IDR WG
IETF 102

Susan Hares and John Scudder
IDR Co-chairs
Jie Dong (WG Secretary)

Note Well

•This is a reminder of IETF policies in effect on various topics such as patents or code of conduct. It is only meant to point you in the right direction. Exceptions may apply. The IETF's patent policy and the definition of an IETF "contribution" and "participation" are set forth in BCP 79; please read it carefully.

•As a reminder:

- By participating in the IETF, you agree to follow IETF processes and policies.
- If you are aware that any IETF contribution is covered by patents or patent applications that are owned or controlled by you or your sponsor, you must disclose that fact, or not participate in the discussion.
- As a participant in or attendee to any IETF activity you acknowledge that written, audio, video, and photographic records of meetings may be made public.
- Personal information that you provide to IETF will be handled in accordance with the IETF Privacy Statement.
- As a participant or attendee, you agree to work respectfully with other participants; please contact the ombudsteam (https://www.ietf.org/contact/ombudsteam/) if you have questions or concerns about this.

•Definitive information is in the documents listed below and other IETF BCPs. For advice, please talk to WG chairs or ADs:

- BCP 9 (Internet Standards Process)
- BCP 25 (Working Group processes)
- BCP 25 (Anti-Harassment Procedures)
- BCP 54 (Code of Conduct)
- BCP 78 (Copyright)
- BCP 79 (Patents, Participation)
- https://www.ietf.org/privacy-policy/ (Privacy Policy)

Thursday (18:10-19:10pm)

0) Agenda bashing (5)

1) Update on merger of RLP and eOTC drafts for route leaks solution [Kotikalapudi Sriram] (8)
draft-ietf-idr-route-leak-detection-mitigation (solution)/
draft-sriram-idr-route-leak-solution-discussion-00 (design discussion)

2) BGP Model for Service Provider Networks [Keyur Patel] (8)
draft-ietf-idr-bgp-model/

3) BGP Extra Extended Community [Jakob Heitz] (8)
draft-heitz-idr-extra-extended-community/

4) BGP Neighbor Autodiscovery [Ketan Talaulikar] (8)
https://tools.ietf.org/html/draft-xu-idr-neighbor-autodiscovery/

5) Requirements for BGP Neighbor Autodiscovery (15)
[Randy Bush] provides LSVR [8]
Discussion [7]

Route leaks solution

- Open questions about semantics, syntax
- Semantics - Space/Information Trade-off
  - Design A – Design option A
  - Design B – Design Option B
  - Sriram’s talk. Chairs believe we are close.
- Syntax – Attribute or Community?
  - Option 1: Proceed with Attribute Approach
  - Option 2: Use (Large) Community Approach
  - Needs development.

Autodiscovery

- Multiple proposals in multiple groups
  - Overlapping functionality
  - Ranging from minimal to maximal
- Clear to chairs that
  - The WG has great interest in the topic
  - There is no consensus on the requirements
- Ideally chairs would have prepared a full comparison of all proposals
  - It’s an imperfect world, we’re going with what we have today
  - Possible interim

BGP Data Model

- NMDA is requirement for all new Models
  - draft-idr-bgp-model-03.txt – is NMDA
  - Replaces the old model
- Going to WG LC at end of today's meeting
- Original draft
  - Authors may publish as historical work product, but little interest

Session II: Friday, 11:50-13:20, 7/20/2018

- 0) Agenda bashing and Chair's slides (10)

- 1) LOCAL_PREF Overloaded = Overwritten [Alexander Azimov] (5)

  2) Updates to BGP Signaled SR Policies [Dhanendra Jain] (8)
  draft-ietf-idr-segment-routing-te-policy

- 3) YANG data model for BGP Segment Routing Extensions [Dhanendra Jain] (8)
  draft-dhjain-spring-bgp-sr-yang

  4) BGP-LS Extend for Inter-AS Topology Retrieval [Aijun Wang] (10)
  draft-wang-idr-bgpls-inter-as-topology-ext

  5) Distribution of Traffic Engineering (TE) Policies and State using BGP-LS [Ketan Talaulikar] (10)
  draft-ietf-idr-te-lsp-distribution

Session II: Friday, 11:50-13:20, 7/20/2018

6) Flexible Algorithm Definition Advertisement with BGP Link-State
[Ketan Talaulikar] (5)
draft-ketant-idr-bgp-ls-flex-algo

BGP Link-State Extensions for Seamless BFD [Ketan Talaulikar]
draft-li-idr-bgp-ls-sbfd-extensions/

7) Applying BGP flowspec rules on a specific interface set [Jeff Haas] (5)
draft-ietf-idr-flowspec-interfaceset/

8) Segment Routing Policies for Path Segment and Bi-directional Path [Cheng Li] (15)
draft-li-idr-sr-policy-path-segment-distribution/

SR Policies for Path Segment and Bi-directional Path in BGP-LS [Cheng Li]
draft-li-idr-bgp-ls-sr-policy-path-segment/

9) BGP-LS Extensions for Advertising Path MTU [Zhibo Hu] (10)
draft-zhu-idr-bgp-ls-path-mtu/

See you at IETF 103

Route Leaks Solution

Merger of RLP and eOTC Drafts

ietf-idr-route-leak-detection-mitigation-09

K. Sriram (Ed.), A. Azimov (Ed.), D. Montgomery, B. Dickson, K. Patel, A. Robachevsky, E. Bogomazov, and R. Bush

IDR Working Group Meeting, IETF-102
July 2018

Acknowledgements: The authors are grateful to many folks in various IETF WGs for commenting, critiquing, and offering very helpful suggestions (see acknowledgements section in the draft.)

Draft Merger Efforts

- Authors from the two drafts met in Chicago (March 2017) and in London (March 2018)
- Support and encouragement from IDR Chairs John and Sue, and Ignas
- Productive authors' meeting in London (IETF 101) followed by substantial discussions via email
- Authors happy to report on convergence to a merged solution and draft

Merged Solution and Design Discussion Drafts

- Merged Solution: https://tools.ietf.org/html/draft-ietf-idr-route-leak-detection-mitigation-09

- Design Discussion: https://tools.ietf.org/html/draft-sriram-idr-route-leak-solution-discussion-00

Format of RLP Attribute

Optional Transitive Attribute

Design A (original RLP)

ASN: N
RLP: N

Most Recently Added

⋮

ASN: 2
RLP: 2
ASN: 1
RLP: 1

Least Recently Added

UP: RLP = 0 DOWN/LATERAL: RLP = 1

Design B

ASN: N

Most Recently Added

⋮

ASN: 2
ASN: 1

Least Recently Added

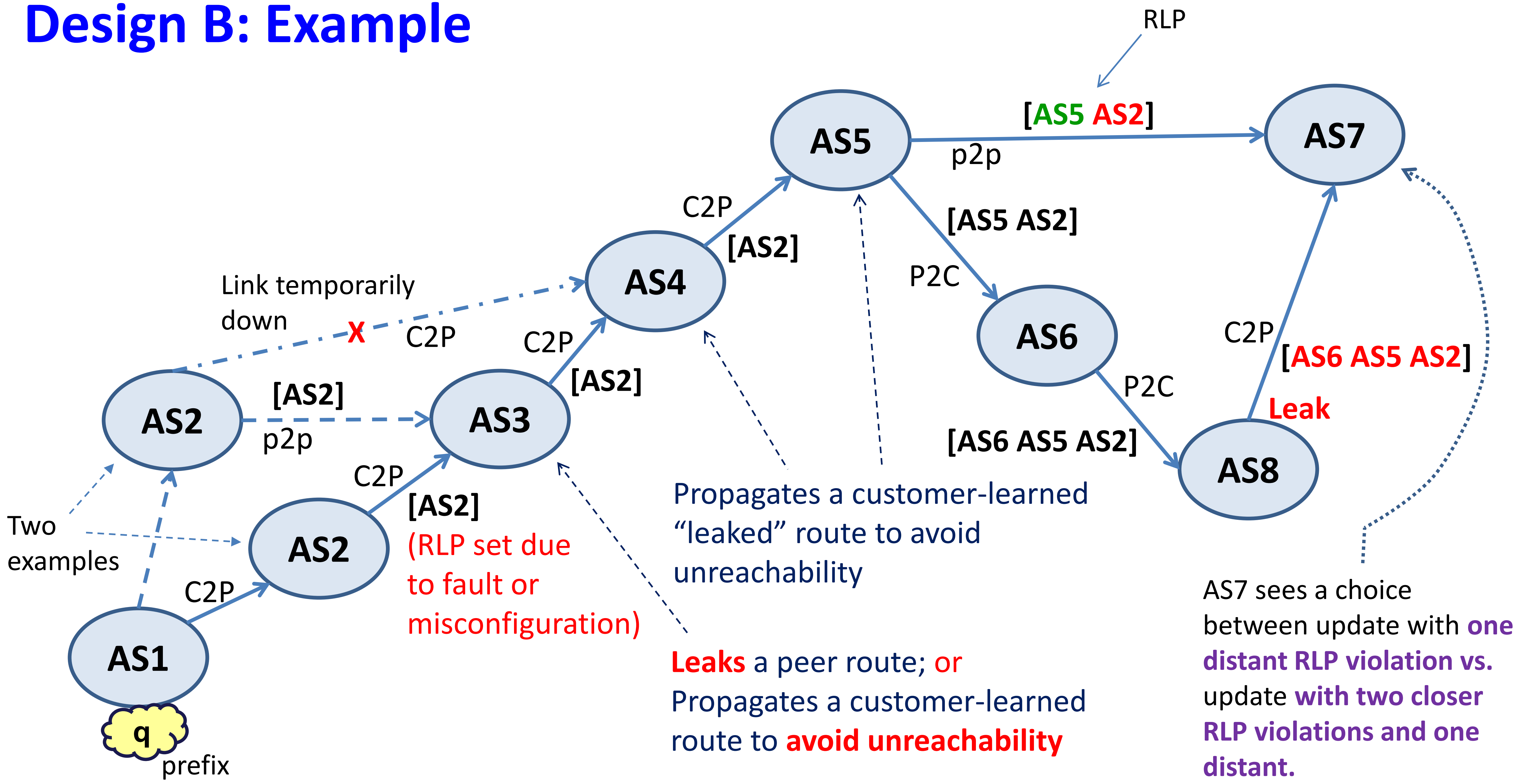
- **eOTC: Design B with only one ASN in the attribute is the original eOTC**

