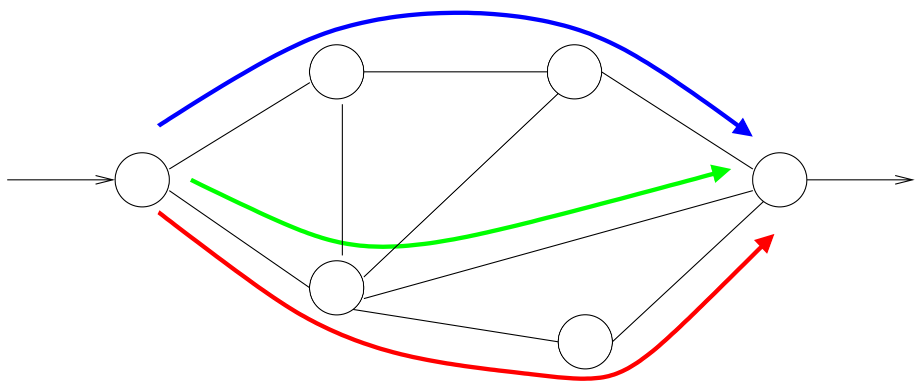
ROUTING

Problem: Given more than one path from source to destination, which one to take?



Features:

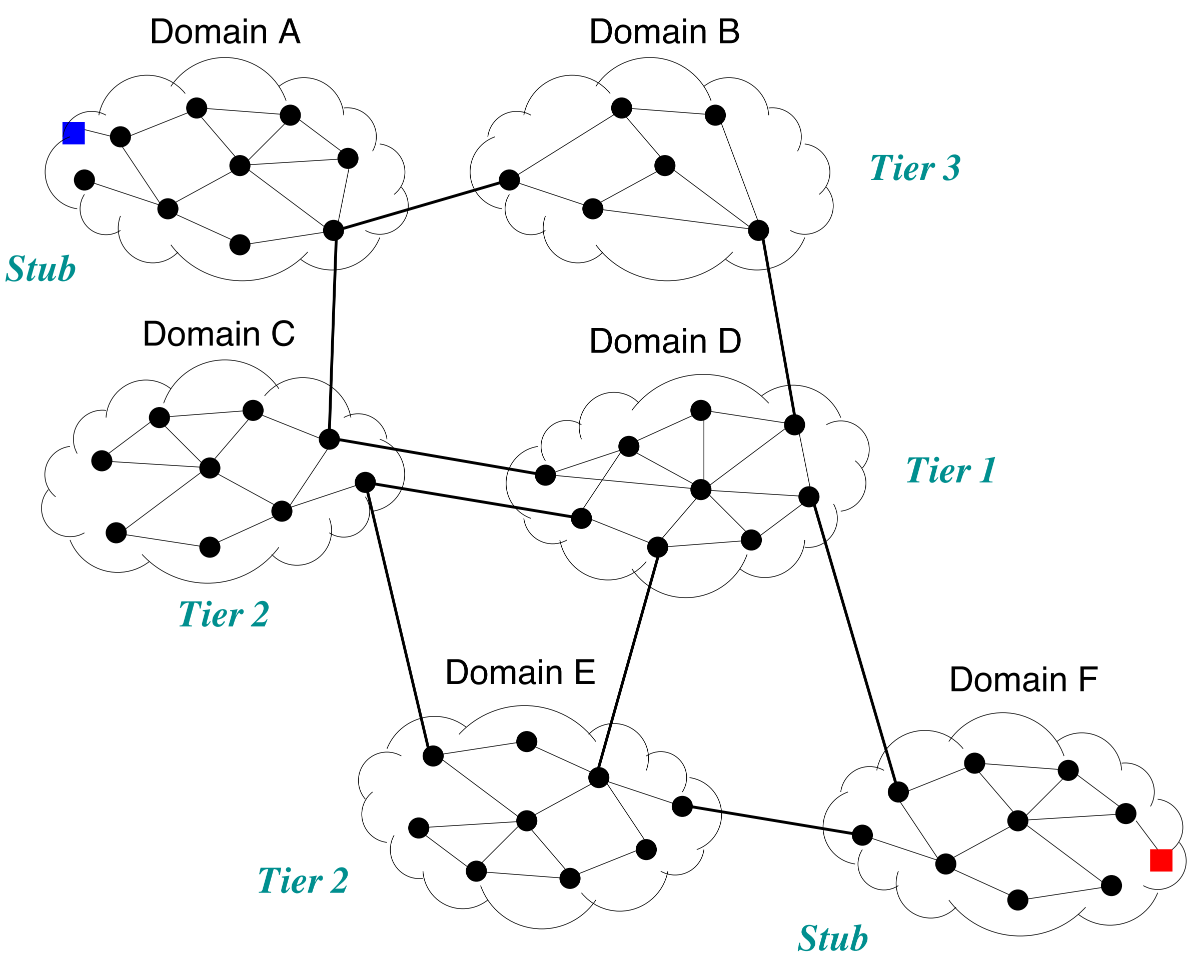
- Architecture
- Algorithms
- Implementation
- Performance

