Lecture 11: BGP policies and issues

Today

- BGP (cont'd)
- Prefix hijacking

Administrivia

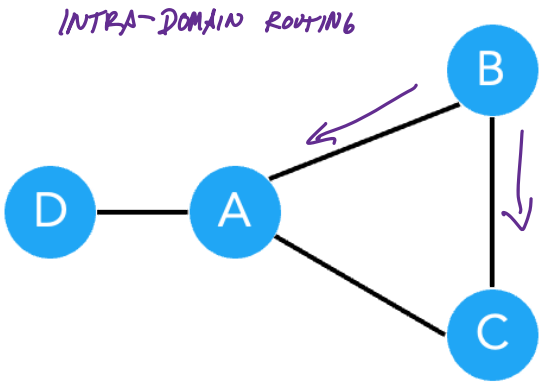
- IP due Thursday
    - Look for announcement today about Gradescope submission, post-project form
    - Sign up for grading meeting next week
        - You can make bugfixes after the deadline, without using late days (see handout/Edstem for details)

- TCP partner form out this week, project out next week
    - You MAY keep the same team, or you can change teams

- HW2 out today (hopefully; will post as soon as I get there, adjust content/deadline accordingly)

Lecture 11: Routing III (BGP Policies and Shenanigans)

Warmup: Split Horizon + Poison reverse



B's routing table:

Dest	Cost	Next Hop
A	1	A
C	1	C
D	2	A

Routers A, B, C, D use RIP. When B sends a periodic update to A, what does it send...

- When using standard RIP?
- When using split horizon + poison reverse?

STANDARD RIP

(A, 1)
(C, 1)
(D, 2)

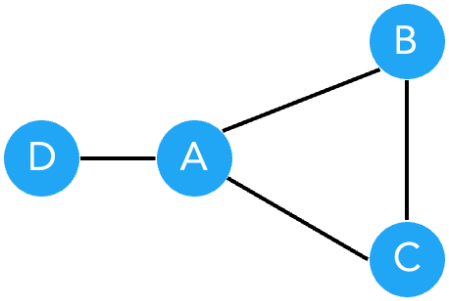
SPLIT HORIZON + POISON REVERSE

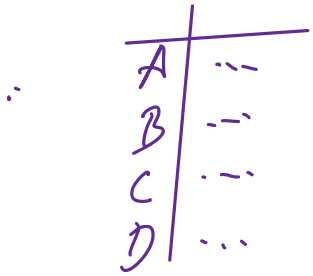
(A, ∞)
(C, 1)
(D, ∞)

Previously: interior routing

All nodes advertise their routes to all other nodes:

- Goal: connect everything to everything
- One administrative domain
- Find optimal path





Now: BGP

Exterior routing: between Autonomous Systems (ASes)

- How networks with different goals/policies/incentives connect to each other (or don't)
- A "path vector protocol"

A BGP update

"I can reach prefix 128.148.0.0/16 through ASes 44444 3356 14325 11078"

A PREFIX @ BROWN

BROWN'S AS

PATH

SENT TO NEIGHBORS

Extra: what would a loop look like in a path announcement??
=> AS number would be repeated => trivial to avoid installing routes that create loops, since this would not make sense

Demo: rviews

Example: here's a snapshot of BGP info from for Brown's prefixes

Each line represents a BGP message received from another router with its path

Lines with a ">" mark which route this router decided was the "best" prefix and installed in its forwarding table

