Interdomain routing and the Border Gateway Protocol

CIS 800/003

12 September 2011

This lecture

1. Brief introduction to the world of BGP
2. BGP operation
3. Route selection

4. Route advertisement
5. Some BGP problems

While this is going on

- Think about this material in the context of the course topics
- Can you identify any desirable **property** of BGP? How might we go about **proving** it?
- How can we reason about the **dynamic**, **unreliable** and **adversarial** environment?

Finding out about BGP

- Tutorials and documentation
  - from router vendors
  - from operators (including at forums like NANOG)
- RFCs
  - 4271 (BGP-4), 2796, 2858, 3065, 4272, ...
- Books
  - “BGP” by Iljitsch van Beijnum (O’Reilly 2002)
  - “Internet routing architectures” by Sam Halabi (Cisco 2000)

The Internet

1. Internet Protocol (IP)
   – hop-by-hop destination-based packet forwarding
2. Internetworking
   – a “network of networks”

The nature of Internet routing is intimately connected to both of these.

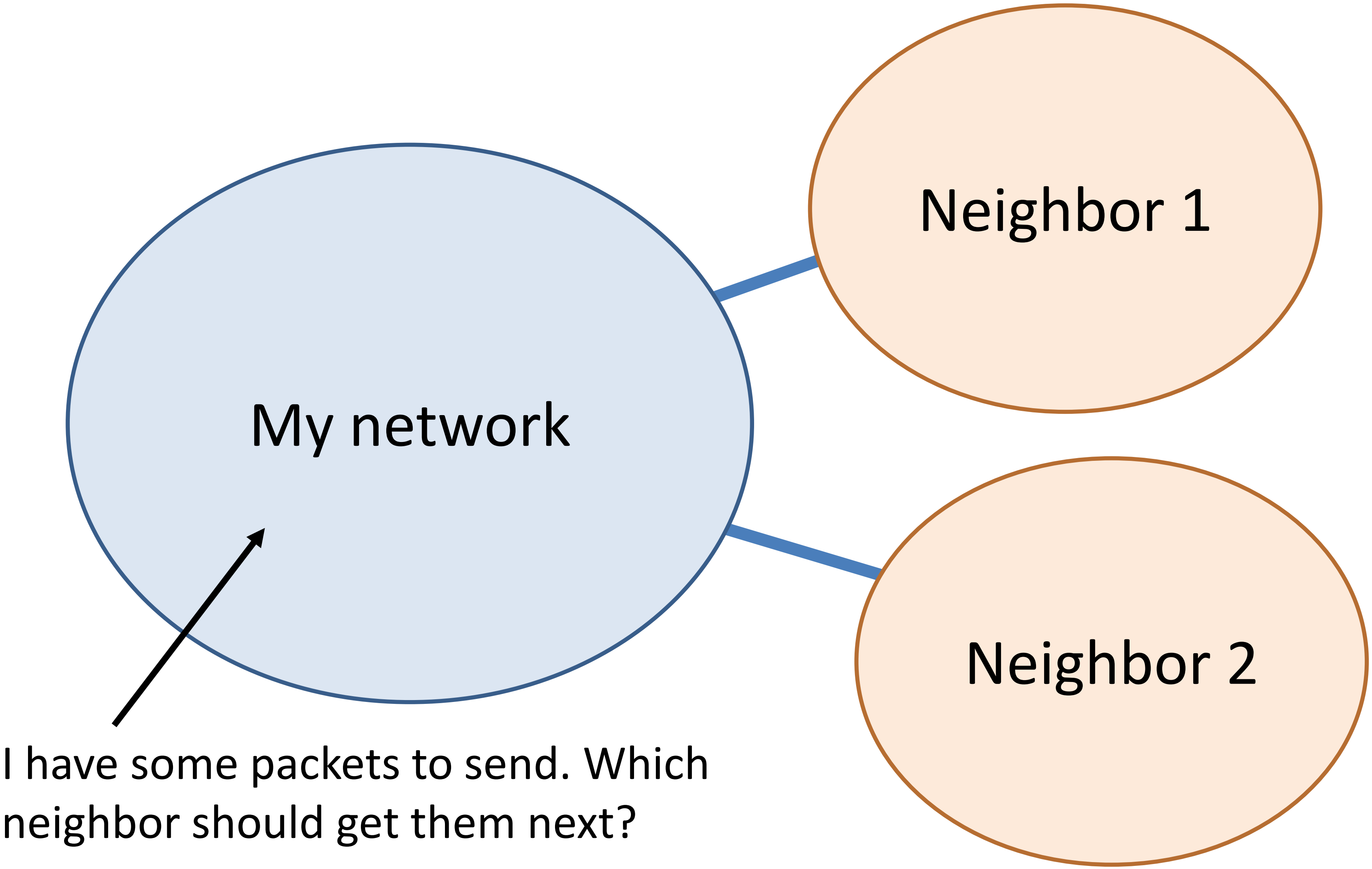
Internetworking

- Technical diversity (below the IP layer)
- Also *organizational* diversity
  - companies, universities, government, military, ISPs, CDNs, ...
- Many competing interests, but still wanting global connectivity

Routing and forwarding

- Receive an IP packet; decide where to send it next, based on its destination
- These decisions are encapsulated in a “forwarding table”
- **Routing** is about filling those tables – ideally ensuring correct and consistent forwarding, in a timely fashion

Routing between networks



Routing between networks

- For destinations within my own network – no problem; my own internal routing system will handle it
- My internal routing is also sufficient to get data to some *egress point*
- But I need **interdomain routing** to tell me which egress point is the right one

My neighbors tell me

- I run the Border Gateway Protocol with each of my neighbors
- Every so often, they send me messages:

Dear Alex,

I have a route to **96.17.168.112**. If you send me traffic for this destination, I will do my best to deliver it. This route has the following characteristics: [... details elided ...].

Your s faithfully,

Comcast (Autonomous System 7922)

I also tell things to my neighbors

- If I am willing to carry some of my neighbor’s traffic, for a particular destination, then I can send out a similar message.
- If not, then I won’t.
- Such decisions are based on economic considerations: What do I get in return for letting other people use my network?

Choosing routes

- It is quite likely that I will hear about lots of different routes for the same destination.
- I have to pick one as “the best” – the route I am going to use, and maybe let others use.
- The BGP standard lays out some rules for how to make these decisions.
- However, there is **a lot of flexibility**.

BGP route attributes

- Every route comes with associated data
- Some data comes from neighbors (and their neighbors, and so on...)
- I can modify any attribute
  - Though it might not be a good idea

Local preference

- The primary means of path selection is *local preference,* a numeric value for each route
- This value is chosen by the recipient
- Only look at other attributes if the local preference is tied

- The **very first step** in route selection is that I get to do **whatever I want**.