Architecture

Internet routing: two separate routing subsystems

→ intra-domain: within an organization

→ inter-domain: across organizations



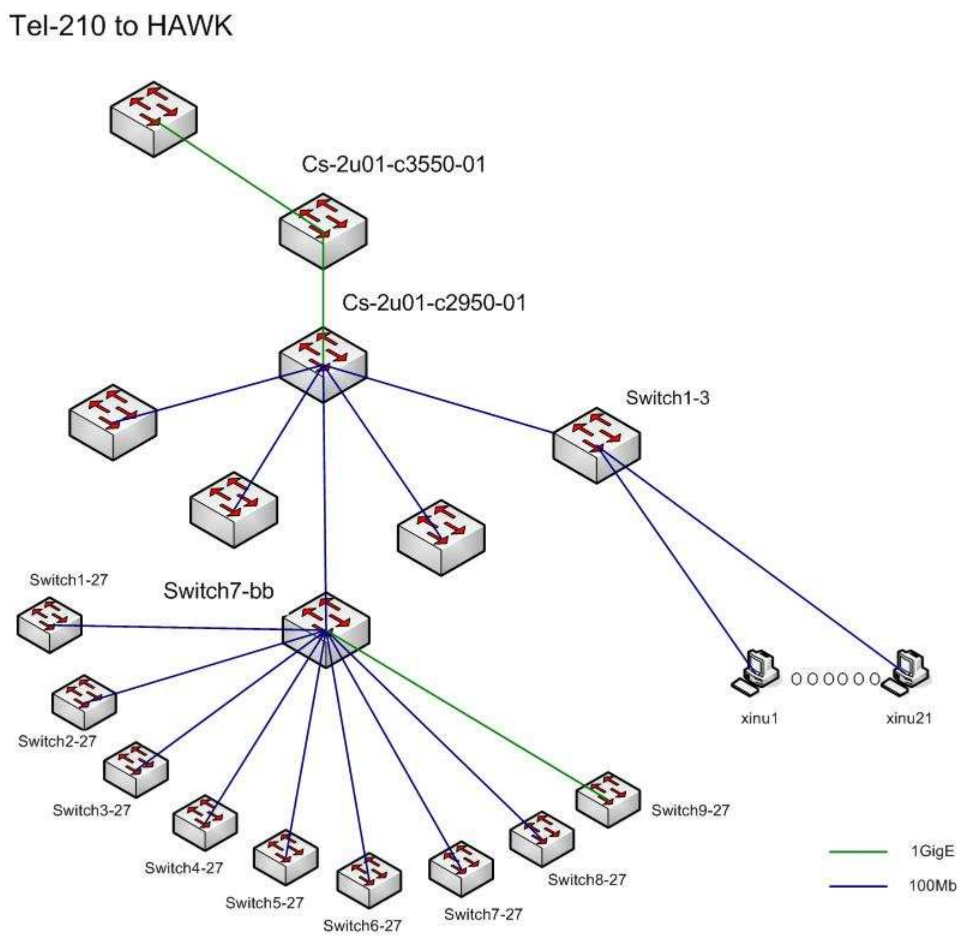
Ex.: Purdue to east coast (BU)

```
[109] infobahn:Routing % traceroute csa.bu.edu
traceroute to csa.bu.edu (128.197.12.3), 30 hops max, 40 byte packets
 1  cisco5 (128.10.27.250)  3.707 ms  0.616 ms  0.590 ms
 2  172.19.60.1 (172.19.60.1)  0.406 ms  0.431 ms  0.520 ms
 3  tel-210-m10-01-campus.tcom.purdue.edu (192.5.40.54)  0.491 ms  0.600 ms  0.510 ms
 4  gigapop.tcom.purdue.edu (192.5.40.134)  9.658 ms  1.966 ms  1.725 ms
 5  192.12.206.249 (192.12.206.249)  1.715 ms  3.381 ms  1.749 ms
 6  chinng-iplsng.abilene.ucaid.edu (198.32.8.76)  5.669 ms  8.319 ms  5.601 ms
 7  nycmng-chinng.abilene.ucaid.edu (198.32.8.83)  25.626 ms  25.664 ms  25.621 ms
 8  noxgs1-PO-6-0-NoX-NOX.nox.org (192.5.89.9)  30.634 ms  30.768 ms  30.722 ms
 9  192.5.89.202 (192.5.89.202)  31.128 ms  31.045 ms  31.082 ms
10  cumm111-cgw-extgw.bu.edu (128.197.254.121)  31.287 ms  31.152 ms  31.146 ms
11  cumm111-dgw-cumm111.bu.edu (128.197.254.162)  31.224 ms  31.192 ms  31.308 ms
12  csa.bu.edu (128.197.12.3)  31.529 ms  31.243 ms  31.367 ms
```

Ex.: Purdue to west coast (Cisco)

```
[112] infobahn:Routing % traceroute www.cisco.com
traceroute to www.cisco.com (198.133.219.25), 30 hops max, 40 byte packets
 1  cisco5 (128.10.27.250)  0.865 ms  0.598 ms  1.282 ms
 2  172.19.60.1 (172.19.60.1)  0.518 ms  0.379 ms  0.405 ms
 3  tel-210-m10-01-campus.tcom.purdue.edu (192.5.40.54)  0.687 ms  0.551 ms  0.551 ms
 4  switch-data.tcom.purdue.edu (192.5.40.34)  3.496 ms  3.523 ms  2.750 ms
 5  so-2-3-0-0.gar2.Chicago1.Level3.net (67.72.124.9)  8.114 ms  20.181 ms  8.512 ms
 6  so-3-3-0.bbr1.Chicago1.Level3.net (4.68.96.41)  11.543 ms  9.079 ms  8.239 ms
 7  ae-0-0.bbr1.SanJose1.Level3.net (64.159.1.129)  62.319 ms as-1-0.bbr2.SanJose1.Level3.net
 8  ge-11-0.ipcolo1.SanJose1.Level3.net (4.68.123.41)  68.180 ms ge-7-1.ipcolo1.SanJose1.Level
 9  p1-0.cisco.bbnplanet.net (4.0.26.14)  75.006 ms  72.557 ms  70.377 ms
10  sjce-dmzbb-gw1.cisco.com (128.107.239.53)  66.075 ms  69.223 ms  68.350 ms
11  sjck-dmzdc-gw1.cisco.com (128.107.224.69)  65.650 ms  74.358 ms  69.952 ms
12 ^C
```