(Try it yourself! Run 'telnet routeviews.route-views.org')

```
route-views>show ip bgp 128.148.0.0/16 longer-prefixes
BGP table version is 1027665629, local router ID is 128.223.51.103
Status codes: s suppressed, d damped, h history, * valid, > best, i - internal,
              r RIB-failure, S Stale, m multipath, b backup-path, f RT-Filter,
              x best-external, a additional-path, c RIB-compressed,
Origin codes: i - IGP, e - EGP, ? - incomplete
RPKI validation codes: V valid, I invalid, N Not found

     Network          Next Hop            Metric LocPrf Weight Path
N*   128.148.0.0/21   77.39.192.30                           0 20912 174 14325 11078 i
N*                    114.31.199.16                          0 4826 6939 14325 11078 i
N*                    37.139.139.17            0             0 57866 5511 174 14325 11078 i
N*                    89.149.178.10           10             0 3257 174 14325 11078 i
N*                    203.181.248.195                        0 7660 11537 14325 11078 i
N*                    217.192.89.50                          0 3303 11164 14325 11078 i
N*                    132.198.255.253                        0 1351 11537 14325 11078 i
N*                    202.232.0.2                            0 2497 174 14325 11078 i
                                                        ...
N*>                   64.71.137.241                          0 6939 14325 11078 i
N*                    209.124.176.223                        0 101 11537 14325 11078 i
N*                    198.32.252.33                          0 20080 11537 14325 11078 i
N*   128.148.8.0/21   77.39.192.30                           0 20912 174 14325 11078 i
N*                    114.31.199.16                          0 4826 6939 14325 11078 i
N*                    37.139.139.17            0             0 57866 5511 174 14325 11078 i
N*                    89.149.178.10           10             0 3257 174 14325 11078 i
N*                    203.181.248.195                        0 7660 11537 14325 11078 i
N*>                   64.71.137.241                          0 6939 14325 11078 i
                                                        ...
N*                    198.32.252.33                          0 20080 11537 14325 11078 i
N*                    132.198.255.253                        0 1351 11537 14325 11078 i
N*                    202.232.0.2                            0 2497 174 14325 11078 i
N*                    217.192.89.50                          0 3303 11164 14325 11078 i
N*                    94.142.247.3                           0 8283 6461 14325 11078 i
N*                    91.218.184.60                          0 49788 12552 6461 14325 11078 i
N*   128.148.16.0/20  77.39.192.30                           0 20912 174 14325 11078 11078 11078 11078 11078 11078 i
N*                    114.31.199.16                          0 4826 6939 14325 11078 11078 11078 11078 11078 11078 i
N*                    37.139.139.17            0             0 57866 5511 174 14325 11078 11078 11078 11078 11078 11078 i
N*                    89.149.178.10           10             0 3257 174 14325 11078 11078 11078 11078 11078 11078 i
N*                    203.181.248.195                        0 7660 11537 14325 11078 11078 11078 11078 11078 i
N*                    64.71.137.241                          0 6939 14325 11078 11078 11078 11078 11078 11078 i
N*                    140.192.8.16                           0 20130 6939 14325 11078 11078 11078 11078 11078 11078 i
N*                    193.0.0.56                             0 3333 1103 11537 14325 11078 11078 11078 11078 11078 i
N*                    12.0.1.63                              0 7018 6461 14325 11078 11078 11078 11078 11078 11078 i
N*                    208.51.134.254           0             0 3549 3356 174 14325 11078 11078 11078 11078 11078 11078 i
N*                    209.124.176.223                        0 101 11537 14325 11078 11078 11078 11078 11078 i
N*                    4.68.4.46                0             0 3356 174 14325 11078 11078 11078 11078 11078 11078 i
N*                    198.32.252.33                          0 20080 11537 14325 11078 11078 11078 11078 11078 i
N*>                   132.198.255.253                        0 1351 11537 14325 11078 11078 11078 11078 11078 i
```

OSHEAN

BROWN

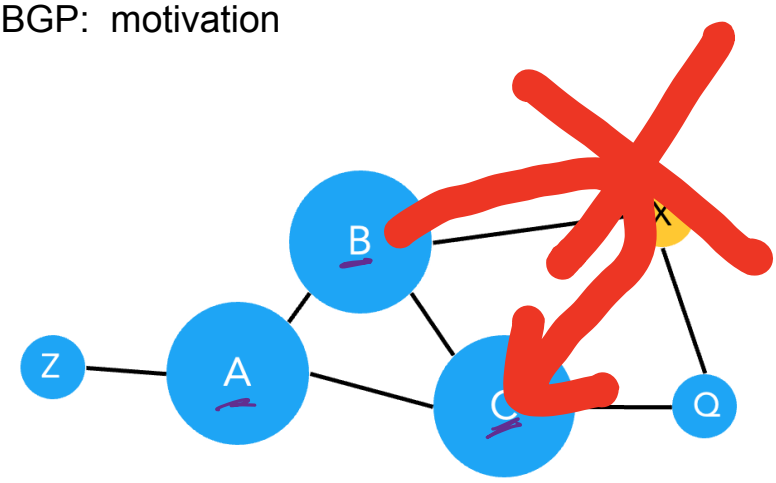
ALL KNOWN ROUTES TO 128.148.0.0/21

ALL KNOWN ROUTES TO 128.148.8.0/21

Why were these routes selected as the "best"? => Shortest AS path!!

Extra: Why does AS11078 appear multiple times? This changes the path length!
This is called "AS path prepending" => used by administrators to make a path less preferred. This is a form of what is called "traffic engineering"

BGP: motivation



X's table (a subset):

Neighbor	Next hop	Path
X	--	(Origin)
B	B	B
C	C	C
Q	Q	Q
A	B	B A
...	...	...

