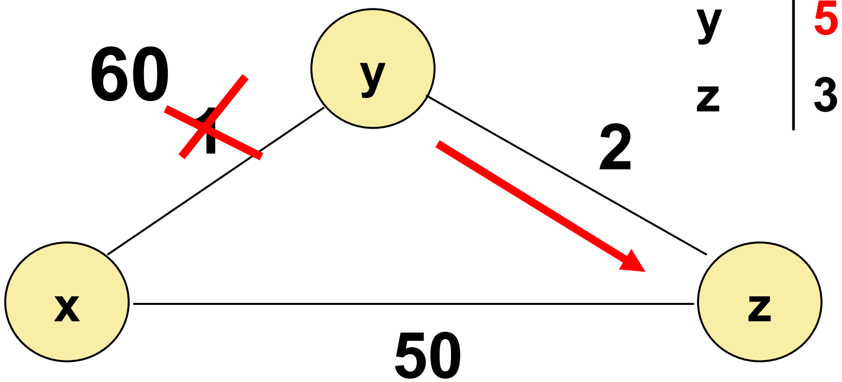
	x	y	z
x	0	1	3
y	1	0	2
z	3	2	0



	x	y	z
x	0	1	3
y	1	0	2
z	3	2	0

- When costs decrease, network converges quickly

Problem: Bad News Travels Slowly

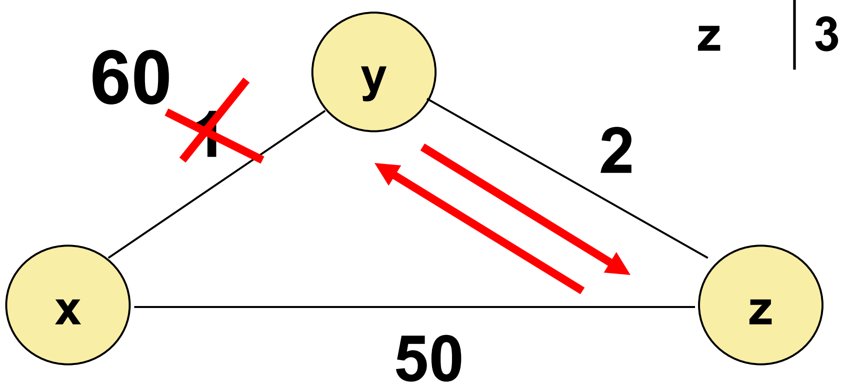


	x	y	z
x	0	60	50
y	5	0	2
z	3	2	0

	x	y	z
x	0	60	50
y	5	0	2
z	7	2	0

Note also that there is a forwarding loop between *y* and *z*.

It Gets Worse



	x	y	z
x	0	60	50
y	5	0	2
z	3	2	0

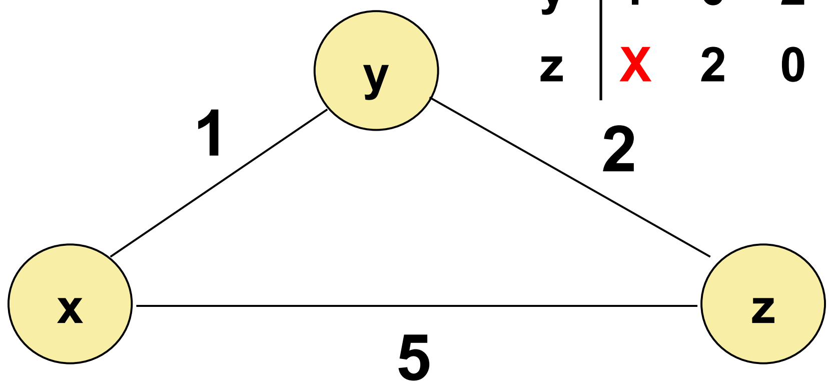
	x	y	z
x	0	60	50
y	5	0	2
z	7	2	0

- **Question:** How long does this continue?
- **Answer:** Until *z's* path cost to *x* via *y* is greater than 50.