Three levels: LAN, intra-domain, and inter-domain



LAN routing:

$\rightarrow$ extended LAN

$\rightarrow$ e.g., internetwork of Ethernet/WLAN switches

$\rightarrow$ bridge functionality

Approaches:

- flooding (i.e., broadcasting)

  $\rightarrow$ inefficient

  $\rightarrow$ must deal with switching loops

  $\rightarrow$ potential vulnerability: broadcast storms

  $\rightarrow$ no TTL field in Ethernet header

  $\rightarrow$ solution: embed logical tree over physical LAN internetwork

First, discover who is where

$\rightarrow$ learning bridges

Discovery procedure:

- switch receives LAN packet on interface i with source MAC address A

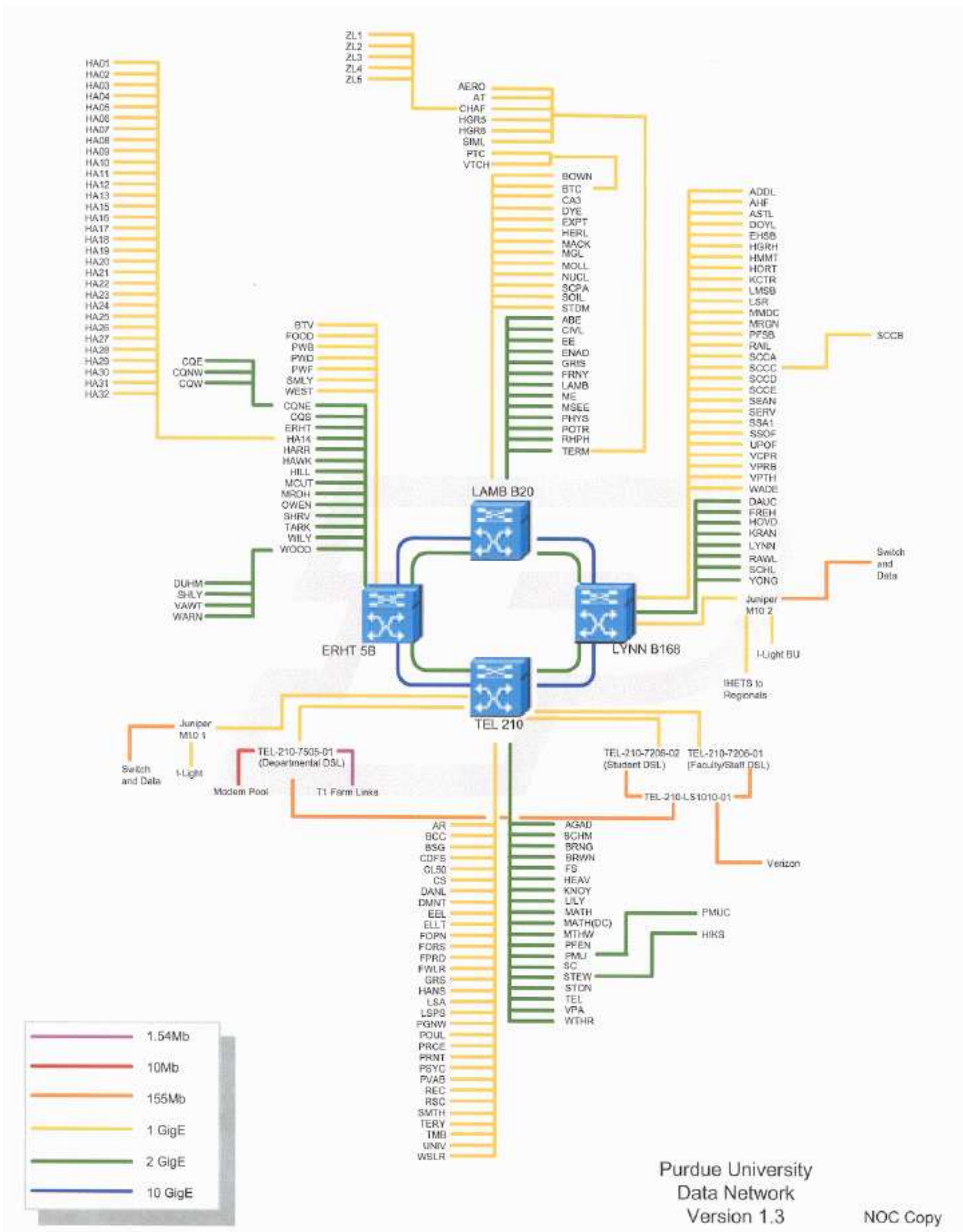
  $\rightarrow$ remember that A can be reached throughput interface i

