BGP Flowspec Interoperability Lab

Christoph Loibl - christoph.loibl@nextlayer.at

Joint work with Martin Bacher (T-Mobile)

Notice

- Joint research project next layer & T-Mobile

  Credits to Martin Bacher from T-Mobile

- Supported by the Manufacturers

  Very cooperative when suggesting changes!

  Special thanks to Nokia and Cisco (provided required hardware for the lab)

- We do not suggest to buy this or that equipment!

  All tested manufacturers have working flow-spec implementations that are RFC5575 compliant as much as possible.

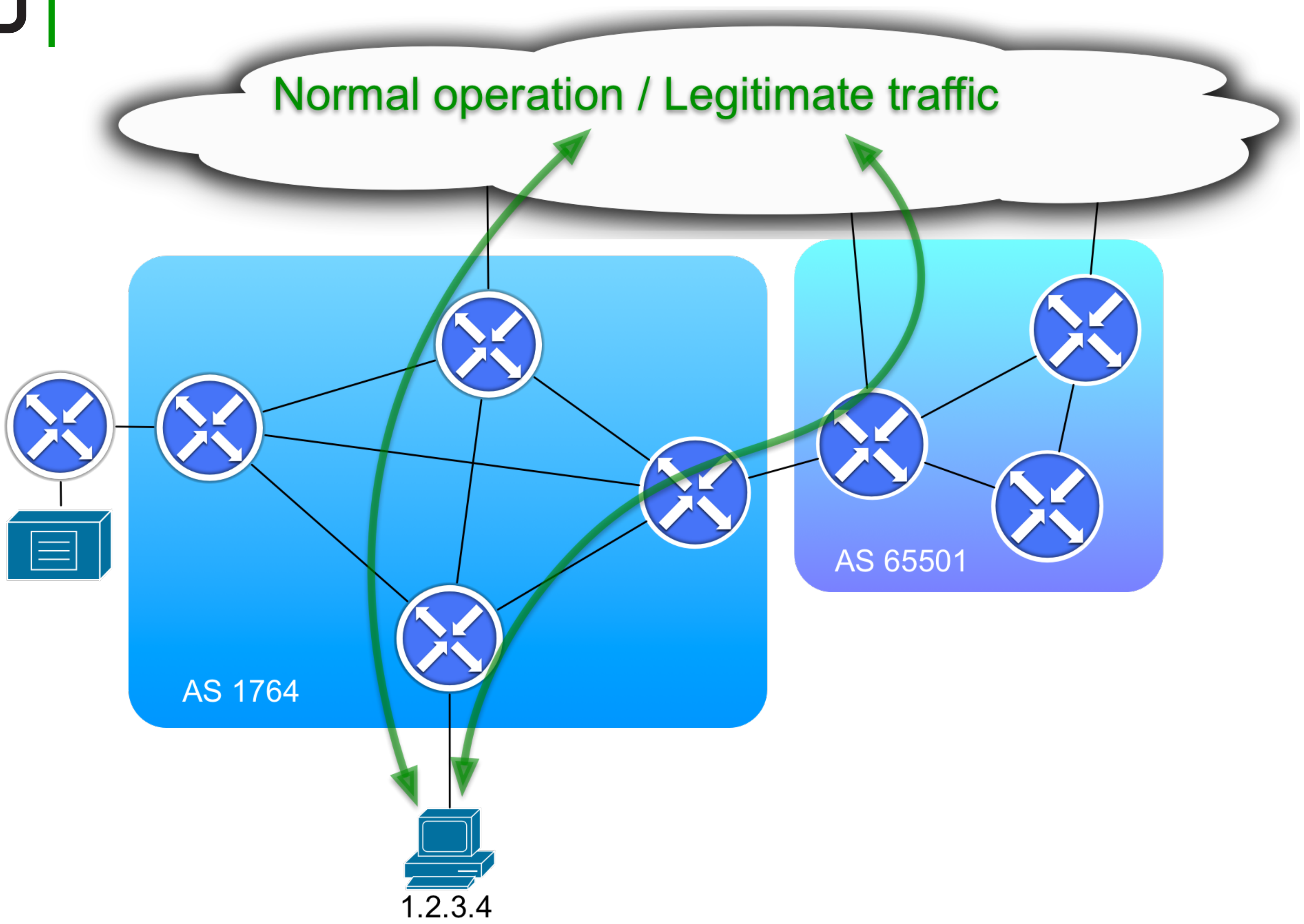
next layer

BGP Flow Specification (RFC5575)