X "originates" a set of prefixes => set of prefixes it wants to advertise to the world

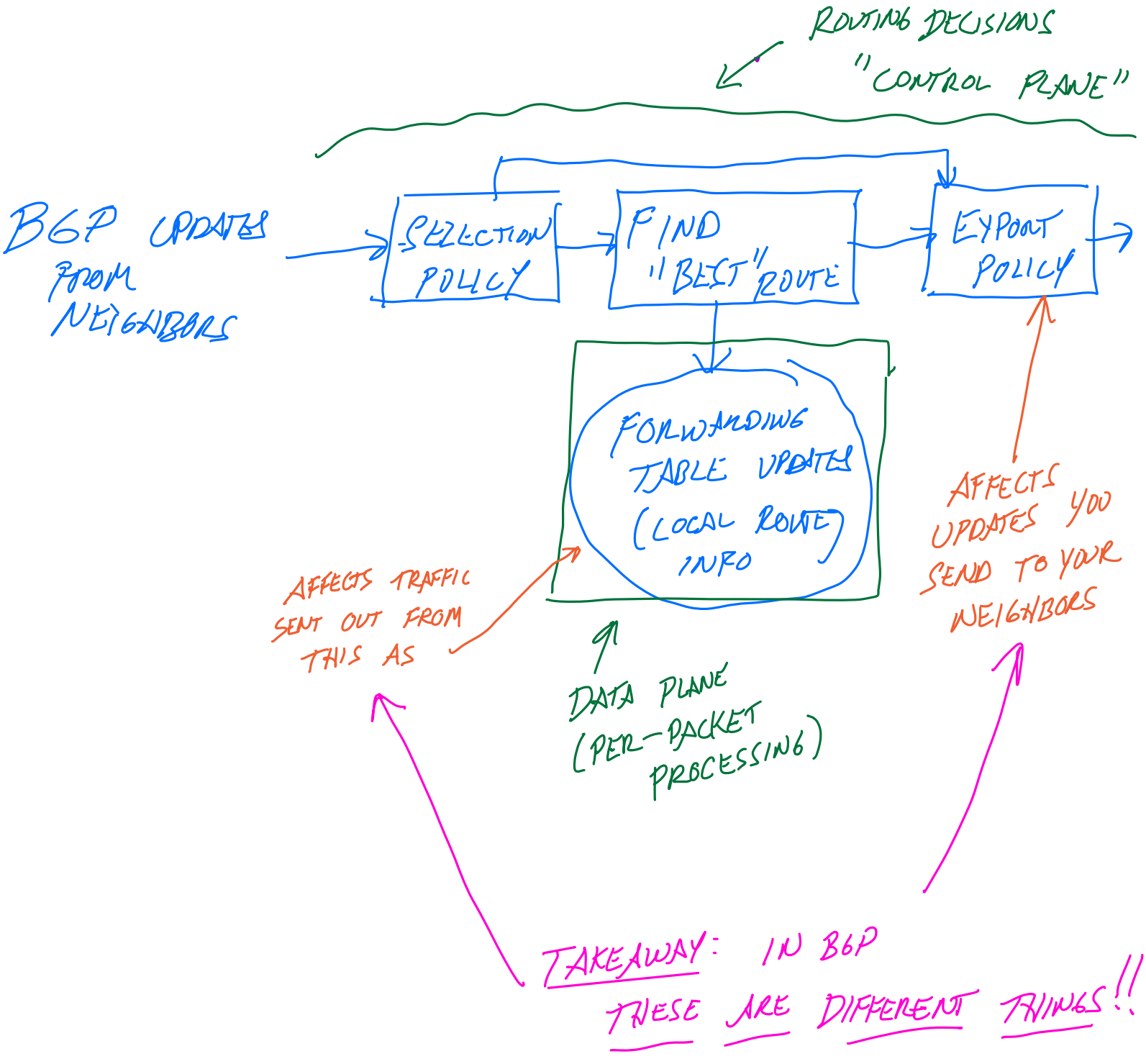
Suppose X has neighbors B, C, Q.
What routes might X NOT want to advertise to B? Why?

If X tells B it has a route to C, B will start sending traffic to X to get to C!
If B is a big network, this probably isn't what we want...

Takeaway: in BGP, there is a difference between:
- What routes you add to YOUR forwarding table
  => Selection policy
- What routes you tell your neighbors about
  => Export policy

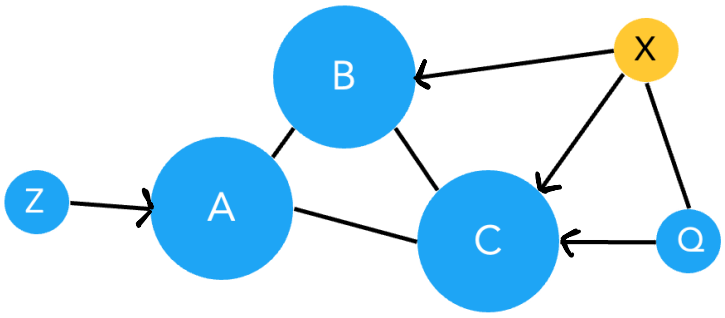
These are different things!

