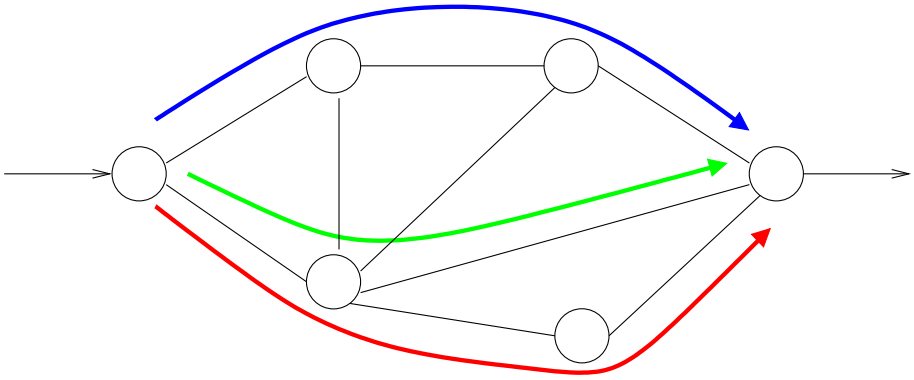
Routing

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



Features:

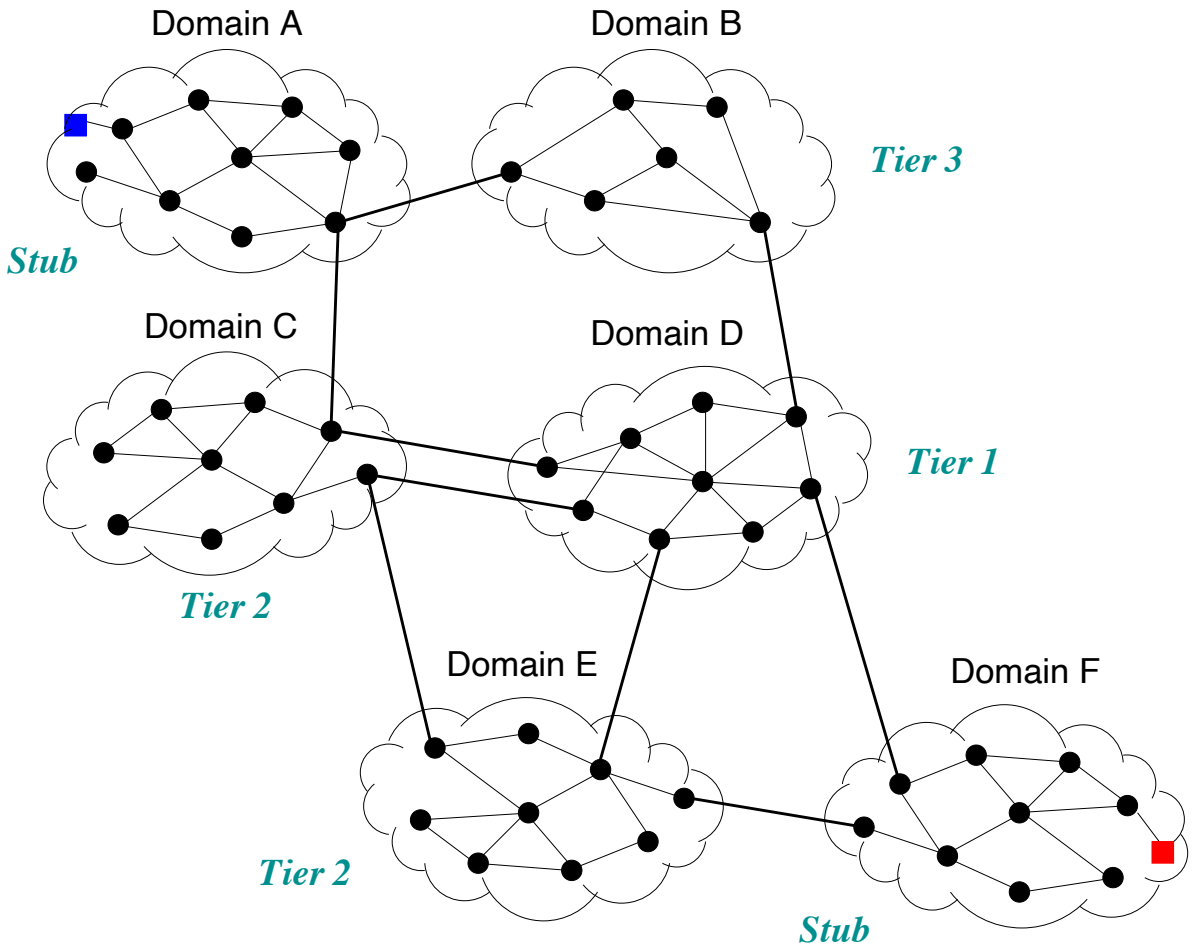
- Architecture
- Algorithms
- Implementation
- Performance