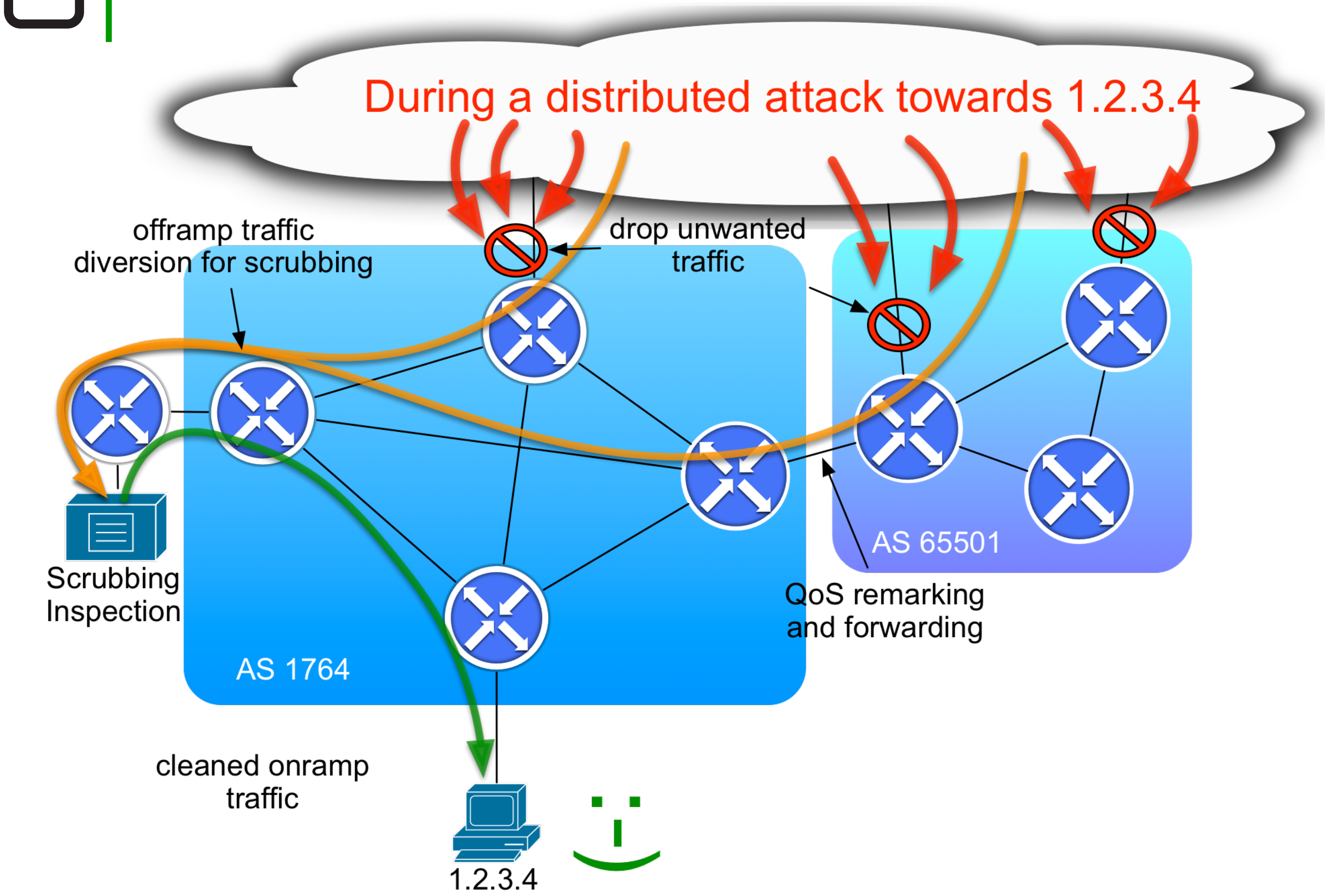
Rapidly deploy access control lists / flow-filters to routers ie. during DDoS mitigation (not limited to that)

- BGP NLRI format to exchange filter rules via BGP
- Set of filter criteria (flow-components) encoded in NLRI
- Set of match-actions encoded as extended BGP communities
- Resulting policies can be applied as ingress policy on the receiving routers
- Intra- and inter-AS distribution of flow-filter rules

BGP Flow Specification Use-Case



Network behaviour during an Attack

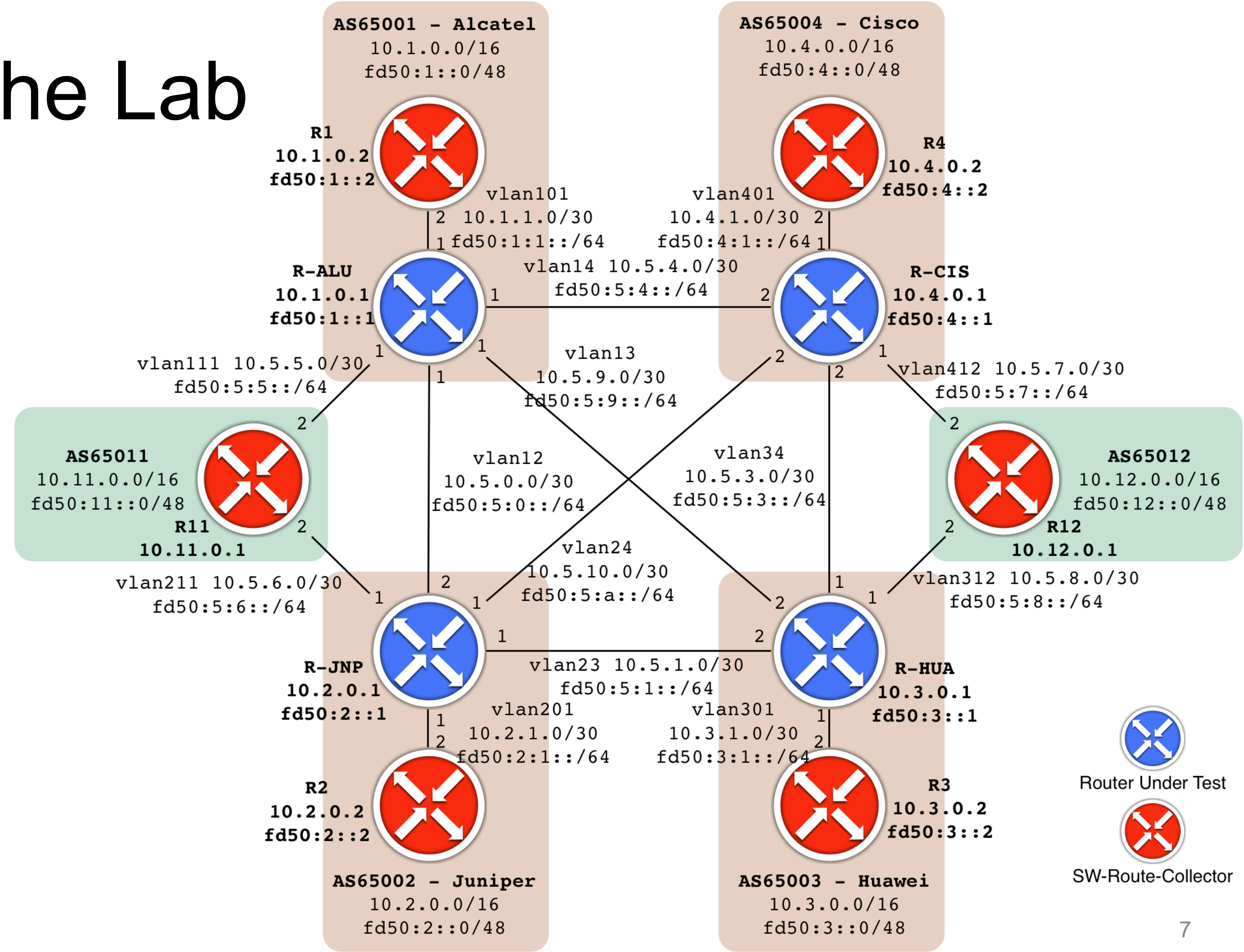

During a distributed attack towards 1.2.3.4
offramp traffic diversion for scrubbing
drop unwanted traffic
Scrubbing Inspection
AS 1764
AS 65501
QoS remarking and forwarding
cleaned onramp traffic
1.2.3.4

BGP Flowspec Interoperability Lab

- Produce a working set of configuration for an inter AS flowspec deployment
- Verify the behavior of the different products
  - Do all products interpret flowspec in the same way?
  - Do they successfully exchange filter rules?
- Identify missing features for inter AS flowspec
- Encourage our customers and peers to use flow-spec and exchange flow filters

The lab was targeted at control-plane (BGP-signaling) ONLY!
NOT at the data-plane (forwarding)!

