Recent BGP Innovations for Operational Challenges

William McCall

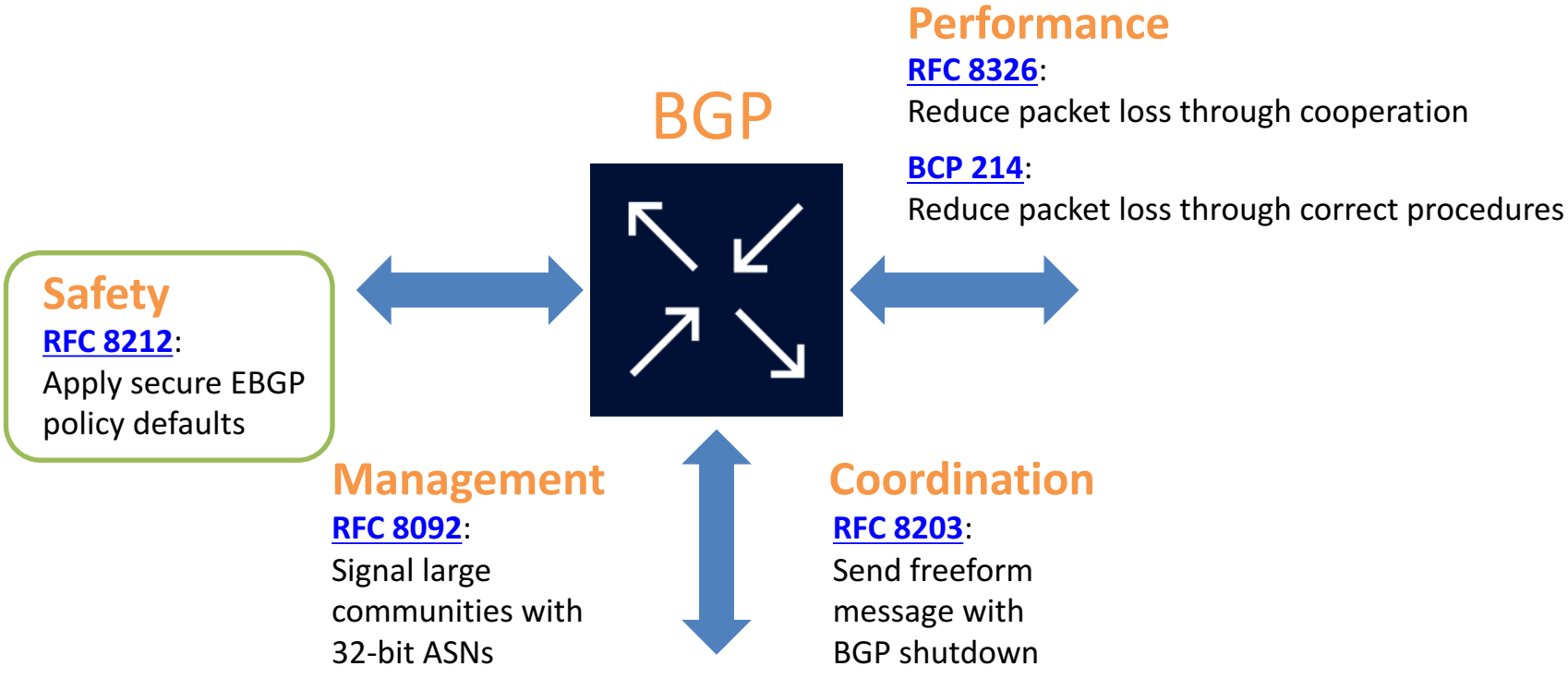
wmccall@gin.ntt.net

NTT Ltd.

Background

- There's been increased participation by operators in the IETF recently to standardize solutions to operational challenges with BGP
  - **IDR** (Inter-Domain Routing) Working Group
  - **GROW** (Global Routing Operations) Working Group
- Several RFCs have been published, and several I-Ds are in the standardization process
  - Operators and implementers are working on solutions together in the WGs
- This presentation provides an overview of some of the recent innovations in BGP
- It's never too late to participate, join the **IDR** and **GROW** mailing lists!
  - https://www.ietf.org/wg/