“Best” paths

- Everybody has their own competing ideas about what a “good” path looks like – because the network owners are commercial competitors.
- These criteria may be expressed in BGP policy.
- The BGP pathfinding process can only deliver some sort of compromise, rather than a globally-agreed optimum.

BGP does not find shortest paths

- Despite the stated intentions of its designers, BGP path selection is *not* just choosing the “shortest” path.

- **So what problem is BGP actually solving?**

- **Could it be solved differently?**

This lecture

1. Brief introduction to the world of BGP
2. **BGP operation**
3. Route selection

4. Route advertisement
5. Some BGP problems

Basic concepts

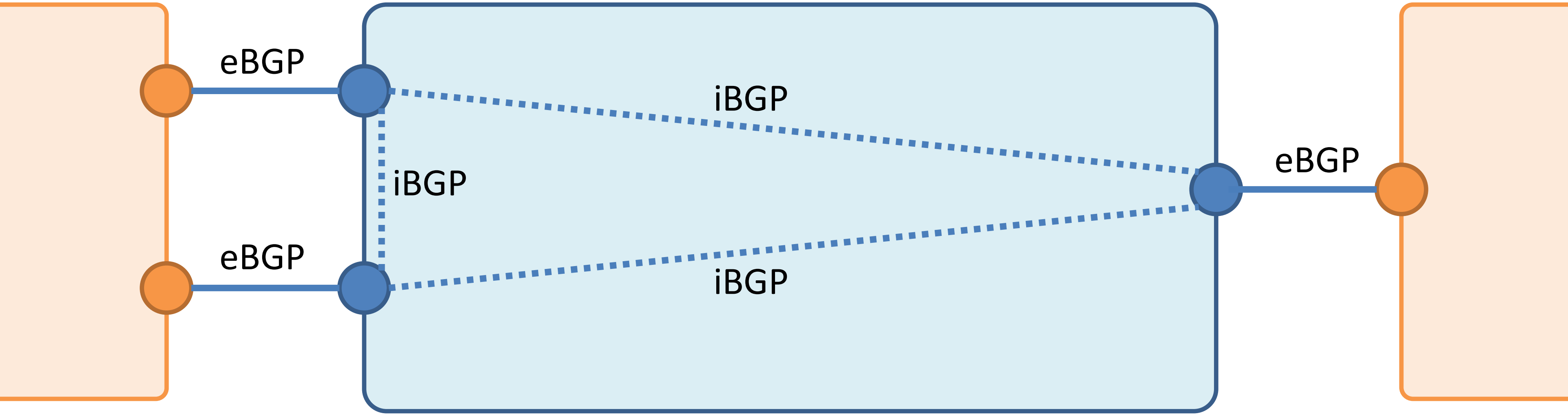
- BGP is a protocol spoken by one router to another.
- Two routers may share a BGP *session* – the context of their conversation.
- The action of each router is governed by
  - the BGP standard, as interpreted by the vendor
  - the operator's configuration

eBGP and iBGP

- Two types of session – external and internal
- eBGP is for talking to other networks
- iBGP is for disseminating route information across the routers of a single network

- Historical concept: the “gateway” that translates from one networking world to another

Original BGP network model



My border routers are the BGP speakers. They are in a full iBGP mesh. This connectivity is virtual – one iBGP link may correspond to many IP-level links.

BGP-learned routes are redistributed into my internal routing protocol, so that hosts within my network can select the correct egress point.

More structure in iBGP

- The full mesh idea does enable complete internal dissemination of route information
- However, it does not scale well as the number of border routers increases
- There are several different ways to resolve the problem

- **Do these change any semantics?**

Route reflectors

- Idea: partition my network into *clusters*.
- Each one has a *reflector*, a special router that talks to the reflectors in other clusters, and to the routers in its own local cluster
- A little like OSPF areas

Confederations

- Basically the same idea
- Have “sub-ASes” within my AS
- Each one is fully meshed internally
  - or could use reflectors, ...

Protocol extensibility

- All this is possible because BGP is extensible
- There are several ways to add new behavior:
  - New route attributes
  - New capabilities
- Backward compatibility is still a pain, but it is frequently possible
- **How can we understand protocol changes?**

Example: 4-byte ASNs

- Every Autonomous System has a number
- Used to be two bytes (and with some values reserved) but we felt the pinch
- Now we have 4-byte numbers as well
- No chance of a flag day – we had to interoperate with “old speakers”

Pretend to be AS23456

- If you have a 4-byte ASN, but need to talk to an old speaker, just say you're AS23456.
- Many things can break now!
- Several neighbors could now look like they are the same network – this affects path selection

- See RFC 4893 (2007)

AS path

- Routes carry a list of the ASes they traverse
- Old speakers expect each entry to be two bytes long
- If we all just say we're AS23456, new speakers will not get useful information from the path attribute

Solution: AS4_PATH

- A new attribute carrying the 4-byte numbers
- Old speakers will pass it along unchanged
- When the route reaches a new speaker, they can *reconstruct* what the AS4_PATH should be

AS_PATH: 17, 23456, 1982, 3287, 440

AS4_PATH: 17, 68901, 1982, **3287, 440**

reconstructed

- Several bugs encountered in practice (hopefully now fixed)

This lecture

1. Brief introduction to the world of BGP
2. BGP operation
3. **Route selection**

4. Route advertisement
5. Some BGP problems

Route selection

- There are several different variations on this
- Cisco supports a “weight” attribute that is applied even before local preference
- Route reflectors and confederations bring their own tweaks
- Not necessarily deterministic(!) though this is strongly recommended

Important route attributes

1. Local preference
2. AS_PATH length
3. Origin type
4. Multi-exit discriminator
5. eBGP vs. iBGP
6. IGP distance to next hop
7. Router ID

Overall impressions

- It’s all quite complicated
- There’s something a bit like shortest paths going on (AS_PATH, IGP distance) but that’s far from the whole story
- I get the last word in route selection (local preference overrides everything)
- But neighbors can give strong hints

Things that are impossible

- *Wouldn’t it be nice* if we had more information about these BGP routes?
  - end-to end delay
  - throughput estimates
  - geographical data
  - shared risk groups
- Alas, we do not. Some of these can be found out in other ways, but are not part of the protocol.

Playing with the AS path (1)

- The AS_PATH attribute is also used for loop avoidance. If I see my own number in the path, I should drop the route straight away.
- That means that somebody can *poison* a route that they don't want me to use, by inserting my own AS number

Playing with the AS path (2)

- Shorter paths are better
- You can discourage me from using a route by *padding* the AS path – artificially inserting many copies of your number, not just one
- After a lot of padding, many BGP implementations will fail due to exceeding various buffer sizes
  - this could be used for evil

Playing with the AS path (3)

- We don’t always have an accurate path
- Aggregation loses some information
  - Merge information for adjacent prefixes into a single set of route data
  - Done for efficiency reasons
  - Two AS lists become one AS set
- **How can we reason about whether optimizations change routing semantics?**

Routing policy considerations

- Basically, money.
- Charges may be based on traffic volume.
- My customers send me lots of data: good.
- I have to send data to my providers: bad.