Comparison / Tradeoffs

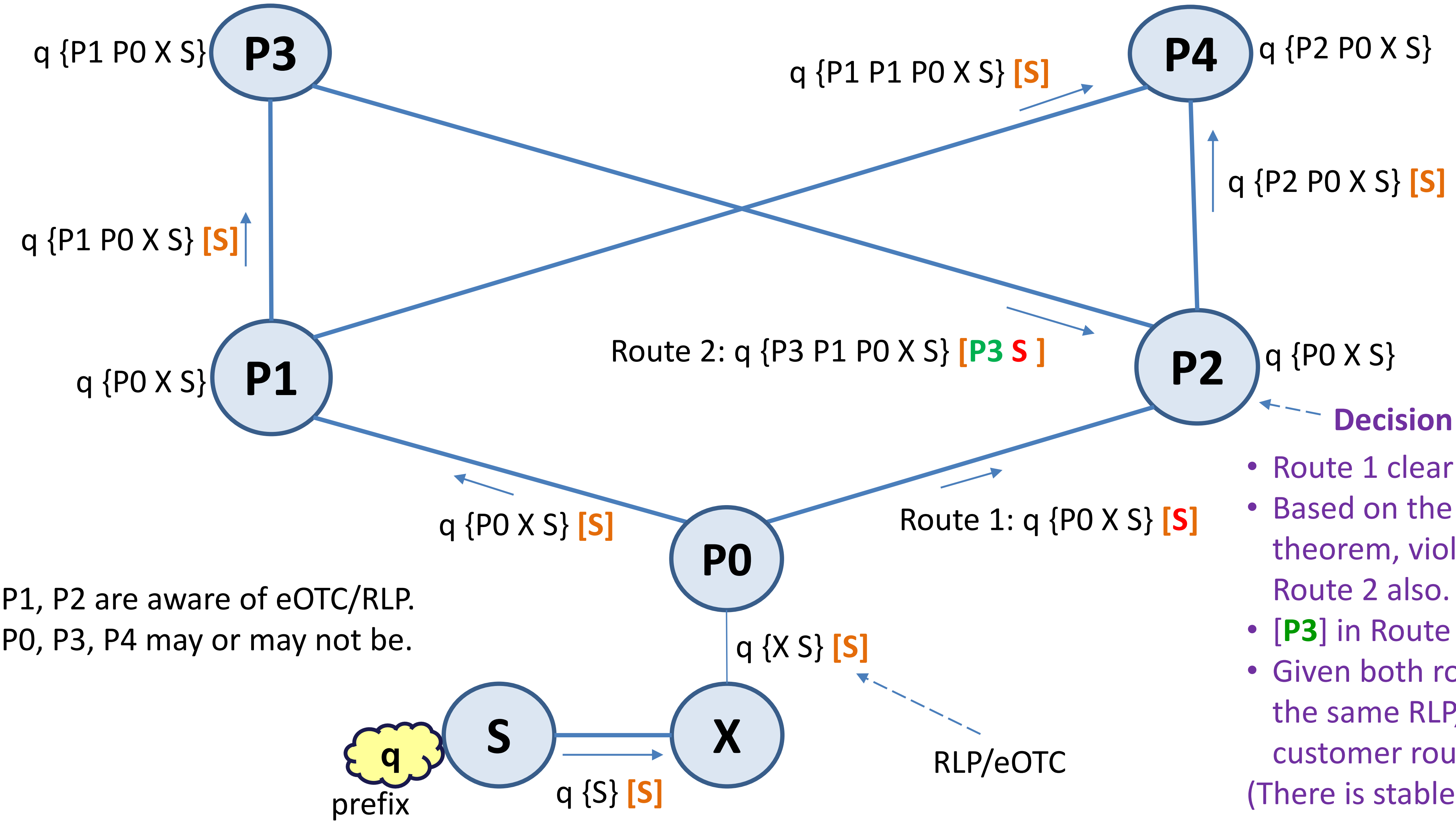
	Design A (Original RLP)	Design B	Original eOTC (Design B with only one ASN)
Functionality	• Detect multiple leaks • Provide up link info also	• Detect multiple leaks • Only down/peer info	• Can't detect multiple leaks • Lack of differentiation in some cases
Detection / mitigation stength	Best	Very good	See above
Memory use* (per update)	~ 136 bytes	~ 72 bytes	~ 32 bytes

* Assume average 4 hop AS path

Design B: Example



Alexander's scenario:
Avoid Persistent Oscillation Possibility



P1, P2 are aware of eOTC/RLP.
P0, P3, P4 may or may not be.

Decision Policy (Algorithm):

- Route 1 clearly violated [S].
- Based on the route-leak detection theorem, violation of [S] applies to Route 2 also.
- [P3] in Route 2 is expected (good).
- Given both routes are in violation of the same RLP, P2 prefers the customer route.

(There is stable convergence.)

Examine Provider Route vis-à-vis Customer's

- If customer route is a leak, and alternative route via provider includes the customer AS in the path, then prioritize customer route over the provider route.

* Stated simply here. See formal statement and explanation in the drafts.

Next Steps

- Request WG feedback on Design A vs. Design B

  - How much utility for the additional information in the RLP attribute in Design A?
    - Indicating when update is sent to transit provider

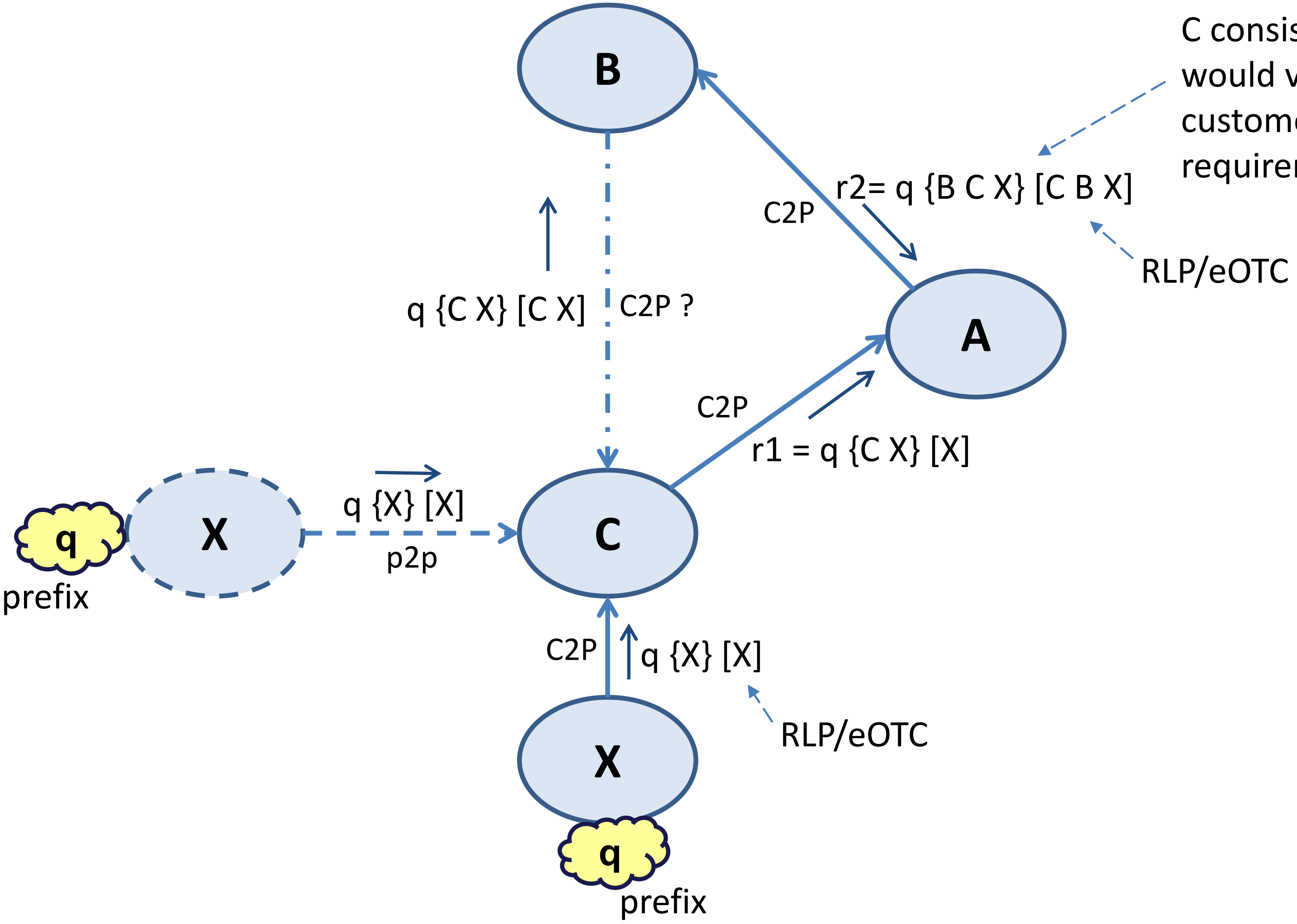
- Request WG feedback on Attribute vs. Community

- Prepare a finalized version for WGLC

Backup slides

Route-Leak Detection Theorem: Illustration

The only possible way that [X] is not violated in r2 is if the path from B to C consists of C2P links only. But that would violate the "No cycle of customer-provider relationships" requirement [Gao-Rexford].



Route-Leak Detection Theorem

The “Gao-Rexford” Stability Conditions

[Gao-Rexford] http://www.cs.princeton.edu/courses/archive/spr11/cos461/docs/lec17-bgp-policy.ppt (slide 27)

- Topology condition (acyclic)
  - No cycle of customer-provider relationships

Route-Leak Detection Theorem: Let it be given that ISP A receives a route r1 from customer AS C and another route r2 from provider AS B (for the same prefix), and both routes r1 and r2 contain AS C and AS X in the path and also contain [X] in their RLP/eOTC. Then, clearly r1 is in violation of [X]. It follows that r2 is also necessarily in violation of [X].