"Solution": Poison Reverse

	x	y	z
x	0	1	X
y	1	0	2
z	X	2	0

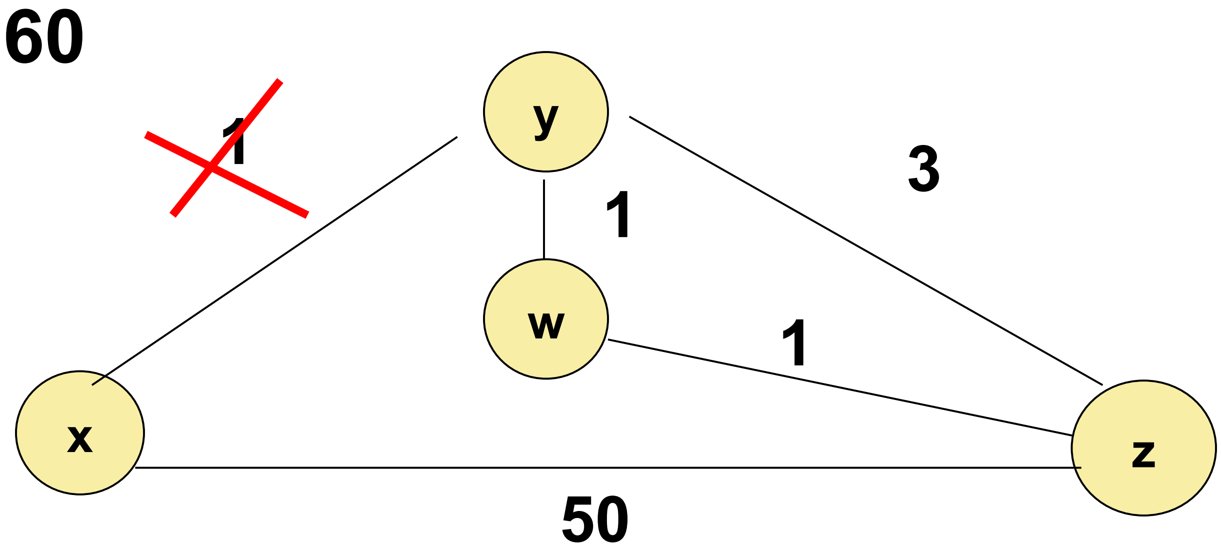
	x	y	z
x	0	1	3
y	1	0	2
z	3	2	0



	x	y	z
x	0	1	3
y	1	0	2
z	3	2	0

- If *z* routes through *y* to get to *x*, *z* advertises infinite cost for *x to* y
- Does poison reverse always work?

Does Poison Reverse Always Work?



Example: Routing Information Protocol

- Earliest IP routing protocol (1982 BSD)
  - Version 1: RFC 1058
  - Version 2: RFC 2453
- Features
  - Edges have unit cost
  - "Infinity" = 16
- Sending Updates
  - Router listens for updates on UDP port 520
  - Message can contain up to 25 table entries

RIP Updates

- Initial
  - When router first starts, asks for copy of table for every neighbor
  - Uses it to iteratively generate own table
- Periodic
  - Table refresh every 30 seconds
- Triggered
  - When every entry changes, send copy of entry to neighbors
    - Except for one causing update (*split horizon* rule)
  - Neighbors use to update their tables

RIP: Staleness and Oscillation Control

- Small value for Infinity
  - Count to infinity doesn't take very long
- Route Timer
  - Every route has timeout limit of 180 seconds
    - Reached when haven't received update from next hop for 6 periods
  - If not updated, set to infinity
  - *Soft-state*
- Behavior
  - When router or link fails, can take minutes to stabilize

