LONAP
London Access Point

IXPs and Robust Configuration

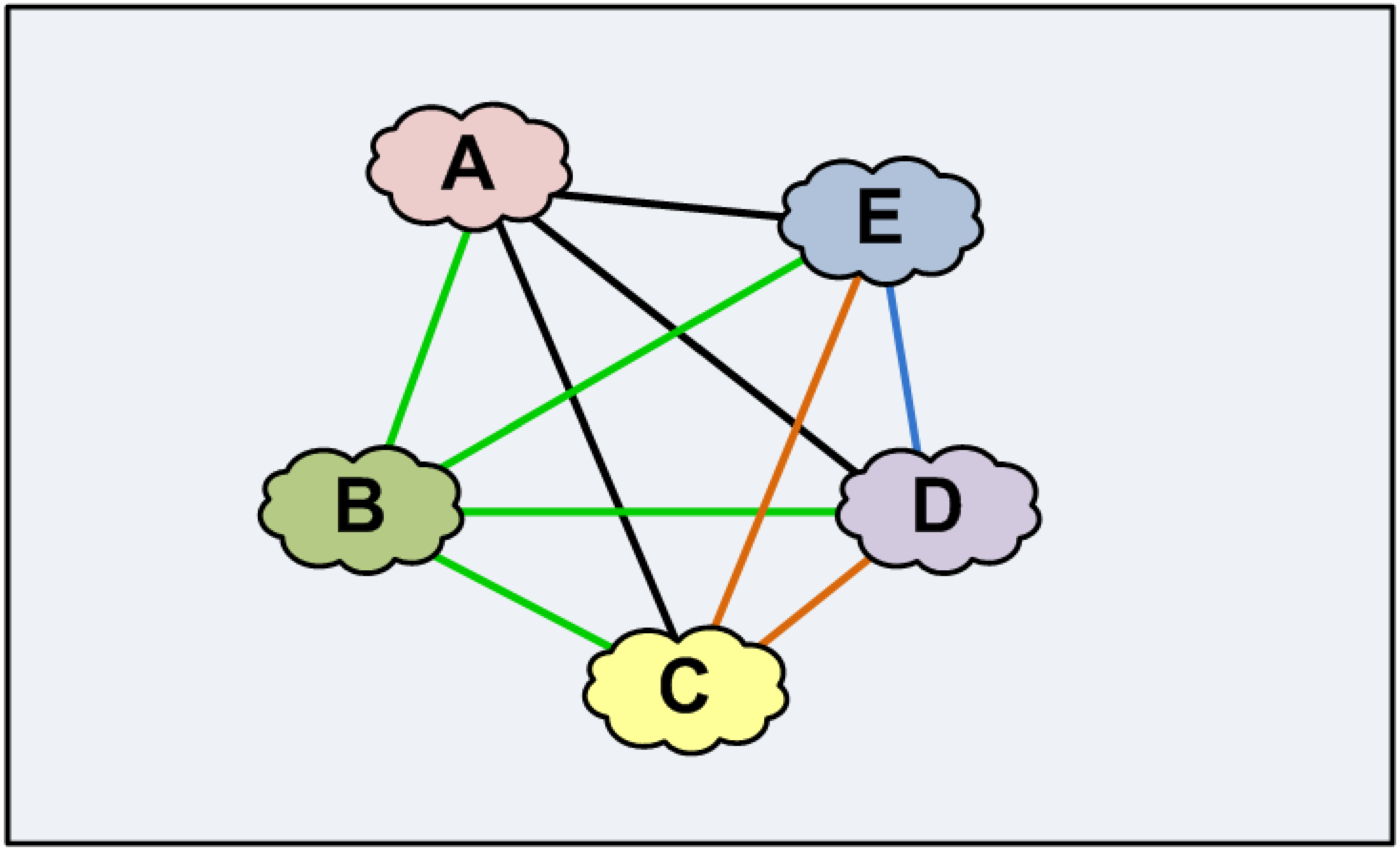
aka. “interesting” configs we have seen…

Rob Lister

UKNOF24

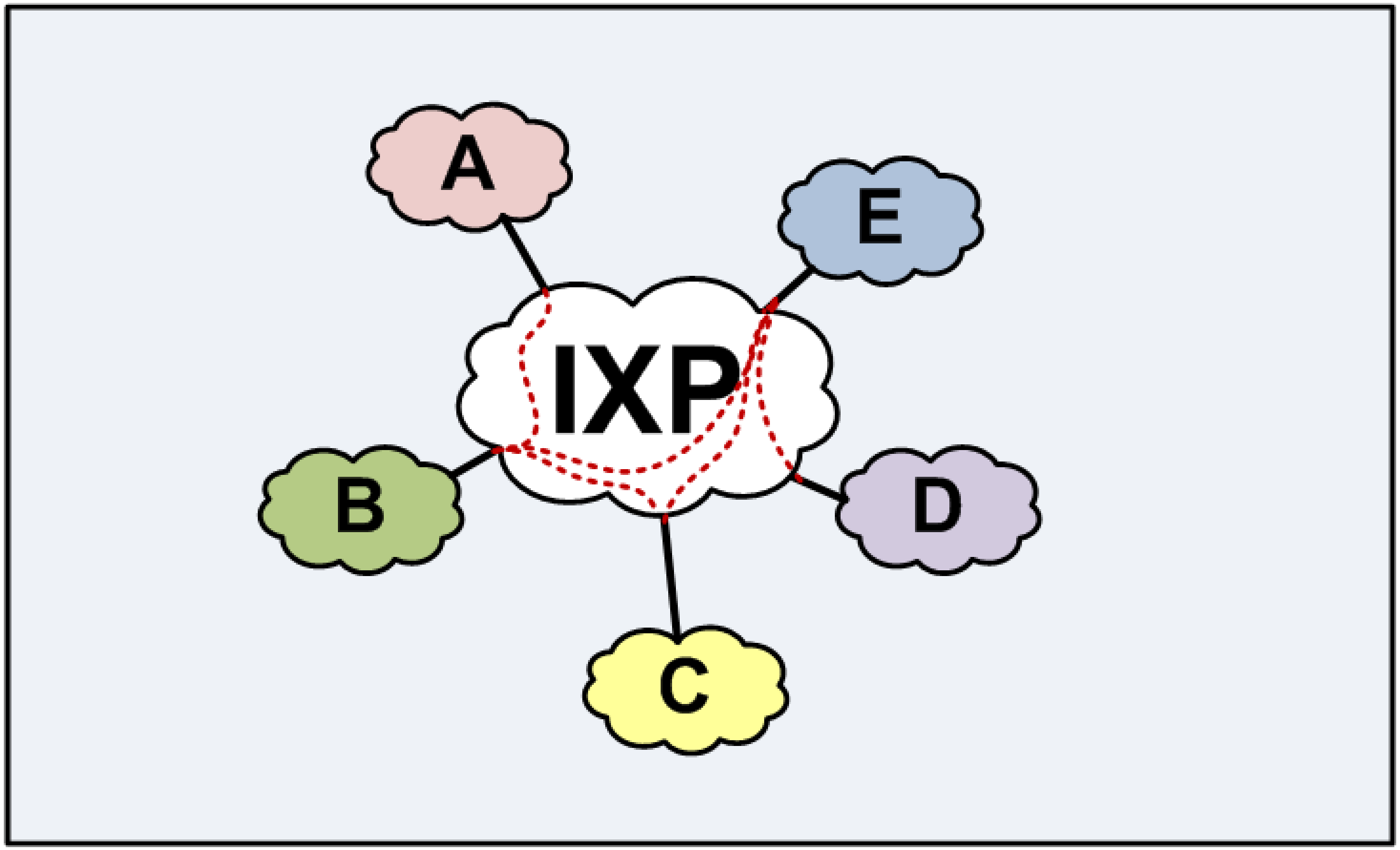
17 January 2013

Original purpose of IXPs...