Architecture

Internet routing: two separate routing subsystems

→ intra-domain: within an organization

→ inter-domain: across organizations



Ex.: Purdue to east coast (BU)

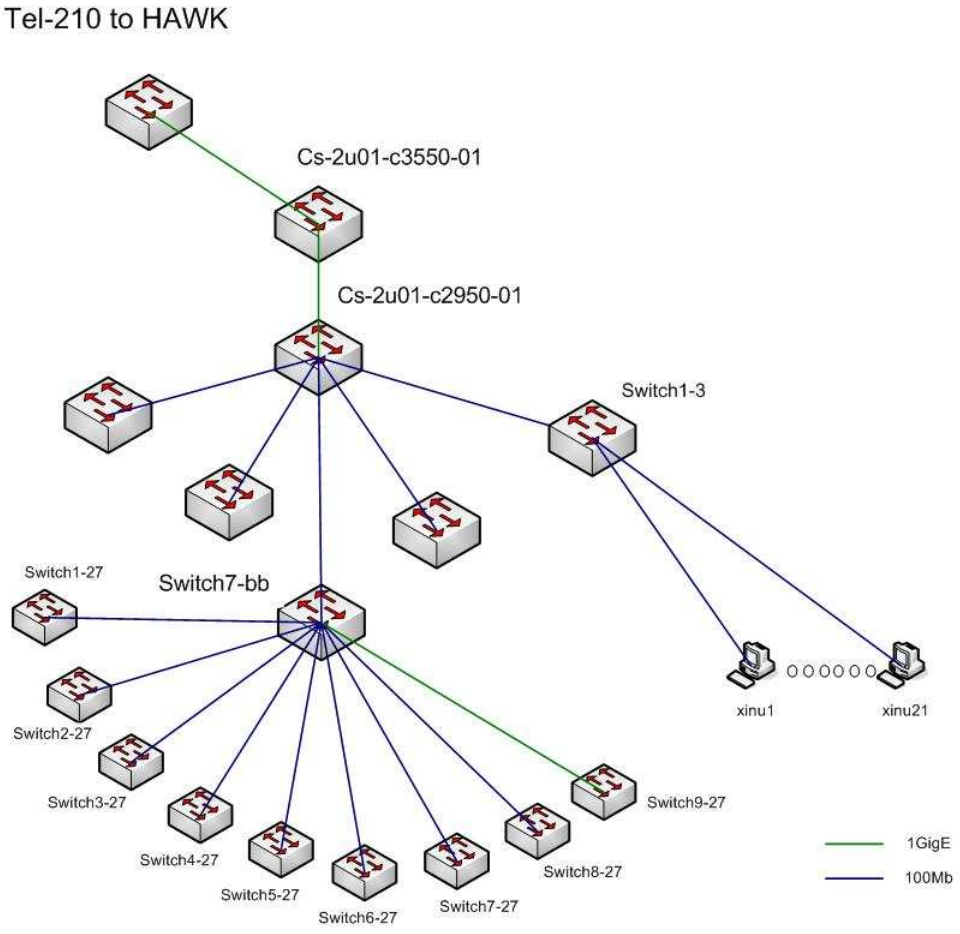
```
[109] infobahn:Routing % traceroute csa.bu.edu
traceroute to csa.bu.edu (128.197.12.3), 30 hops max, 40 byte packets
 1  cisco5 (128.10.27.250)  3.707 ms  0.616 ms  0.590 ms
 2  172.19.60.1 (172.19.60.1)  0.406 ms  0.431 ms  0.520 ms
 3  tel-210-m10-01-campus.tcom.purdue.edu (192.5.40.54)  0.491 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
```