How to think about selection vs. export policies (visually)



AS *relationships*

NOTATION
→ CUSTOMER
— PEER



Customer

Pays provider to advertise its routes to the network
=> Q pays C
=> X pays B, C ("multihomed")

Ex. Q, X, Z

Examples:
- "B is transit [provider] for X"
- X is NOT transit for B => X isn't paying for that! ...and it likely doesn't have the capacity

Provider

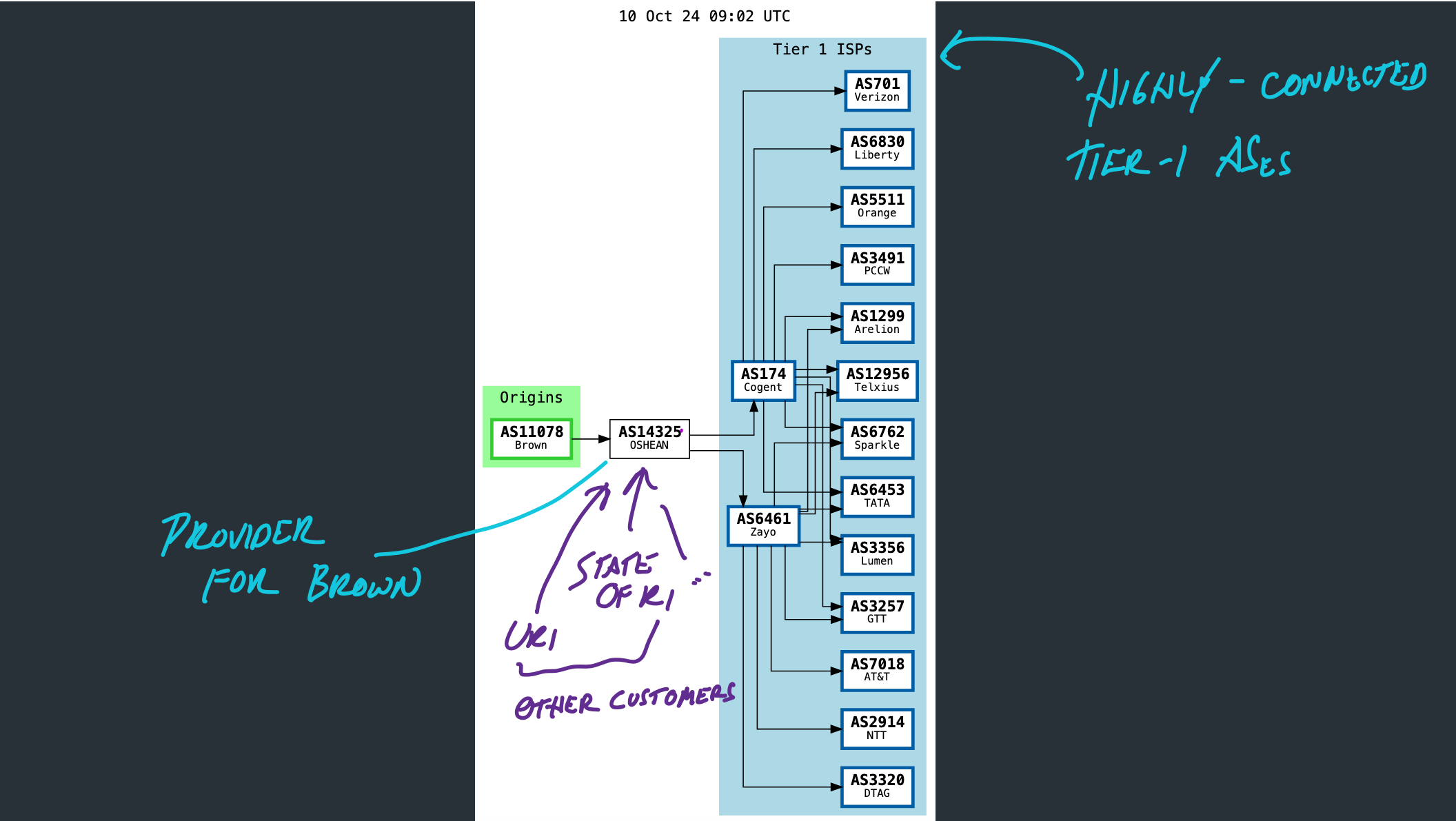
Highly connected ISPs
=> Most connected called "tier 1" ISPs => no default route!
=> "tier 2" is customer of tier 1