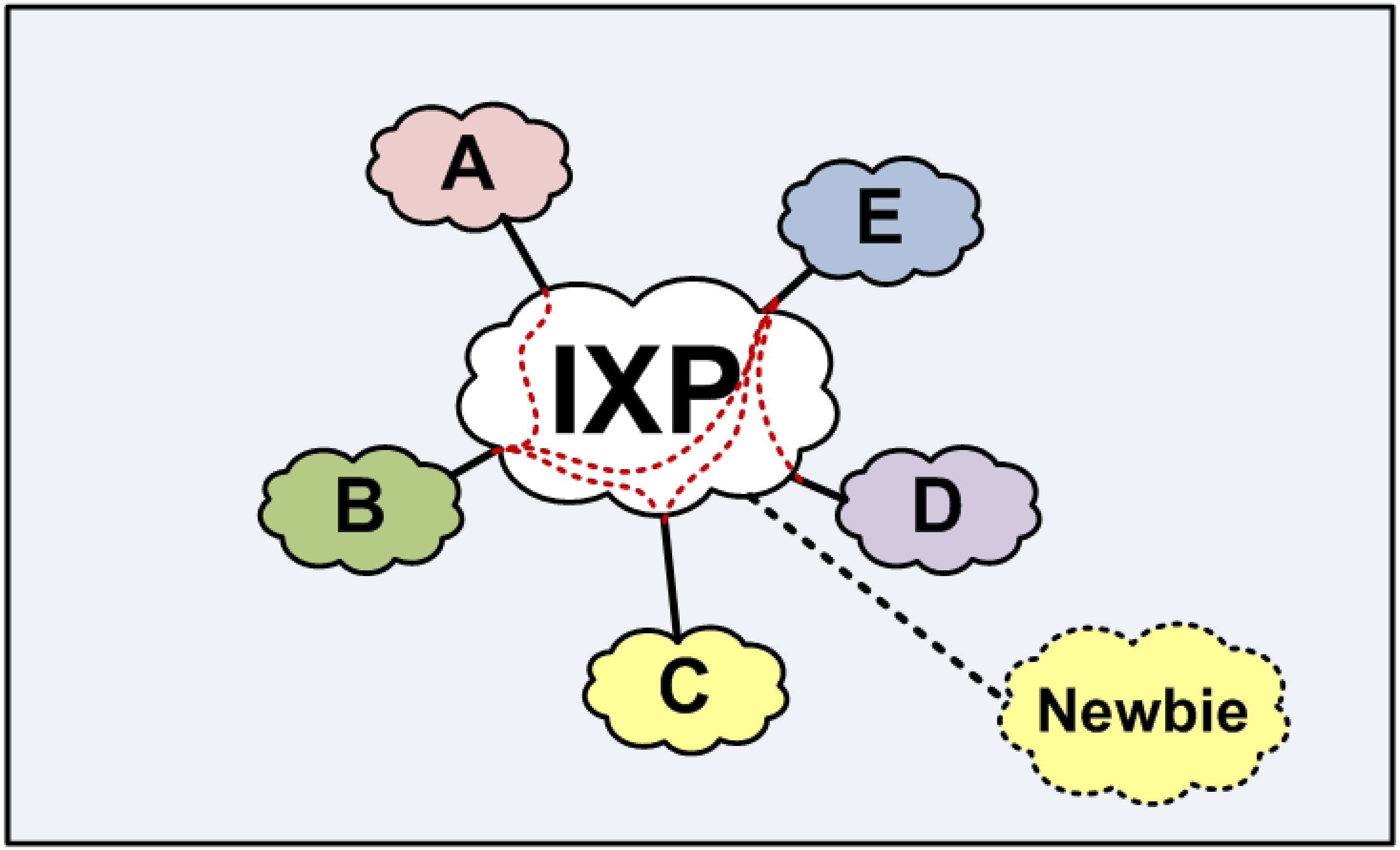


In the years Before Exchanges (BE)

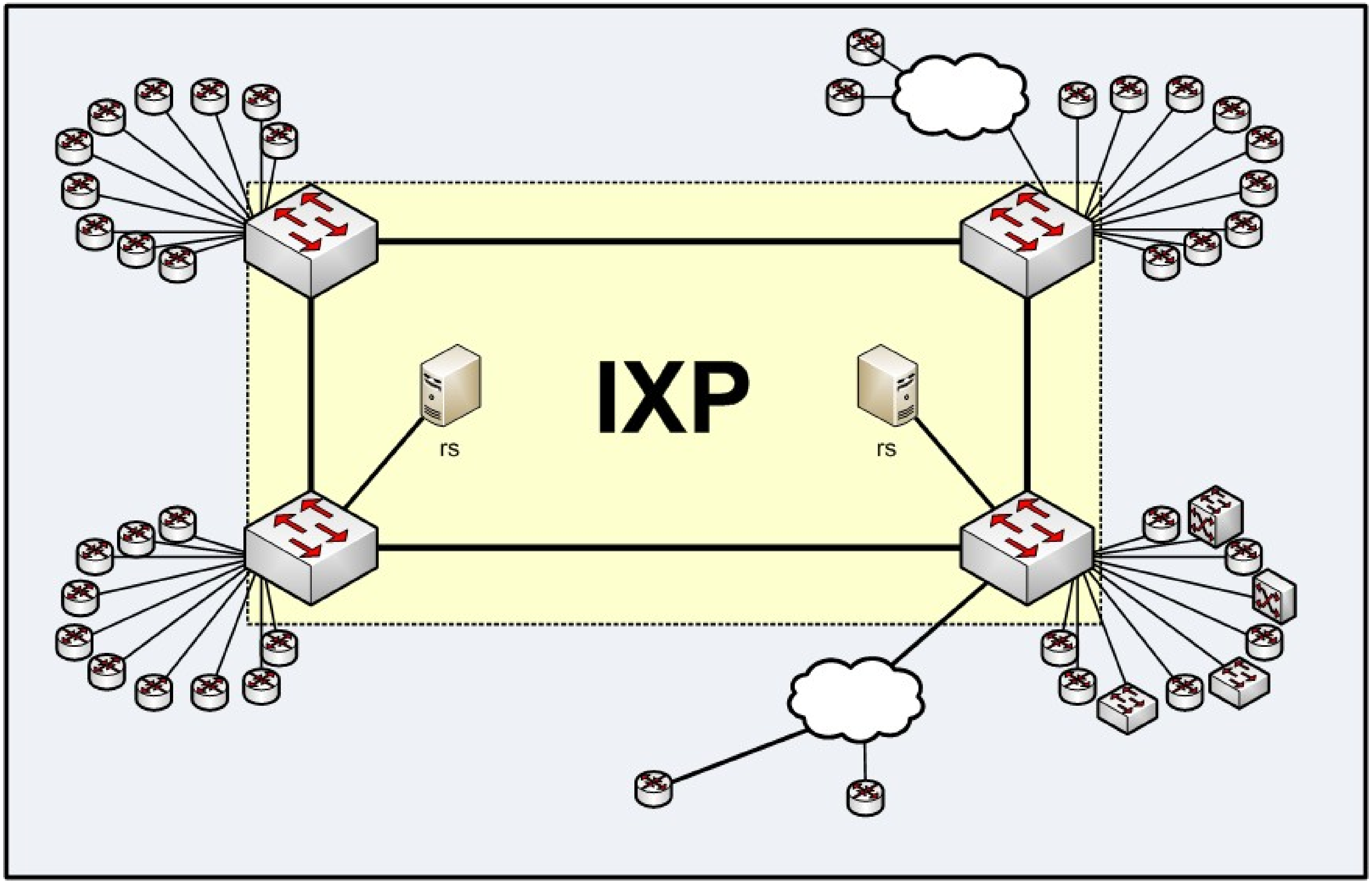
..enter the IXP...



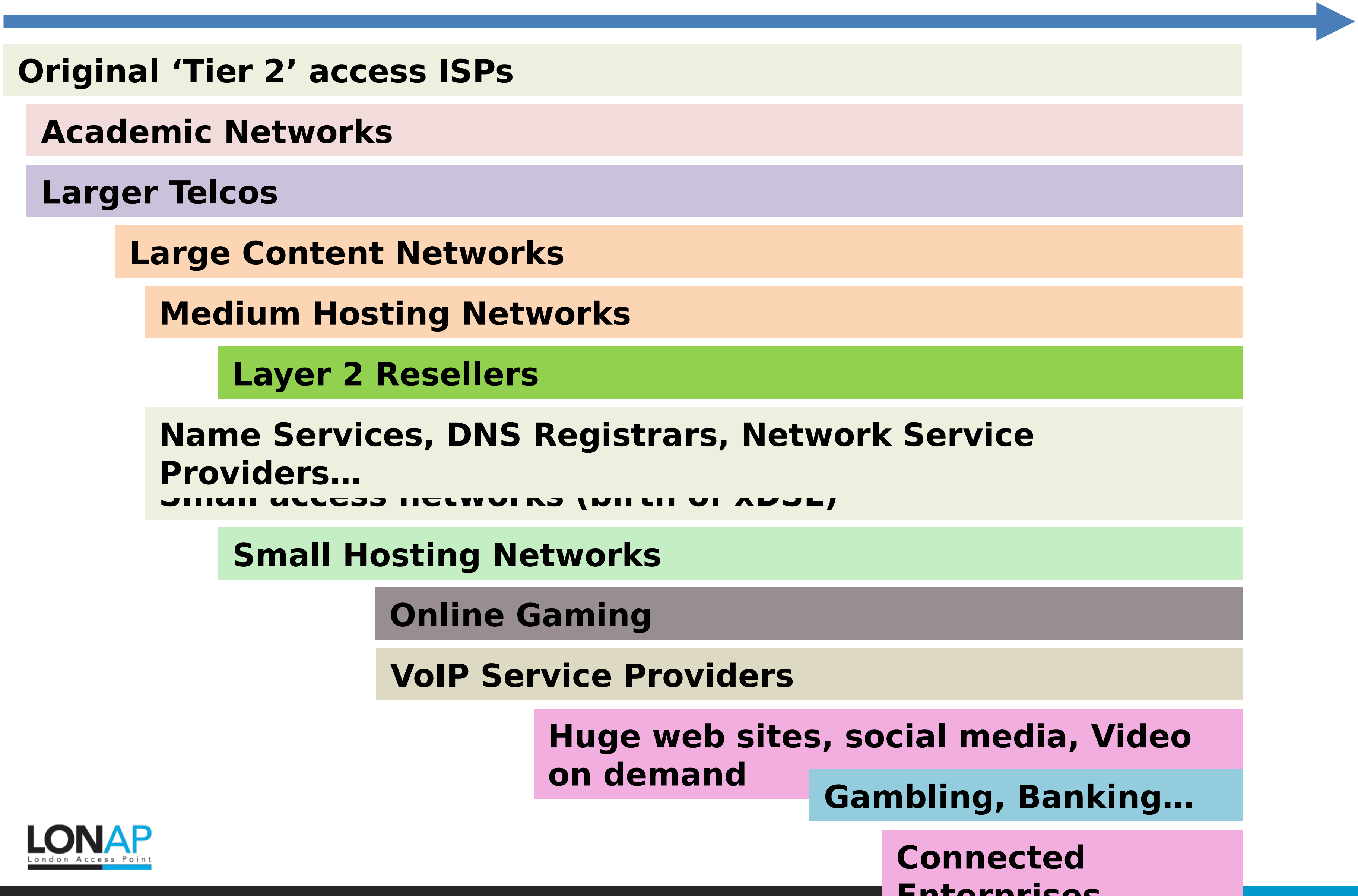
..enter the IXP…



The IXP today



From ISPs to… *$thing* provider



Original ‘Tier 2’ access ISPs

Academic Networks

Larger Telcos

Large Content Networks

Medium Hosting Networks

Layer 2 Resellers

Name Services, DNS Registrars, Network Service Providers…

Small access networks (birth of xDSL)