Link-State Routing

- ***Idea:*** distribute a network map
- Each node performs shortest path (SPF) computation between itself and all other nodes
- Initialization step
  - Add costs of immediate neighbors, *D(v)*, else infinite
  - Flood costs *c(u,v)* to neighbors, *N*
- For some *D(w)* that is not in *N*
  - $D(v) = min(c(u,w) + D(w), D(v))$

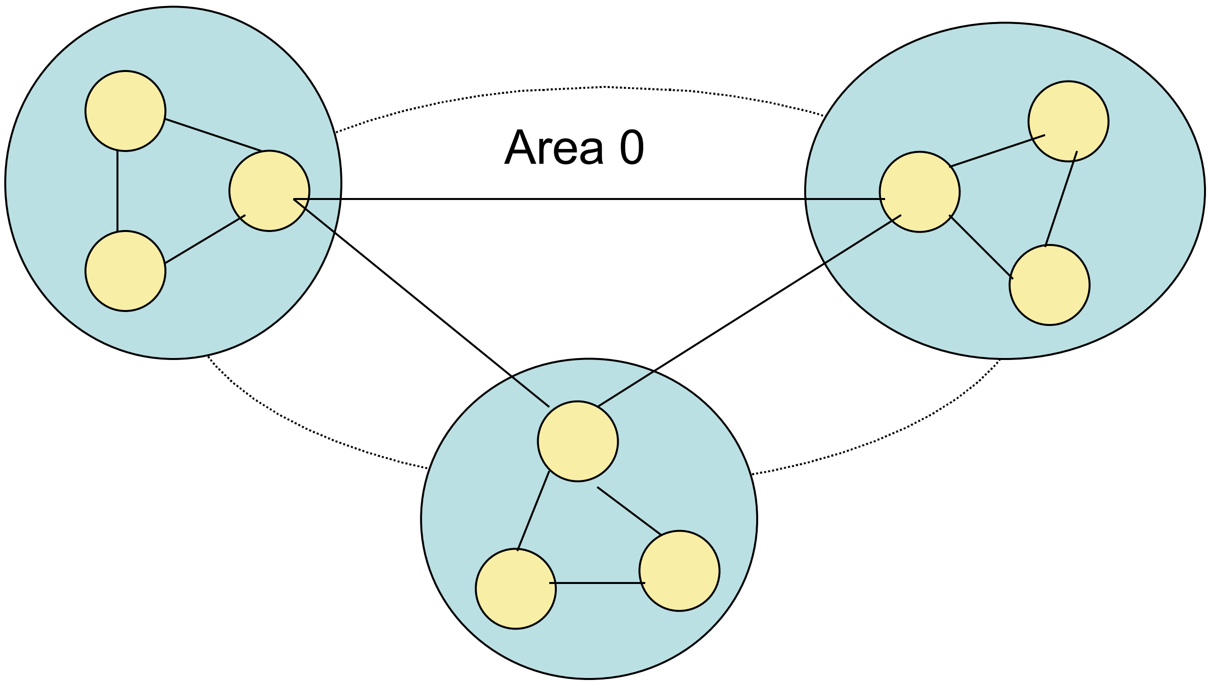
Link-State vs. Distance-Vector

- **Convergence**
  - DV has count-to-infinity
  - DV often converges slowly (minutes)
  - Odd timing dependencies in DV
- **Robustness**
  - Route calculations a bit more robust under link-state.
  - DV algorithms can advertise incorrect least-cost paths
- **Bandwidth Consumption** for Messages
- **Computation**
- **Security**

OSPF: Salient Features

- Dijkstra, plus some additional features
- Equal-cost multipath
- Support for hierarchy: Inter-Area Routing

Example: Open Shortest Paths First (OSPF)