A, B, C

Peer

ASes may share routes at no cost, for mutual benefit

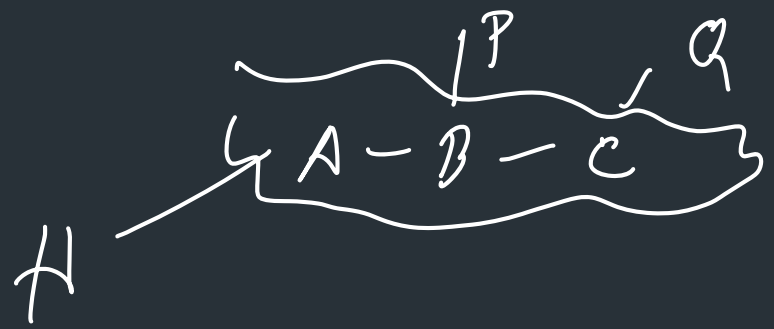
Ex. A–B
B–C
A–C
X–Q


10 Oct 24 09:02 UTC
Tier 1 ISPs
AS701 Verizon
AS6830 Liberty
AS5511 Orange
AS3491 PCCW
AS1299 Arelion
AS12956 Telxius
AS174 Cogent
AS6762 Sparkle
AS6453 TATA
AS6461 Zayo
AS3356 Lumen
AS3257 GTT
AS7018 AT&T
AS2914 NTT
AS3320 DTAG
Origins
AS11078 Brown
AS14325 OSHEAN
HIGHLY - CONNECTED TIER-1 ASES
PROVIDER FOR BROWN
STATE OF R1 ...
R1
OTHER CUSTOMERS

Typical route selection policy

In decreasing priority order:

1. Make or save money (send to customer > peer > provider)
2. Try to maximize performance (smallest AS path length)
3. Minimize use of my network bandwidth ("hot potato routing"
4. …



How to think about selection routes

PAYS YOU :)

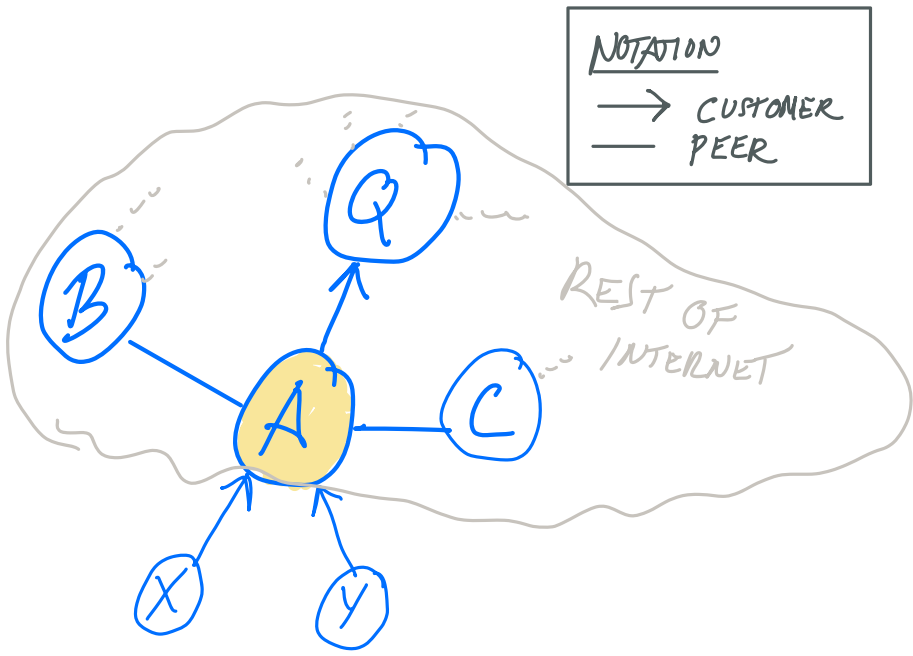
YOU PAY THEM :)

In order by priority:

1. Make or save money (send to customer > peer > provider)
2. (other admin-defined parameters)
3. Try to maximize performance (e.g., AS path length)
4. ..

"Technical" policies (like shortest AS path) are less of a priority than economic, legal, political policies

How to think about export policies
(Gao-Rexford principles)



- Example: For some ISP A
  - A has customers X and Y (i.e., X and Y pay A)
  - A peers with B and C
  - A is customer of Q (i.e., A pays Q)

If prefix is advertised by...	Export prefix to...
CUSTOMER (EG. X, Y)	EVERYONE! (X, Y, C, B, Q)
PEER (EG. B)	CUSTOMERS ONLY (X, Y) (NOT C, Q)
PROVIDER (Q)	CUSTOMERS ONLY (X, Y)

Goal: don't become transit when there's no gain!

"Don't all these advertisments from different ASes lead to really big routing tables?"
=>Yes, though there are some tricks that can help (to an extent)...