- switch receives LAN packet destined to MAC address A

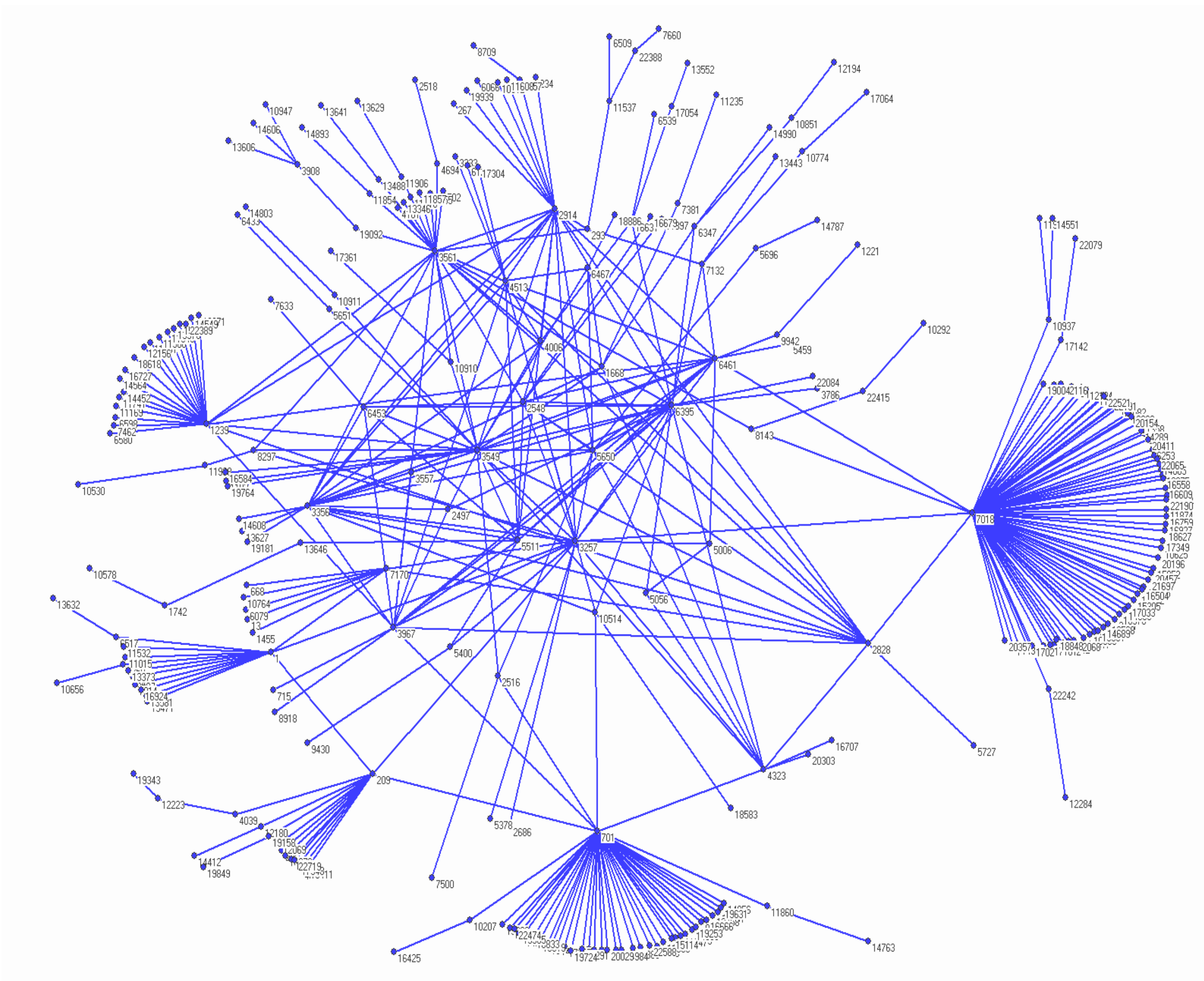
  $\rightarrow$ forward on interface i

Build logical spanning tree

- Perlman's algorithm: spanning tree protocol (STP)
- prune links to be loop-free
- other protocols


LAMB B20
ERHT 5B
LYNN B168
TEL 210
Juniper M10 1
Juniper M10 2
Switch and Data
I-Light
I-Light BU
IHETS to Regionals
TEL-210-7505-01 (Departmental DSL)
Modem Pool
T1 Farm Links
TEL-210-7206-02 (Student DSL)
TEL-210-7206-01 (Faculty/Staff DSL)
TEL-210-LS1010-01
Verizon
1.54Mb
10Mb
155Mb
1 GigE
2 GigE
10 GigE
Purdue University
Data Network
Version 1.3
NOC Copy

Inter-domain topology:



⟶ each dot (or node) is a domain (e.g., Purdue)

⟶ called autonomous system (AS): 16- or 32-bit ID

Inter-domain connectivity of Purdue:

- Level3 (AS 3356) → INDIANAGIGAPOP (AS 19782) → Purdue (AS 17)
- Internet2/Abilene (AS 11537) → INDIANAGIGAPOP (AS 19782) → Purdue (AS 17)

→ changes over time (e.g., economic reasons)