Small Hosting Networks

Online Gaming

VoIP Service Providers

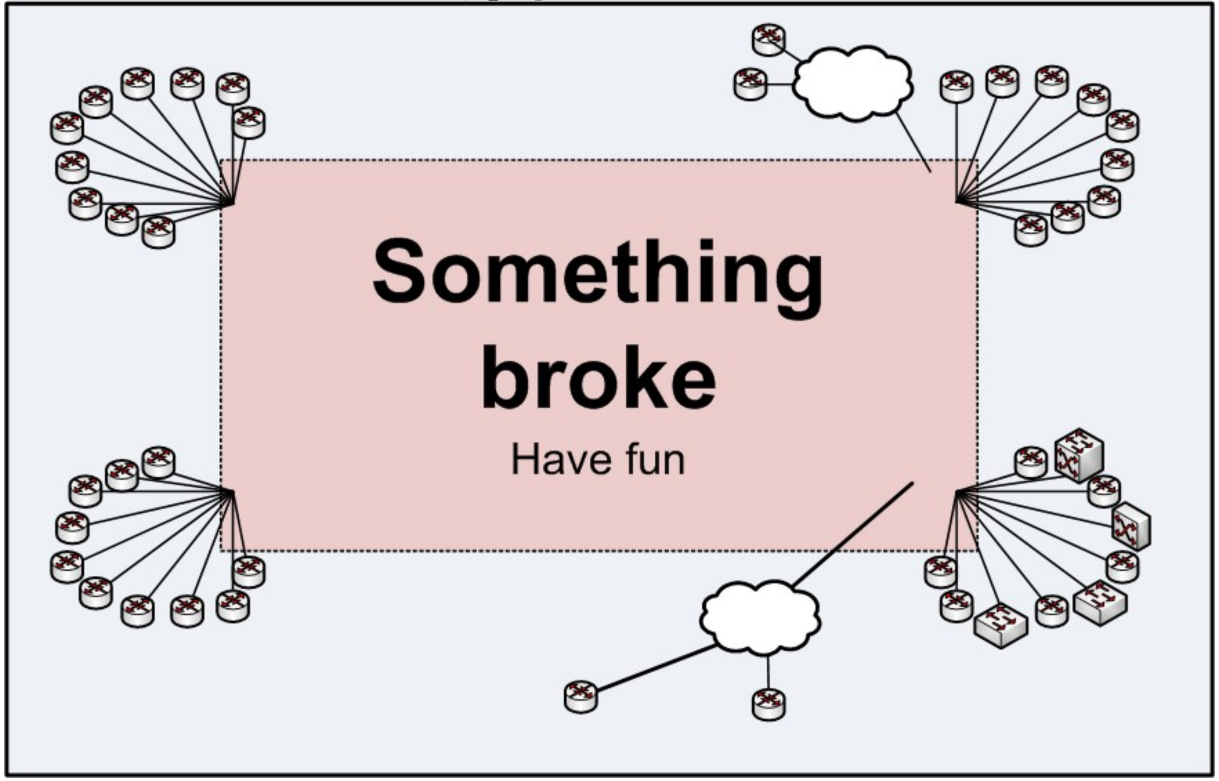
Huge web sites, social media, Video on demand

Gambling, Banking…

Connected Enterprises

IXP Protection

- "Layer-2 blob" works great most of the time!
- Until something / someone screws up
- Can



IXP Protection (1)

- IXPs have allowed traffic policies
  - We ask you nicely not to do evil things!
  - If we find evil things, we will unplug you…
  - “Router only” policy didn’t break much □

- This was enough for a long time….

IXP Protection (2)

- IXPs have allowed traffic policies
  - We ask you nicely not to do evil things!
  - If we find evil things, we will unplug you...
  - “Router only” policy didn’t break much □
- This was enough for a long time….
- Some evil happens too fast!

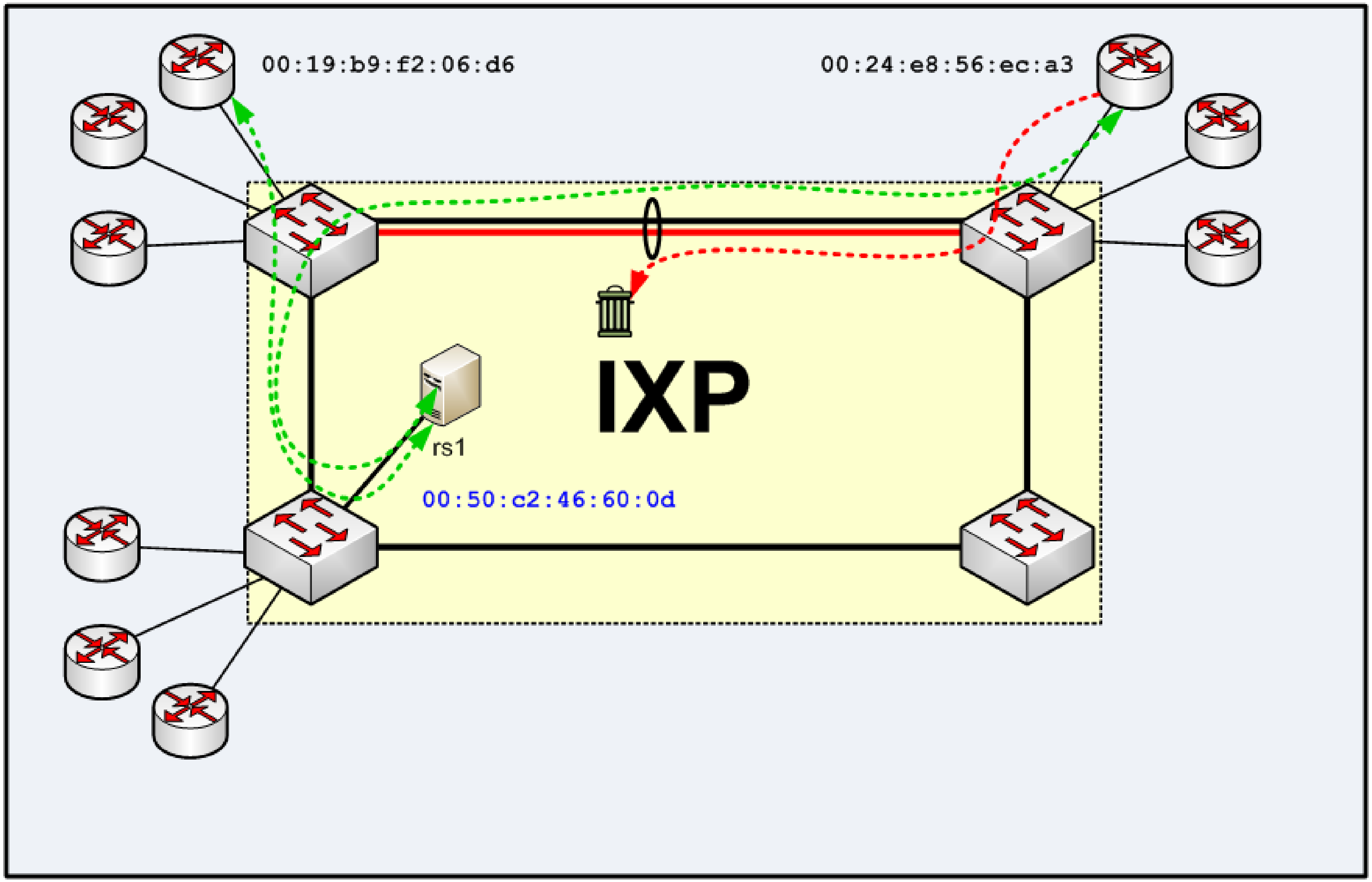
IXPs protecting against evil...

- IXPs enable various **port security** mechanisms
  - Limit to particular MAC
  - Restrict to 1 MAC address
  - Shut port down if > 1 MAC
  - (Hopefully) stop loops!
  - Limit ethertype?

- IXPs enable various **rate limiting** features
  - Limit broadcast traffic
  - Limit unknown unicast

- Quickly stops ***most*** (but not ***all)*** evil..

IXP Forwarding Path Failure



What goes wrong with members?

- “Magic” protocols like VTP/DTP/STP
- Proxy ARP / IPv6 ND etc.. DHCP…
- Internal routing issues
- BGP configuration
- Other configuration
- “Interesting” network designs…

DTP/VTP/STP (Cisco…)

- DTP/VTP: Automatic trunk configuration and VLAN distribution….
- STP: Fail-over/loop resolution
- Harmless… until another device starts sending you these frames.. then bad things happen!
  - STP topology change issues
  - Port shuts down due to VLAN config mismatch

Proxy ARP

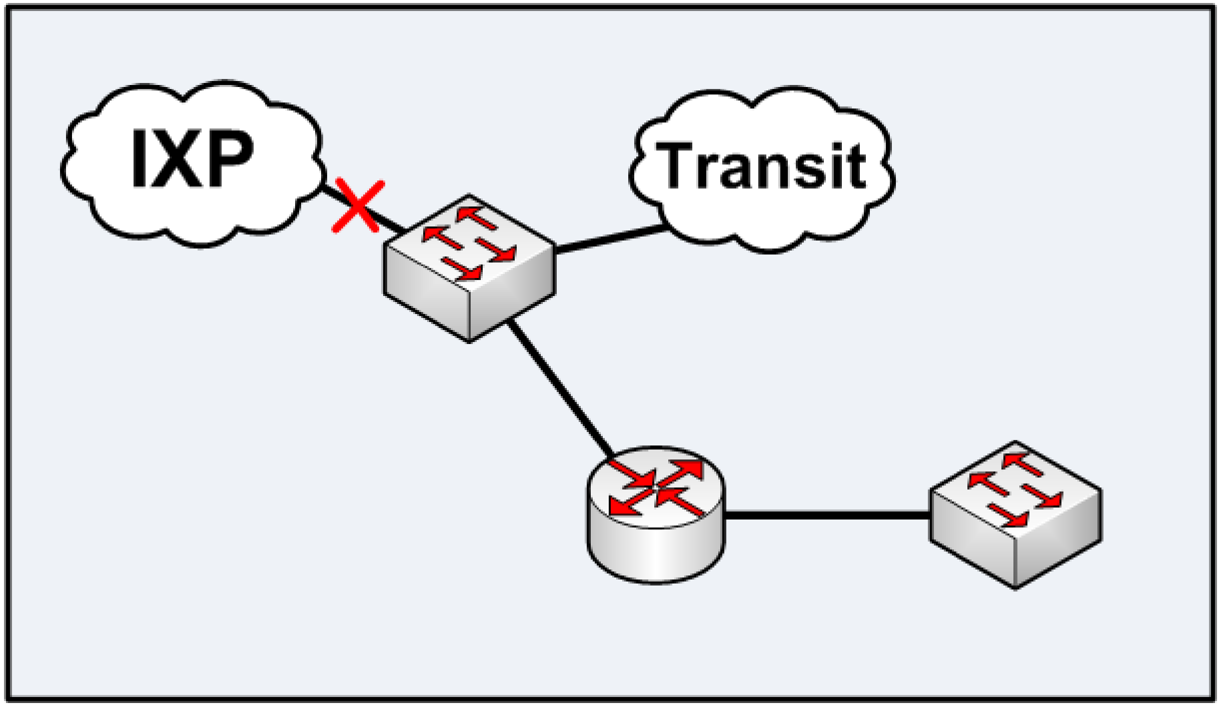
- Somebody starts responding to ARP that’s not for its interface’s IP address.
- Commonly caused by mask misconfiguration
- Cisco – often enabled **by default** on interfaces