Agenda

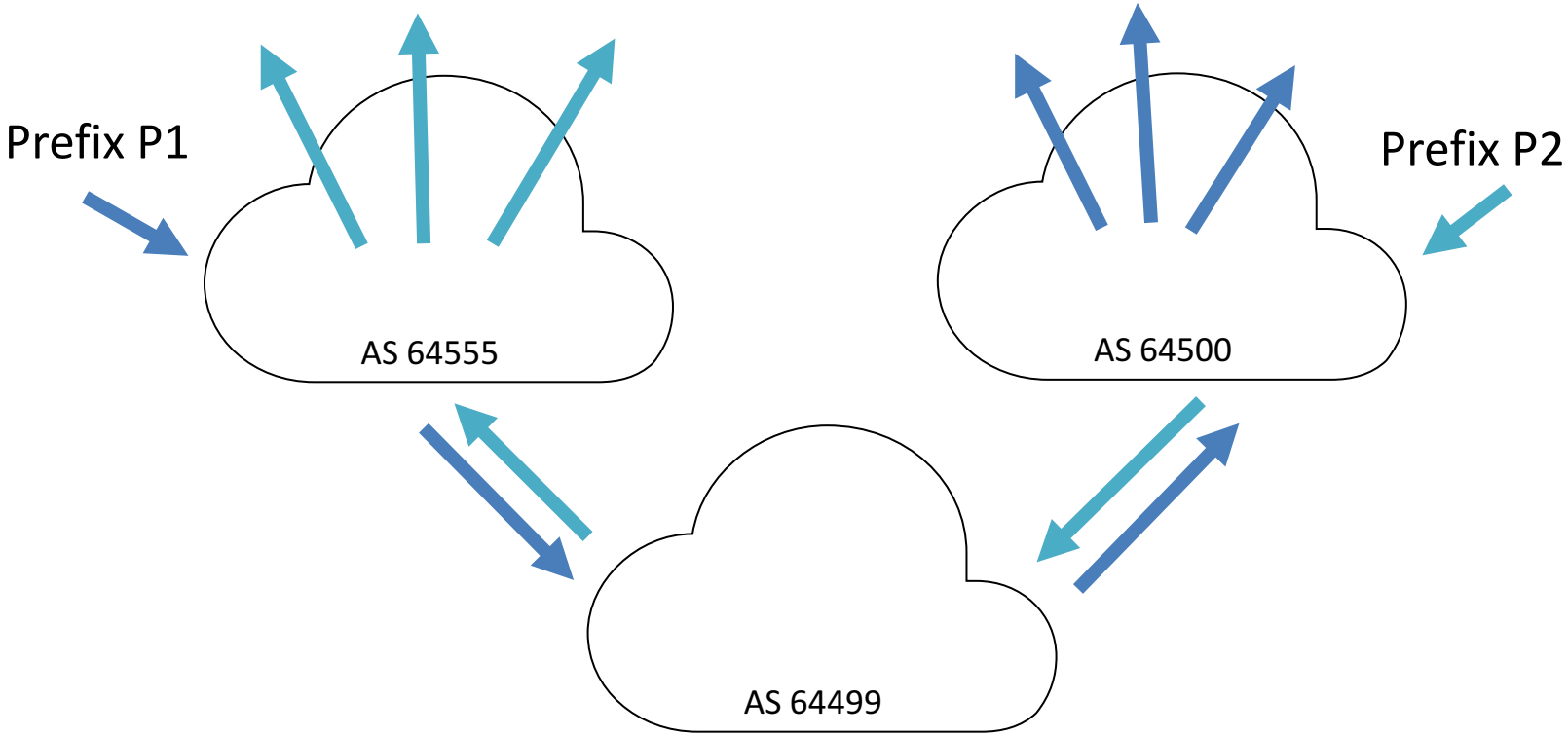

Performance
RFC 8326:
Reduce packet loss through cooperation
BCP 214:
Reduce packet loss through correct procedures
BGP
Safety
RFC 8212:
Apply secure EBGP policy defaults
Management
RFC 8092:
Signal large communities with 32-bit ASNs
Coordination
RFC 8203:
Send freeform message with BGP shutdown

RFC 8212 “Default External BGP (EBGP) Route Propagation Behavior without Policies”

What does this configuration do?

```
router bgp 64499
   !
   neighbor 192.0.2.1 remote-as 64555
   neighbor 192.0.2.1 description Upstream 1
   !
   neighbor 192.0.2.5 remote-as 65444
   neighbor 192.0.2.5 description Upstream 2
!
```

Lateral AS-AS-AS Leak



RFC 8212 in a Nutshell

Opponents Argued

- “We can’t change defaults”
- “It can’t be done”
- ”It will break everything we love and know”
- Customers don’t read release notes
  - And don’t test whether the software boots
    - And deploy new software absolutely everywhere at once
      - And don’t follow community mailing lists
        - » And don’t talk to each other
          - …..