- If a route is visible through a customer and a provider network, I will typically prefer the customer version.

Traffic engineering

- I can use local preference (or other attributes) to balance traffic across my neighbors
  - at least, for outgoing traffic
- I can also take advantage of *hot-potato* rules to reduce internal load – get the traffic out of my network as quickly as possible

This lecture

1. Brief introduction to the world of BGP
2. BGP operation
3. Route selection

4. **Route advertisement**
5. Some BGP problems

Route advertisement

- The other side of policy is control over what goes *out*.
- As previously mentioned, I can use tools like AS path padding to influence my neighbors' route selection.
- Traffic engineering for inbound traffic is, generally, a bit more difficult.

BGP communities

- Yet another extensibility mechanism
- Defined in RFC 1997; now very widely used
- Embed additional instructions in the route advertisement, scoped to a particular receiving AS.
- Also can be used to carry extra information about a route.

Typical community policy (1)

- This is from Comcast (AS 7922)
- Customer routes get local preference 300
- Routes tagged 7922:290 get local preference 290 instead (used for backup routes)
- Similarly, 7922:250, 7922:150, 7922:100.

Typical community policy (2)

- 7922:999 – don’t tell anyone about this
- 7922:888 – only tell Comcast customers, not its peers

- If you receive a route from Comcast, tagged with 7922:3000, it’s from one of their peers.
- Send 65100:(peer ASN) to suppress advertisement to that specific peer

Typical community policy (3)

- Ask Comcast to pad routes when announcing to a specific peer
- Example: 65103:(peer ASN) means to prepend three copies of 7922 when announcing this route to that peer

- Send the community of the beast, 7922:666, to activate blackholing (an emergency measure)

Communities

- Some are defined in the standard
- Most are left up to the discretion of the implementing AS
- There are some common patterns, but no standard vocabulary