How BGP detects deadness

- Next hop for a route must be in routing table
- When an interface goes down, BGP tears down all (eBGP) sessions reached via that interface
- BGP sends keepalives to peers every 30 seconds
- When3 keepalives are not received, the BGP session is torn down.

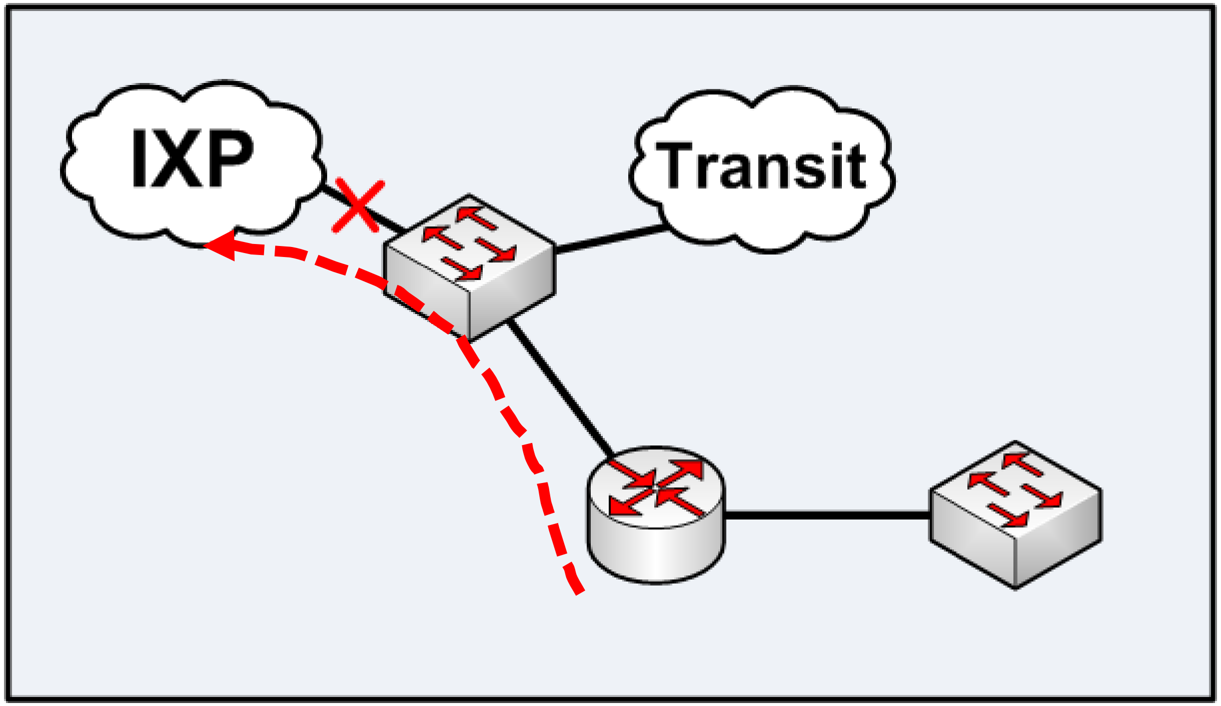
Intermediate Switches / BGP

- Member connects a switch to the IXP
- Not many good mitigations…



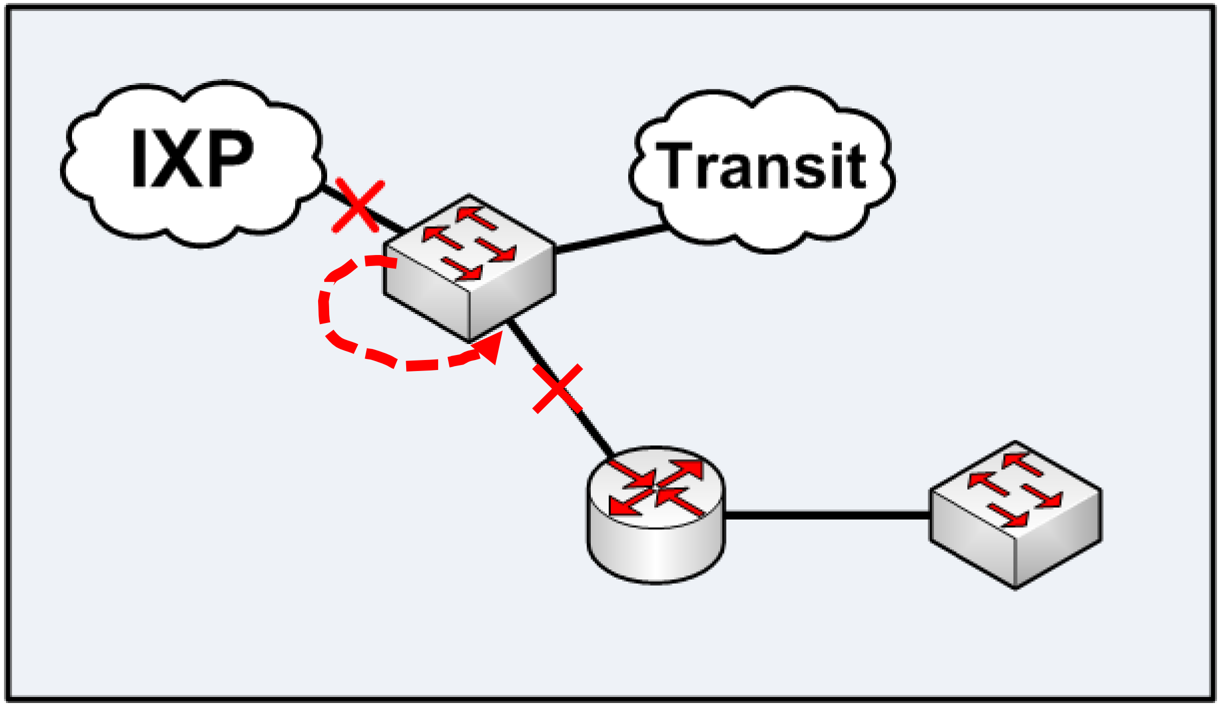
‘SLA’ Features

- Pings some external IP and shuts down interface or withdraw route if unpingable
- Pick the destination carefully…
- Maybe not much faster than BGP timeout… ▯



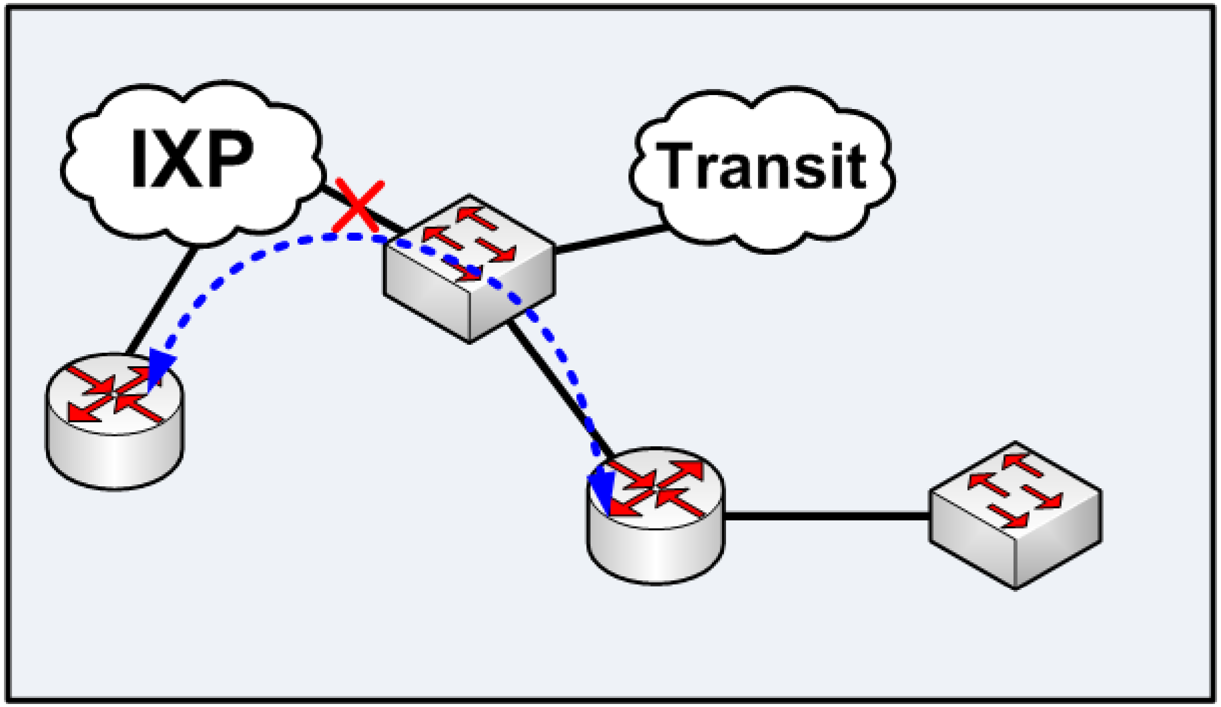
Switch Features

- “Uplink group”/LFS type feature shuts down ports when another port goes down…
- Not practical for tagged links