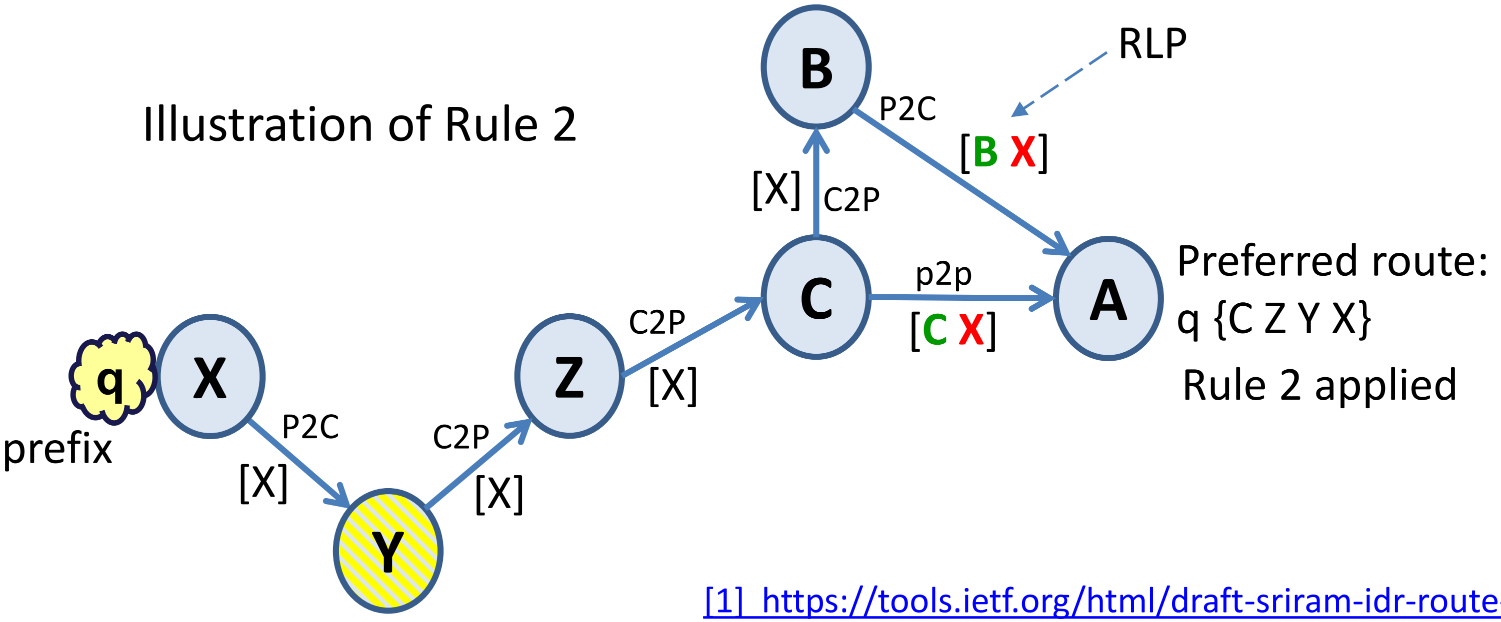
Proof: Let us suppose that r2 is not in violation of [X]. That implies that r2’s path from C to B to A included only P2C links. That would mean that there is a cycle of customer-provider relationships involving the ASes in the AS path in r2. However, any such cycle is ruled out in practice as a necessary stability condition [Gao-Rexford]. QED.

Route-Leak Mitigation Rules

Rule 1: If ISP A receives a route r1 from customer AS C and another route r2 from provider (or peer) AS B (for the same prefix), and both routes r1 and r2 contain AS C and AS X (any X not equal to C) in the path and also contain [X] in their RLP, then prioritize the customer (AS C) route over the provider (or peer) route.
(Rationale: This rule is based on the theorem (slide 8). See detailed rationale in Section 3.1 in [1].)

Rule 2: If ISP A receives a route r1 from peer AS C and another route r2 from provider AS B (for the same prefix), and both routes r1 and r2 contain AS C and AS X (any X not equal to C) in the path and also contain [X] in their RLP, then prioritize the peer (AS C) route over the provider (AS B) route.
(Rationale: See illustration below. See detailed rationale in Section 3.1 in [1].)

Illustration of Rule 2



[1] https://tools.ietf.org/html/draft-sriram-idr-route-leak-solution-discussion-00

Default Route-Leak Mitigation Policy

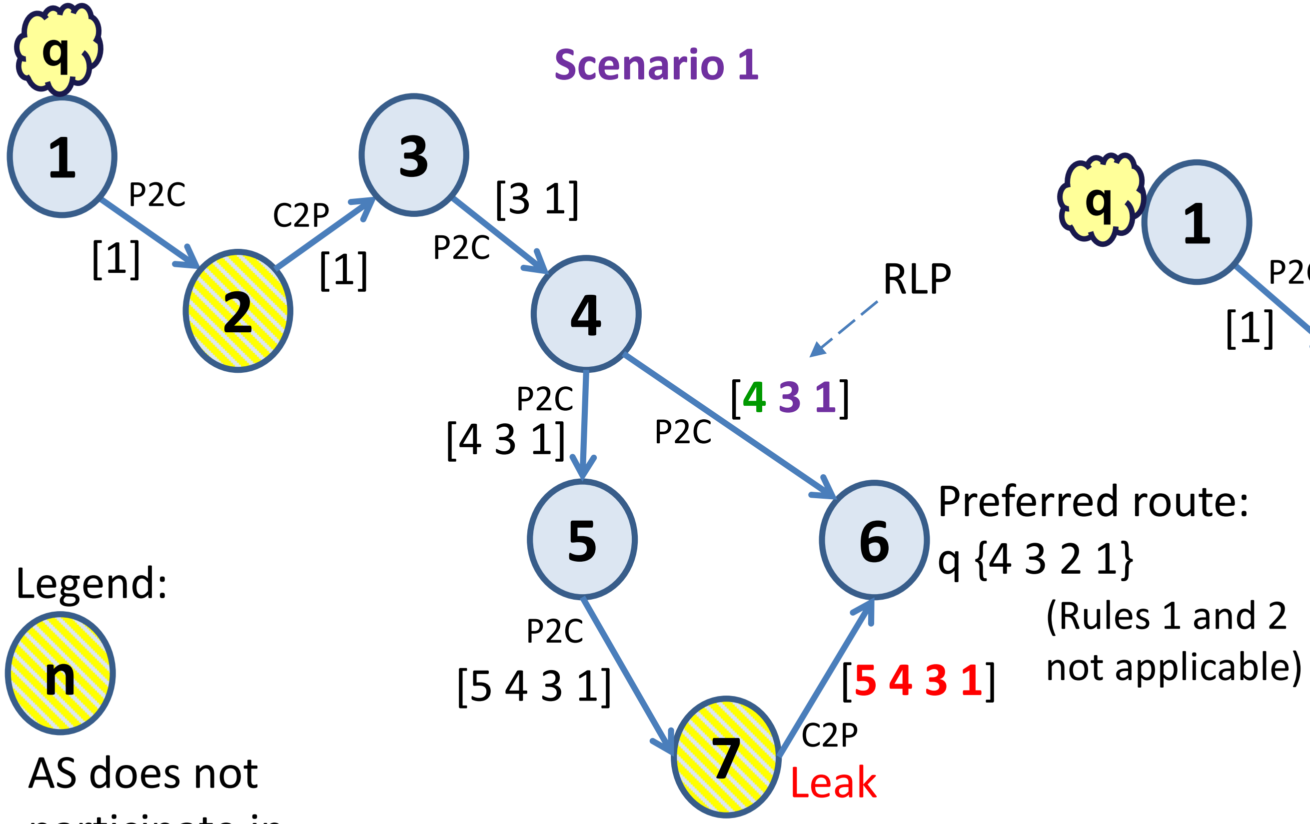
- **Given a choice between a customer route versus a provider (or peer) route,**
  - if no route leak is detected in the customer route, then prioritize the customer over the provider (or peer);
  - else (i.e., when route leak is detected in the customer route) and the conditions of Rule 1 apply, then too prioritize the customer over the provider (or peer);
  - else (i.e., when route leak is detected in the customer route and the conditions of Rule 1 DO NOT apply), then prioritize the provider (or peer) over the customer.

- **Given a choice between a peer route versus a provider route*,**
  - if no route leak is detected in the peer route, then prioritize the peer over the provider;
  - else (i.e., when route leak is detected in the peer route) and the conditions of Rule 2 apply, then too prioritize the peer over the provider;
  - else (i.e., when route leak is detected in the peer route and the conditions of Rule 2 DO NOT apply), then prioritize the provider over the peer.

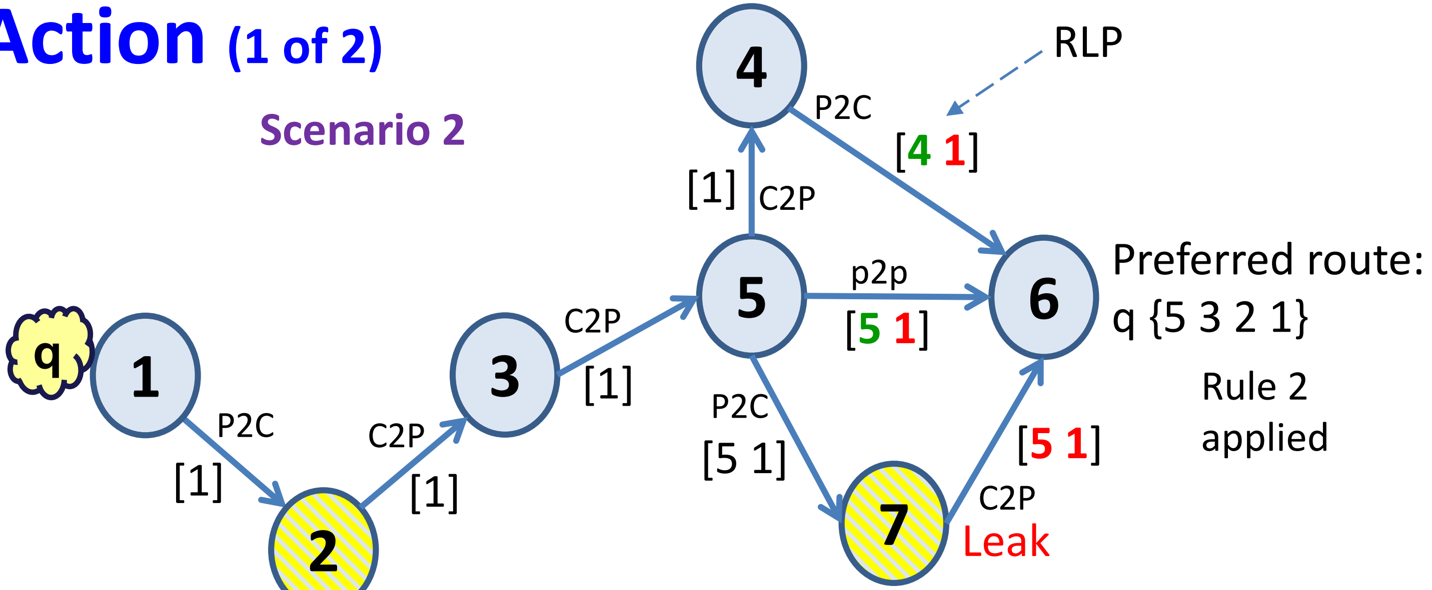
* Operator MAY override (the second bullet) to prefer provider route over peer route.