- **How can this kind of thing be modelled?**

This lecture

1. Brief introduction to the world of BGP
2. BGP operation
3. Route selection

4. Route advertisement
5. **Some BGP problems**

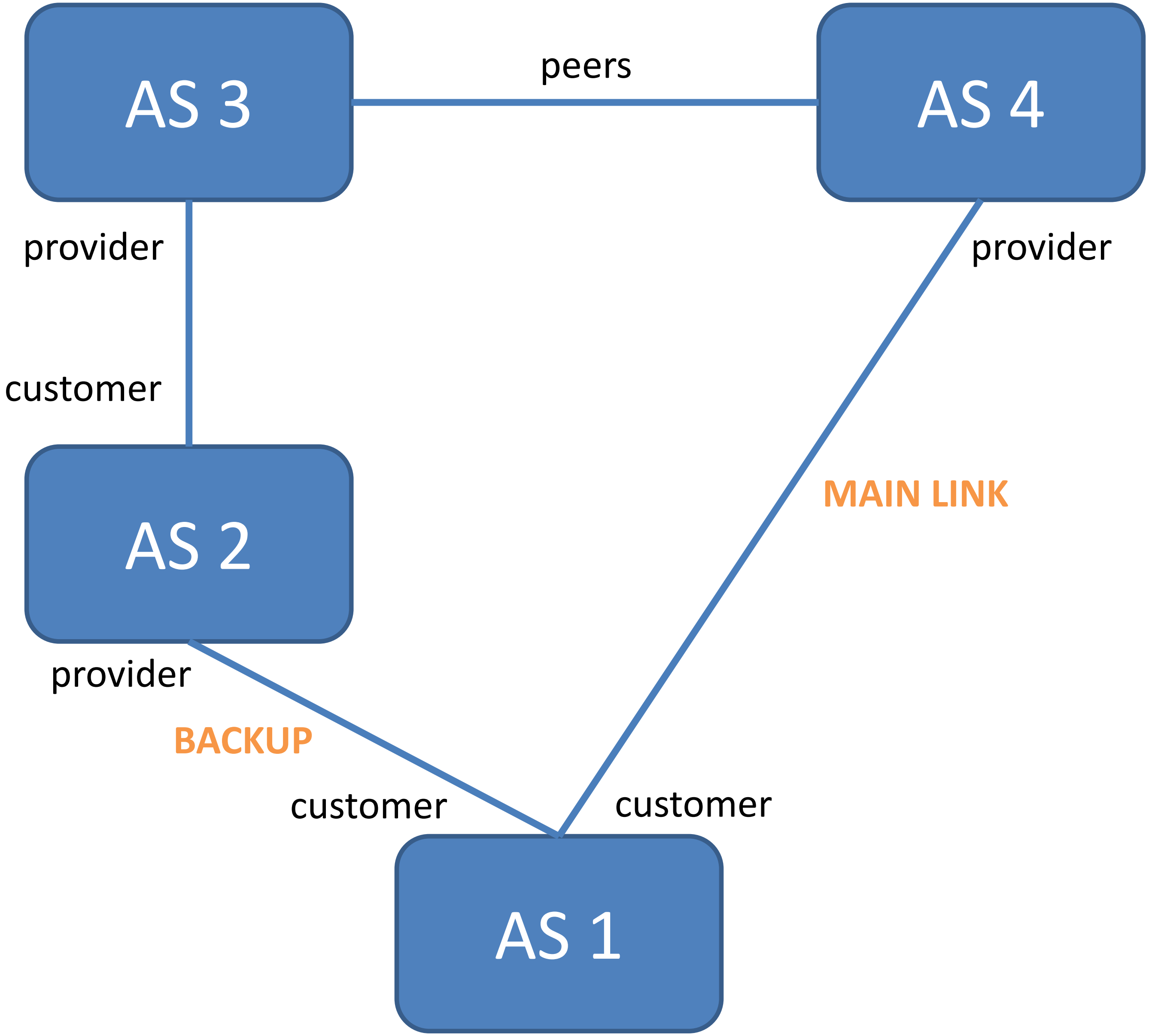
Thinking about BGP problems

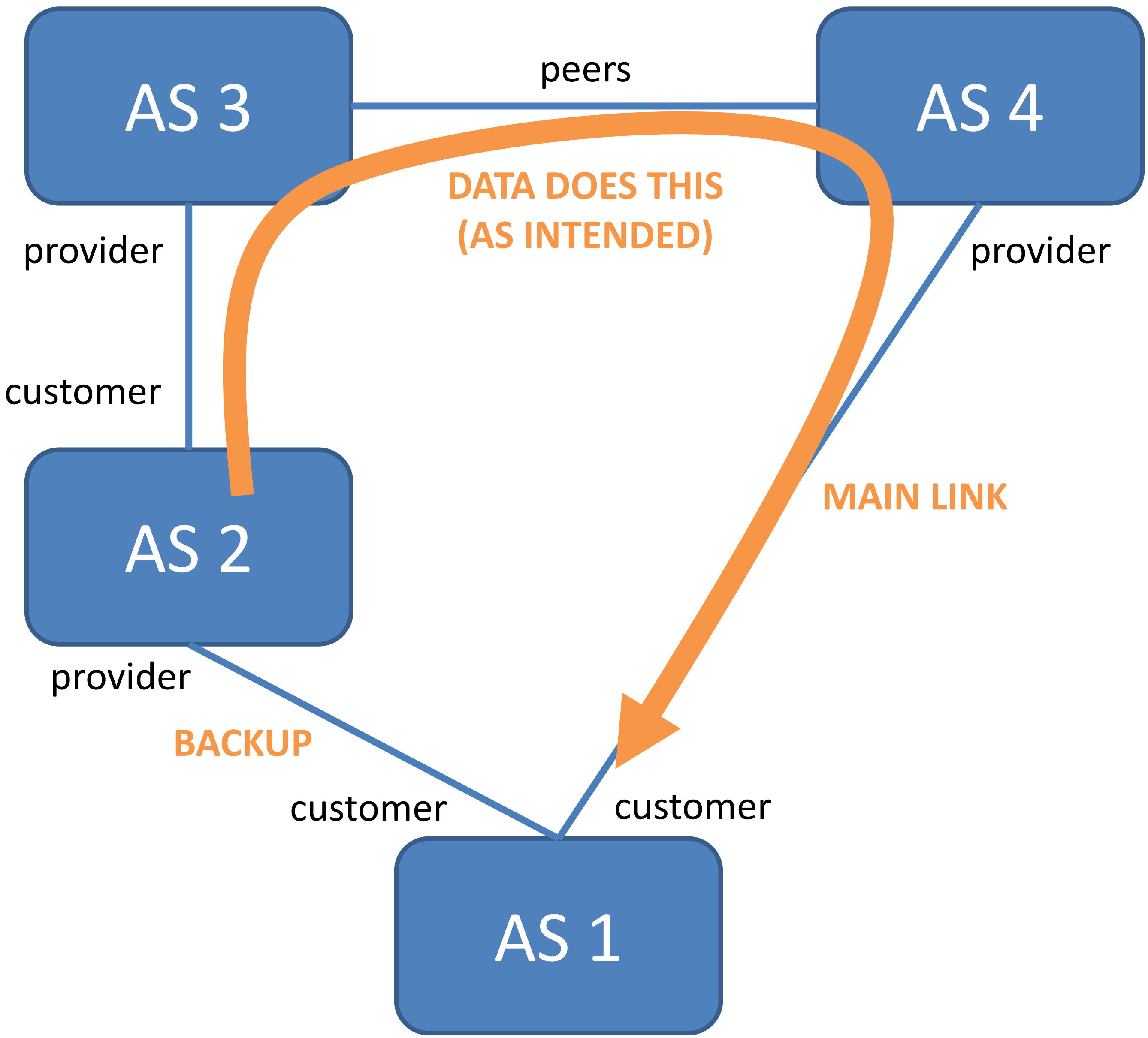
- To what can we attribute each failure?
  - A bug in the implementation?
  - An error by an operator?
  - A design flaw in the protocol?

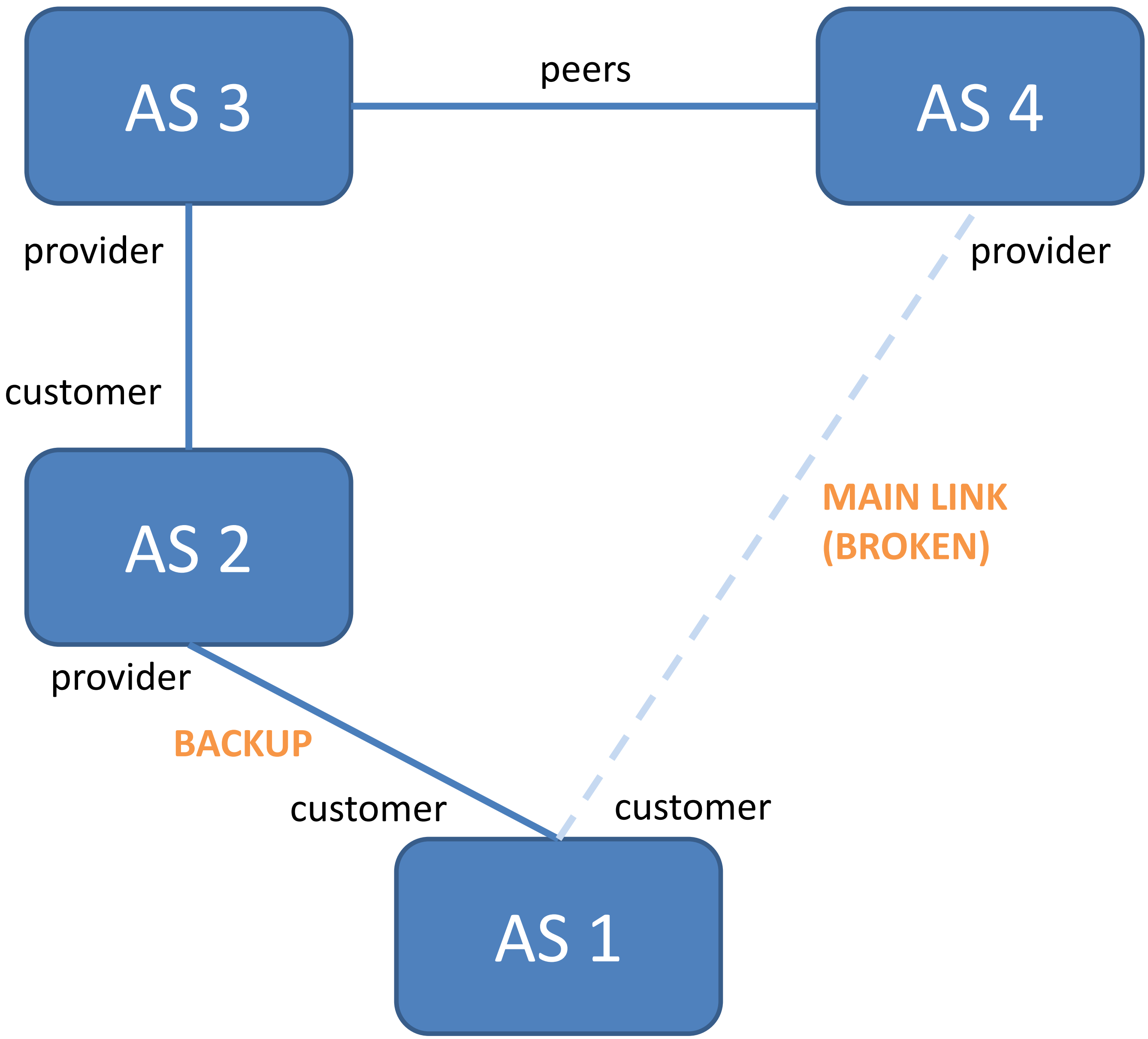
- How can these problems be fixed for good?