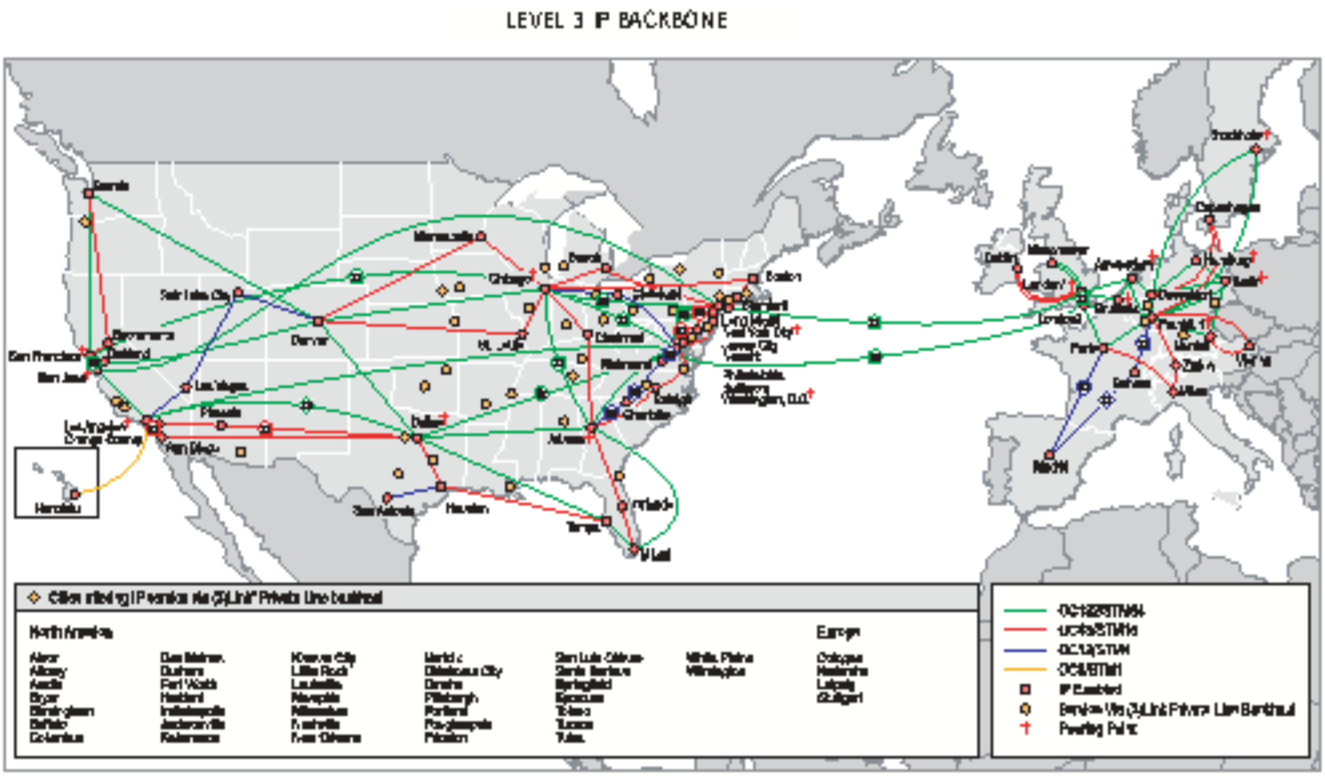
The Indy GigaPoP has its own AS number (19782).

→ part of I-Light (Indiana state-wide project)