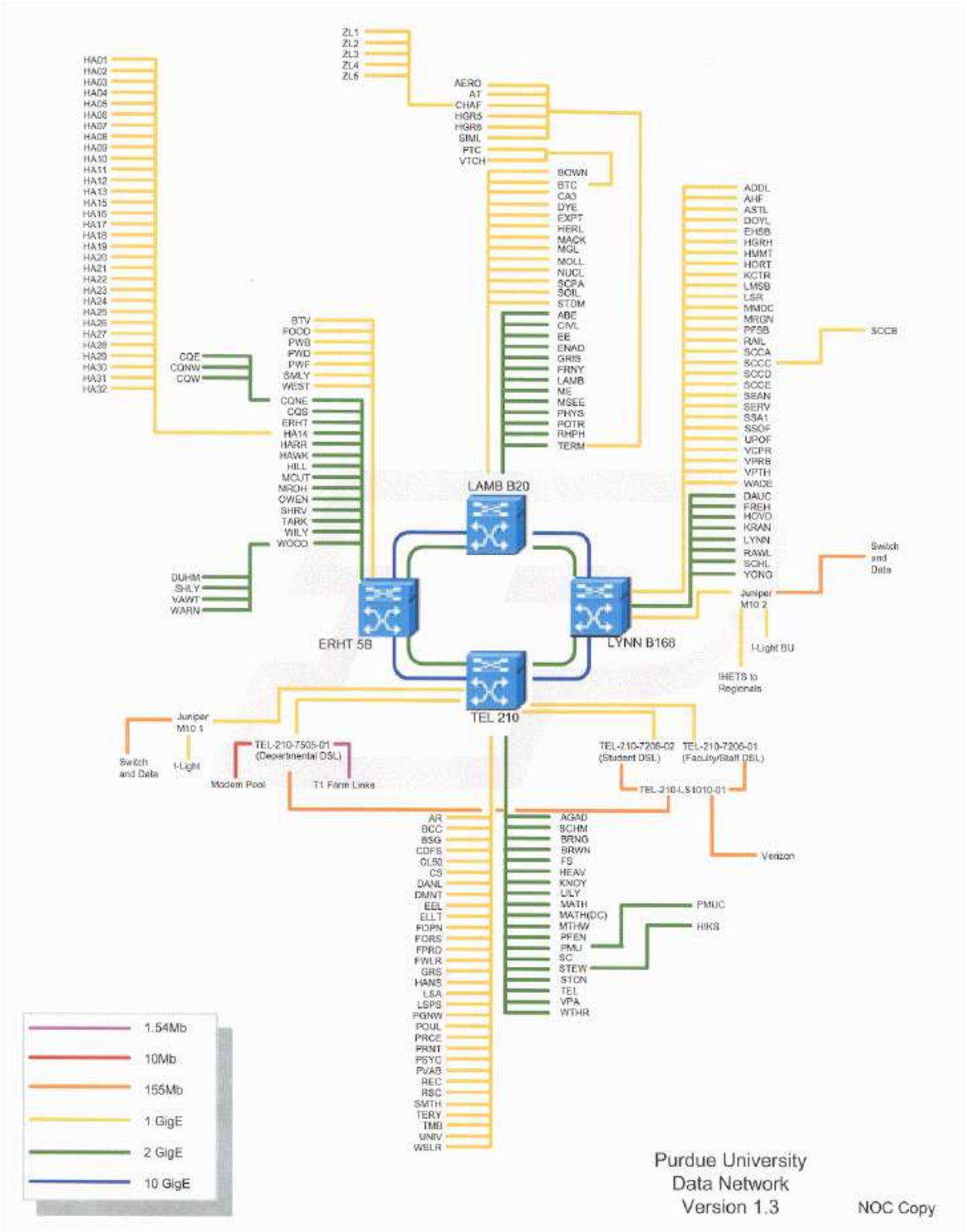
Ex.: Purdue to west coast (Cisco)

```
[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
```