The Wedgie

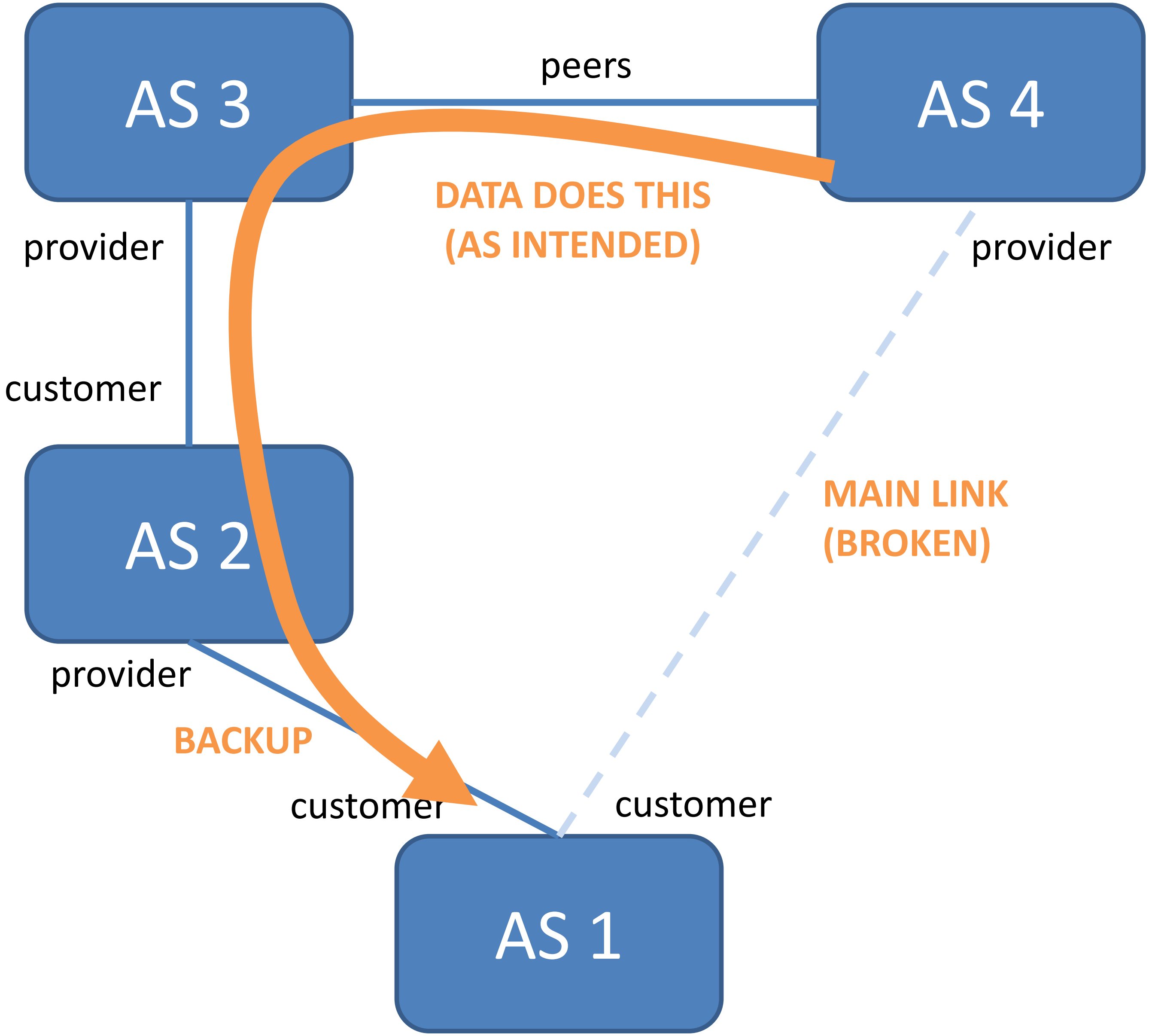
- RFC 4264 observed a BGP problem related to backup semantics
- It is called a “wedgie” because the routing system gets stuck in a bad state
- And it can be difficult to get unstuck (requiring non-local knowledge and action)