Post-RFC 8212 implication (hypothetical)

```
route-map implicit-deny-all deny 1
!
router bgp 64499
  !
  neighbor 192.0.2.1 remote-as 64555
  neighbor 192.0.2.1 description Upstream 1
  neighbor 192.0.2.1 route-map implicit-deny-all in
  neighbor 192.0.2.1 route-map implicit-deny-all out
  !
  neighbor 192.0.2.5 remote-as 65444
  neighbor 192.0.2.5 description Upstream 2
  neighbor 192.0.2.5 route-map implicit-deny-all in
  neighbor 192.0.2.5 route-map implicit-deny-all out
```

Advantages of RFC 8212

- Protects your network
  - Consistency across platforms & vendors
  - Explicit configuration (‘`grep`’ suddenly is useful again)
  - Handover between personnel is easier as we don’t have to guess
- Protects the Default-Free Zone (EBGP is a shared resource)
  - Because relying on others will break someday

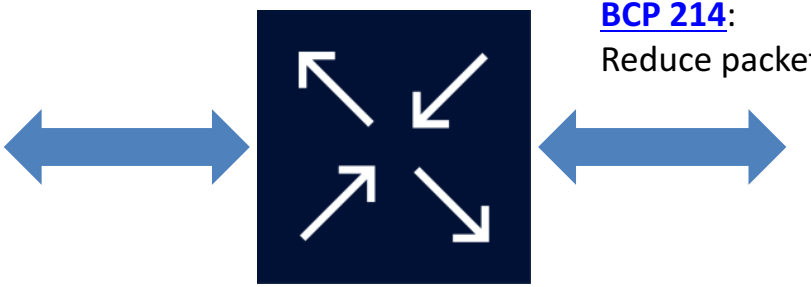
What This Means

- BGP speakers that announce routes and/or accept routes, without explicitly being configured to do so, **are no longer compliant with the core BGP specification**
- Fixed Vendors
  - BIRD
  - OpenBGPD
  - Nokia SR OS
- Current list of vendors that need to do some work
  - Cisco IOS
  - Cisco IOS XE
  - Cisco NX-OS
  - Arista EOS
  - Juniper Junos OS
  - Brocade Ironware
- We're keeping track here, with workarounds https://github.com/bgp/RFC8212

Usage Guidelines

- Start to implement a routing policy with secure EBGP defaults now
  - It's the right thing to do and now is a good time to start
- Keep an eye out for when your BGP implementations change their default behavior
  - Check release notes and documentation
- Following these steps will ensure you are prepared in advance

Agenda



Performance

RFC 8326:
Reduce packet loss through cooperation

BCP 214:
Reduce packet loss through correct procedures

Safety

RFC 8212:
Apply secure EBGP policy defaults

Management

RFC 8092:
Signal large communities with 32-bit ASNs

Coordination

RFC 8203:
Send freeform message with BGP shutdown

Needed RFC 1997 Style Communities, but Larger

- ASN:Community form
- We knew we'd run out of 16-bit ASNs eventually and came up with 32-bit ASNs
- RIRs started allocating 32-bit ASNs by request in 2007, no distinction between 16-bit and 32-bit ASNs now
- However, you can't fit a 32-bit value into a 16-bit field
- Can't use native 32-bit ASNs with RFC 1997 communities
- Needed an Internet routing communities solution for 32-bit ASNs for almost 10 years
- Parity so everyone can use their globally unique ASN

RFC 8092 “BGP Large Communities Attribute”

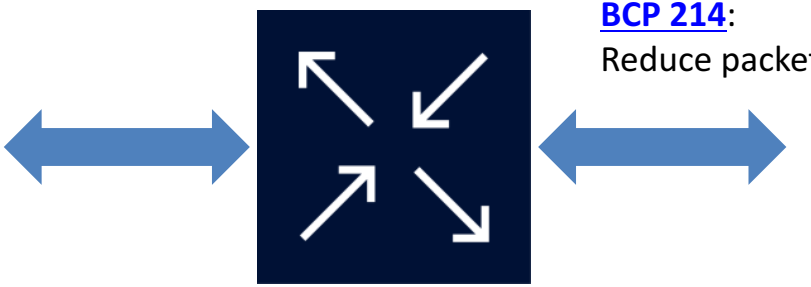
- Idea progressed rapidly from inception in March 2016
- First I-D in September 2016 to RFC publication on February 16, 2017 in just seven months
- Final standard, plus a number of implementation and tools developed as well
- Network operators can test and deploy the new technology now

Cake and photo courtesy of the NTT Communications NOC.