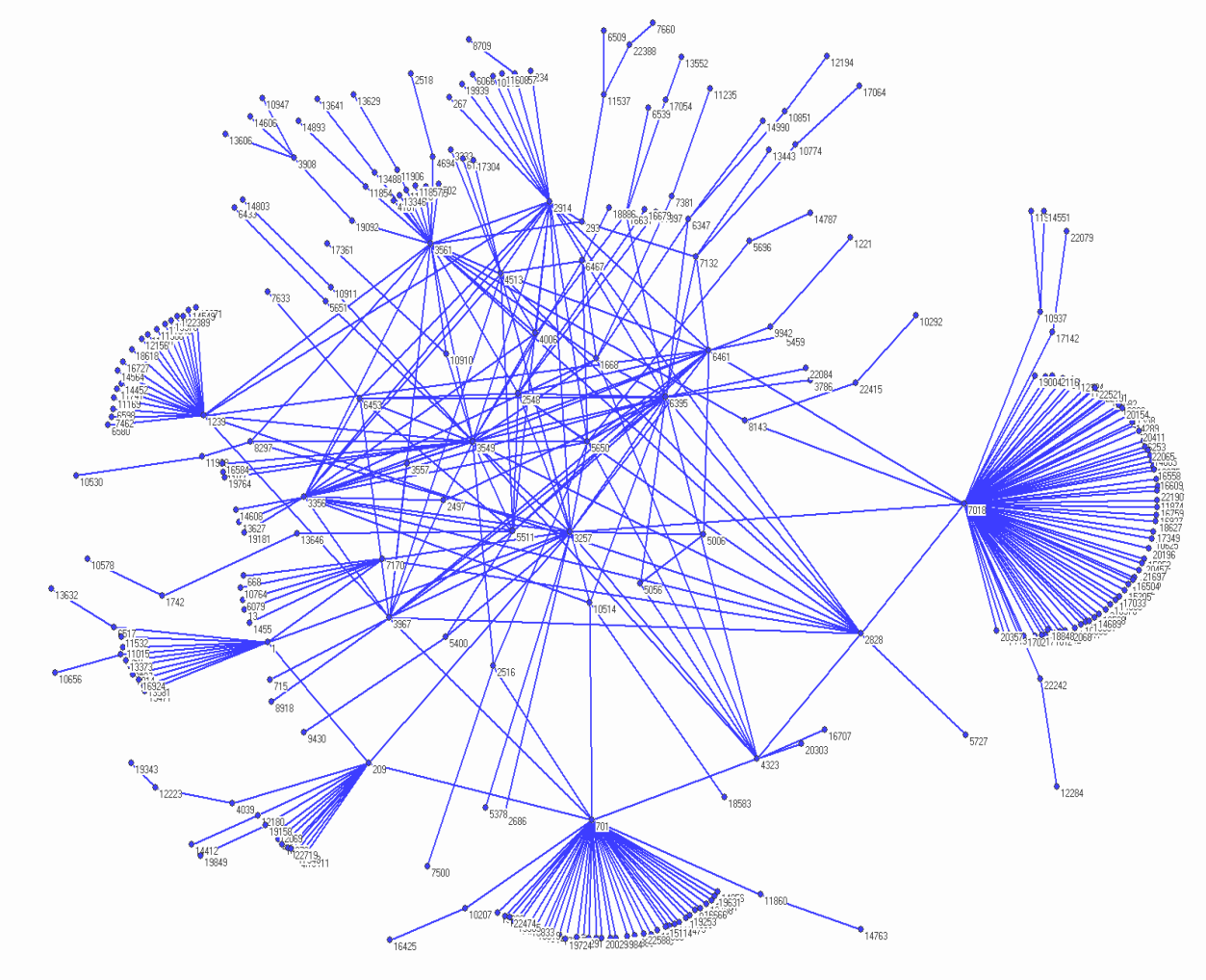
Three levels: LAN, intra-domain, and inter-domain