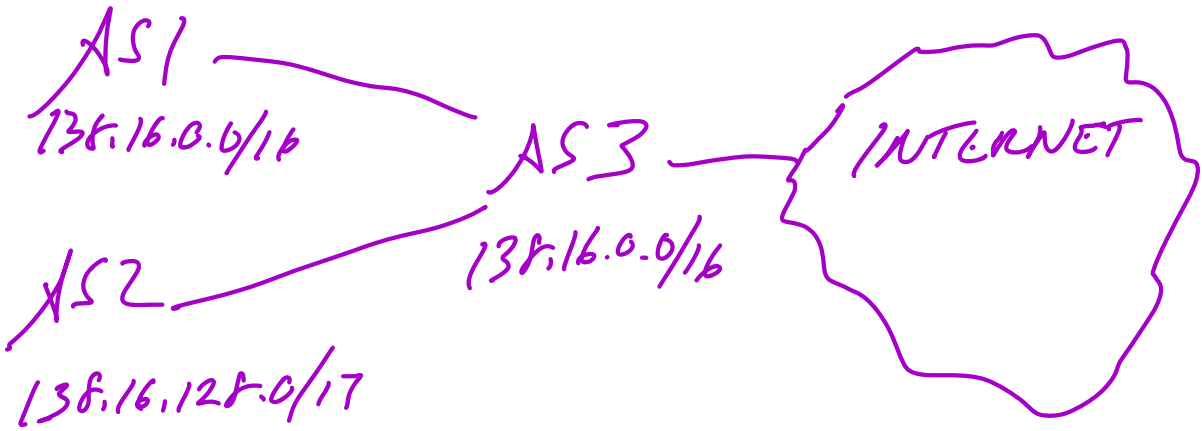
IP PREFIXES / ROUTE AGGREGATION

138.16.0.0/16 138.16.X.X

IDEA: ALLOCATE SMALLER NETWORKS FROM ONE PREFIX

EX. COULD DIVIDE INTO TWO NETWORKS

① 138.16.0.0/17 0000 0000
② 138.16.128.0/17 1000 0000



IDEA: AS3 COMBINES, OR AGGREGATES, PREFIXES FOR ITS CUSTOMERS
=> LEVERAGE HIERARCHY OF ADDRESSES!

HOWEVER, NOT SO EASY IN PRACTICE...

BGP stats (as of today)

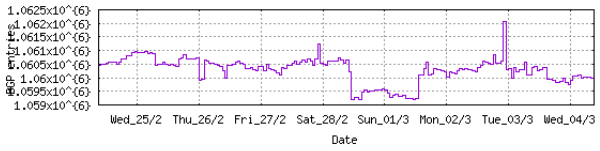
(from cidr-report.org)

Status Summary

Table History

Date	Prefixes	CIDR Aggregated
25-02-26	1060947	584607
26-02-26	1060721	585414
27-02-26	1060400	586101
28-02-26	1060547	585655
01-03-26	1059585	586435
02-03-26	1060213	586938
03-03-26	1060285	587323
04-03-26	1059765	587334

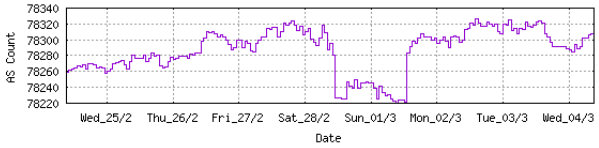
Plot: BGP Table Size



AS Summary

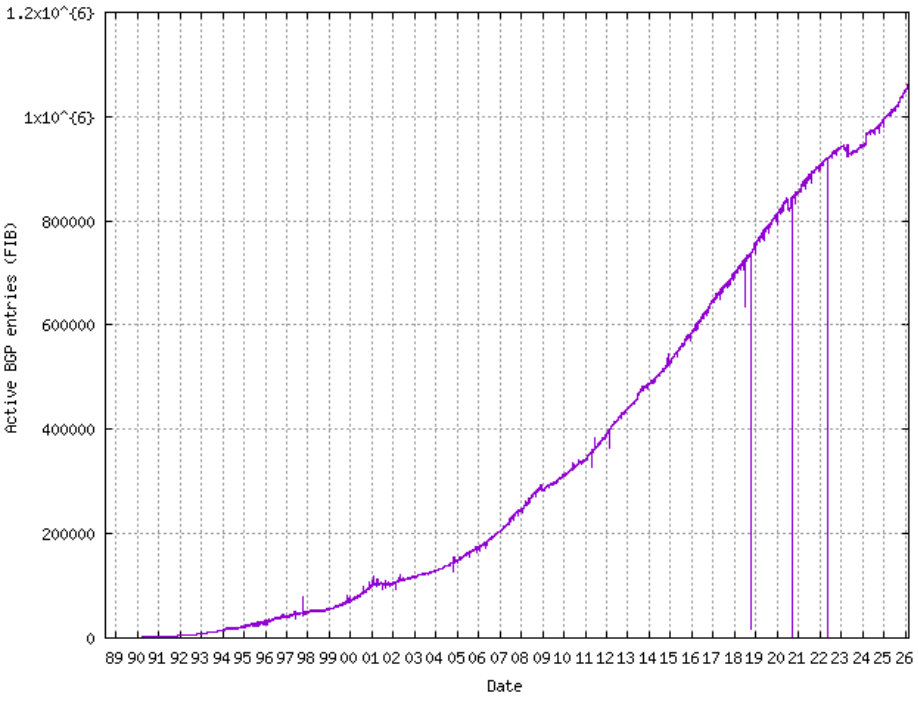
78307	Number of ASes in routing system
27168	Number of ASes announcing only one prefix
14432	Largest number of prefixes announced by an AS AS16509: AMAZON-02 - Amazon.com, Inc., US
227731712	Largest address span announced by an AS (/32s) AS749: DNIC-AS-00749 - United States Department of Defense DoD, US