Examples Showing Policy in Action (1 of 2)

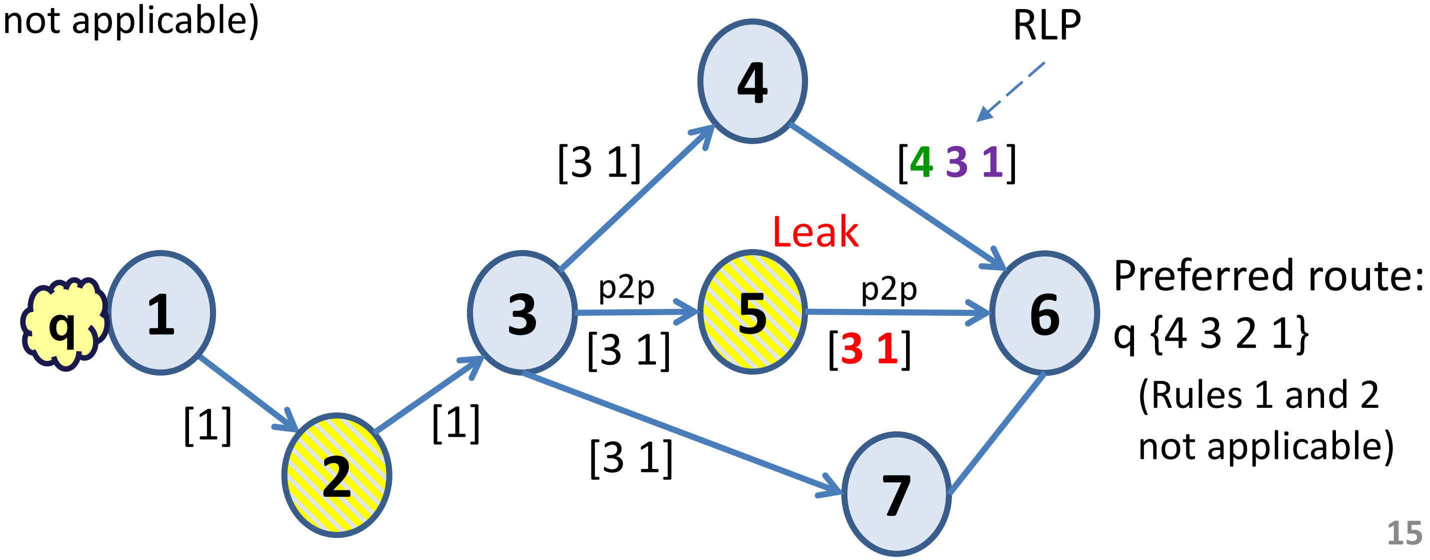
Scenario 1



Scenario 2



Scenario 3



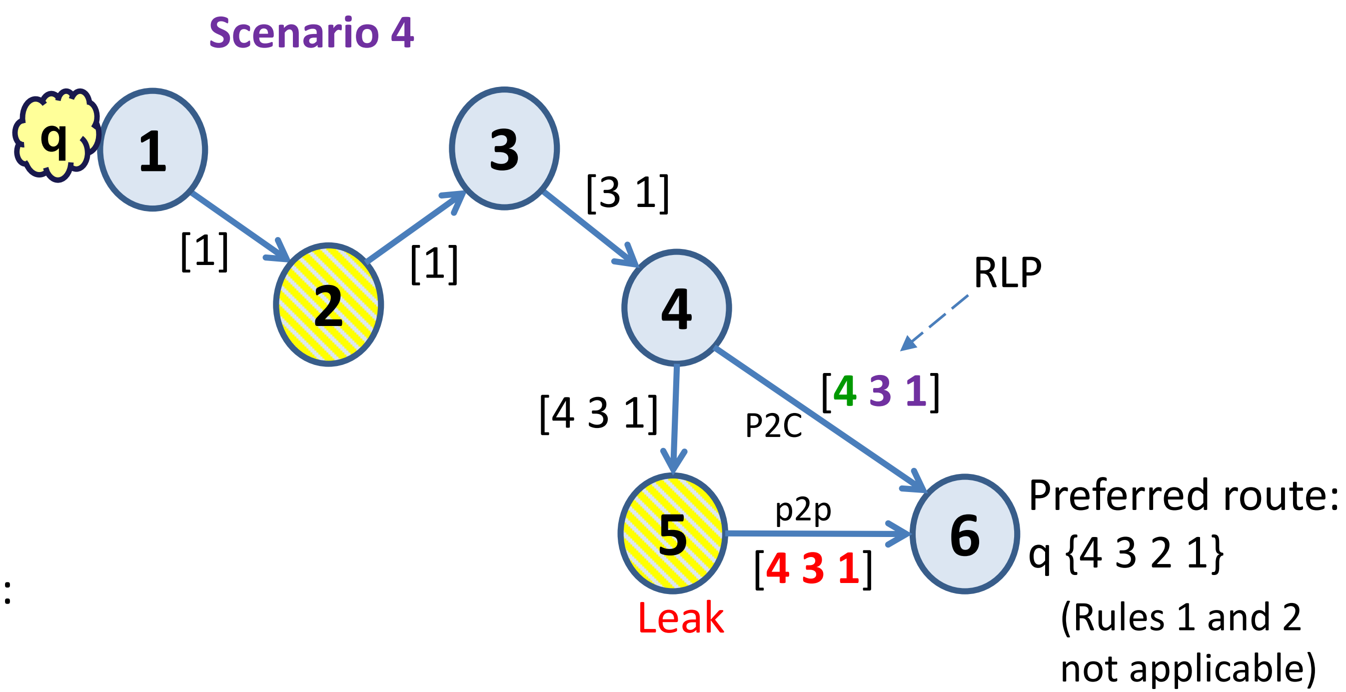
Legend:

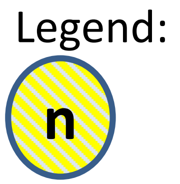
n

AS does not participate in RLP and/or Propagates a route that violates RLP

Green – not violation
Red – violation
Purple – can't tell

Examples Showing Policy in Action (2 of 2)