→ LAN: spanning tree, broadcast/flooding




ZL1
ZL2
ZL3
ZL4
ZL5
HA01
HA02
HA03
HA04
HA05
HA06
HA07
HA08
HA09
HA10
HA11
HA12
HA13
HA15
HA16
HA17
HA18
HA19
HA20
HA21
HA22
HA23
HA24
HA25
HA26
HA27
HA28
HA29
HA30
HA31
HA32
AERO
AT
CHAF
HGR5
HGR6
SIML
PTC
VTCH
BOWN
BTC
CA3
DYE
EXPT
HERL
MACK
MGL
MOLL
NUDL
SCPA
SOIL
STDM
ABE
CIVL
EE
ENAD
GRIS
FRNY
LAMB
ME
MSEE
PHYS
POTR
RHPH
TERM
ADDL
AHF
ASTL
DOYL
EHSB
HGRH
HMMT
HORT
KCTR
LMSB
LSR
MMDC
MRGN
PFSB
RAIL
SCCA
SCCC
SCCD
SCCE
SEAN
SERV
SSA1
SSOF
UPOF
VCPR
VPRB
VPTH
WADE
DAUC
FREH
HOYD
KRAN
LYNN
RAWL
SCHL
YONG
SCCB
Switch and Data
Juniper M10 2
I-Light BU
IHETS to Regionals
BTV
FOOD
PWB
PWD
PWF
SMLY
WEST
CQE
CQNW
CQW
CQNE
CQS
ERHT
HA14
HARR
HAWK
HILL
MCUT
MROH
OWEN
SHRV
TARK
WILY
WOOD
DUHM
SHLY
YAWT
WARN
LAMB B20
ERHT 5B
LYNN B168
TEL 210
Juniper M10 1
Switch and Data
I-Light
TEL-210-7505-01 (Departmental DSL)
Modem Pool
T1 Form Links
TEL-210-7206-02 (Student DSL)
TEL-210-7206-01 (Faculty/Staff DSL)
TEL-210-LS1010-01
Verizon
AR
BCC
BSG
CDFS
CL50
CS
DANL
DMNT
EEL
ELLT
FOPN
FORS
FPRD
FWLR
GRIS
HANS
LSA
LSPS
PGNW
POUL
PRCE
PRNT
PSYC
PVAB
REC
RSC
SMTH
TERY
TMB
UNIV
WSLR
AGAD
BCHM
BRNG
BRWN
FS
HEAV
KNOY
LILY
MATH
MATH(DC)
MTHW
PFEN
PMU
SC
STEW
STON
TEL
VPA
WTHR
PMUC
HIKS
1.54Mb
10Mb
155Mb
1 GigE
2 GigE
10 GigE
Purdue University
Data Network
Version 1.3
NOC Copy

Inter-domain topology:



$\longrightarrow$ each dot (or node) is a domain (e.g., Purdue)

$\longrightarrow$ autonomous system (AS): 16- or 32-bit ID

Inter-domain connectivity of Purdue:

- AS 6939 (Hurricane), Internet 2 (AS11164), etc. → INDIANAGIGAPOP (AS 19782) → Purdue (AS 17)
- AS 3356 (Level3), AS 6939 (Hurricane), AS 1299 (Arelion), AS 6461 (Zayo) → WINTEK (AS 11114) → Purdue (AS 17)

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

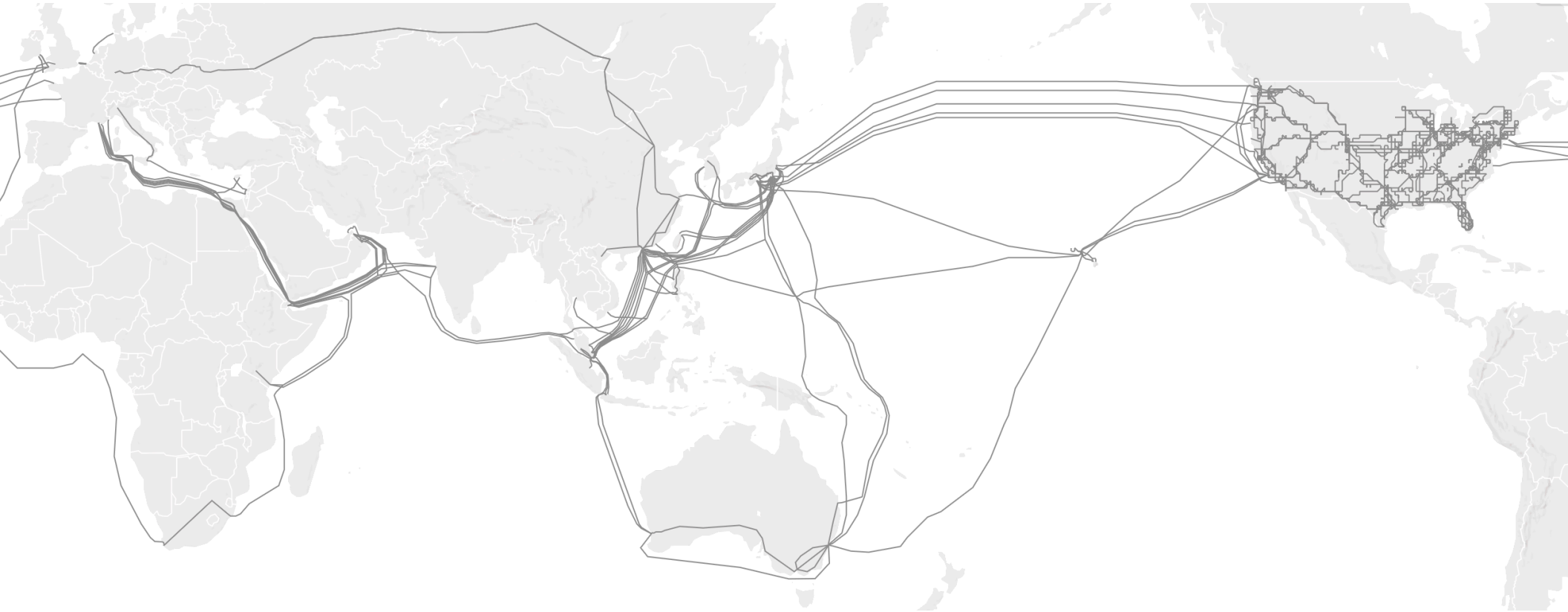
Indy GigaPoP (19782):