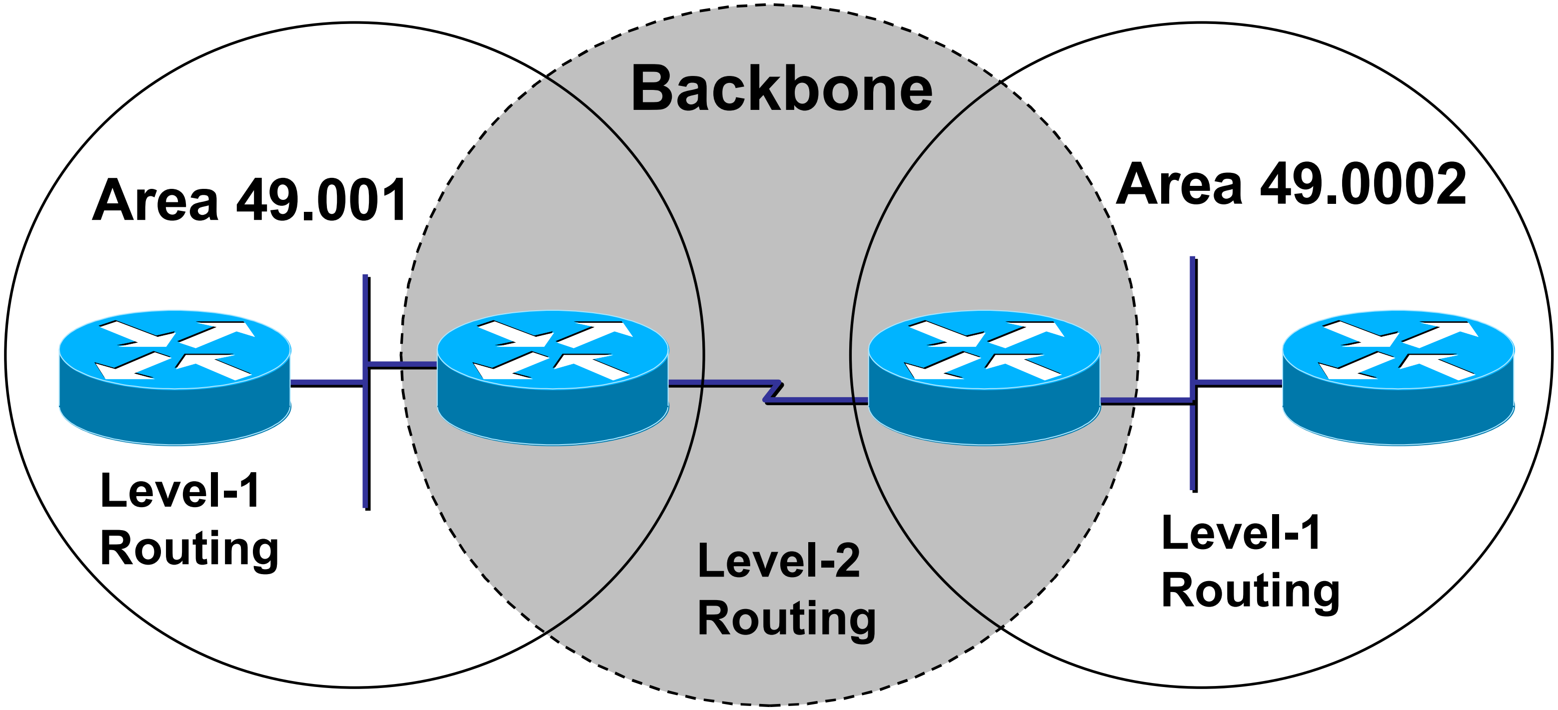


- Key Feature: hierarchy
- Network's routers divided into *areas*
- *Backbone area* is area 0
- Area 0 routers perform SPF computation
  - All inter-area traffic travles through Area 0 routers ("border routers")

Example: IS-IS

- Originally: ISO Connectionless Network Protocol (CLNP) .
  - CLNP: ISO equivalent to IP for datagram delivery services
  - ISO 10589 or RFC 1142
- Later: Integrated or Dual IS-IS (RFC 1195)
  - IS-IS adapted for IP
  - Doesn't use IP to carry routing messages
- OSPF more widely used in enterprise, IS-IS in large service providers