→ located at IUPUI

→ provides state-level connectivity including Purdue and IU

Level3 backbone network: www.level3.com

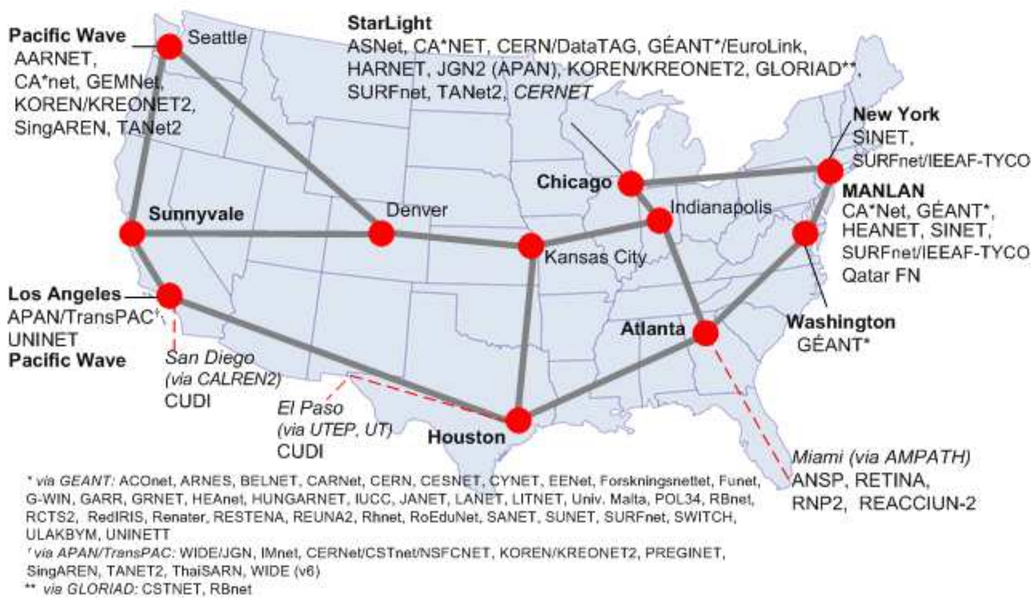


→ multi-Gbps backbone

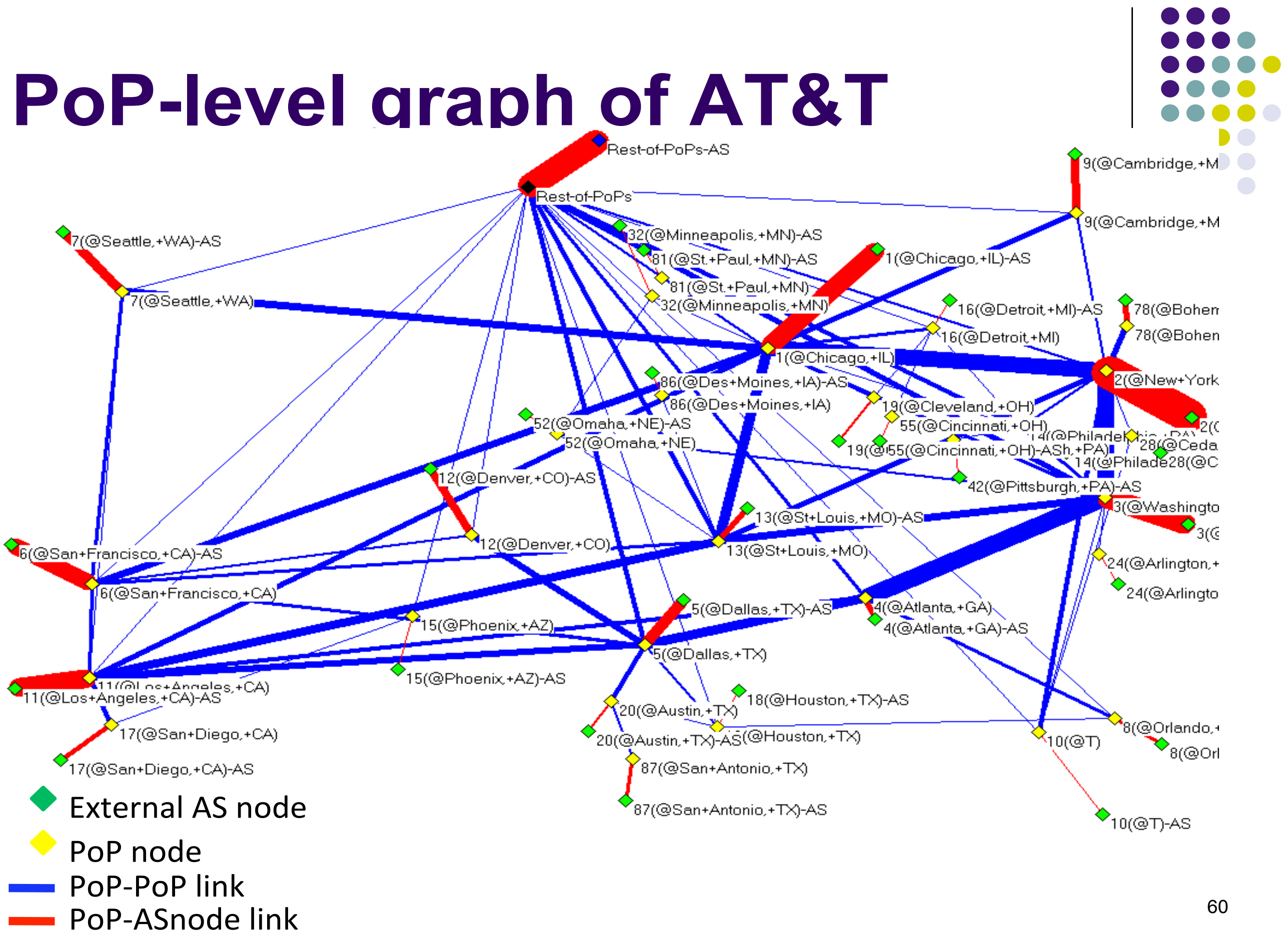
→ e.g., 1 Gbps, 10 Gbps, multiples of 10 Gbps