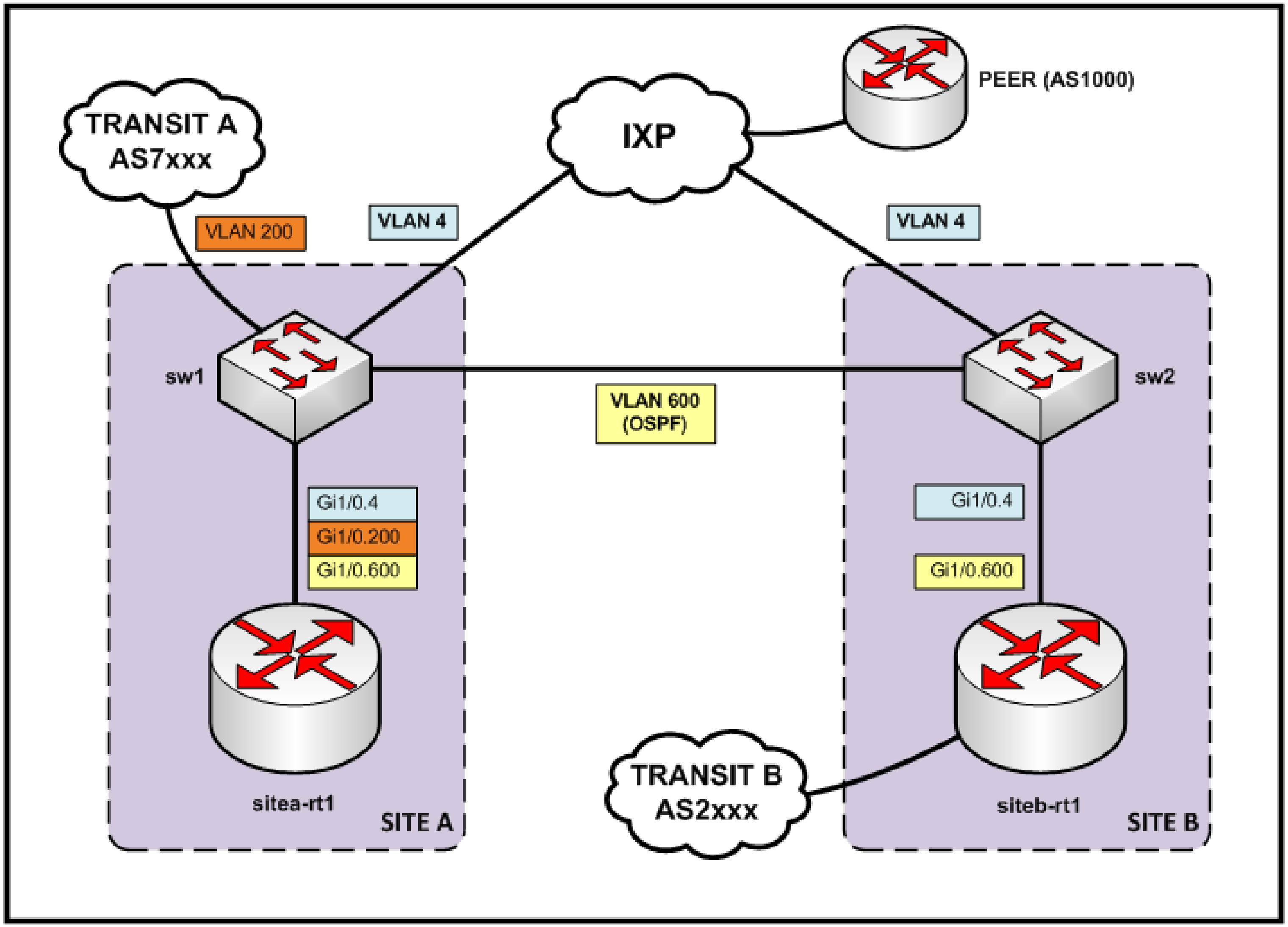


BFD - Bidirectional Forwarding Detection

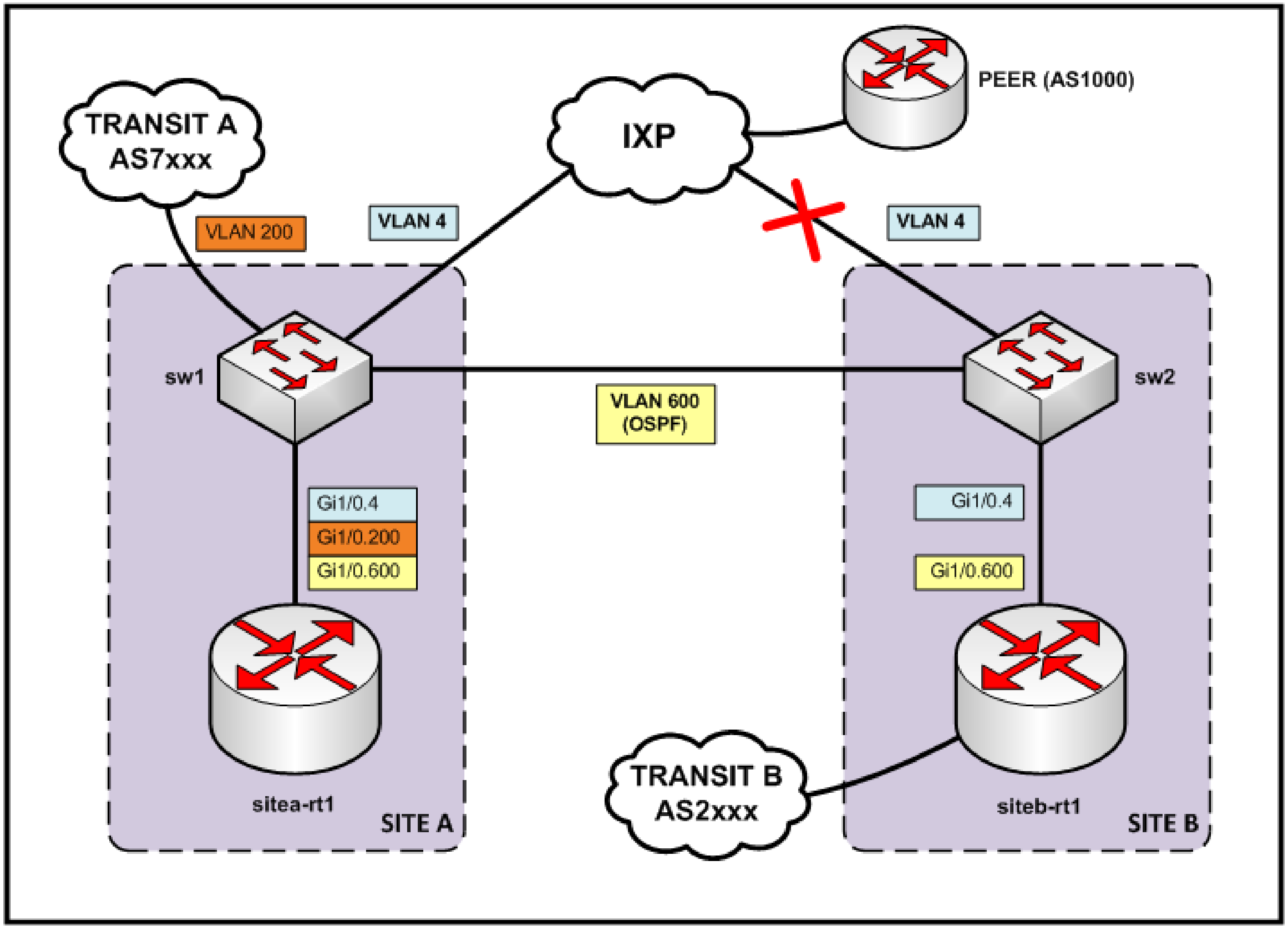
- [RFC 5881] – BFD. Detects failures in the forwarding path between routers
- Good – not widely used inter-AS (yet)



Multiple IXP Connections



Multiple IXP Connections



Multiple IXP Connections

First, verify connectivity to an IP in our peer’s network:

```
sitea-rt1>ping 10.10.0.1
Type escape sequence to abort.
Sending 5, 100-byte ICMP Echos to 10.10.0.1, timeout is 2 seconds:
!!!!!
Success rate is 100 percent (5/5), round-trip min/avg/max = 8/15/20 ms
```

Then, show the route going out via the directly connected interface to LONAP:

```
sitea-rt1>sh ip bgp 10.10.0.0/16 subnets
BGP table version is 899, local router ID is 192.168.20.1
Status codes: s suppressed, d damped, h history, * valid, > best, i - internal,
              r RIB-failure, S Stale
Origin codes: i - IGP, e - EGP, ? - incomplete

   Network          Next Hop            Metric LocPrf Weight Path
*> 10.10.0.0/16     193.203.5.10             0             0 1000 i
* i                 193.203.5.10             0    100      0 1000 i
```

Multiple IXP Connections

Now we shut down the interface (or the BGP goes down somehow…)

```
sitea-rt1#conf t
Enter configuration commands, one per line.  End with CNTL/Z.
sitea-rt1(config)#int Gi1/0.4
sitea-rt1(config-subif)#shut
Mar 12 20:34:40.186: %BGP-5-ADJCHANGE: neighbor 193.203.5.10 Down .. .
```