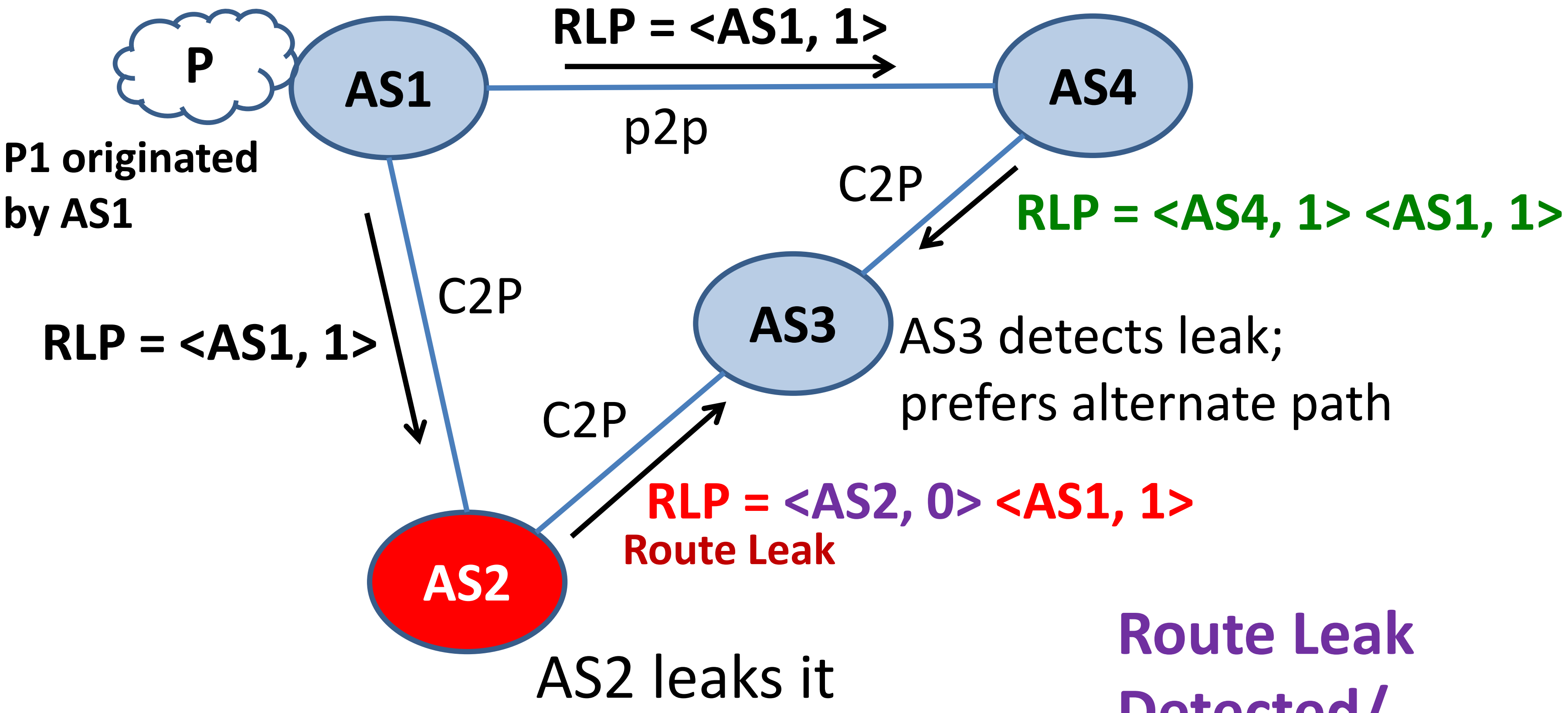


AS does not participate in RLP and/or Propagates a route that violates RLP

Green – not violation
Red – violation
Purple – can't tell

Design A – RLP Attribute

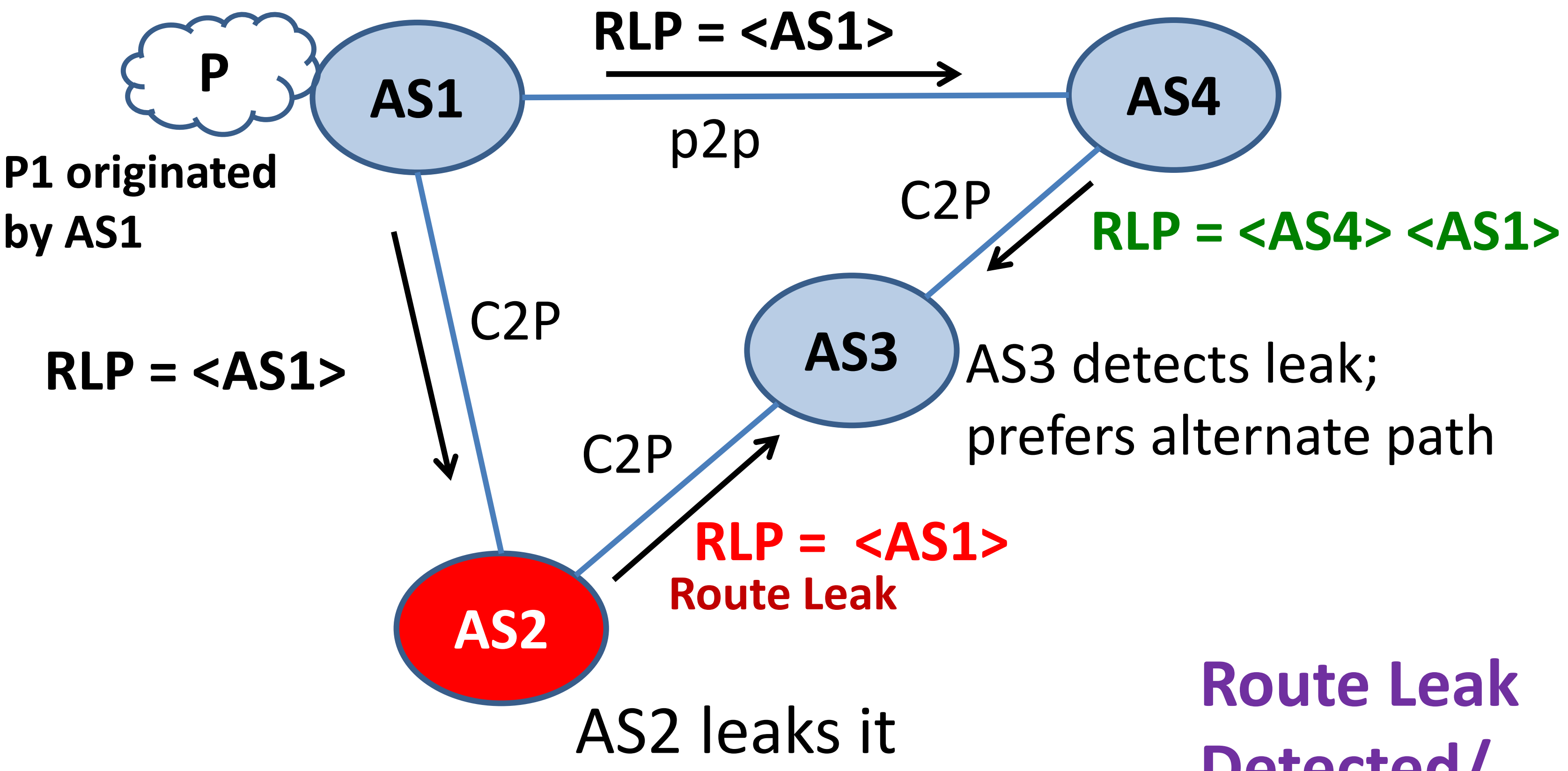
- Insert <ASN, 1> if sending to Customer or Peer
- else, insert <ASN, 0>



Route Leak Detected/ Mitigated ☺

Design B – RLP Attribute

- Insert <ASN> if sending to Customer or Peer
- else, insert nothing



Route Leak Detected/ Mitigated ☺

BGP Yang Model
draft-ietf-idr-bgp-model

Keyur Patel
Mahesh Jethanandani
Susan Hares

Status

- BGP Yang Model is now NMDA compatible
- Removed dependencies on OpenConfig models

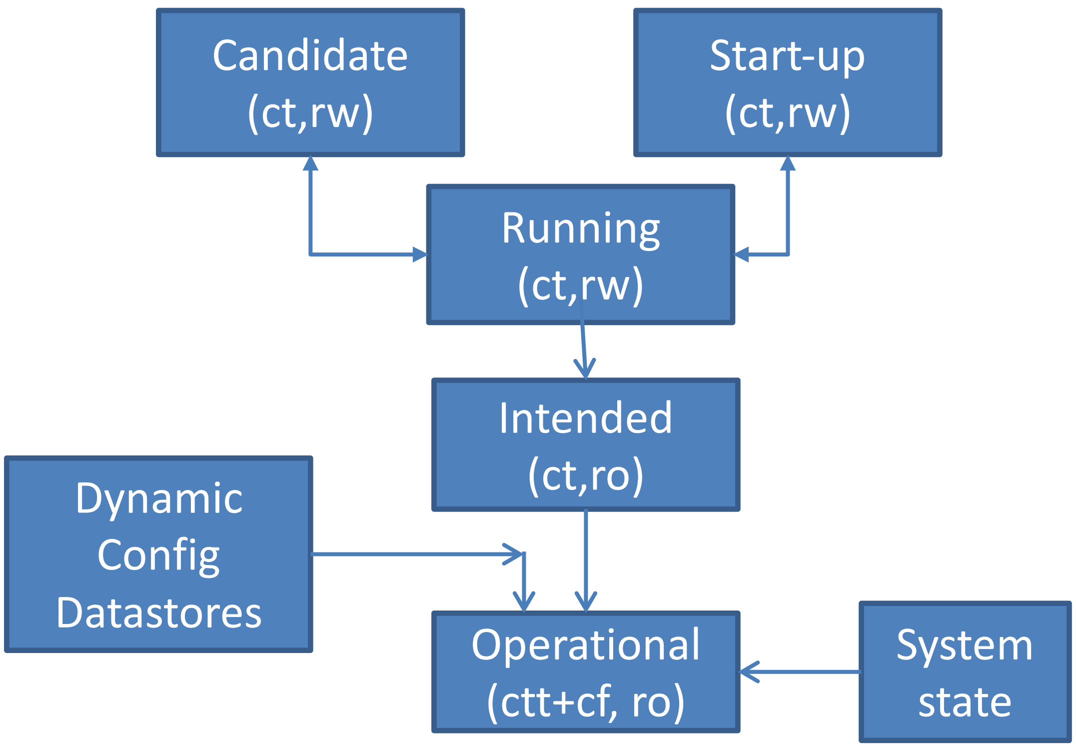
IETF NMDA

- RFC 8342

+ --config

  | rw mtu (intended datastore)

  | r mtu (in operational state datastore)



Extensions

- Provide extensions for additional features
  - In draft-keyupate-bgp-extensions-00
  - BGP signaled VPLS
  - BGP EVPN
  - options for L2VPN address families

Next Step

- WGLC

Feedback and questions

BGP Extra Extended Community
draft-heitz-idr-extra-extended-community

IETF 102
July 2018

Jakob Heitz
Ali Sajassi
Cisco
Ignas Bagdonas
Equinix

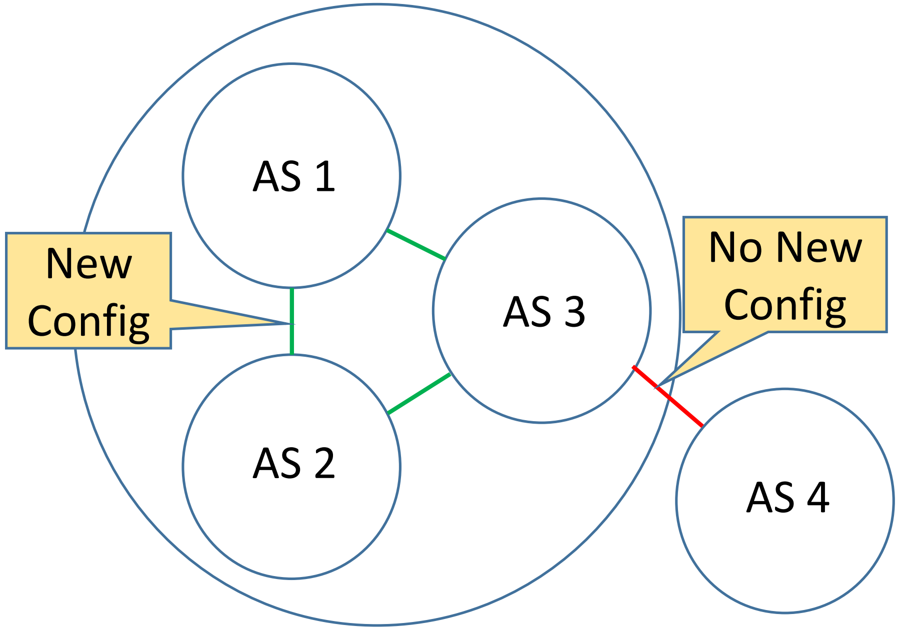
eXtra eXtended Community (XXC)

- Why Extended Community?
  - Easier to enhance than to invent brand new.
- 24 octets. Why fixed length?
  - Easier to enhance Extended Community code.
- Why bigger?
  - Easier to auto-derive by combining multiple existing identifiers:- reduce configuration.