Abilene/Internet2 backbone: www.internet2.edu

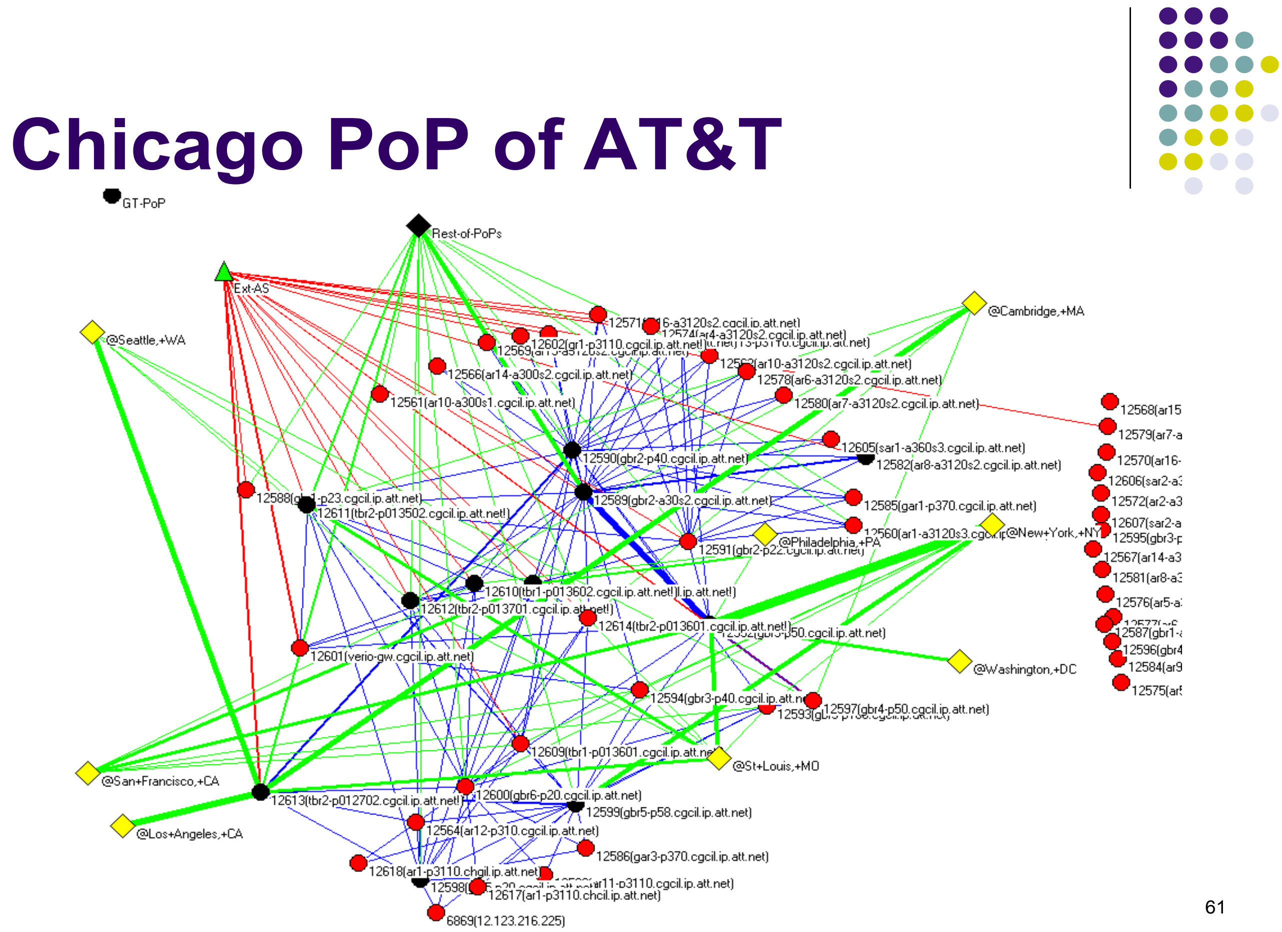
Abilene International Network Peers



AT&T (AS 7018)'s U.S. PoP topology (inferred):



AT&T's Chicago PoP connectivity (inferred):



Granularity of routing network:

- router

  → IP routing

  → note: LAN routing is invisible

- domain: autonomous system

  → 16- or 32-bit identifier ASN

  → extended to 32-bit in 2007

  → assigned by IANA along with IP prefix block (CIDR)

  → e.g., Purdue ASN: 17

Network topology

→ i.e., connectivity

- router graph

  → node: router

  → edge: physical link between two routers

- AS graph

  → node: AS

  → edge: physical link between 2 or more border routers

  → sometimes at exchange point/network

Router type:

- access router

  → collects traffic from devices of a domain/network

  → distributes traffic to devices of a domain/network

- border router

  → interface between two or more domains

  → packet crosses administrative boundary

- backbone router

  → routers that form intradomain network

  → e.g., Purdue's backbone routers (ring)

AS type:

- stub AS: customer AS

  $\rightarrow$ no forwarding

  $\rightarrow$ may be multi-homed (more than one provider)

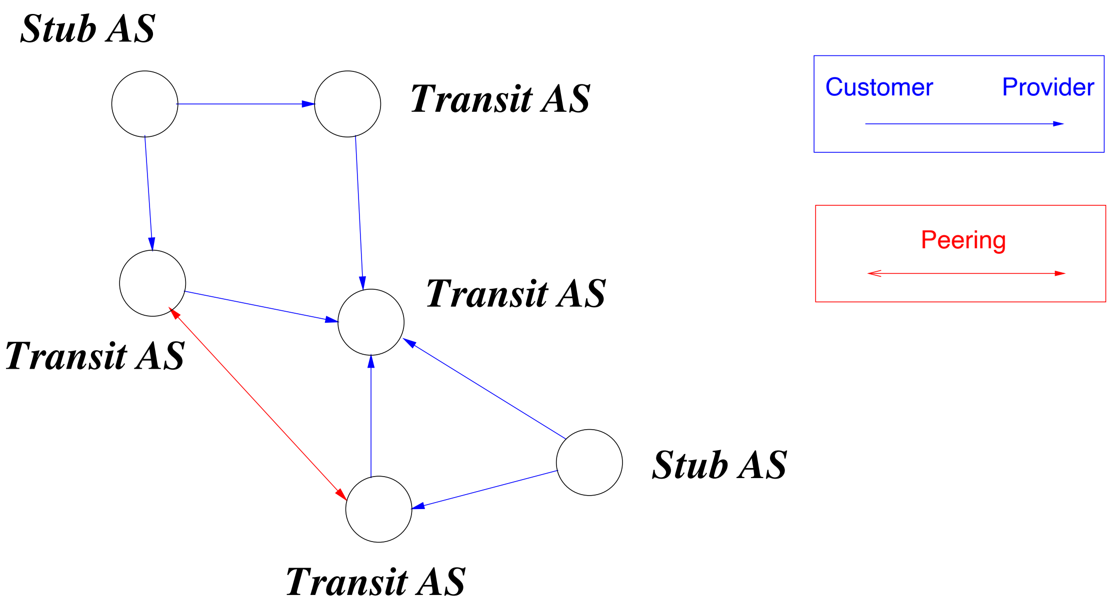
- transit AS

  $\rightarrow$ provide connectivity to stub AS's and smaller transit AS's

  $\rightarrow$ tier-1: global reachability and no provider above

  $\rightarrow$ tier-2 or tier-3: regional providers as well as customers of tier-1 AS's

AS graph:



Inter-AS relationship: bilateral

- customer-provider: customer subscribes bandwidth from provider

  → customer can reach provider's reachable IP space

- peering:

  → only the peer's IP address and below

  → the peer's provider's address space: invisible

Common peering:

- among tier-1 providers

  → ensures global reachability

  → exclusive club

  → less regulated than telephony

- among tier-2 providers

  → regional providers

  → economic factors

- among stubs

  → economic factors

  → e.g., content provider and access ("eyeball") provider

  → e.g., Time Warner and AOL

Route or path: criteria of goodness

- hop count
- delay
- bandwidth
- loss rate

Composition of goodness metric:

$\longrightarrow$ quality of end-to-end path

- additive: hop count, delay
- min: bandwidth
- multiplicative: loss rate

Goodness of routing:

$\longrightarrow$ assume N users or sessions

$\longrightarrow$ suppose path metric is delay

Two approaches:

- system optimal routing

  $\rightarrow$ choose paths to minimize $\frac{1}{N}\sum_{i=1}^{N} D_i$

  $\rightarrow$ good for the system as a whole

- user optimal routing

  $\rightarrow$ each user i chooses path to minimize D_i

  $\rightarrow$ selfish route selections by each user

  $\rightarrow$ end result may not be good for system as a whole

Pros/cons:

- system optimal routing:
  - good: minimizes delay for the system as a whole
  - bad: complex and difficult to scale up
- user optimal routing:
  - good: simple
  - bad: may not make efficient use of resources

    $\rightarrow$ low utilization

    $\rightarrow$ recall "tragedy of commons" in congestion control

Two pitfalls of user optimal routing:

- fluttering or ping pong effect

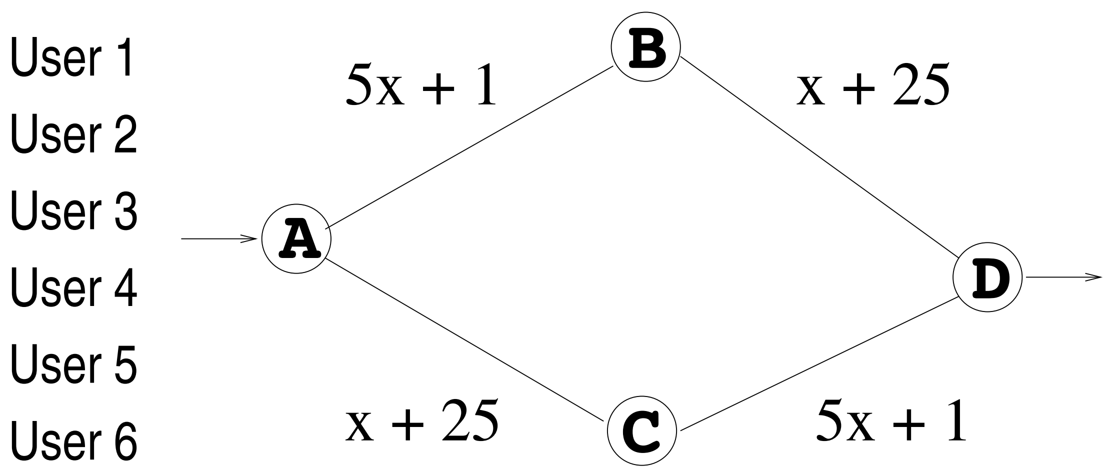
  $\rightarrow$ induced synchronization

- Braess paradox

  $\rightarrow$ adding more resources (extra link) can make things worse

Braess paradox example:

- 6 users sending 1 Mbps traffic
- delay on shared link increases with traffic volume x

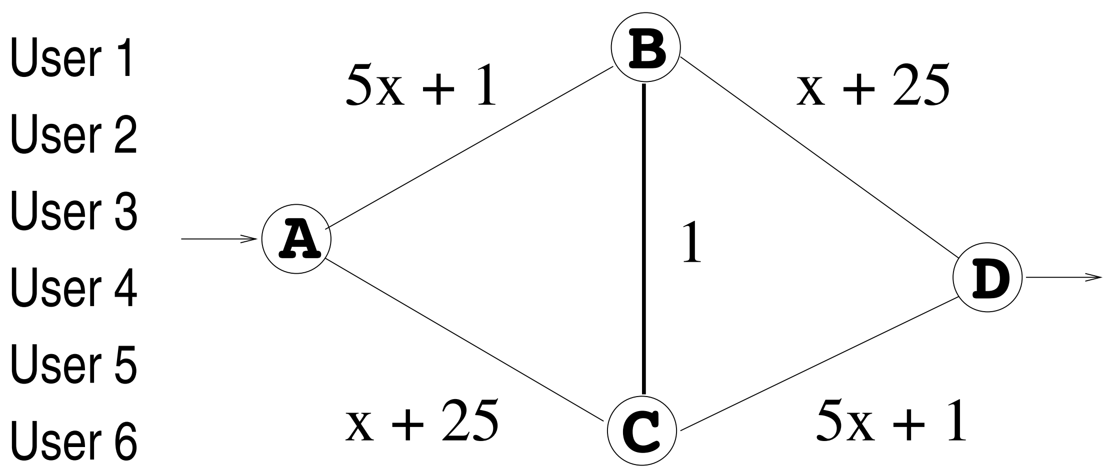


- 3 users will take $A \rightarrow B \rightarrow D$
- 3 users will take $A \rightarrow C \rightarrow D$
- delay experienced per user:

  $\rightarrow (5 \cdot 3 + 1) + (3 + 25) = 44$

Resource provisioning:

$\longrightarrow$ high bandwidth link is added between B and C



- User 1: $A \rightarrow B \rightarrow C \rightarrow D$ (13)
- User 2: $A \rightarrow B \rightarrow C \rightarrow D$ (23)
- User 3: $A \rightarrow B \rightarrow C \rightarrow D$ (33)
- User 4: $A \rightarrow B \rightarrow C \rightarrow D$ (43)
- User 5: $A \rightarrow B \rightarrow D$ (52)
- User 6: $A \rightarrow C \rightarrow D$ (52)

Note:

- delay of link $A \to B$ has increased to $5 \cdot 5 + 1 = 26$
- same for delay of link $C \to D$

  $\to$ user 1's cost has increased from 13 to 53

  $\to$ users 2, 3, 4: same cost increase to 53

Higher than per user cost 44 without high bandwidth link.

$\longrightarrow$ why did adding link degrade performance?

Increasing resource should improve things but has the opposite effect

- $\longrightarrow$ D. Braess (1969)
- $\longrightarrow$ paradox possible due to user optimal routing
- $\longrightarrow$ cannot arise in system optimal routing

Modus operandi of the Internet: user optimal routing

- $\longrightarrow$ simplicity wins the day

Conceptually related problem in operating systems?