Plot: AS count
Plot: Average announcements per origin AS
Report: ASes ordered by originating address span
Report: ASes ordered by transit address span
Report: Autonomous System number-to-name mapping (from Registry WHOIS data)



RECENTLY: NOW OVER 1M PREFIXES!

Active BGP entries (FIB)



Plot Range: 30-Jun-1988 1430 to 04-Mar-2026 0904

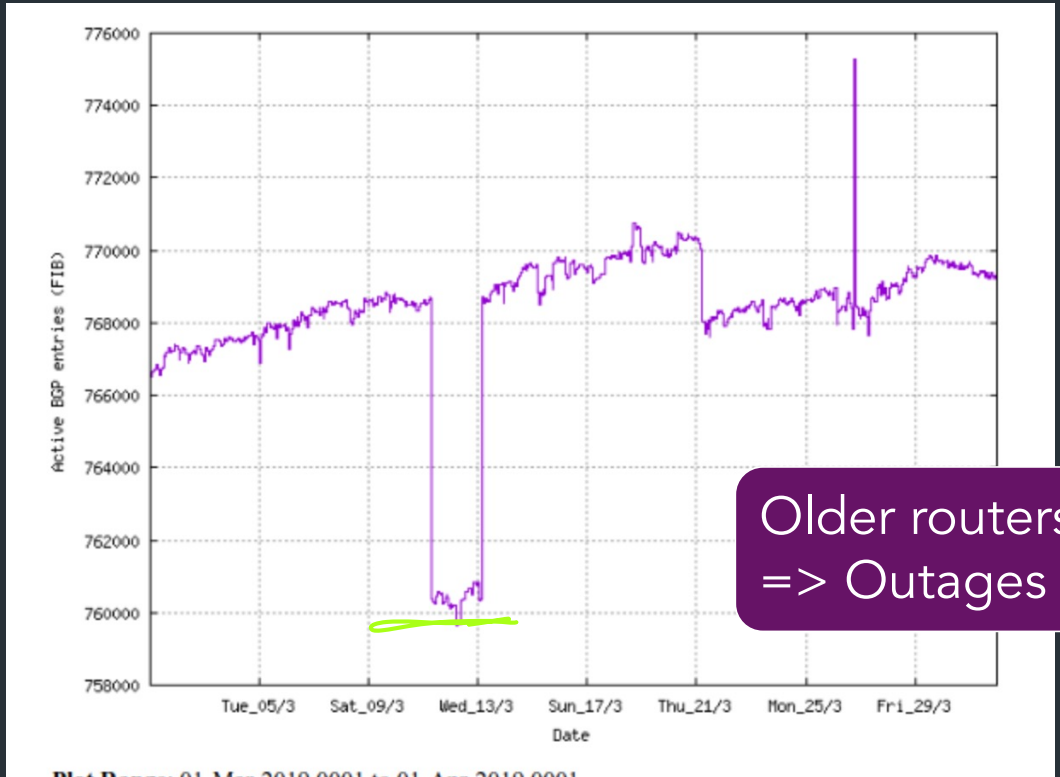
Why is this a problem?

Routers need to store forwarding tables in very fast memory called TCAM, which is designed to do longest-prefix matching very quickly. However, TCAM is very expensive, so routers only have a very limited amount

This has caused problems when the table has grown larger than routers' typical amount of TCAM!

How big can the table get?

- August 12, 2014: the full IPv4 BGP table reached 512k prefixes
- March 5, 2019: 768k prefixes



Older routers run out of space => Outages

BGP can be fragile!