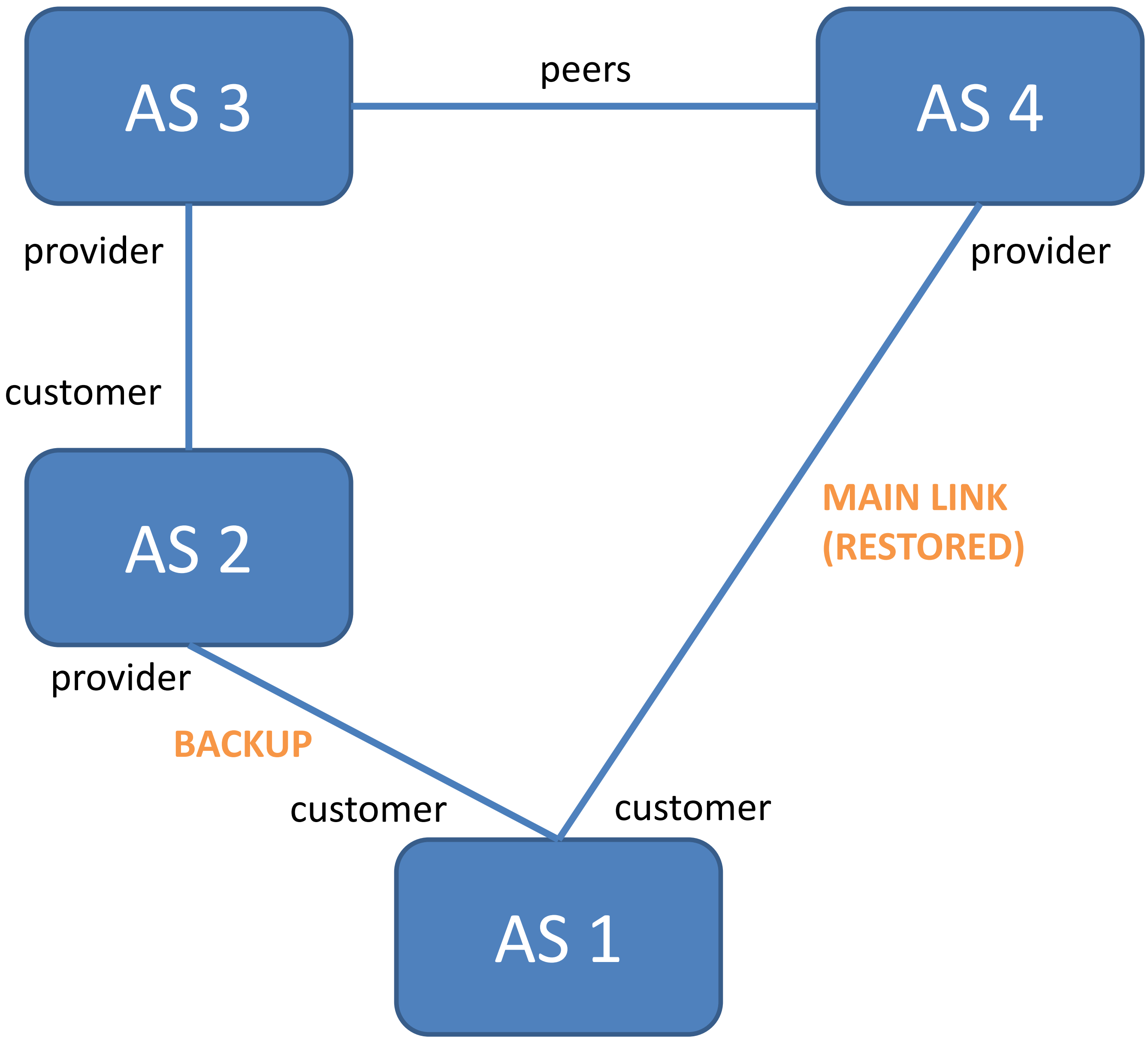

AS 3
peers
AS 4
provider
provider
customer
AS 2
MAIN LINK
provider
BACKUP
customer
customer
AS 1

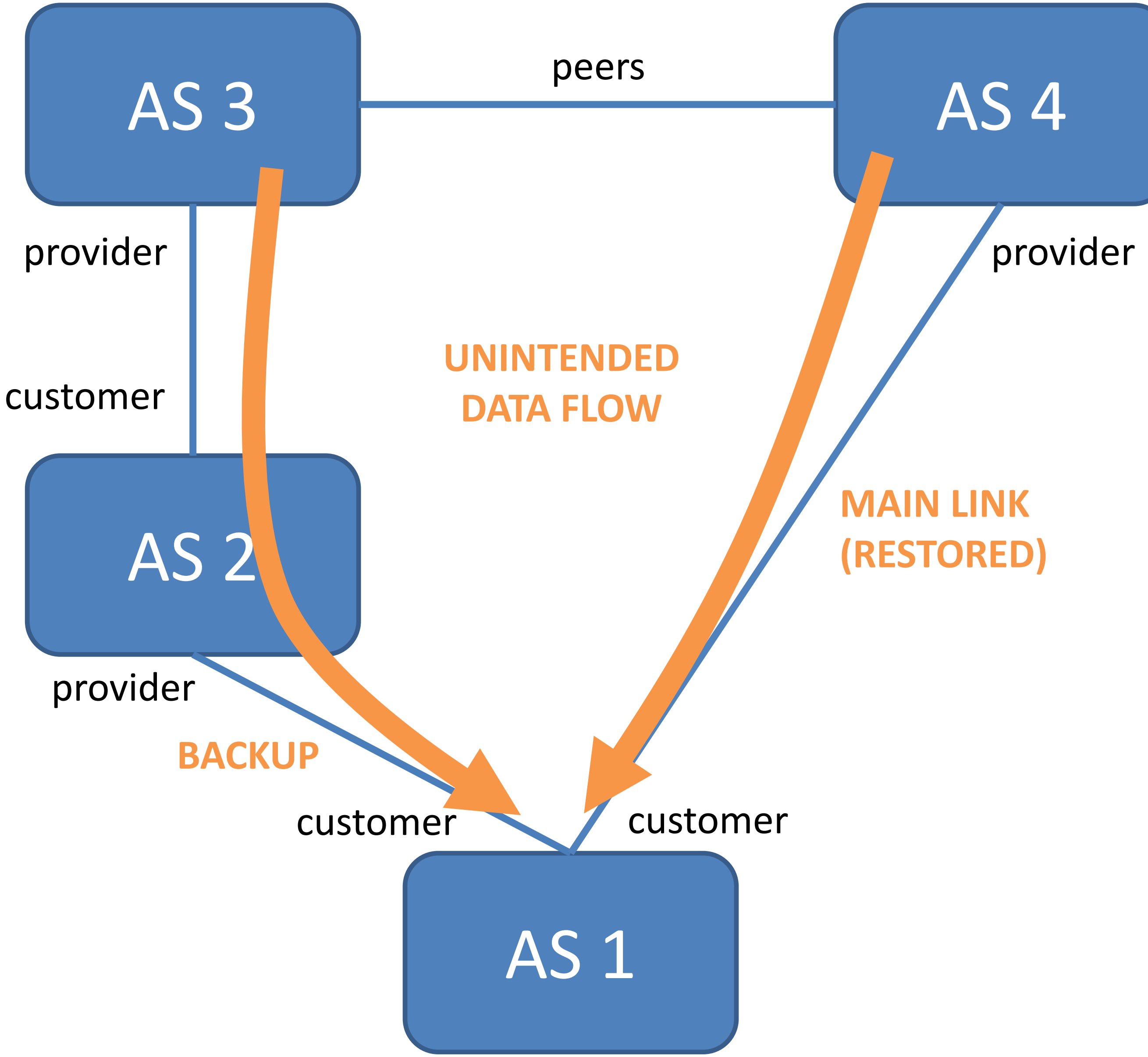

AS 3
peers
AS 4
DATA DOES THIS
(AS INTENDED)
provider
provider
customer
AS 2
MAIN LINK
provider
BACKUP
customer
customer
AS 1



AS 3
peers
AS 4
provider
provider
customer
MAIN LINK
(BROKEN)
AS 2
provider
BACKUP
customer
customer
AS 1


AS 3
AS 4
peers
DATA DOES THIS
(AS INTENDED)
provider
provider
customer
AS 2
MAIN LINK
(BROKEN)
provider
BACKUP
customer
customer
AS 1


AS 3
peers
AS 4
provider
provider
customer
MAIN LINK
(RESTORED)
AS 2
provider
BACKUP
customer
customer
AS 1


AS 3
peers
AS 4
provider
provider
UNINTENDED
DATA FLOW
customer
AS 2
MAIN LINK
(RESTORED)
provider
BACKUP
customer
customer
AS 1

Basic problem

- AS 2 does not “pass on” the backup semantics to AS 3
- There might not even be any mechanism for it do so! AS 3 is free to choose the path via AS 2.