Hierarchical Routing in IS-IS



- Like OSPF, 2-level routing hierarchy
  - Within an area: level-1
  - Between areas: level-2
  - Level 1-2 Routers: Level-2 routers may also participate in L1 routing

Level-1 vs. Level-2 Routing

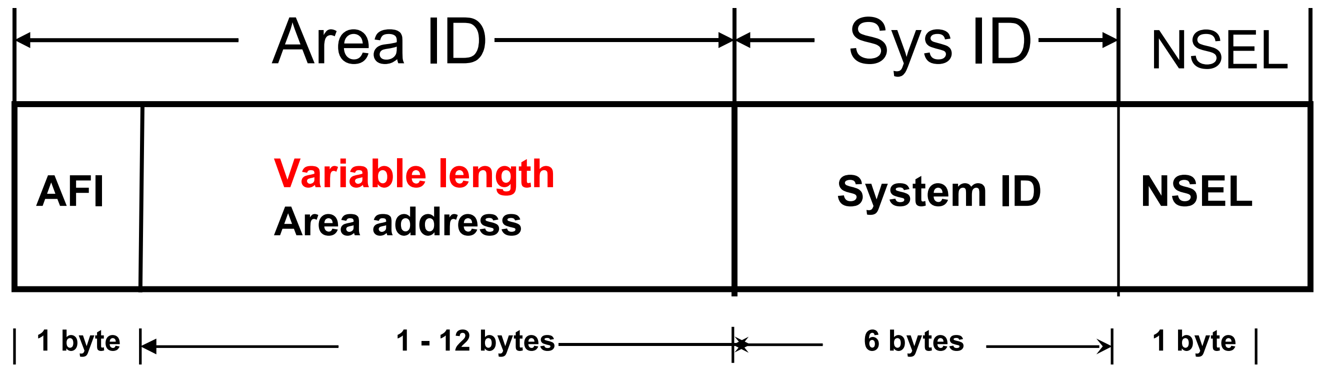
Level 1 routing

- Routing *within* an area
- Level 1 routers track links, routers, and end systems within L1 area
- L1 routers do not know the identity of destinations outside their area.
- A L 1 router forwards all traffic for destinations outside its area to the nearest L2 router within its area.