The Lab



AS65001 – Alcatel
10.1.0.0/16
fd50:1::0/48

R1
10.1.0.2
fd50:1::2

vlan101
10.1.1.0/30
fd50:1:1::/64

R-ALU
10.1.0.1
fd50:1::1

AS65004 – Cisco
10.4.0.0/16
fd50:4::0/48

R4
10.4.0.2
fd50:4::2

vlan401
10.4.1.0/30
fd50:4:1::/64

R-CIS
10.4.0.1
fd50:4::1

vlan14 10.5.4.0/30
fd50:5:4::/64

vlan111 10.5.5.0/30
fd50:5:5::/64

vlan13
10.5.9.0/30
fd50:5:9::/64

vlan412 10.5.7.0/30
fd50:5:7::/64

AS65011
10.11.0.0/16
fd50:11::0/48
R11
10.11.0.1

vlan12
10.5.0.0/30
fd50:5:0::/64

vlan34
10.5.3.0/30
fd50:5:3::/64

AS65012
10.12.0.0/16
fd50:12::0/48
R12
10.12.0.1

vlan211 10.5.6.0/30
fd50:5:6::/64

vlan24
10.5.10.0/30
fd50:5:a::/64

vlan312 10.5.8.0/30
fd50:5:8::/64

R-JNP
10.2.0.1
fd50:2::1

vlan23 10.5.1.0/30
fd50:5:1::/64

R-HUA
10.3.0.1
fd50:3::1

vlan201
10.2.1.0/30
fd50:2:1::/64

vlan301
10.3.1.0/30
fd50:3:1::/64

R2
10.2.0.2
fd50:2::2

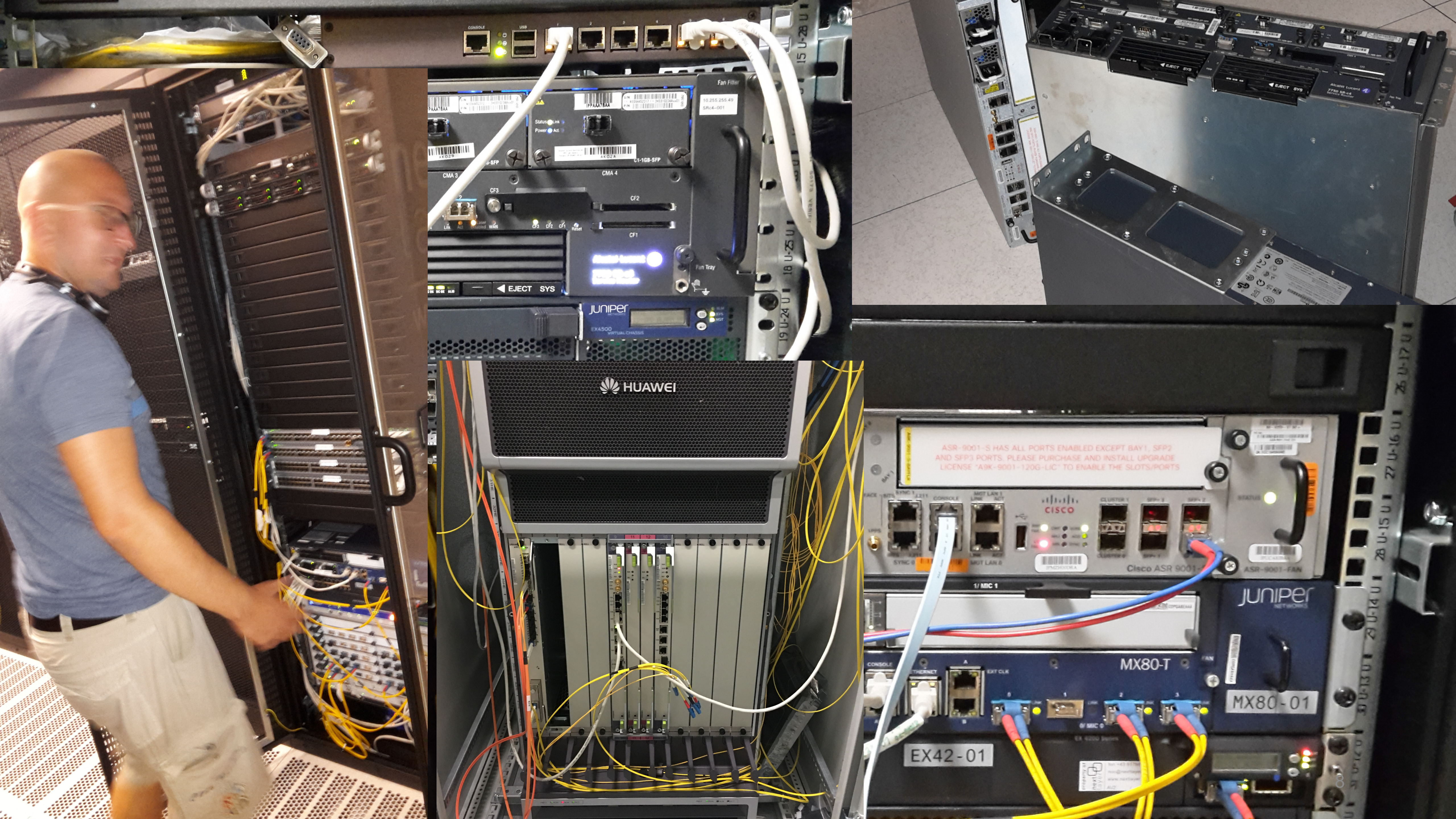
R3
10.3.0.2
fd50:3::2

AS65002 – Juniper
10.2.0.0/16
fd50:2::0/48

AS65003 – Huawei
10.3.0.0/16
fd50:3::0/48

Router Under Test

SW-Route-Collector


CONSOLE
USB
Fan Filter
CMA 3
CMA 4
CF3
CF2
CF1
Fan Tray
EJECT SYS
JUNIPER
EX4500
VIRTUAL CHASSIS
EJECT SYS
HUAWEI
ASR-9001-S HAS ALL PORTS ENABLED EXCEPT BAY1, SFP2 AND SFP3 PORTS. PLEASE PURCHASE AND INSTALL UPGRADE LICENSE "A9K-9001-120G-LIC" TO ENABLE THE SLOTS/PORTS
CISCO
Cisco ASR 9001
ASR-9001-FAN
1/ MIC 1
JUNIPER
MX80-T
MX80-01
EX42-01