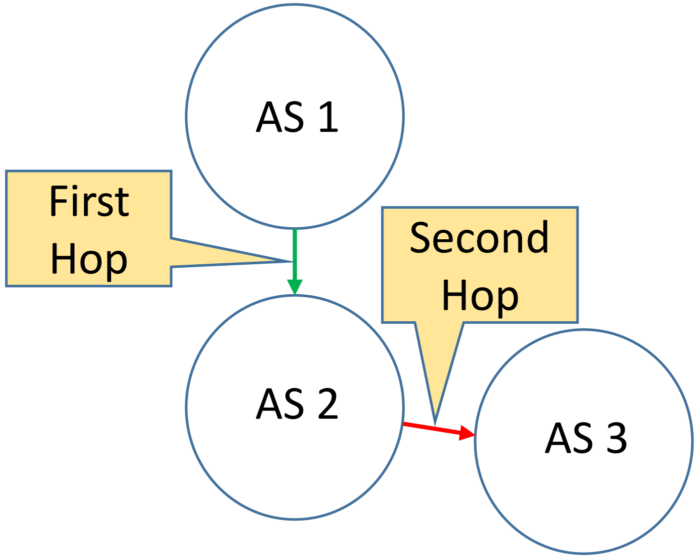
New Transitivity

Coarse grained, to prevent accidental distribution to the entire Internet, but still covers major use cases. Use route-policy for fine grained distribution.

- Administration Transitive
  - Non-Transitive, except when session is configured as “Same-Admin”



- One Time Transitive
  - For your neighbor only
  - Link-Bandwidth and LLGR_STALE could use this.



RT Constraint

```
+--------------------------------+
| AFI                (2 octets)  |
+--------------------------------+
| SAFI               (1 octets)  |
+--------------------------------+
| origin AS          (4 octets)  |
+--------------------------------+
| XXC value         (24 octets)  |
+                                +
|                                |
+--------------------------------+
```

Applies to routes with only this AFI/SAFI

Not necessarily a Route Target

XXC only. Not for Extended Community

XXC Types

- AS-Specific (4 octet AS only)
- IPv4-Address-Specific
- IPv6-Address-Specific
- EVPN

Type/sub-type copied from Extended communities.
Just a suggestion. Can structure it differently.
Sub-Type not optional, unlike in RFC 4360.

EVPN XXC Sub-Types

- EVI Route Target
- ES-Import Route Target
- ESI-EVI Route Target
- Overlay Route Target

New size allows the use of the complete Ethernet Tag ID and ESI.

The new EVPN Route Targets are to be used in addition to the existing Route Targets, not as a replacement.

IETF 102 – Montreal
July 2018
IDR Working Group

draft-xu-idr-neighbour-autodiscovery-09

Xiaohu Xu, Chao Huang, Guixin Bao (Alibaba)
Ketan Talaulikar, Satya Mohanty (Cisco Systems)
Kunyang Bi, Shunwan Zhuang (Huawei)
Jeff Tantsura (Nuage Networks)
Nikos Triantafillis
Jinghui Liu (Ruijie Networks)
Zhichun Jiang (Tencent)
Shaowen Ma (Juniper Networks)

Problem Statement

- BGP is used as the only routing protocol in DCs using RFC7938 design
- Operational complexity involved in provisioning of hop-by-hop per link eBGP peering between BGP nodes
- When doing peering using loopbacks (e.g. due to ECMP links or when using IPv6 link-local addresses or unnumbered links) need to also provision static route for reachability

Requirements

- Need a neighbour discovery mechanism that runs on top of IPv4/IPv6
  - Is media independent; works on IPv4, IPv6 and dual stack
  - Needs to support authentication mechanism for security purposes
- Keep it simple and focus on current BGP requirements
  - We have LLDP and BFD widely deployed; leverage them
  - Make mechanism extensible for signalling of information for BGP
- Auto-discovery and bootstrap for BGP TCP Sessions between directly connected nodes
- Separate discovery and liveness for BGP neighbours
  - Discovery and maintenance of adjacency is the core part
  - Use of liveness mechanism is optional; continue to leverage BGP KA, BFD and Fast External Failover features
- Minimal changes for integration with BGP Peer FSM and no changes in BGP protocol operations

What does this draft propose?

- Automated neighbour discovery using UDP Hello Messages on a per link basis for directly connected neighbours only
- Signalling of peering address and ASN so that BGP Peering session can be automatically initiated with discovered neighbour
- BGP session can be setup using loopbacks and reachability established via peering route setup that points over the links over which neighbour is discovered
- Minimal changes to the BGP Peer FSM and no change to BGP route processing