Getting Started With Large Communities

- 2018 is the year of large BGP communities
  - Preparation, testing, training and deployment can take weeks, months or even over a year
  - Start the work now, so you are ready when customers want to use large communities
- Lots of resources are available to help network operators learn about large communities at http://largebgpcommunties.net/
  - BGP speaker implementations
  - Analysis and ecosystem tools
  - Presentations (http://largebgpcommunities.net/talks/)
  - Documentation for each implementation
  - Configuration examples (http://largebgpcommunities.net/examples/)
  - RFC 8195 provides examples and inspiration for network operators to use large communities

Agenda



Performance

RFC 8326:
Reduce packet loss through cooperation

BCP 214:
Reduce packet loss through correct procedures

Safety

RFC 8212:
Apply secure EBGP policy defaults

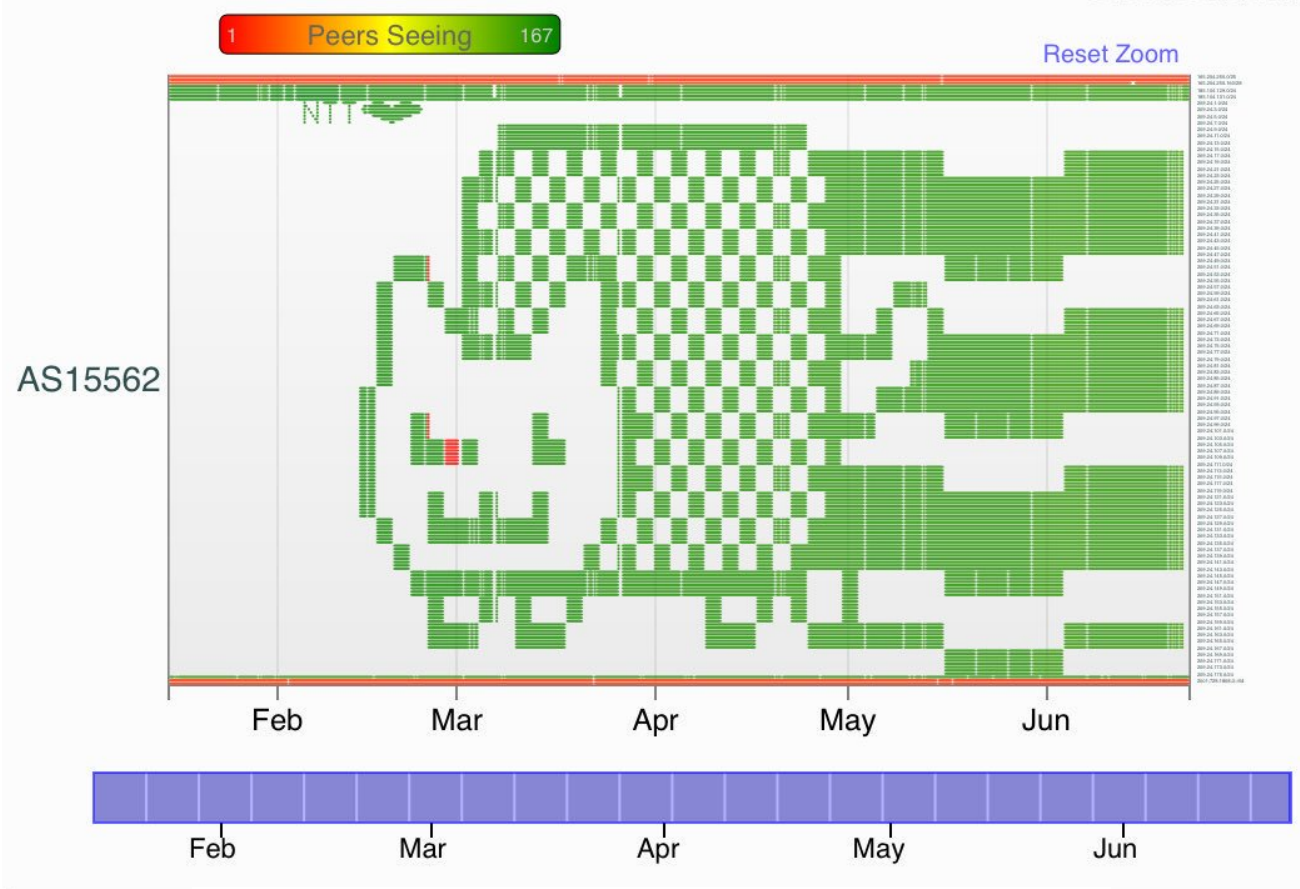
Management

RFC 8092:
Signal large communities with 32-bit ASNs

Coordination

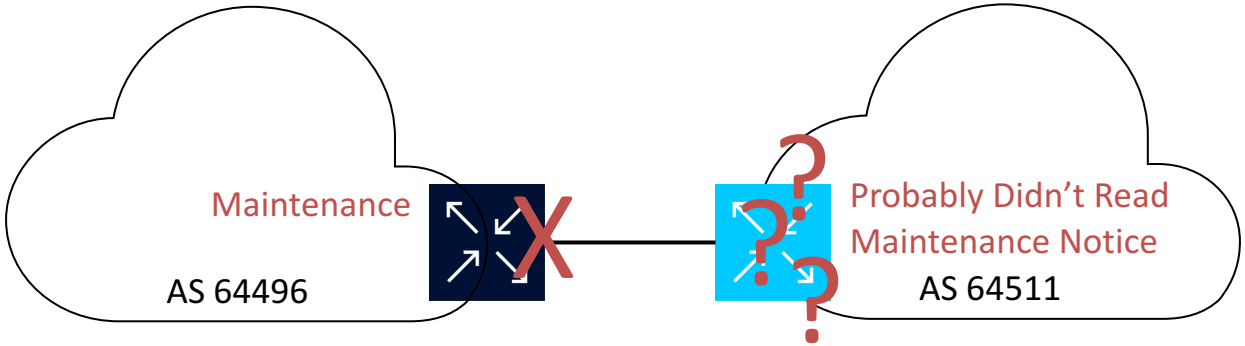
RFC 8203:
Send freeform message with BGP shutdown

Communication can be a Challenge



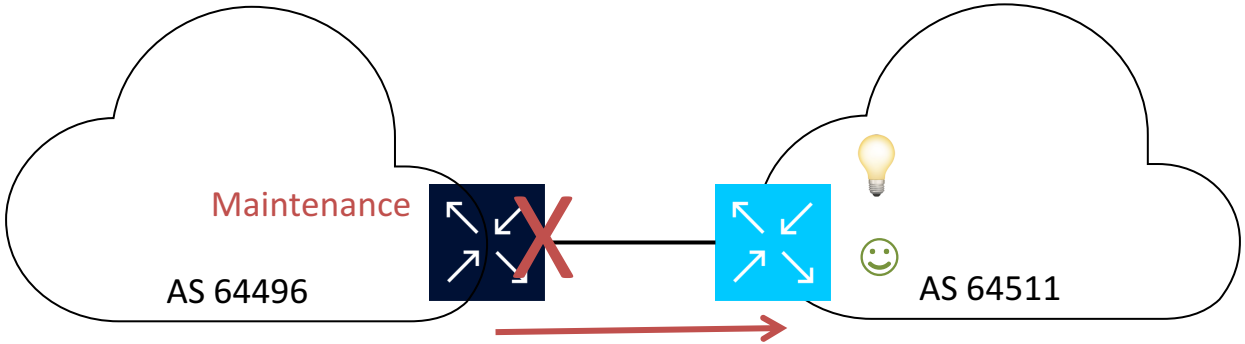
https://labs.ripe.net/Members/cteusche/bgp-meets-cat

RFC 8203
"BGP Administrative Shutdown Communication"



- Coordination problem: you shutdown your BGP session and your peers don't know why
- Solution: add a freeform message embedded in the BGP NOTIFICATION message when the session is shutdown

RFC 8203
“BGP Administrative Shutdown Communication”