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

→ located at IUPUI

→ provides state-level fiber connectivity including Purdue and IU

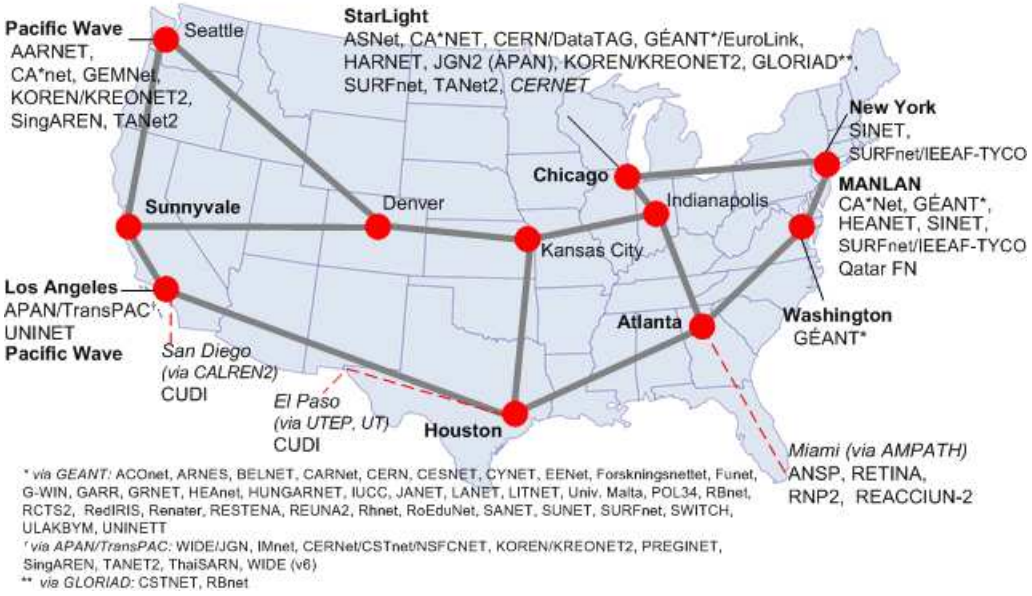
Lumen Technologies (AS 3356): tier-1



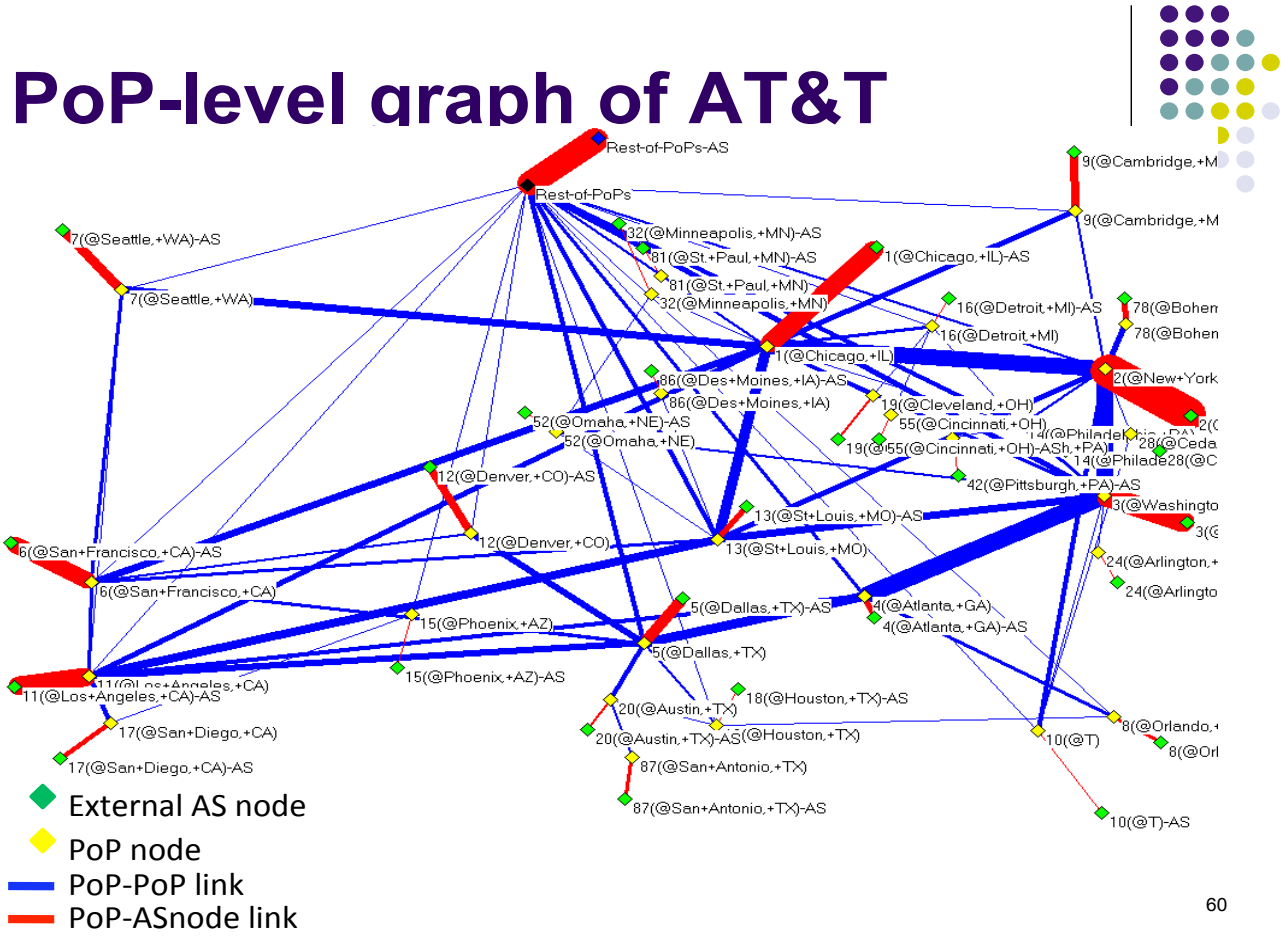
from lumen.com

Abilene/Internet2 backbone: www.internet2.edu

Abilene International Network Peers

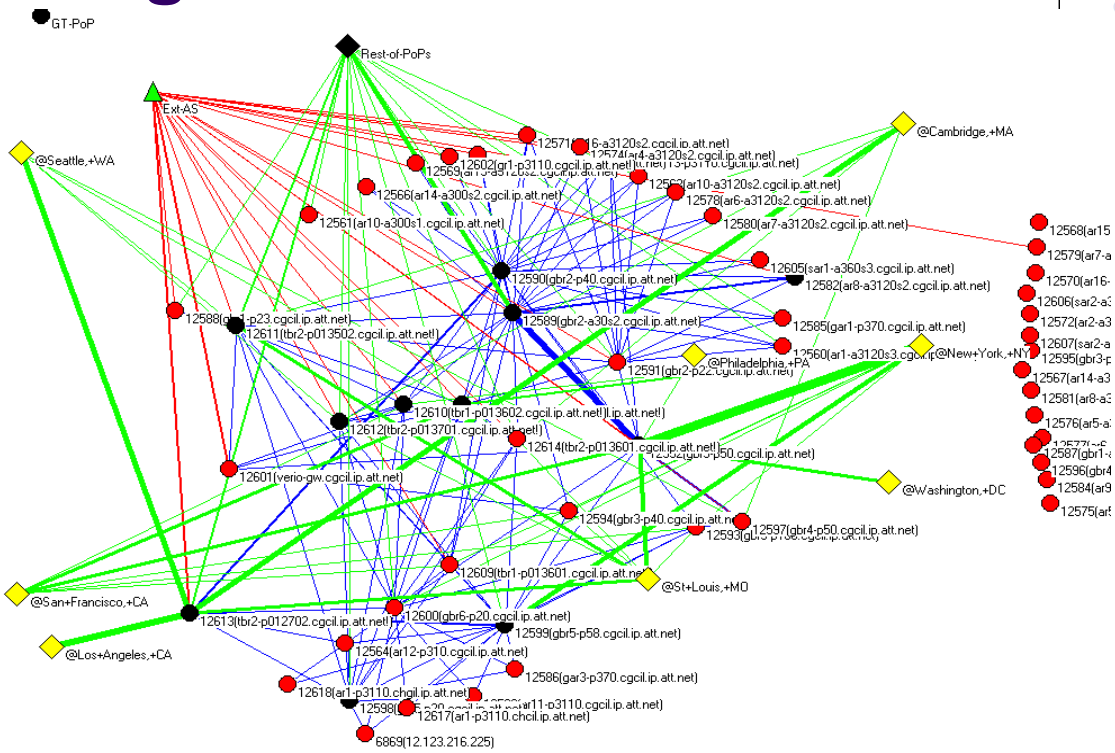


AT&T (AS 7018)'s U.S. PoP topology (inferred):



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

Chicago PoP of AT&T



Granularity of routing network:

- router
  - → IP routing (also MPLS routing in large ISPs)
  - → 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)