- **Is there a general model for such problems?**
- **Can we come up with a better design for backup routing?**

Protocol oscillation

- We hope that the overall routing system will *converge*, delivering the desired routes
- This does not always happen. Sometimes we see oscillations – routers bouncing back and forth between several routes
- Sometimes the problem goes away on its own
- **What causes this?**

Example

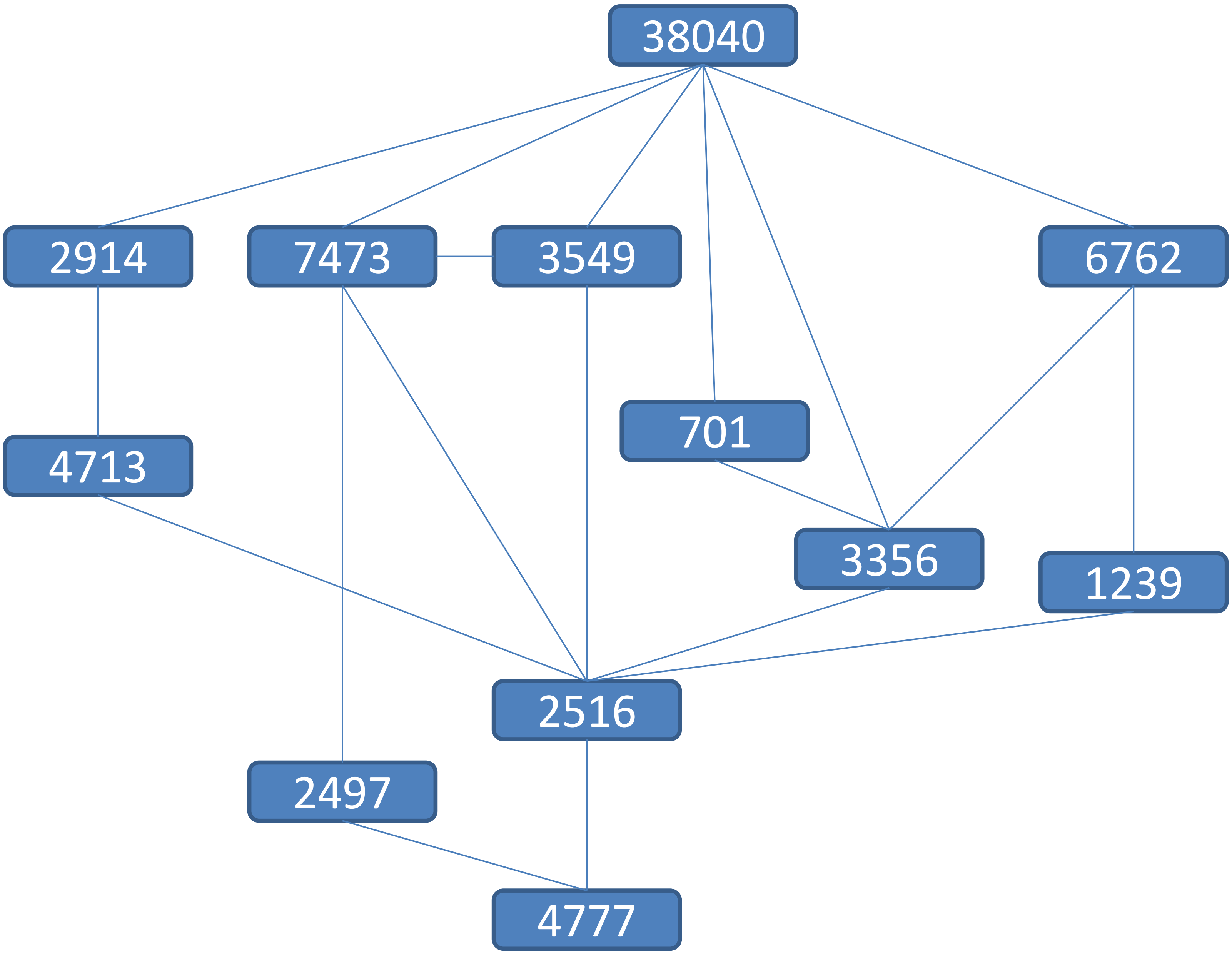
- Route from Reach Global Services Ltd. (AS 4637) to a prefix of TOT IIG (AS 38040), 180.180.240.0/24.

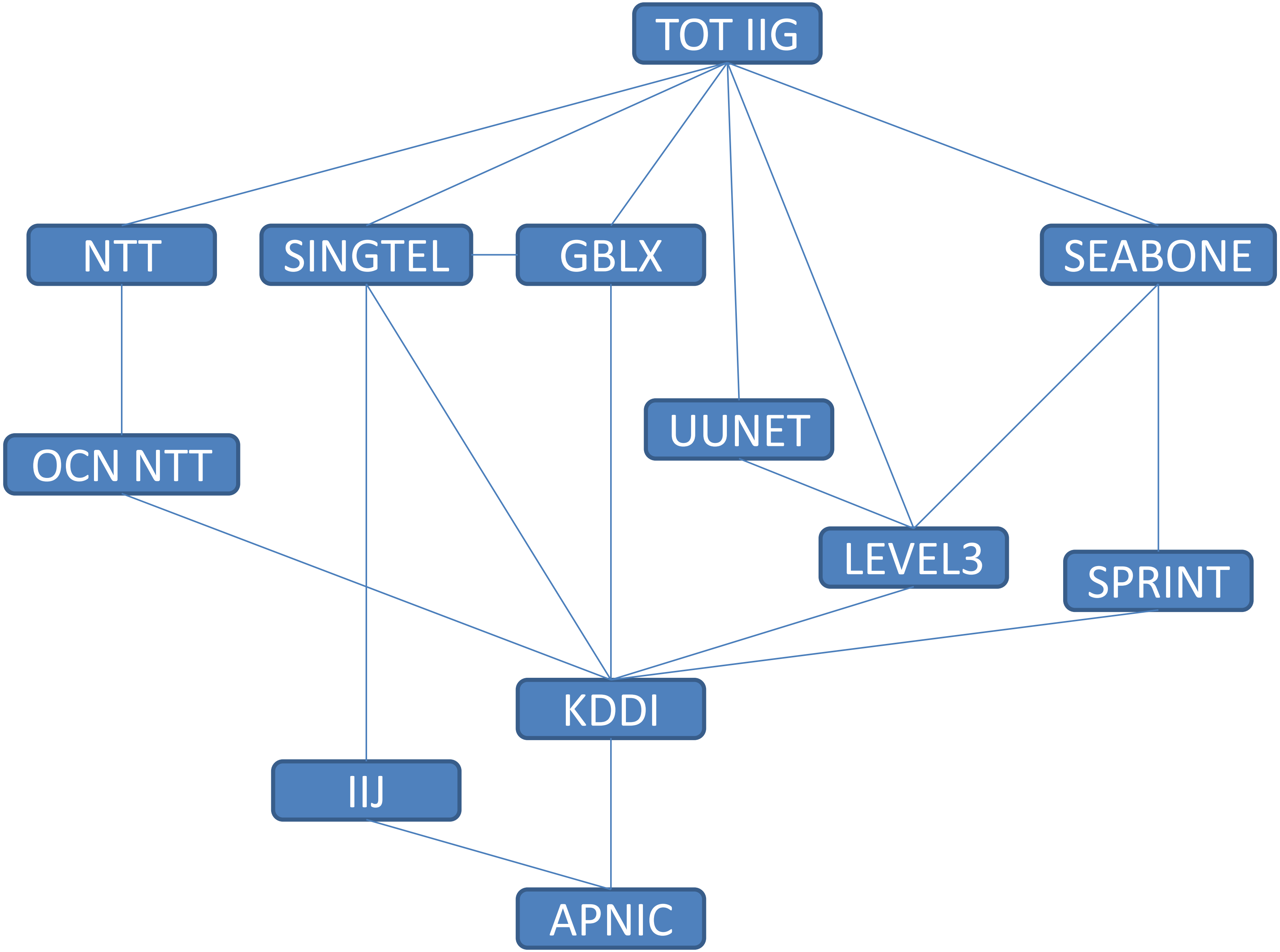
- All this data comes from Geoff Huston's site:

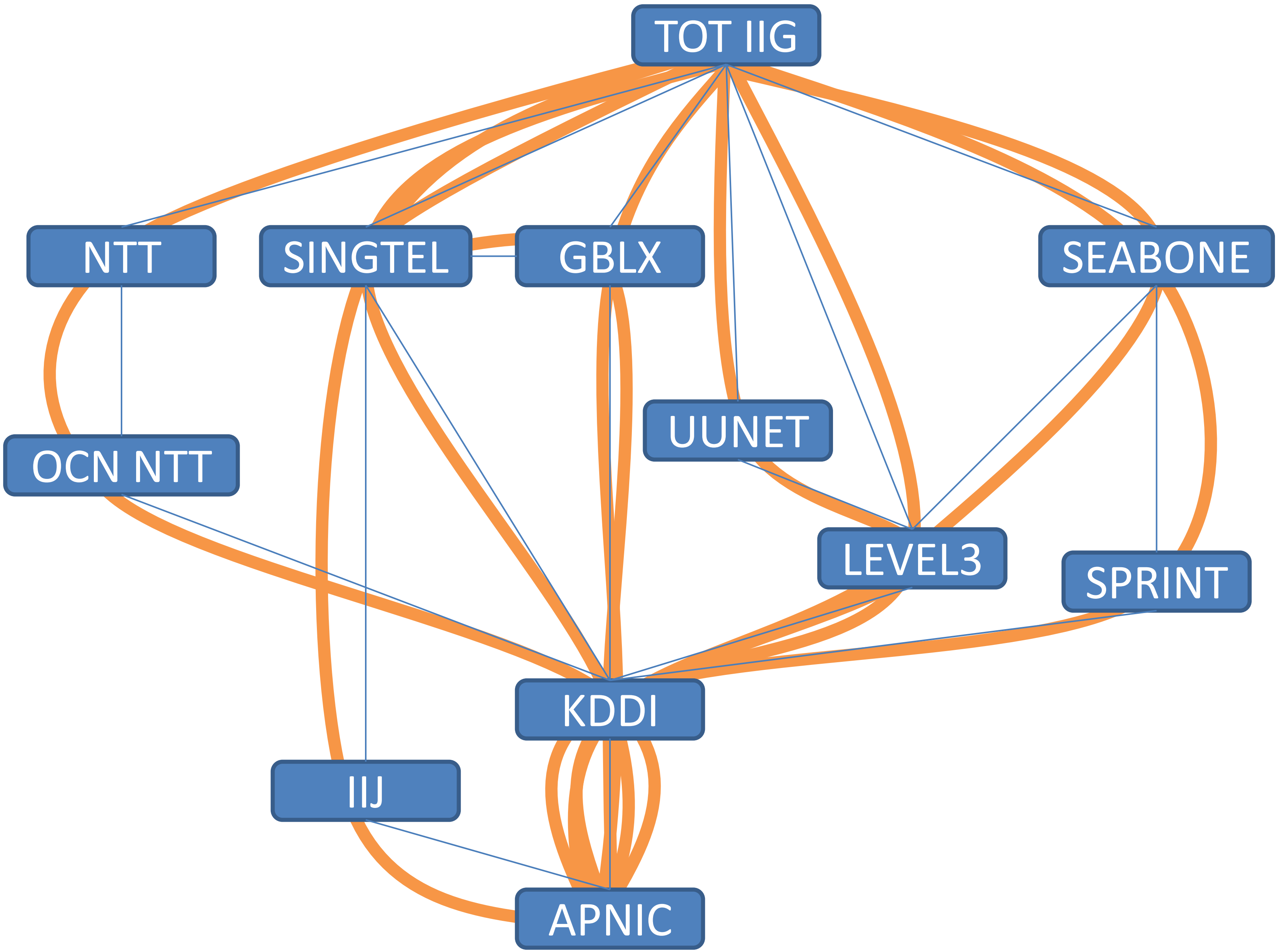
  http://bgpupdates.potaroo.net

In the past week

- 6461 updates
  - about one every 98 seconds
- 83 withdrawals (about 2 hours withdrawn)
- 5535 changes to next-hop
- 43 different paths
  - 9 were active for an hour or more
  - 2 were active for a day or more


38040
2914
7473
3549
6762
701
4713
3356
1239
2516
2497
4777


TOT IIG
NTT
SINGTEL
GBLX
SEABONE
UUNET
OCN NTT
LEVEL3
SPRINT
KDDI
IIJ
APNIC


TOT IIG
NTT
SINGTEL
GBLX
SEABONE
UUNET
OCN NTT
LEVEL3
SPRINT
KDDI
IIJ
APNIC

Analyzing oscillation

- For a *particular* oscillation, it is hard to figure out who to blame.
- If indeed any one party *is* to blame.
- In the wild, oscillations tend to be complicated – involving many networks and routes

Analyzing oscillation

- In *general* – why does this happen?
- Some possible causes:
  - A physical fault somewhere
  - Router bugs
  - Improper configuration (violating protocol expectations)
  - Something inherent to the protocol (but what?)

Persistent or transient?

- Some oscillations disappear on their own
- Not very satisfying
- Impossible(?) to predict from the log how long it might take

- BGP convergence can take a long time and many oddities can occur along the way. Why?

Conclusion

- BGP is actually quite difficult to analyze
- It is “syntactically” simple
- But it’s tricky to characterize exactly what it is trying to do, independently of its definition
- Extensible semantics make analysis hard
- Mysterious behavior lacks an obvious explanation in the design and configuration

Next lectures

- Mathematical models of BGP – its algorithm and decision-making process
- Understand where oscillations come from
- Relate that to routing policy
- Suggest protocol changes

On Wednesday

- Sangeetha will present “The stable paths problem and interdomain routing” (ToN 2002) by Griffin, Shepherd and Wilfong

- See you all there!