Hello Message Format

- Uses UDP port 179 and sent to link-local multicast address
- Can be used over either IPv4 or IPv6 addresses

```
 0                   1                   2                   3
 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|    Version    |      Type     |        Message Length         |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                           AS number                           |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                        BGP Identifier                         |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|    Adjacency Hold Time        |           Reserved            |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                             TLVs                              |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Important TLVs

- Peering Address TLV
  - Indicates one or more IPv4 and/or IPv6 peering address(es) to be used
  - Optionally can indicate which AFI/SAFI to be used for which Peering
- Link Attributes TLV
  - Indicates link addresses and link identifiers for describing the link endpoint (so information is learnt for exporting via BGP-LS)
  - Can perform subnet and other policy checking before session setup
- Neighbour TLV
  - Signals discovered neighbours and their adjacency status (1-way, 2-way, reject and established)
  - Used to indicate to neighbour whether the BGP TCP session can be initiated (i.e. when both sides have accepted each other)

Optional TLVs

- Local Prefix TLV
  - Indicates the prefix route to be programmed after neighbour discovery goes to 2-way state to ensure reachability for the neighbour’s peering address
  - Required when peering is to be done using loopback interface; not required when doing peering with interface addresses
- Accepted ASN TLV
  - Indicates the list of ASNs to which peering session would be established – local policy
- Cryptographic Authentication TLV
  - Carries the SA ID and authentication information

Adjacency State Machine

- Initial State
  - Initial state when a neighbour is detected
- 1-way State
  - When router accepts the peer and includes it in its own hello message
- Reject State
  - When router rejects the peer due to detection of some config mismatch or violation of local policy
- 2-way State
  - When router detects itself in the neighbour’s hello; now ready for TCP session establishment step
  - Adds peering route for the neighbour over the link (i.e. when using loopbacks for peering)
  - Creates the BGP Peer State context for discovered peer and triggers the BGP Peer FSM
- Established
  - When the BGP TCP session is established

Session Management

- Once established, session management is performed as per BGP FSM
- Liveness detection via Keepalives & Hold timer
  - BFD and Fast External Failover also works when enabled
- Established BGP session is NOT brought down due to adjacency hold timer expiry by default
  - This may be optionally enabled in cases where required
- Adjacency hold timer expiry used to clean-up BGP Peer state after the session goes down for auto-discovered peer

Peering Route

- Required only when peering is done using loopback interfaces
- Route programmed with higher Admin Distance than normal BGP routes to prevent oscillation (in case the peering route is also learnt via BGP itself)
- When there are multiple links between neighbours then peering route will have ECMP paths over each of them
- BGP NH for the neighbour resolved over this peering route for reachability
- No need for programming static route or running another protocol when doing Peering over loopback addresses

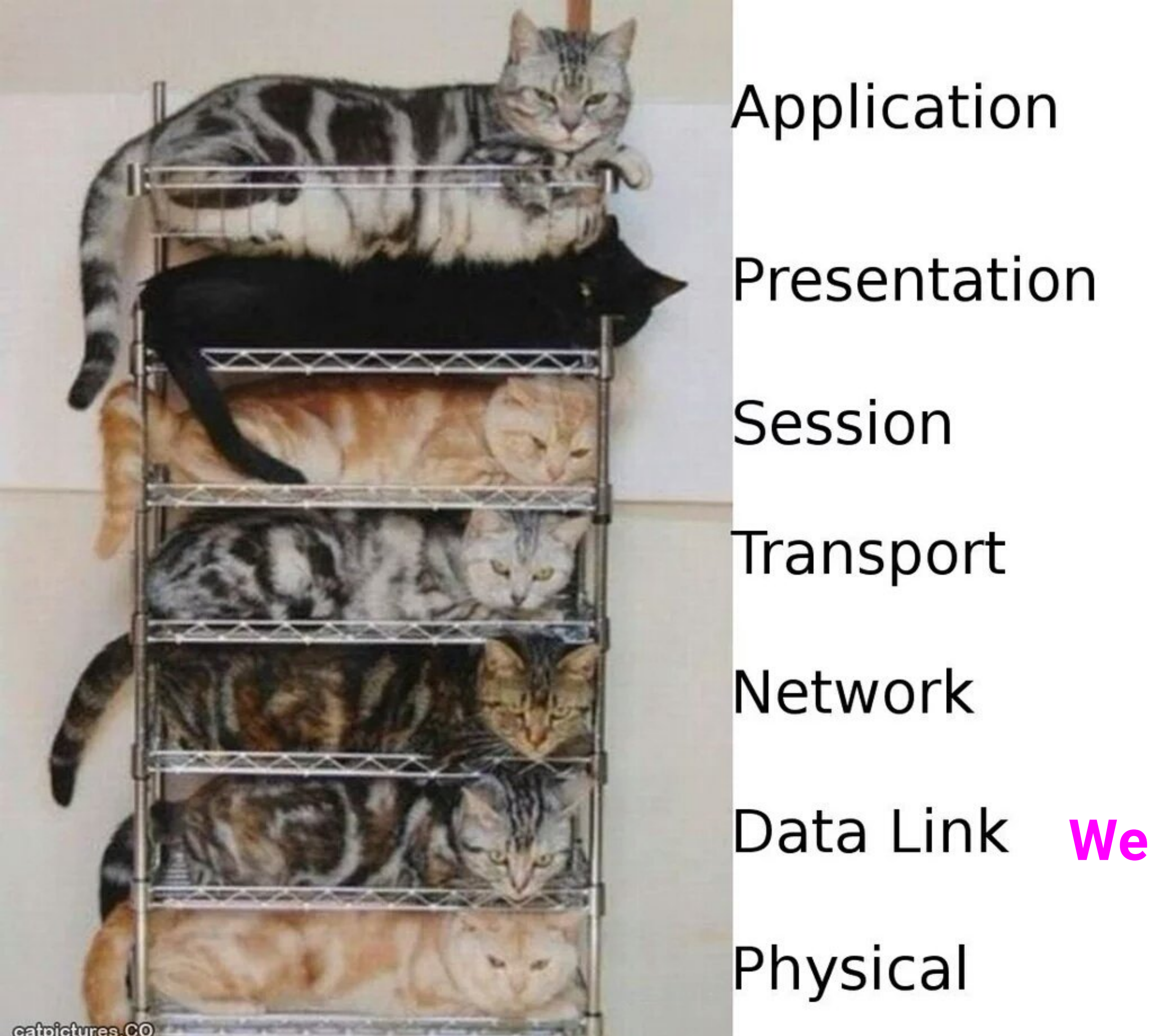