Testcases

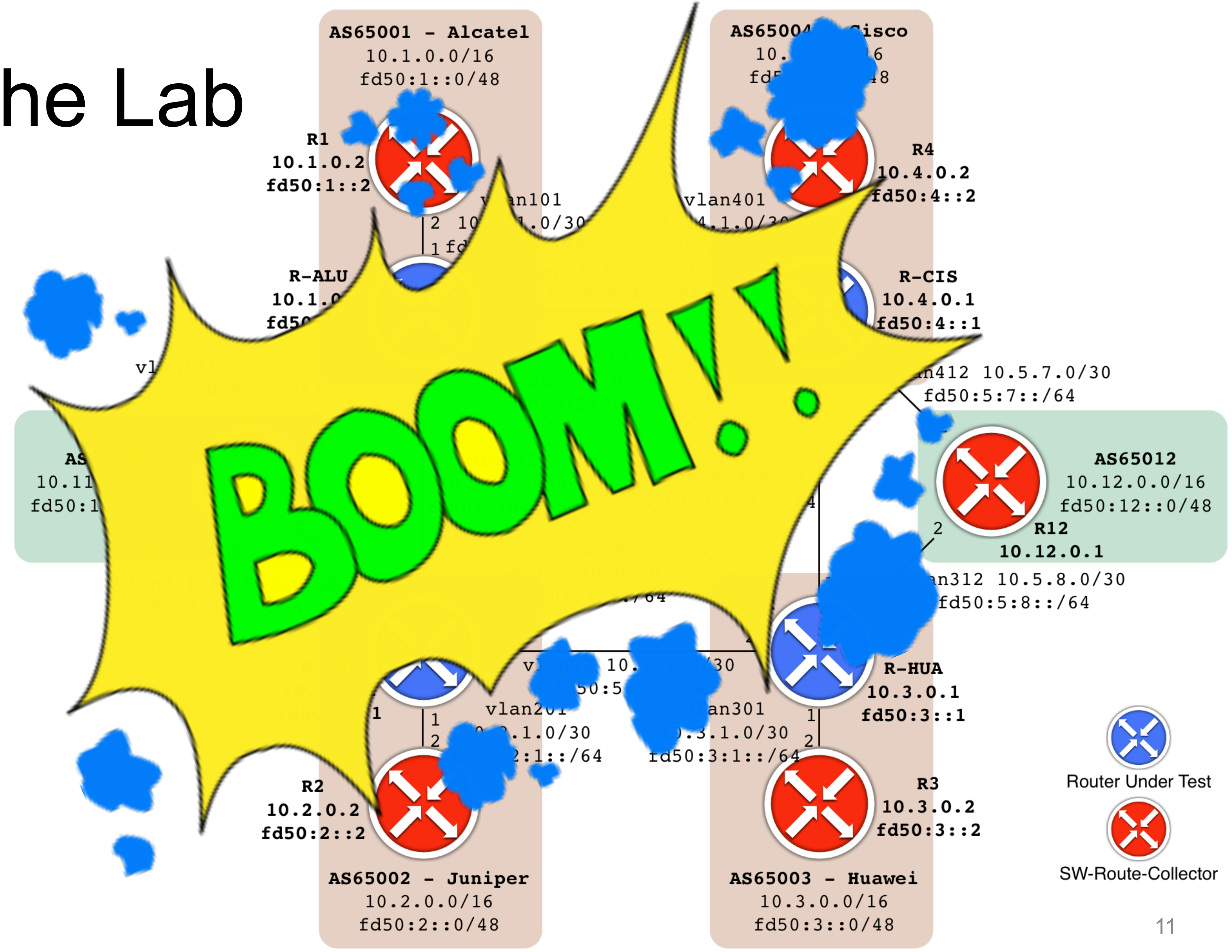
- General Match Patterns
- Action Community Combinations
- Transitivity of Action Communities
- Policy-Frameworks / Update Filtering
- Flow Specification Validation
- Term Ordering
- IPv6 Flow-Spec
- VRF Flow-Spec

General Match Pattern
R11 (ExaBGP)