Level 2 routing

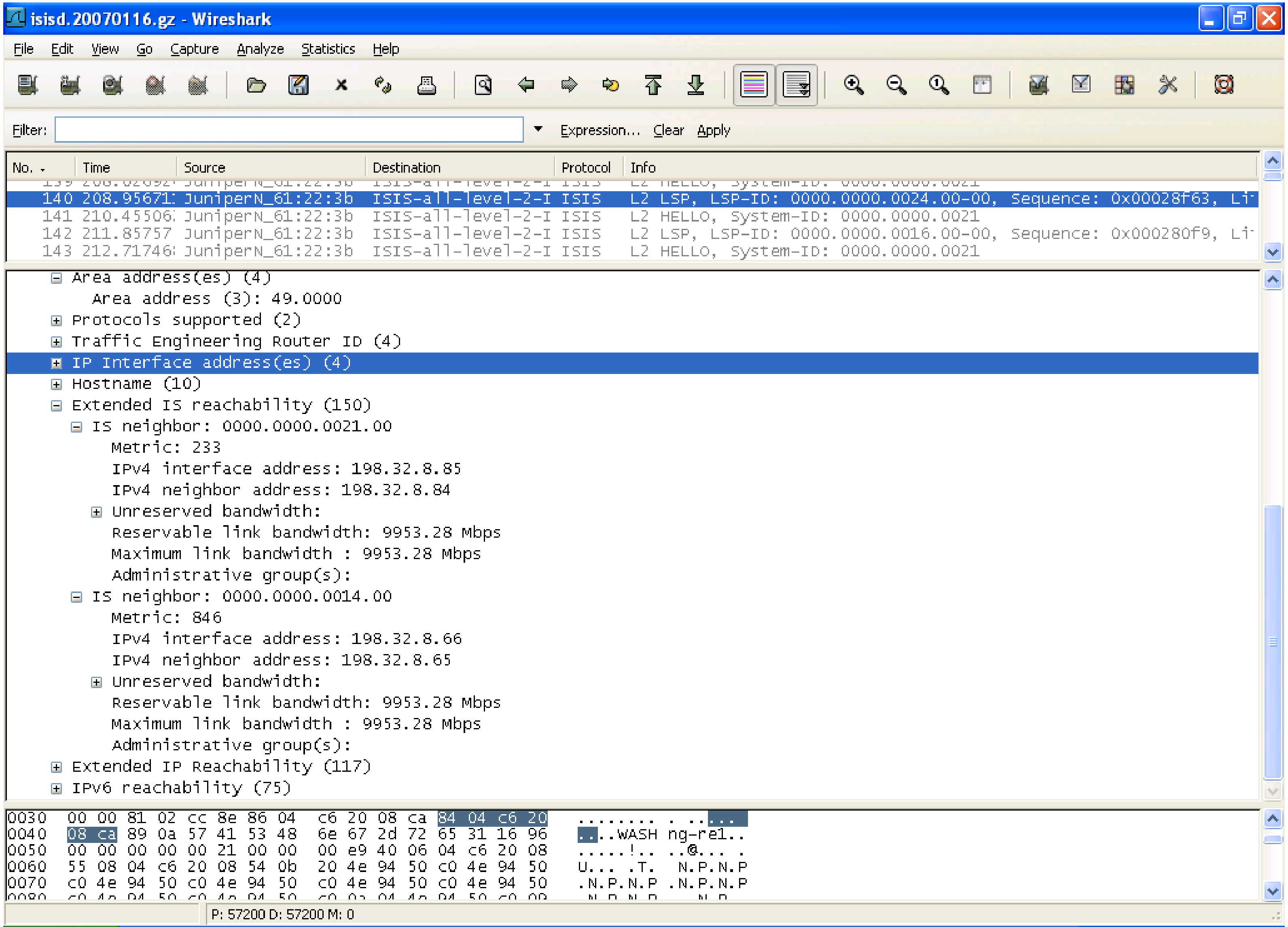
- Routing *between* areas
- Level 2 routers know the level 2 topology and know which addresses are reachable via each level 2 router.
- Level 2 routers track the location of each level 1 area.
- Level 2 routers are not concerned with the topology within any level 1 area (for example, the details internal to each level 1 area).
- Level 2 routers can identify when a level 2 router is also a level 1 router within the same area.
- Only a level 2 router can exchange packets with external routers located outside its routing domain.

CLNS Addressing: “NSAPs”



- **NSAP:** Network-Service Attachment Point (a network-layer address)
- All routers in the same area must have a common Area ID
- System ID constraints
  - Each node in an area must have a unique System ID
  - All level 2 routers in a domain must have unique System IDs
  - All NSAPs on the same router must have the same system ID.
  - All systems belonging to a given domain must have System IDs of the same length in their NSAP addresses

ISIS on the Wire…



IS-IS Configuration on Abilene (atlang)

```
lo0 {
    unit 0 {
        ….
        family iso {
         address 49.0000.0000.0000.0014.00;
        }
    ….
}

isis {
    level 2 wide-metrics-only;

    /* OC192 to WASHng */
    interface so-0/0/0.0 {
        level 2 metric 846;
        level 1 disable;
    }
}
```

ISO Address Configured on Loopback Interface

Only Level 2 IS-IS in Abilene

IS-IS vs. OSPF

- Cisco ships OSPF in 1991
- Cisco ships dual IS-IS in 1992
- *Circa 1995:* ISPs need to run IGPs, IS-IS is recommended due to the recent rewrite
- IS-IS became very popular in late 1990s
  - Deployed in most large ISPs (also Abilene)
  - Some ISPs (*e.g.*, AOL backbone) even switched