- Individual router configurations and policy can affect whole network
- Consequences sometimes disastrous…

Peering Drama

- Cogent vs. Level3 were peers
- In 2003, Level3 decided to start charging Cogent
- Cogent said no
- Internet partition: Cogent's customers couldn't get to Level3's customers and vice-versa
  - Other ISPs were affected as well
- Took 3 weeks to reach an undisclosed agreement

Facebook DNS outage

- October 2021: Misconfiguration causes Facebook to withdraw routes for its DNS servers
- DNS: core service that translates domain names to IPs

  facebook.com => 1.2.3.6
- All services dependent on Facebook services go offline

Some more examples after this (for further reading)
(we'll talk about a subset next lecture)

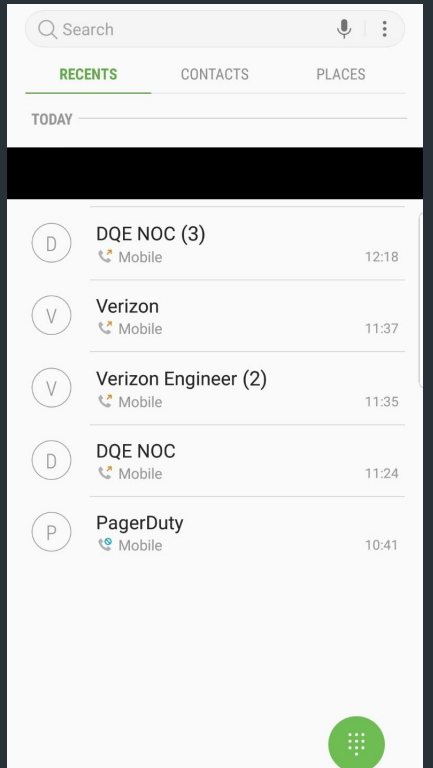
Some Notable incidents

June 24, 2019: Misconfigured small customer router accepted lots of transit traffic

Jérôme Fleury

[URGENT] Route-leak from your customer

To: CaryNMC-IP@one.verizon.com, peering@verizon.com, help4u@verizon.com,

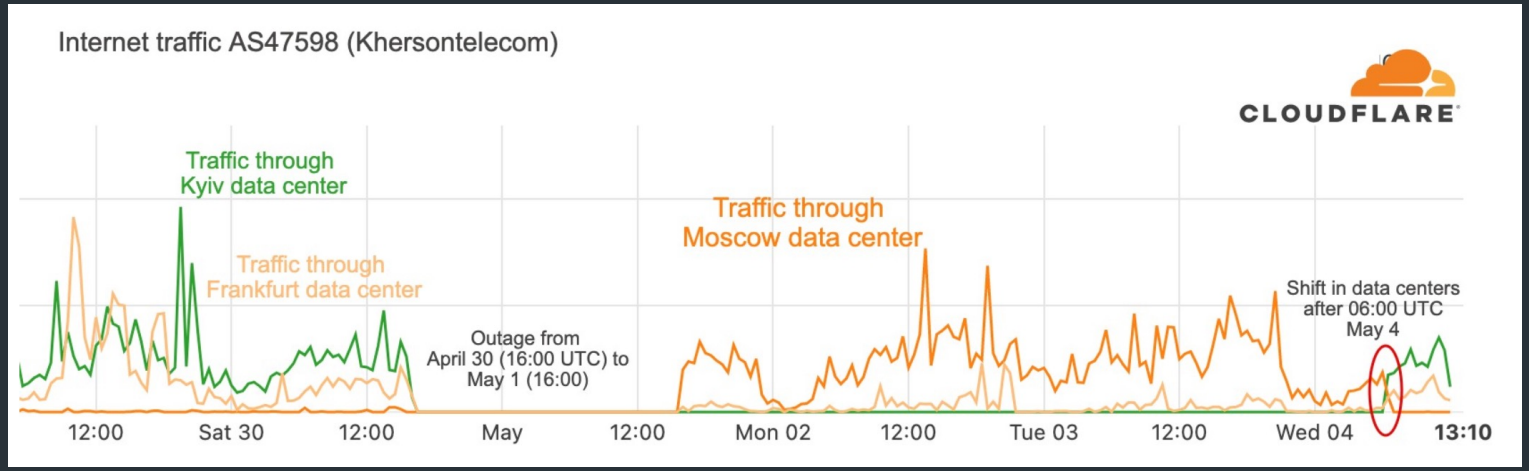


At this level, solving problems involves a lot of human expertise!

Pakistan Youtube incident

- Youtube's has prefix 208.65.152.0/22
- Pakistan's government order Youtube blocked
- Pakistan Telecom (AS 17557) announces 208.65.153.0/24 in the wrong direction (outwards!)
- Longest prefix match caused worldwide outage
- http://www.youtube.com/watch?v=IzLPKuAOe50

- ISP outage in Russian-occupied city of Kherson, Ukraine
- Comes back several days later… with traffic routed through a Russian ISP



https://blog.cloudflare.com/tracking-shifts-in-internet-connectivity-in-kherson-ukraine/