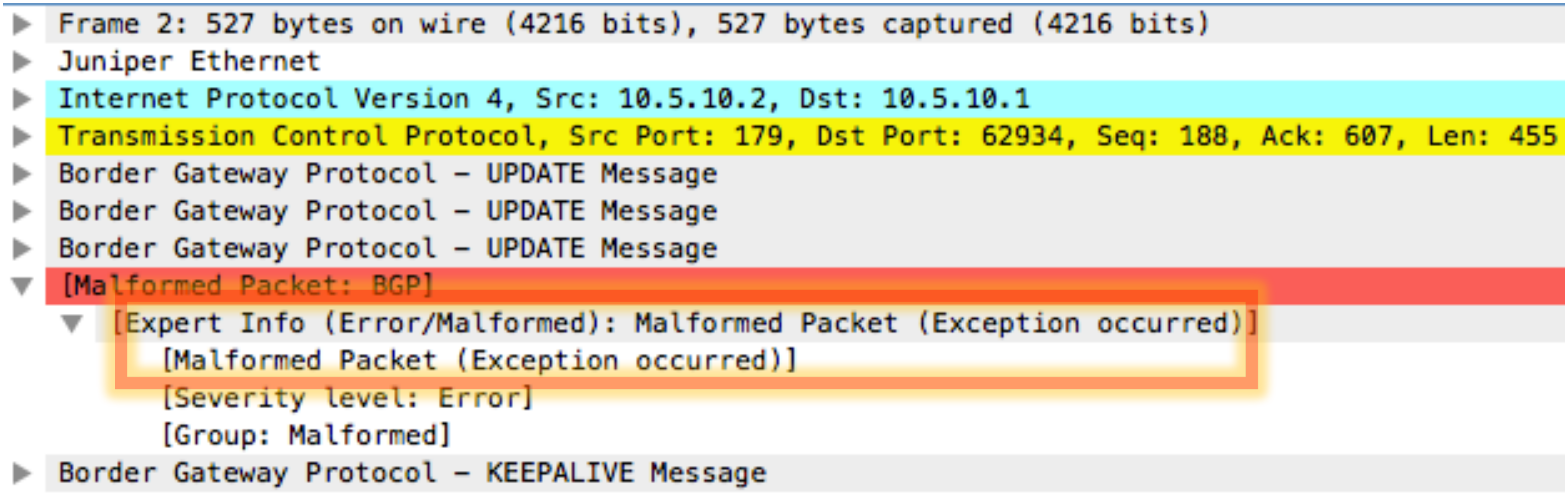
```
static { route 10.11.0.0/16 self; }
flow {
  route {
    match {
      destination 10.11.255.1/32;
      source 10.12.255.0/24;
      protocol =0 =1 =3 =5 =6 =7 >=10&<=12
        >=13&<=15 >=17&<=19 =255;
      port =0 =21 =23 =25 =26 =27 >=30&<=32
        >=33&<=35 >=37&<=39 =65535;
      destination-port =0 =41 =43 =45 =46 =47
        >=50&<=52 >=53&<=55 >=57&<=59 =65535;
      source-port =0 =61 =63 =65 =66 =67
        >=70&<=72 >=73&<=75 >=77&<=79 =65535;
      icmp-type =0 =1 =3 =5 =6 =7 >=10&<=12
        >=13&<=15 >=17&<=19 =255;
      icmp-code =0 =10 =21 =23 =25 =26 =27
        >=30&<=32 >=33&<=35 >=37&<=39 =255;
      tcp-flags [fin syn rst push ack urgent];
      packet-length =0 =40 =46 =201 =203 =205
        =206 =207 >=300&<=302 >=303&<=305
        >=307&<=309 =65535;
      dscp =0 =1 =3 =5 =6 =7 >=10&<=12
        >=13&<=15 >=17&<=19 =48 =63;
      fragment [ not-a-fragment dont-fragment
        is-fragment first-fragment
        last-fragment ];
    }
    then { accept; }
  }
}
```

next layer

The Lab



Wireshark / Packet Analysis



```
▶ Frame 2: 527 bytes on wire (4216 bits), 527 bytes captured (4216 bits)
▶ Juniper Ethernet
▶ Internet Protocol Version 4, Src: 10.5.10.2, Dst: 10.5.10.1
▶ Transmission Control Protocol, Src Port: 179, Dst Port: 62934, Seq: 188, Ack: 607, Len: 455
▶ Border Gateway Protocol - UPDATE Message
▶ Border Gateway Protocol - UPDATE Message
▶ Border Gateway Protocol - UPDATE Message
▼ [Malformed Packet: BGP]
  ▼ [Expert Info (Error/Malformed): Malformed Packet (Exception occurred)]
       [Malformed Packet (Exception occurred)]
       [Severity level: Error]
       [Group: Malformed]
▶ Border Gateway Protocol - KEEPALIVE Message
```

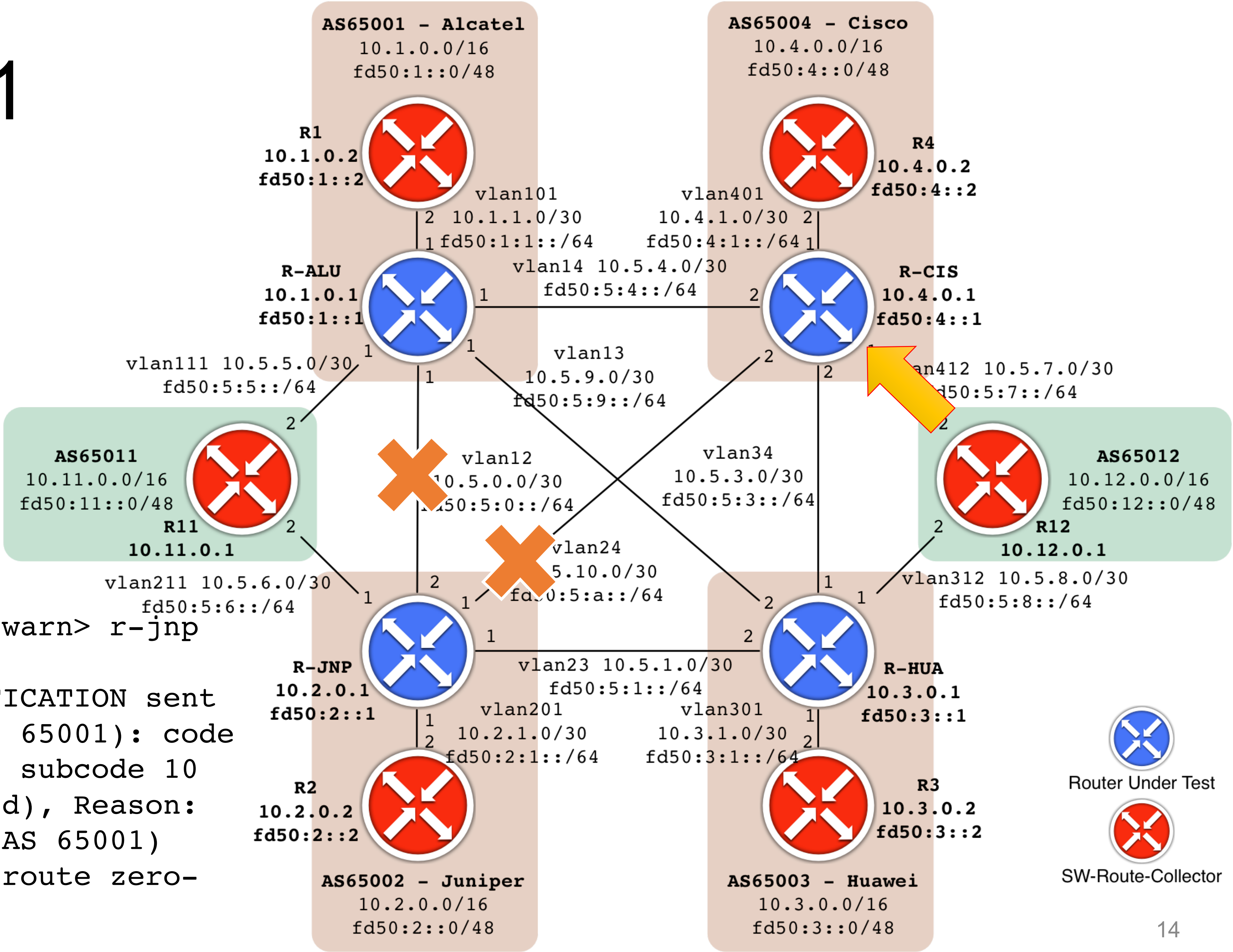
Wireshark BGP Dissector NLRI extended-length field

from RFC 5575 Section 4:

```
+-------------------------------+
| length (0xnn or 0xfn nn)      |
+-------------------------------+
| NLRI value (variable)         |
+-------------------------------+
         flow-spec NLRI
```