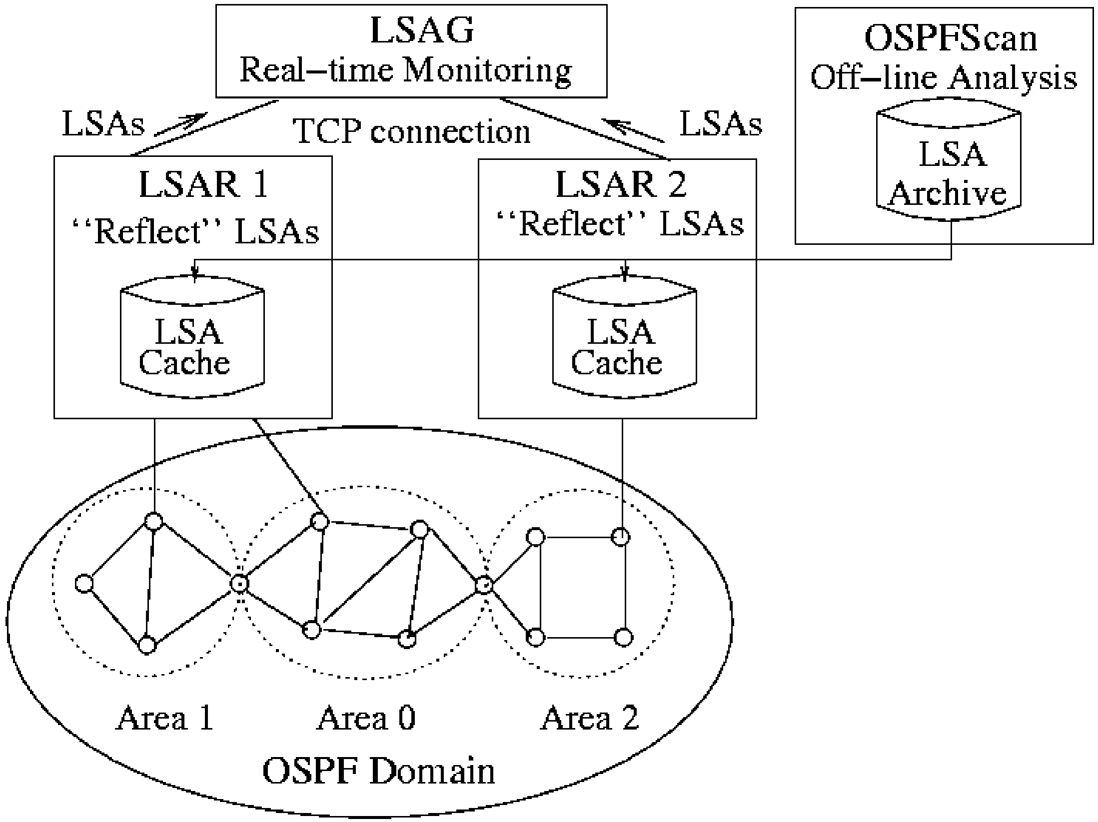
Monitoring OSPF

- **Challenge:** How to get the OSPF Link State Advertisements (LSAs)?

Challenge #1: Capturing LSAs

- Wire-tap mode
  - Invasive
  - Dependent on Layer-2
- Host mode
  - Distribute LSAs over multicast
  - LSAR joins multicast group
- Full adjacency mode
  - Form high-cost adjacency with network
- Partial adjacency mode

Challenge #2: Dealing with Areas



- **Problem:** OSPF LSAs not advertised across area boundaries.