Network topology

→ i.e., connectivity

- router graph
  - → node: router/switch
  - → 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)

Note: traditional intra-domain/enterprise router topology emphasizes vertical (i.e., north-south) traffic

→ vertical aggregation/deaggregation (e.g., fat tree)

Recent data center development: surge of horizontal (i.e., east-west) traffic.

→ driven by workload demand (e.g., AI)

→ Clos (or leaf-spine) network

In general, data center intranet connectivity is related to interconnection networks studied in parallel computing.

- multi-stage networks: $\log n$ stages

  → e.g., shuffle-exchange, omega, butterfly networks

- non-blocking: satisfy input/output permutation without collision

Other commonly used connectivity:

- crossbar switch

  → overhead quadratic

  → suited for router switching fabric, SoC

- Clos network

  → 2-stage variant: provides multiple parallel paths between leaf nodes

  → L2 leaf switches connect switches

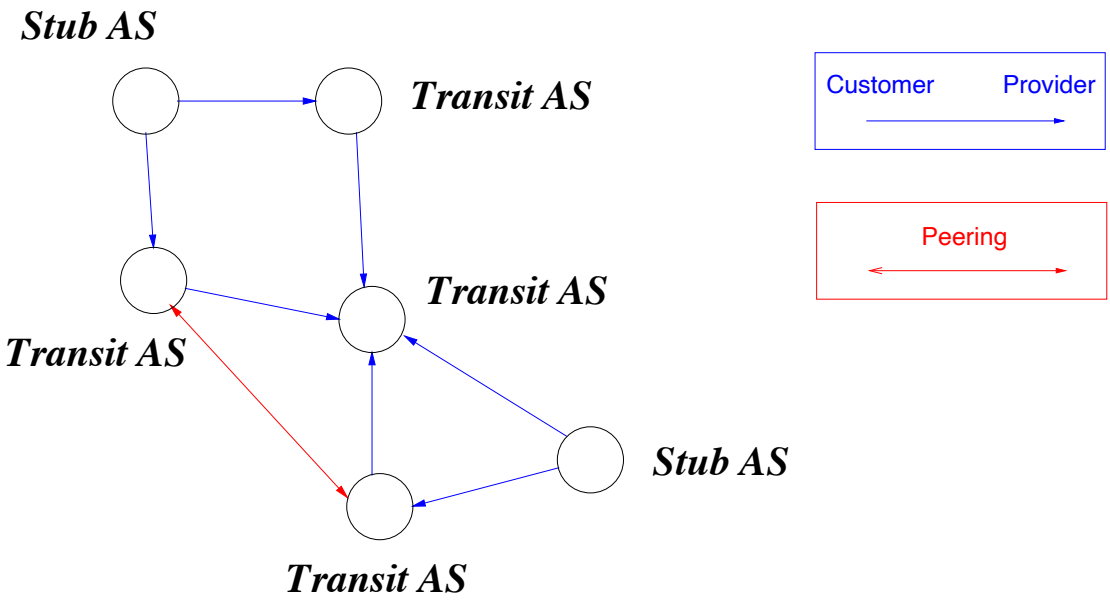
  → L3 spine switches connect leaf switches

  → variations

AS type:

- stub AS: customer AS
  - → no forwarding
  - → may be multi-homed (more than one provider)
- transit AS
  - → provide connectivity to stub AS's and smaller transit AS's
  - → tier-1: global reachability and no provider above
  - → 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