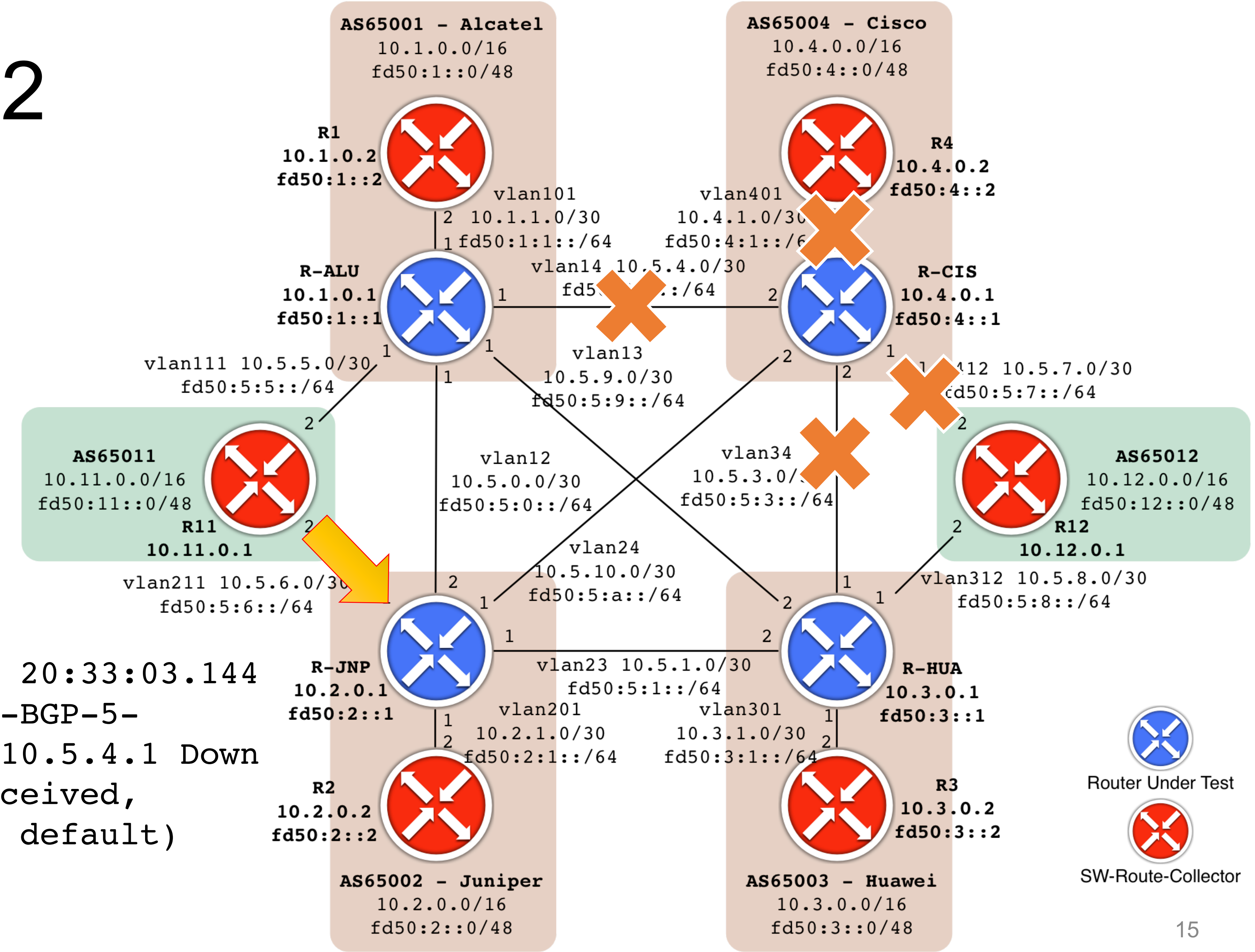
If the NLRI length value is smaller than 240 (0xf0 hex), the length field can be encoded as a single octet. Otherwise, it is encoded as an **extended-length 2-octet** value in which the most significant nibble of the first byte is all ones.