We see that we have one remaining route via iBGP:

```
sitea-rt1>sh ip bgp 10.10.0.0/16 subnets
BGP table version is 901, local router ID is 192.168.20.1
Status codes: s suppressed, d damped, h history, * valid, > best, i - internal,
              r RIB-failure, S Stale
Origin codes: i - IGP, e - EGP, ? - incomplete

   Network          Next Hop            Metric LocPrf Weight Path
*>i10.10.0.0/16     193.203.5.10             0    100      0 1000 i
```

Multiple IXP Connections

However, when we try to ping 10.10.0.1 again, it doesn’t work…

```
sitea-rt1>ping 10.10.0.1
Type escape sequence to abort.
Sending 5, 100-byte ICMP Echos to 10.10.0.1, timeout is 2 seconds:
.....
Success rate is 0 percent (0/5)
```

Multiple IXP Connections

Verify how we reach the next hop address:

```
sitea-rt1>sh ip route 193.203.5.10
Routing entry for 193.203.5.0/24
  Known via "bgp 65009", distance 20, metric 0
  Tag 7xxx, type external
  Last update from 203.0.113.14 00:03:31 ago
  Routing Descriptor Blocks:
  * 203.0.113.14, from 203.0.113.14, 00:03:31 ago
      Route metric is 0, traffic share count is 1
      AS Hops 1
      Route tag 7xxx
```

```
sitea-rt1>sh ip bgp 193.203.5.10
BGP routing table entry for 193.203.5.0/24, version 901
Paths: (1 available, best #1, table default)
  Advertised to update-groups:
     1          2
  7xxx
    203.0.113.14 from 203.0.113.14 (203.0.113.1)
      Origin incomplete, metric 0, localpref 100, valid, external, best
```

Routing Fun…

- Default Administrative Distance for eBGP
- Appropriate in service provider netw

Why is this?

Protocol	Administrative Distance
Directly Connected	0
Static Route	1
External BGP (eBGP)	**20**
OSPF	110
IS-IS	115
RIP	120
Internal BGP (iBGP)	200

Multiple IXP Connections

One way…

```
sitea-rt1#conf t
Enter configuration commands, one per line.  End with CNTL/Z.
sitea-rt1(config)#router bgp 65009
sitea-rt1(config-router)#distance bgp 150 200 200

sitea-rt1#clear ip bgp 203.0.113.14
Mar 12 20:48:28.122: %BGP-5-ADJCHANGE: neighbor 203.0.113.14 Down User reset
Mar 12 20:48:36.870: %BGP-5-ADJCHANGE: neighbor 203.0.113.14 Up
```

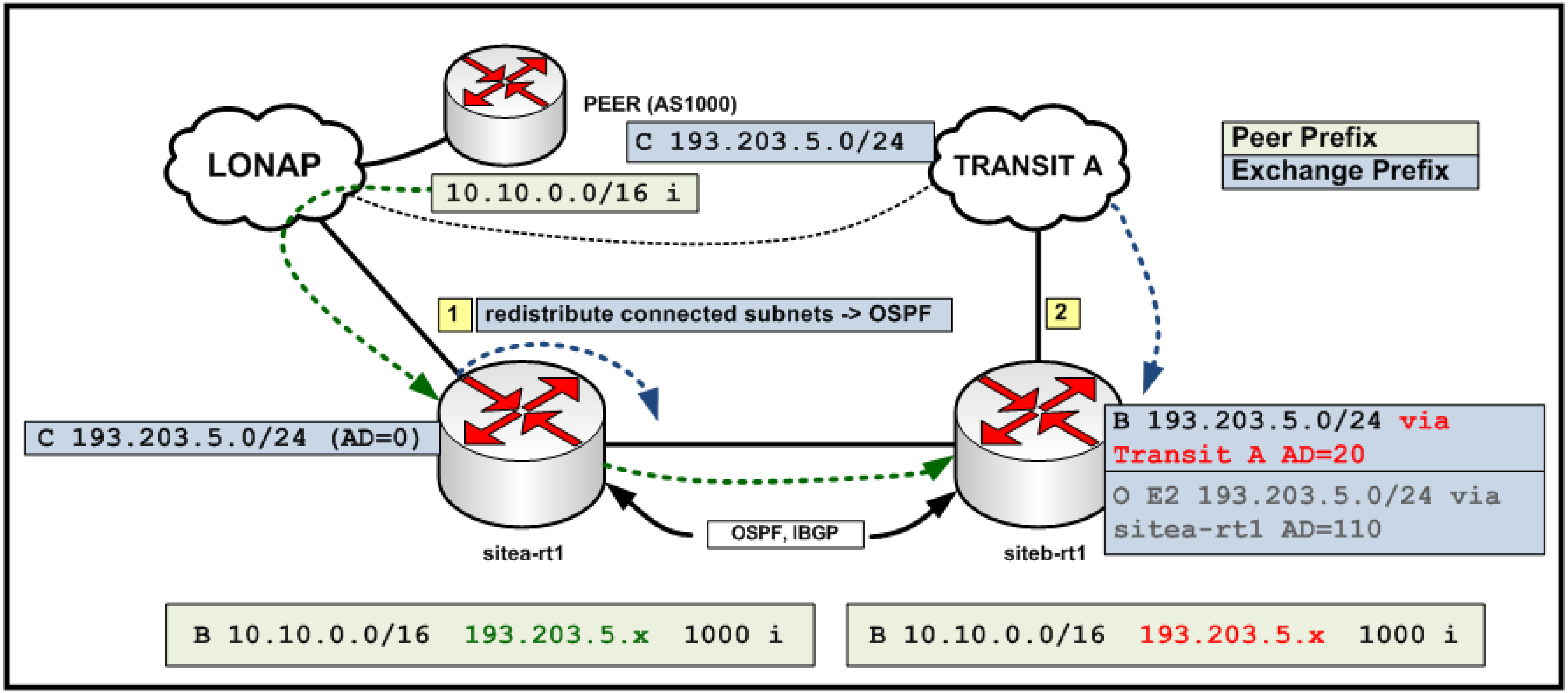
```
sitea-rt1>sh ip route 193.203.5.10
Routing entry for 193.203.5.0/24
  Known via "ospf 1", distance 110, metric 20, type extern 2, forward metric 1
  Last update from 172.16.1.2 on GigabitEthernet1/0.600, 00:00:05 ago
  Routing Descriptor Blocks:
  * 172.16.1.2, from 192.168.20.2, 00:00:05 ago, via GigabitEthernet1/0.600
      Route metric is 20, traffic share count is 1
```

```
sitea-rt1>ping 10.10.0.1
Type escape sequence to abort.
Sending 5, 100-byte ICMP Echos to 10.10.0.1, timeout is 2 seconds:
!!!!!
Success rate is 100 percent (5/5), round-trip min/avg/max = 8/32/56 ms
```

Fixes for the blackhole

- We could change the odd eBGP Admin Distance
- And/or **filter out connected IXP prefixes**
- Fiddle with next-hop-self? Maybe…

Routing Fun…


PEER (AS1000)
C 193.203.5.0/24
LONAP
10.10.0.0/16 i
TRANSIT A
Peer Prefix
Exchange Prefix
1
redistribute connected subnets -> OSPF
2
C 193.203.5.0/24 (AD=0)
B 193.203.5.0/24 via Transit A AD=20
O E2 193.203.5.0/24 via sitea-rt1 AD=110
OSPF, IBGP
sitea-rt1
siteb-rt1
B 10.10.0.0/16 193.203.5.x 1000 i
B 10.10.0.0/16 193.203.5.x 1000 i

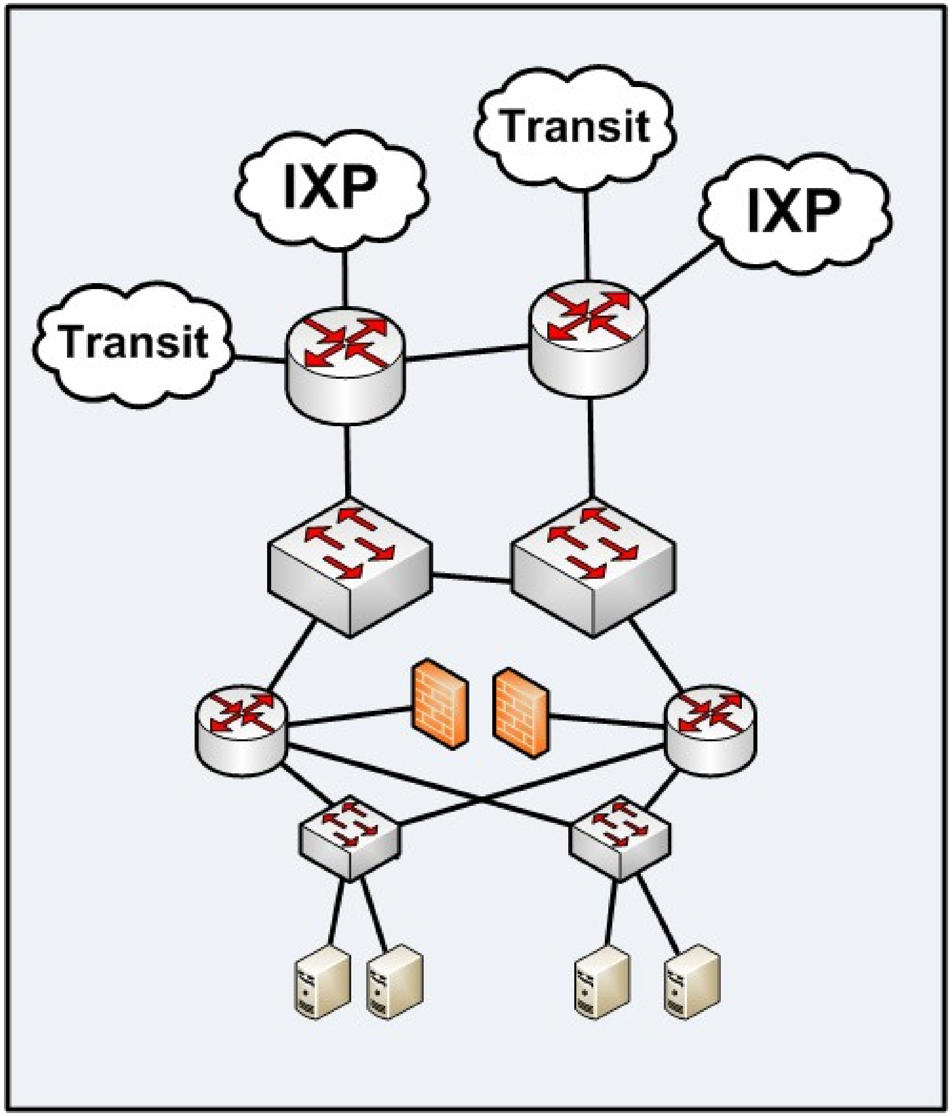
Possible solutions…

- **Filter out IXP prefixes and more specifics**
- **Tweak Administrative Distance…**