Next Steps …

- WG adoption call ongoing in IDR
- Solicit WG review and comments/inputs/feedback

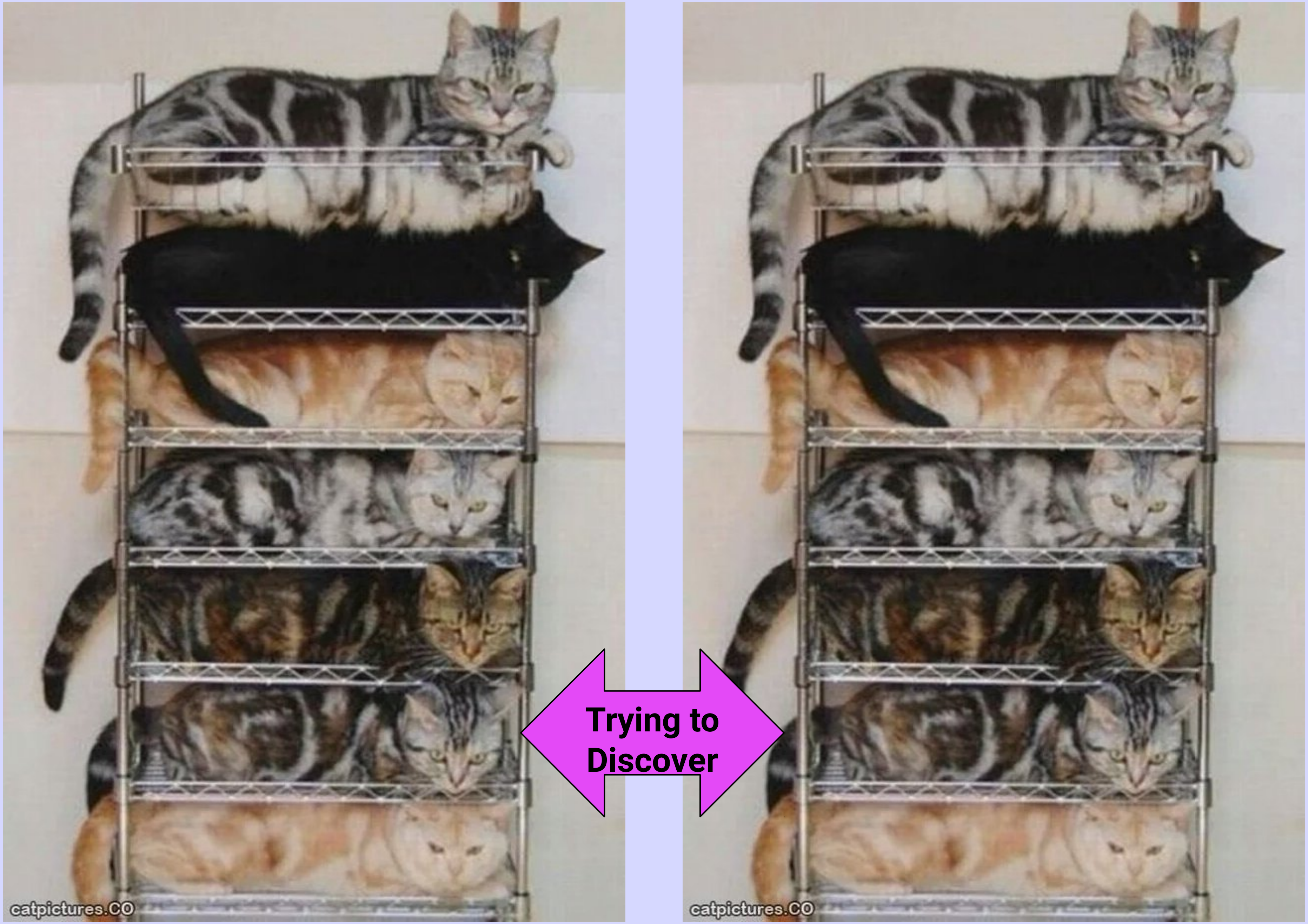
Link Discovery and Liveness

What do we really need?

Randy Bush <randy@psg.com>


Application
Presentation
Session
Transport
Network
Data Link
Physical
catpictures.CO

We Are Here


catpictures.CO
Trying to Discover
catpictures.CO

IIJ is Building a Second <u>Medium</u> Scale Data Center (MSDC) in Shiroi/Chiba Capacity of 6k Racks

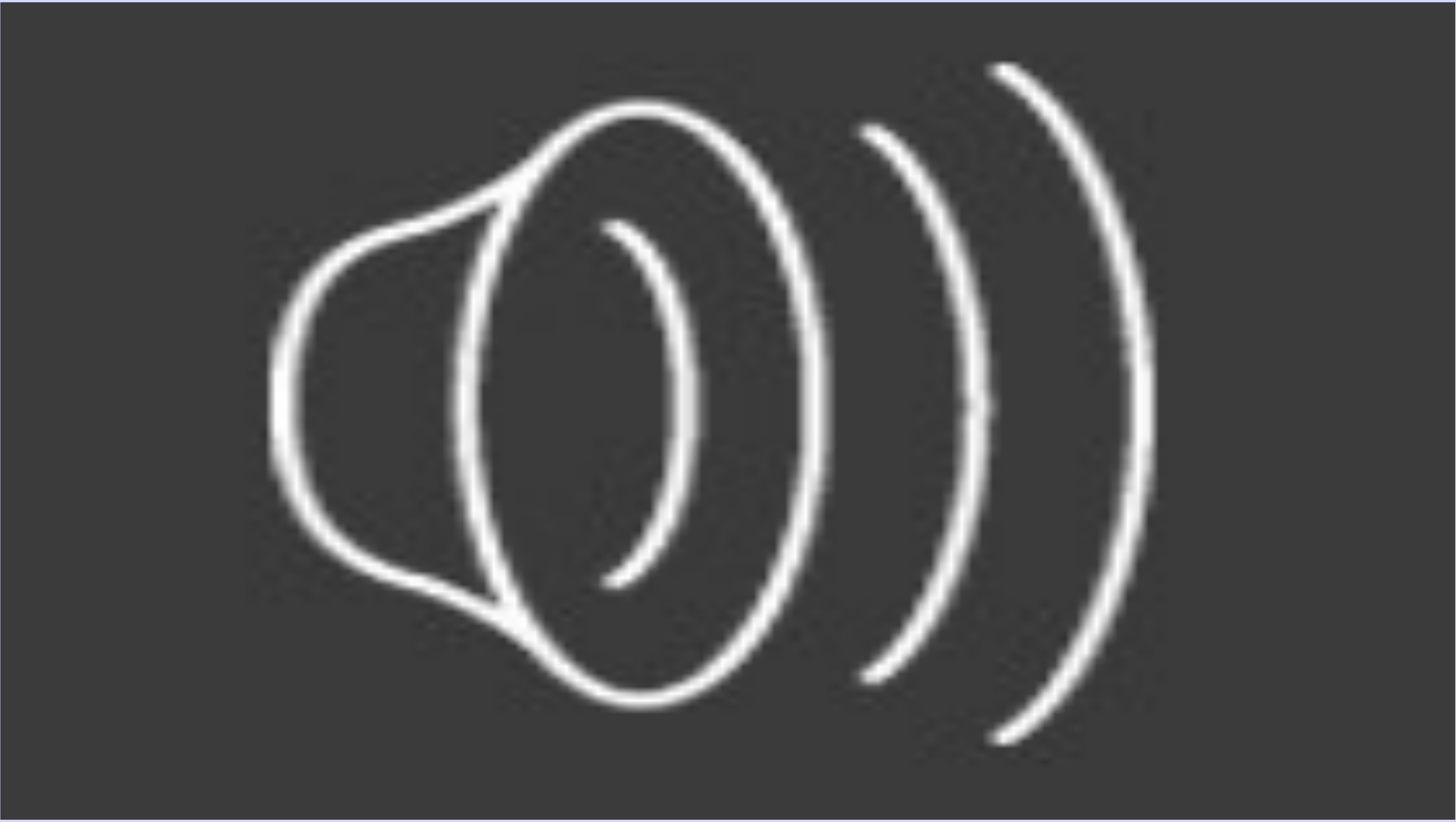
How Can We Route In Something of This Scale?

OSPF OK to 500 Nodes

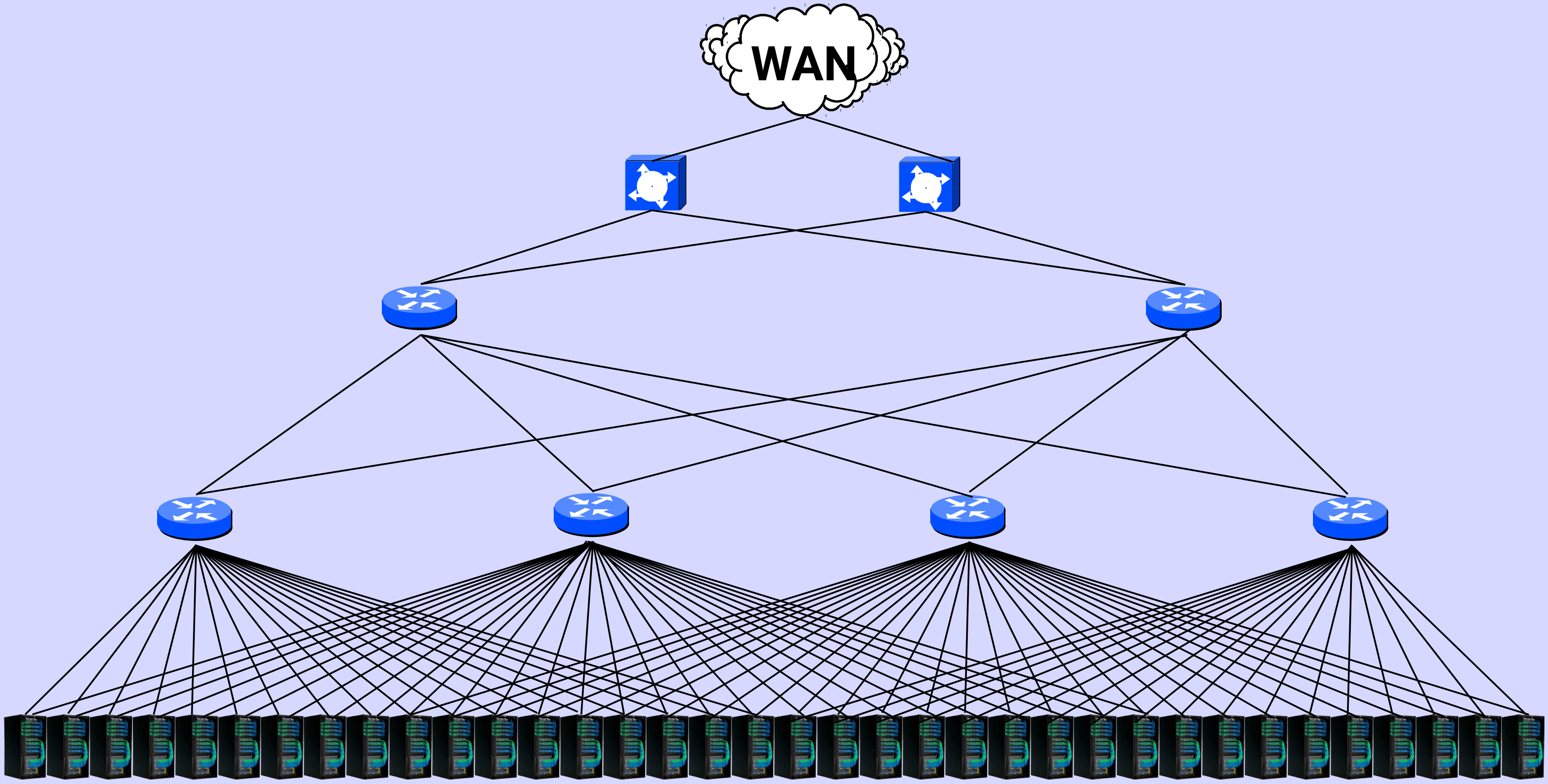
IS-IS good to 1,000

Limited Because They Repeatedly Flood Everything

Your Clos on IS-IS or OSPF



BGP Is Great as Updates are Infrequent

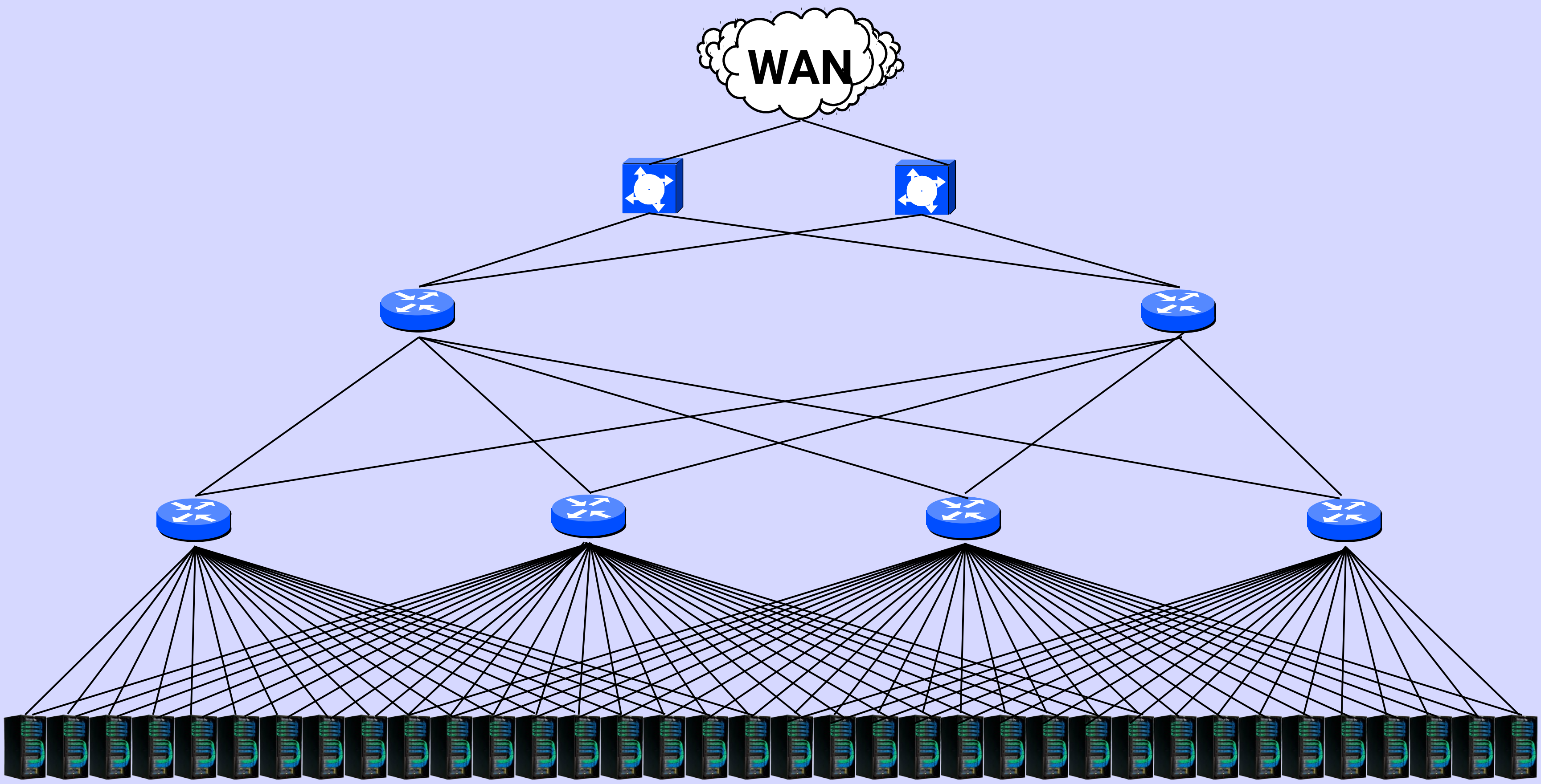


BGP Scales Because It Signals Only Changes

So BGP has become common in MSDCs

ECMP can be Very Wide

32, 64, even 128



The Problem is Topology Discovery

Two Kinds of Standards

Union – the accumulation of all the features anybody wanted

Intersection – only those things everybody absolutely had to have

Either Tony Hoare or Klaus Wirth – I can not find the quote <blush>

IETF asks the ITU

Q: So you add features until the "NO"s stop

A: We don't like to think of it that way

Must Haves

- Discover Nodes and Links
- Discover Link Encapsulations:
  - IPv4, IPv6, MPLS4/6, ...
- Maintain Layer-2 Liveness
- Northbound API to BGP-SPF

Security?

- Datacenter Ops seem not to think of security at this layer (or any!)
- We need Authentication. Maybe Integrity?
- One of the things which are likely to drive PDU size over 1,500

Non-Features

- Routing Data, BGP-SPF does that
- Access to IGP Databases, This is discovery and liveness, not routing
- Just want the Link
- Transport, not our job

Desiderata

- Discovery & Liveness for BGP-SPF
- Simple but usable in Massively Scalable networks of >10,000 nodes
- May be useful for other applications
- Simple
- Extensible (e.g. authentication, cost)
- Simple
- No IPR

Why Simple?

We are here to produce easily understood, implementable, and securable standards, not build résumés.

Why Simple?

A high goal of software engineering is to remove the need for features. It's a vital part of designing for simplicity, even invisibility. -- Rob Pike

Candidates?

- LLDP and its children
- IS-IS link discovery
- Edge Control Protocol (Alvaro)
- BGP Neighbor Autodiscovery
- Link State Over Ether

LLDP

- IEEE Protocol
- IPR over 1,500 bytes
- A bit complex
- Won’t go through a switch (feature or bug?)
- Beacons, not KeepAlives
- Viable but

IS-IS Discovery

- IETF now has control
- Complex enough that BGP-LS was invented so normals could get the link state database
- IS-IS not commonly implemented on MSDC devices, so would need to profile and develop

Edge Control Protocol

- It is a transport controlled by IEEE
- A Reliable layer two transport, on top of LLC
- Has flow control, reliable, non-reorder, ... transport
- used for EVP and PD/CSP
- Reinventing TCP over 802.1

BGP Neighbor Autodiscovery

- IETF protocol
- Very new
- Needs the peering address to get the peering address
- AS Based, can not use other idents
- Not really discovery at all, configuration
- No liveness

Link State Over Ether

- Custom made for the job
- Very bare bones, brutally simple
- Only does discovery and liveness
- New, therefore risky
- But so is BGP-SPF
- No measurement or monitoring tools

	LLDP	IS-IS	ECP	BNA	LSOE
Who Owns	IEEE	IETF	IEEE	IETF	IETF
Maturity	Mature	Mature	Recent	New	New
Complexity	Somewhat	Very	Rather	Somewhat	Almost too Simple
Discovery	Yes	Yes	Yes	Configure	Yes
Liveness	Beacons	Yes	No	No	Yes
IPR	IPR	No	?	?	No

Discussion