#1



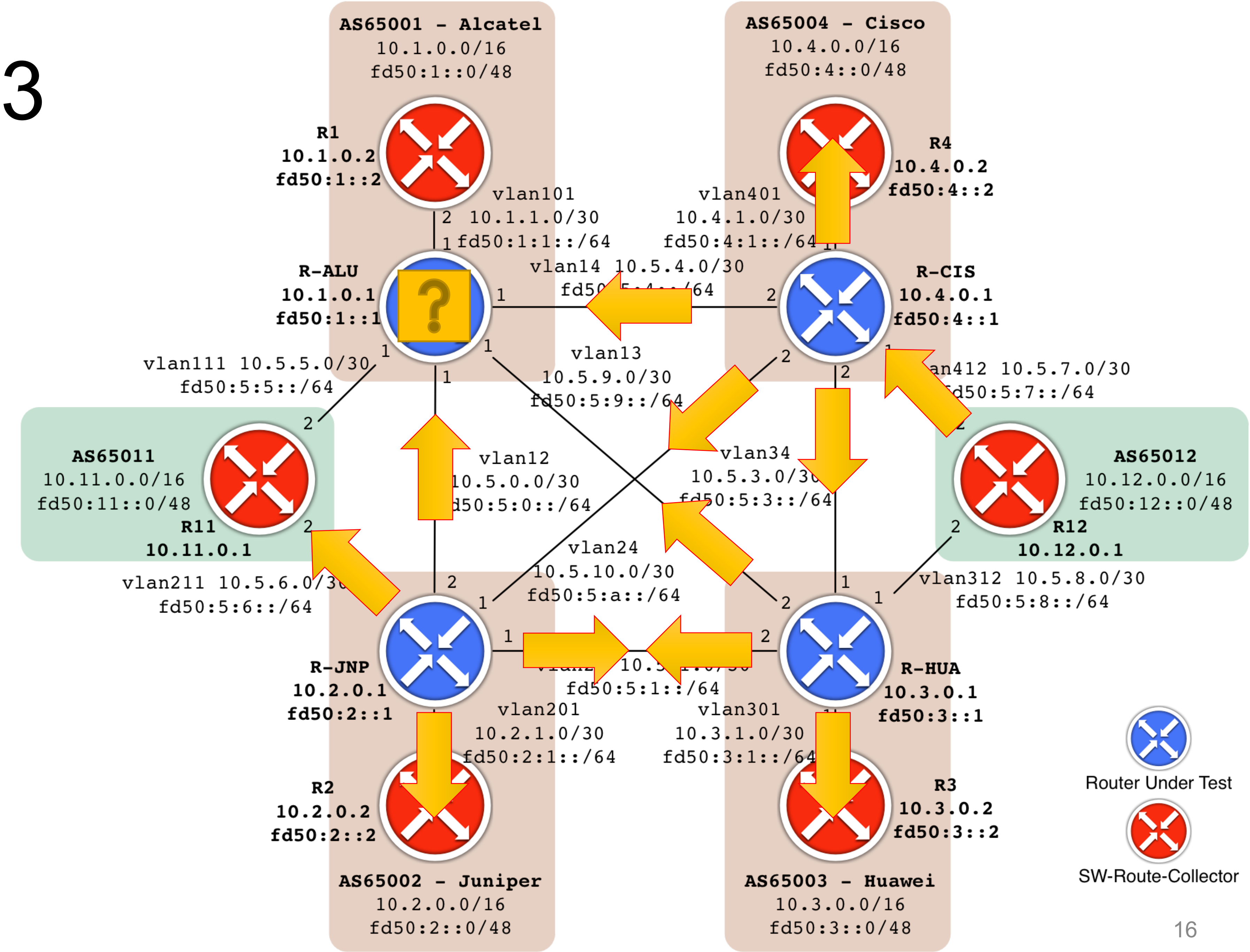
```
Jun 28 10:41:58 <daemon.warn> r-jnp
mx480-01-re1 rpd[14661]:
bgp_rcv_nlri:9989: NOTIFICATION sent
to 10.5.0.1 (External AS 65001): code
3 (Update Message Error) subcode 10
(bad address/prefix field), Reason:
peer 10.5.0.1 (External AS 65001)
update included invalid route zero-
len/0 (0 of 47)
```

#2

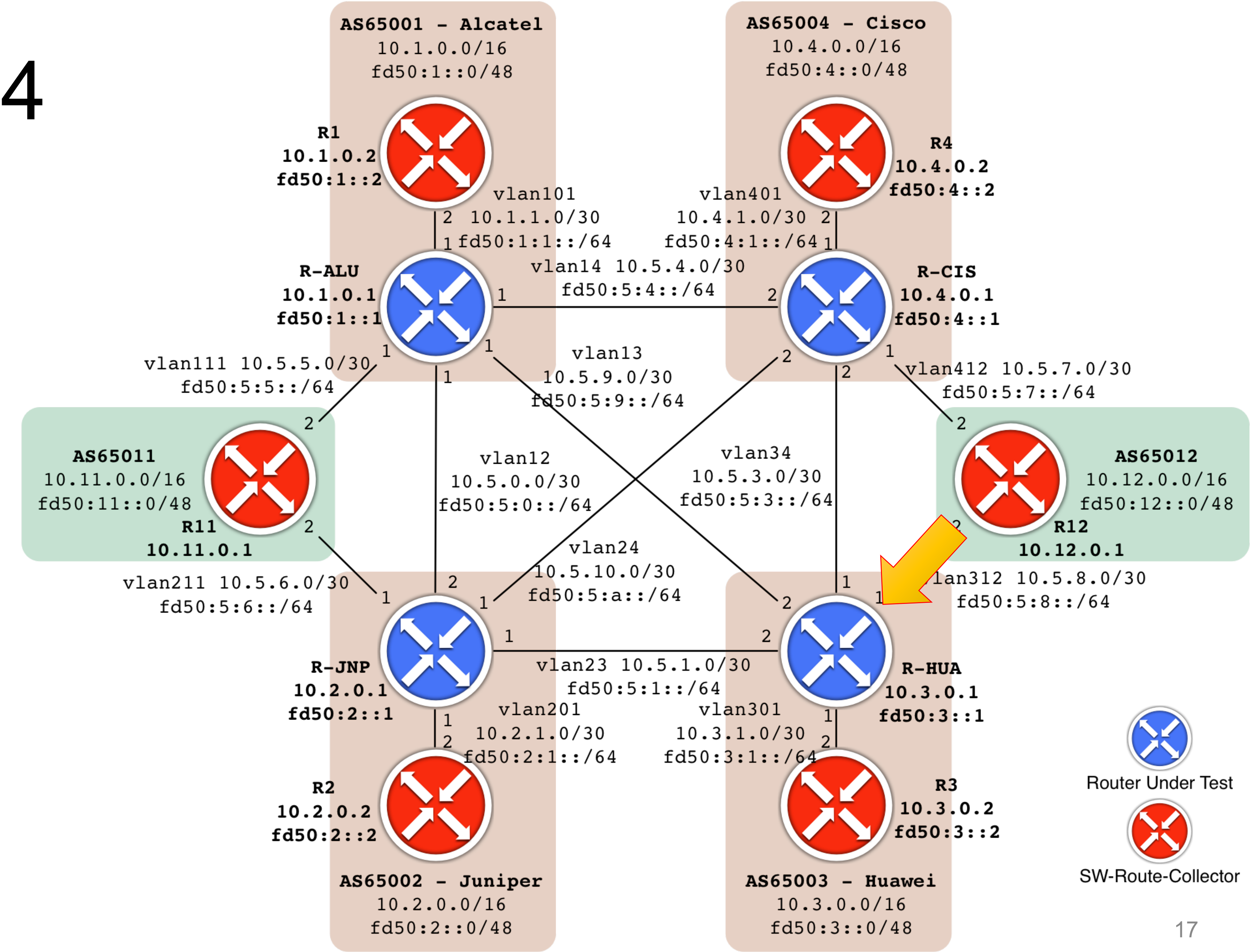


```
RP/0/RSP0/CPU0:Jul  5 20:33:03.144
: bgp[1058]: %ROUTING-BGP-5-
ADJCHANGE : neighbor 10.5.4.1 Down
- BGP Notification received,
illegal network (VRF: default)
(AS: 65001)
```