Network Design

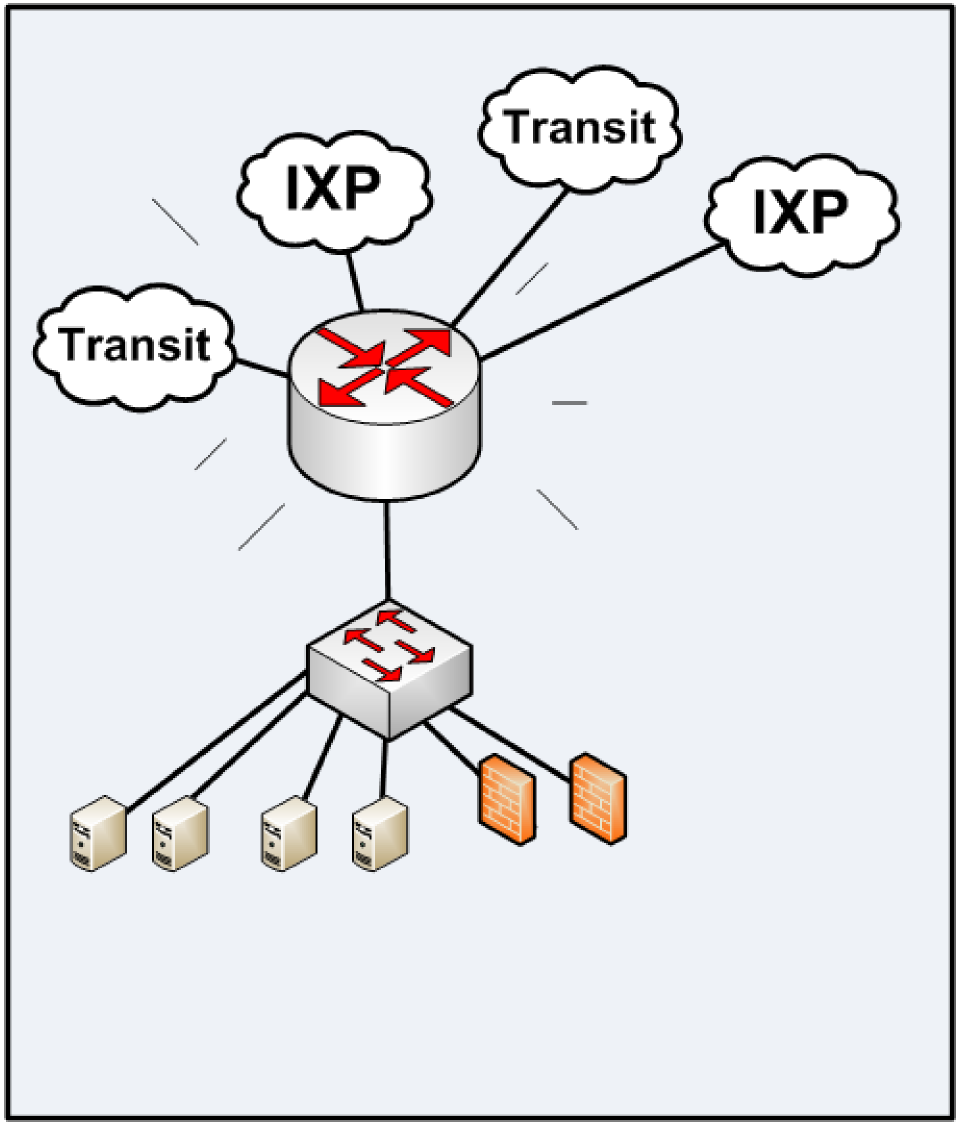
- Some thought
- Layered design
- Redundancy
- Failover

- There are others!



Network Design

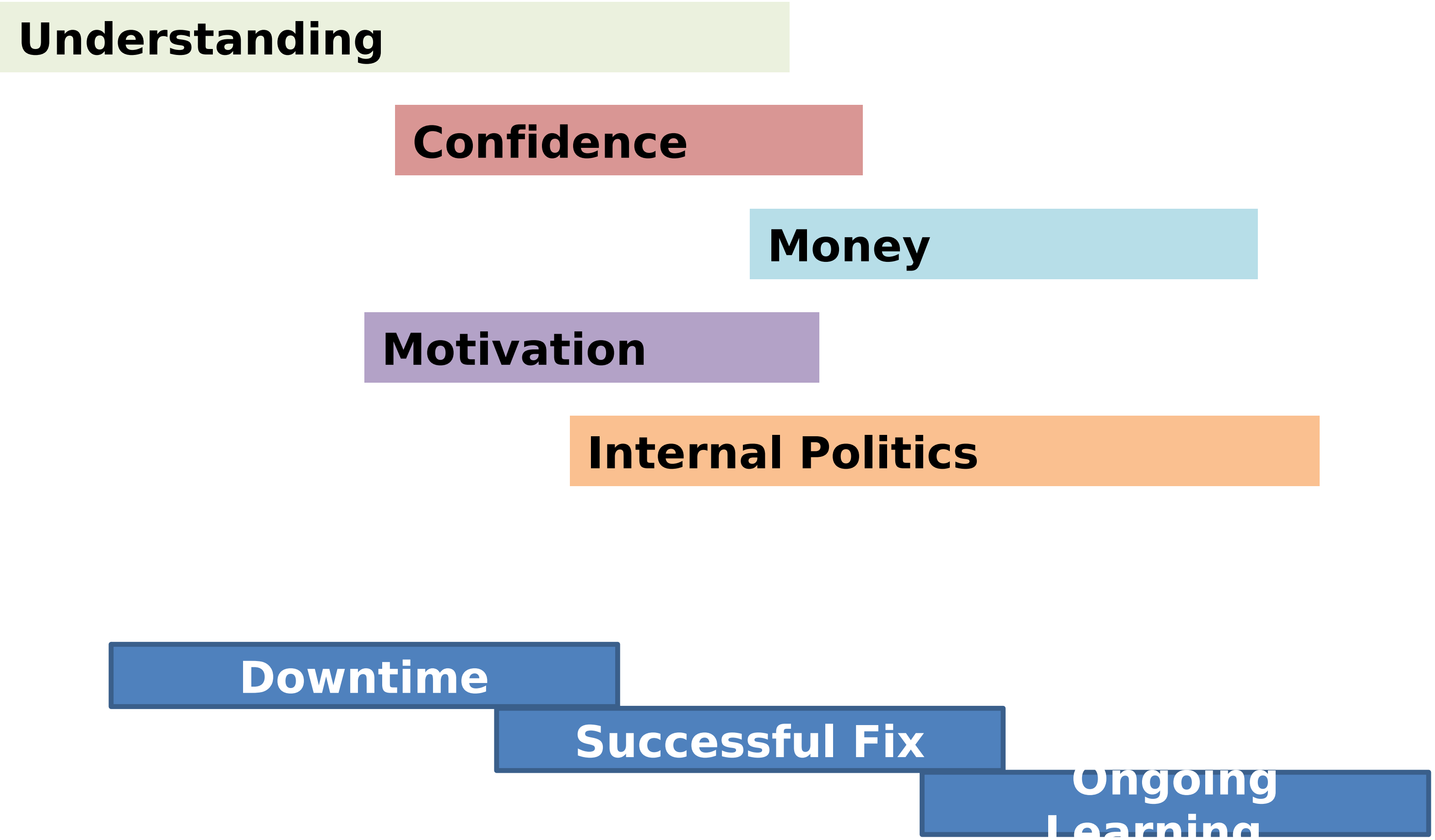
- **BIG SCARY BOX**
- "We paid $$$ for it"
- "Temporary"
- Use it for *everything*
- Never touch it
- Never document it
- Run away!!!



So..

- Do these things get fixed?

It depends...