Today's Papers: Alternative Intradomain Routing Mechanisms

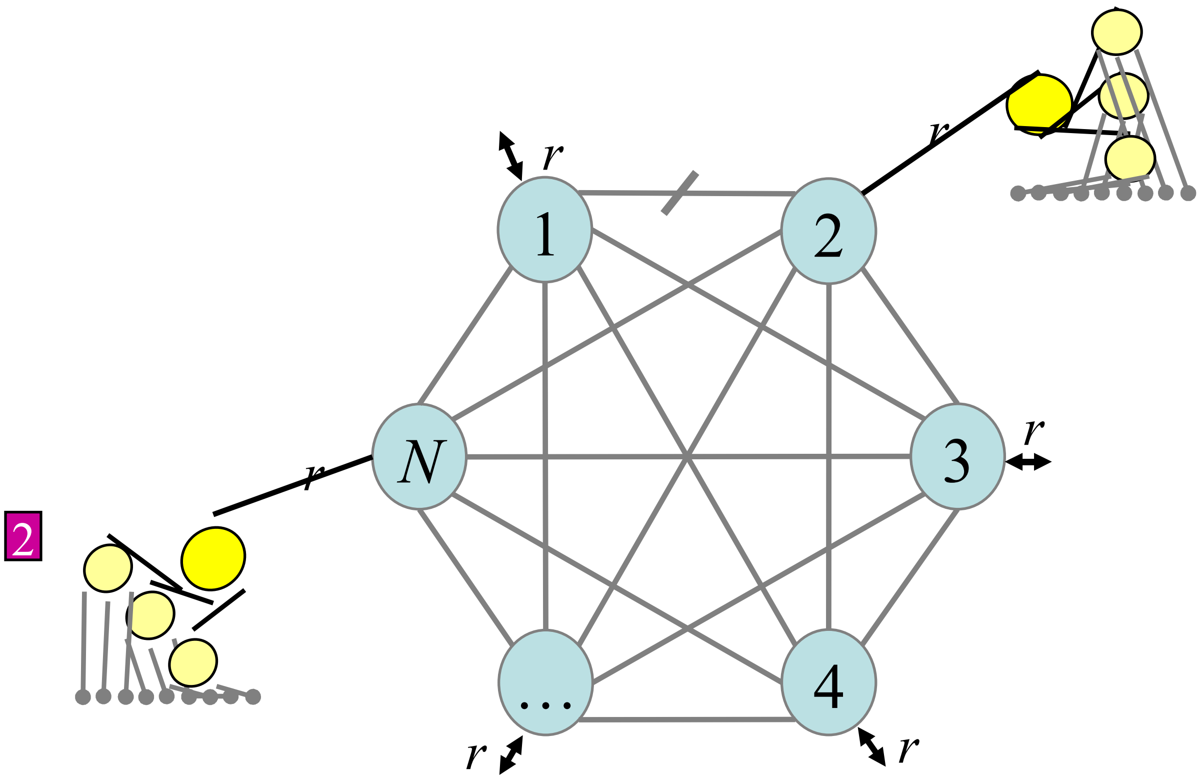
- A Key Question: How to set weights in a shortest-path routing protocol?
- Problem: Link cost becomes a protocol knob, not a reflection of the topology
- Options:
  - Link-weight tuning
  - Set up circuits (MPLS, and route on different circuits)
  - Random perturbations on link weights
  - …

Valiant Load Balanced Networks

- **Problem:** Impossible to have the perfectly tuned network
  - Traffic matrix hard to estimate
    - …and it's always changing
  - Links and nodes fail, and the failure mode scenario may not be desirable
  - Networks continually growing, changing, etc.

- **Idea:** Valiant load-balanced networks

Valiant Load-Balancing



- Suppose each node has capacity r
- How much capacity for each link?
- What if a node fails?

Thought Questions

- How might you use VLB types of routing to reduce per-router routing table state?
- Is there an alternate constrained VLB design that might put better bounds on latency increases?
- What would an internet of all VLB-routed ISPs look like? (How might traffic flow, etc.?)
- What other ways can you think of to design an intradomain routing protocol that handles traffic dynamism and failures and yet still scales well?