#3


AS65001 – Alcatel
10.1.0.0/16
fd50:1::0/48
R1
10.1.0.2
fd50:1::2
vlan101
10.1.1.0/30
fd50:1:1::/64
AS65004 – Cisco
10.4.0.0/16
fd50:4::0/48
R4
10.4.0.2
fd50:4::2
vlan401
10.4.1.0/30
fd50:4:1::/64
R-ALU
10.1.0.1
fd50:1::1
?
vlan14 10.5.4.0/30
fd50:5:4::/64
R-CIS
10.4.0.1
fd50:4::1
vlan13
10.5.9.0/30
fd50:5:9::/64
vlan111 10.5.5.0/30
fd50:5:5::/64
vlan412 10.5.7.0/30
fd50:5:7::/64
AS65011
10.11.0.0/16
fd50:11::0/48
R11
10.11.0.1
vlan12
10.5.0.0/30
fd50:5:0::/64
vlan34
10.5.3.0/30
fd50:5:3::/64
AS65012
10.12.0.0/16
fd50:12::0/48
R12
10.12.0.1
vlan211 10.5.6.0/30
fd50:5:6::/64
vlan24
10.5.10.0/30
fd50:5:a::/64
vlan312 10.5.8.0/30
fd50:5:8::/64
R-JNP
10.2.0.1
fd50:2::1
vlan23 10.5.1.0/30
fd50:5:1::/64
R-HUA
10.3.0.1
fd50:3::1
vlan201
10.2.1.0/30
fd50:2:1::/64
vlan301
10.3.1.0/30
fd50:3:1::/64
R2
10.2.0.2
fd50:2::2
R3
10.3.0.2
fd50:3::2
AS65002 – Juniper
10.2.0.0/16
fd50:2::0/48
AS65003 – Huawei
10.3.0.0/16
fd50:3::0/48
Router Under Test
SW-Route-Collector

#4



AS65001 – Alcatel
10.1.0.0/16
fd50:1::0/48

R1
10.1.0.2
fd50:1::2

vlan101
10.1.1.0/30
fd50:1:1::/64

R-ALU
10.1.0.1
fd50:1::1

AS65004 – Cisco
10.4.0.0/16
fd50:4::0/48

R4
10.4.0.2
fd50:4::2

vlan401
10.4.1.0/30
fd50:4:1::/64

R-CIS
10.4.0.1
fd50:4::1

vlan14 10.5.4.0/30
fd50:5:4::/64

vlan13
10.5.9.0/30
fd50:5:9::/64

vlan111 10.5.5.0/30
fd50:5:5::/64

vlan412 10.5.7.0/30
fd50:5:7::/64

AS65011
10.11.0.0/16
fd50:11::0/48
R11
10.11.0.1

vlan12
10.5.0.0/30
fd50:5:0::/64

vlan34
10.5.3.0/30
fd50:5:3::/64

AS65012
10.12.0.0/16
fd50:12::0/48
R12
10.12.0.1

vlan211 10.5.6.0/30
fd50:5:6::/64

vlan24
10.5.10.0/30
fd50:5:a::/64

lan312 10.5.8.0/30
fd50:5:8::/64

R-JNP
10.2.0.1
fd50:2::1

vlan23 10.5.1.0/30
fd50:5:1::/64

R-HUA
10.3.0.1
fd50:3::1

vlan201
10.2.1.0/30
fd50:2:1::/64

vlan301
10.3.1.0/30
fd50:3:1::/64

R2
10.2.0.2
fd50:2::2

R3
10.3.0.2
fd50:3::2

AS65002 – Juniper
10.2.0.0/16
fd50:2::0/48

AS65003 – Huawei
10.3.0.0/16
fd50:3::0/48

Router Under Test

SW-Route-Collector

Issue #5 – Unclear Specification Transitivity of Action Communities

All firmwares tested implemented all action communities as transitive.

- IANA assigned the extend communities from a transitive pool
- RFC 5575 defines the traffic-rate action as non-transitive
- Transitivity of the other actions not defined in RFC 5575
- All implementation violate RFC 5575

Test Summary

- Found some bugs (unlikely that we found all of them)
  - Goal was not a complete feature test, but to come up with stable/usable inter AS configuration
- Found different interpretations of RFC 5575
  - Ranging from unpredictable flow-spec propagation, to BGP flaps
- Discussed all bugs and problems with manufacturers
  - Many bugs/problems already fixed or on a roadmap
  - Very cooperative even though RFC 5575 sometimes unclear

nextlayer

Missing Features

- BGP import / export policies
  - (policy-statement, route-map)
  - Match on flow-spec components
  - Modify/delete/filter actions
  - Filter updates
- Flow-spec for IPv6 Flowspec
  - only an IETF draft available!
- Flowspec in a VRF
  - RFC 5575 based

Conclusion

- Testing took longer than expected!
- Incompatible NLRI decoding
  - Leading to major network instabilities (BGP notification)
  - High risk in inter AS setting – no filtering possible!
- Absence BGP export/import filters
  - showstopper for inter AS deployments
  - remote network may redirect packets in any VRF or modify QoS
- RFC 5575 unclear sections
  - Implementations follow RFC with their own interpretation
  - Hardly any multi manufacturer testing results available
- If you exchange BGP Flowspec with external peers, be careful!

draft-ietf-idr-rfc5575bis

- Clarify unclear sections
  - Encoding of flow types
  - Traffic redirect community encoding
- Redefines all flow action communities as transitive
- New section on flow action interference
- Adding traffic-rate-packets action

  May be out of scope and removed (other draft available that specifies that action)
- Adopted by IETF IDR WG

  Inter Domain Routing – Working Group
- Patches in GoBGP, ExaBGP

Questions?

christoph.loibl@nextlayer.at

https://www.nextlayer.at/flowspec-paper.pdf
https://datatracker.ietf.org/doc/draft-ietf-idr-rfc5575bis/