Understanding

Confidence

Money

Motivation

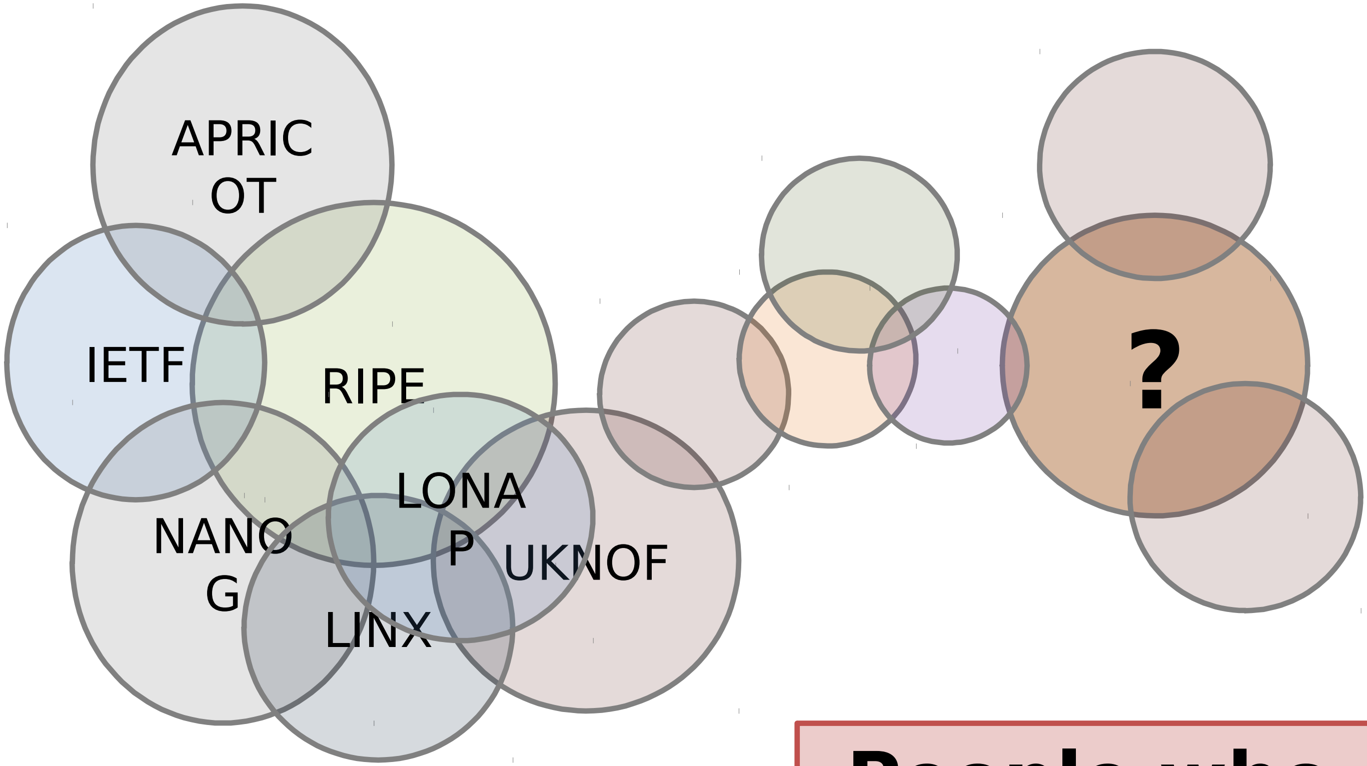
Internal Politics

Downtime

Successful Fix

Ongoing Learning

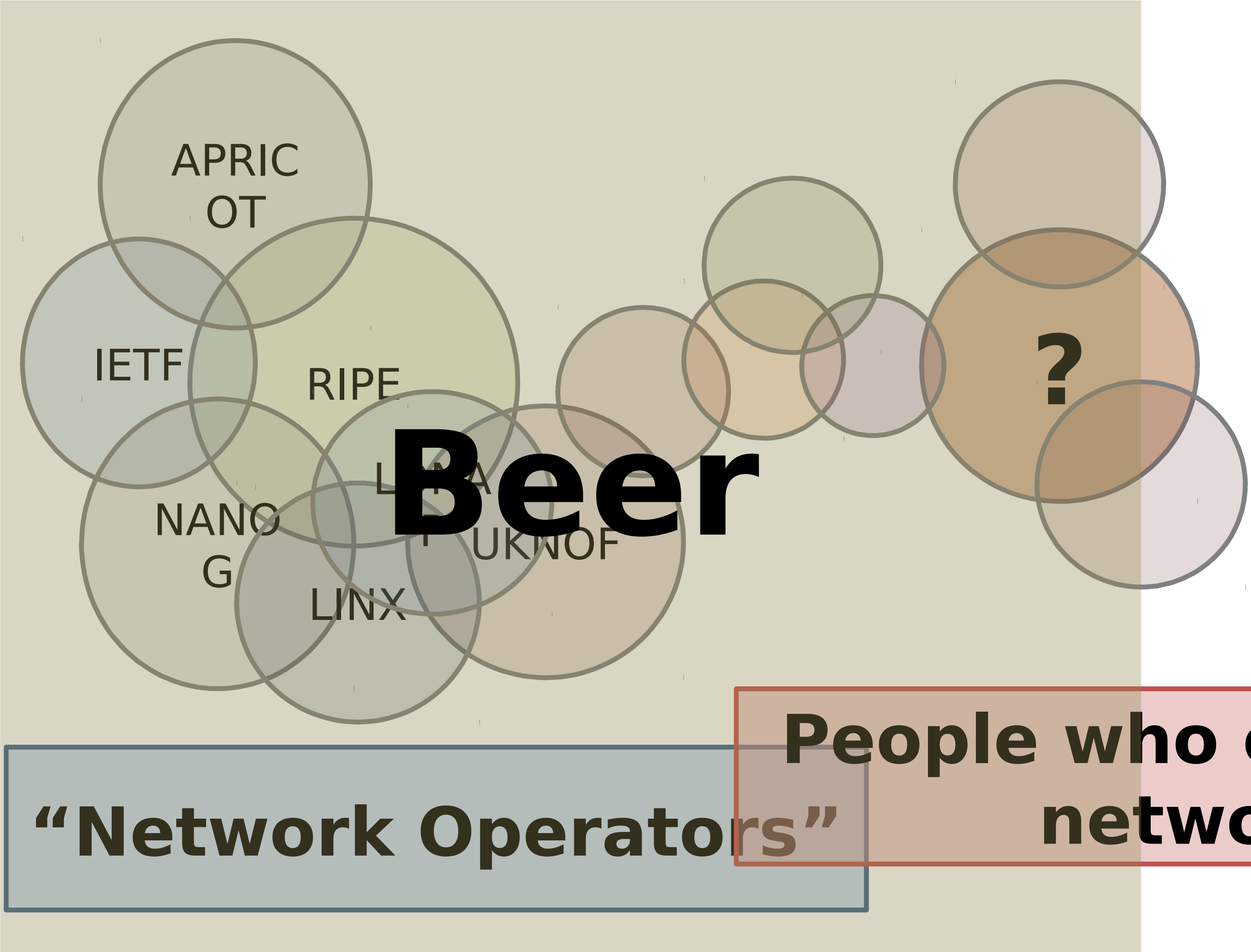
It’s a community…



“Network Operators”

People who operate a network

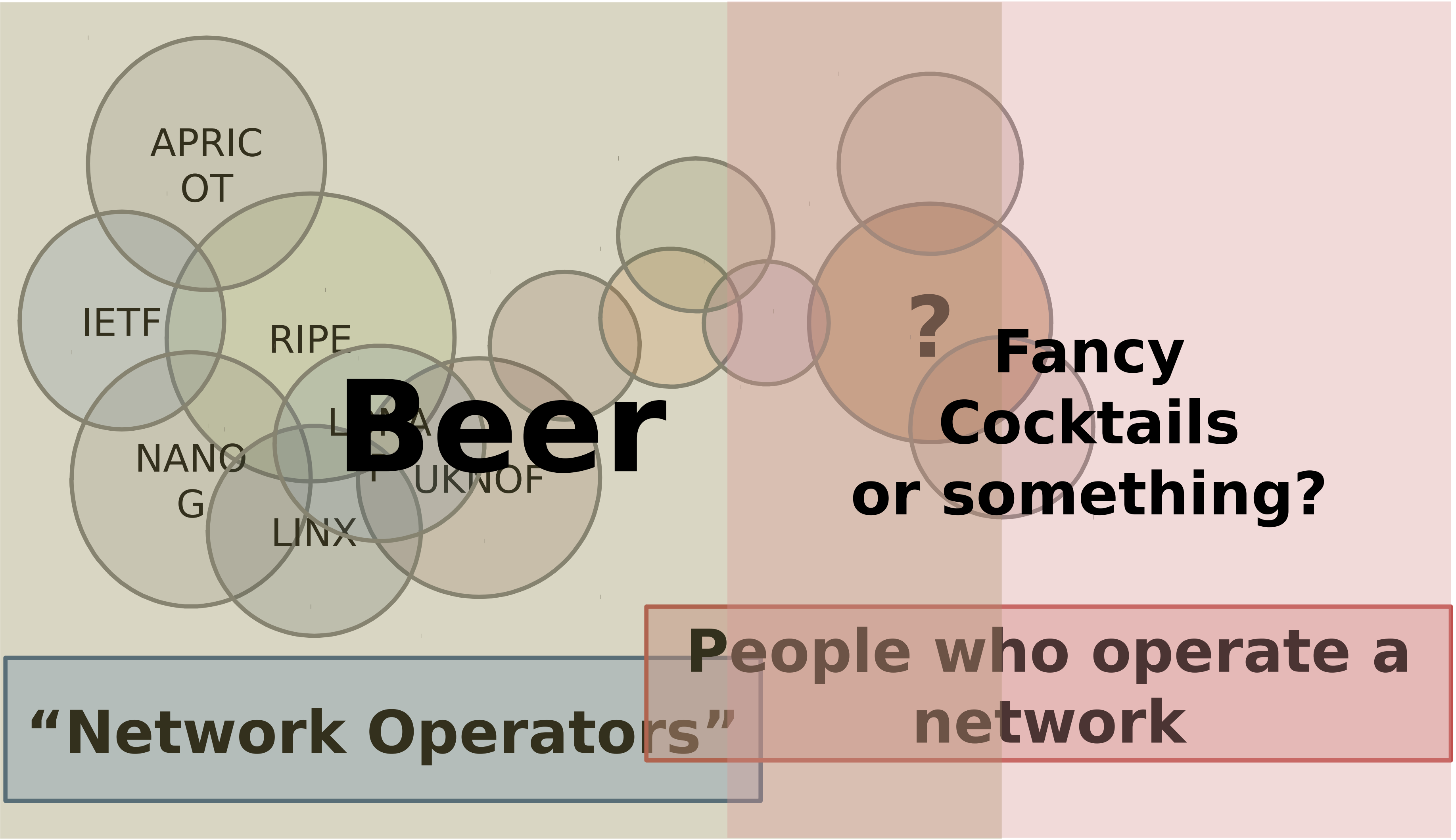
It’s a community…



“Network Operators”

People who operate a network

It’s a community…



Beer

Fancy Cocktails or something?

People who operate a network

“Network Operators”