- Message can be up to 128 bytes long
- UTF-8 is supported too: 💩🦄😍😡👫🐱👬

Usage Guidelines

Sender

- Send “Administrative Shutdown” message for maintenance that is going to take some period of time
- Send “Administrative Reset” message for maintenance that is for a short time, for example to reset a peer or to reboot a router
- Include a ticket or reference number and make the message as informative as possible

Receiver

- Log messages to logging systems
- Reference ticket number in email or other notifications for more details

OpenBGPD Example

Sender:

```
[job@kiera ~]$ bgpctl neighbor 165.254.255.24 down "[TICKET-1-
1438367390] we are upgrading to openbsd 6.1, be back in 30 minutes"
[job@kiera ~]$
```

Receiver:

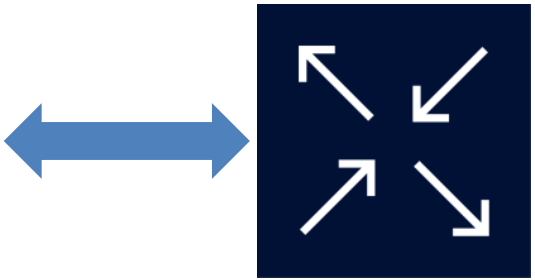
```
Jan  8 19:28:54 shutdown bgpd[50719]: neighbor 165.254.255.26:
received notification: Cease, administratively down

Jan  8 19:28:54 shutdown bgpd[50719]: neighbor 165.254.255.26:
received shutdown reason: "[TICKET-1-1438367390] we are upgrading to
openbsd 6.1, be back in 30 minutes"
```

Implementation Status

Implementation	Software	Status
cz.nic	BIRD	Unknown
Cisco	IOS XR	Unknown
ExaBGP	ExaBGP	✔ Done!
FreeRangeRouting	frr	✔ Done!
OSRG	GoBGP	✔ Done!
Juniper	Junos OS	Unknown
Nokia	SR OS	Unknown
OpenBSD	OpenBGPD	✔ Done!
OSRG	GoBGP	✔ Done!
pmacct.net	pmacct	✔ Done!
tcpdump.org	tcpdump	✔ Done!
Wireshark	Dissector	✔ Done!

Agenda



Performance

RFC 8326:
Reduce packet loss through cooperation

BCP 214:
Reduce packet loss through correct procedures

Safety

RFC 8212:
Apply secure EBGP policy defaults

Management

RFC 8092:
Signal large communities with 32-bit ASNs

Coordination

RFC 8203:
Send freeform message with BGP shutdown

Two Types of Maintenance

Voluntary Shutdown (YOU)

- You take action before maintenance to reroute traffic and minimize the impact
- ➢ You use BGP shutdown communication
- ➢ You use graceful BGP session shutdown

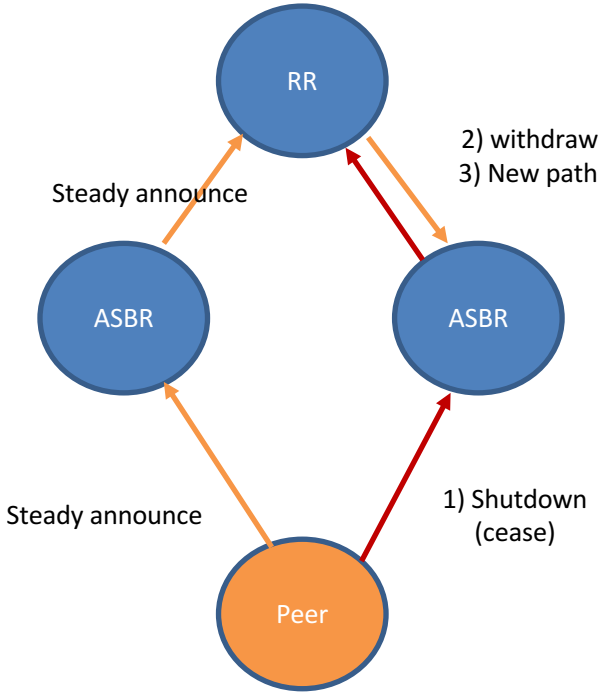
Involuntary Shutdown (other folks)

- Maintenance on lower layer network breaks end-to-end path, but link stays up
- BGP sessions only go down after hold timer expires
- Could blackhole traffic during this time until traffic is rerouted
- ➢ Your network provider uses BGP culling

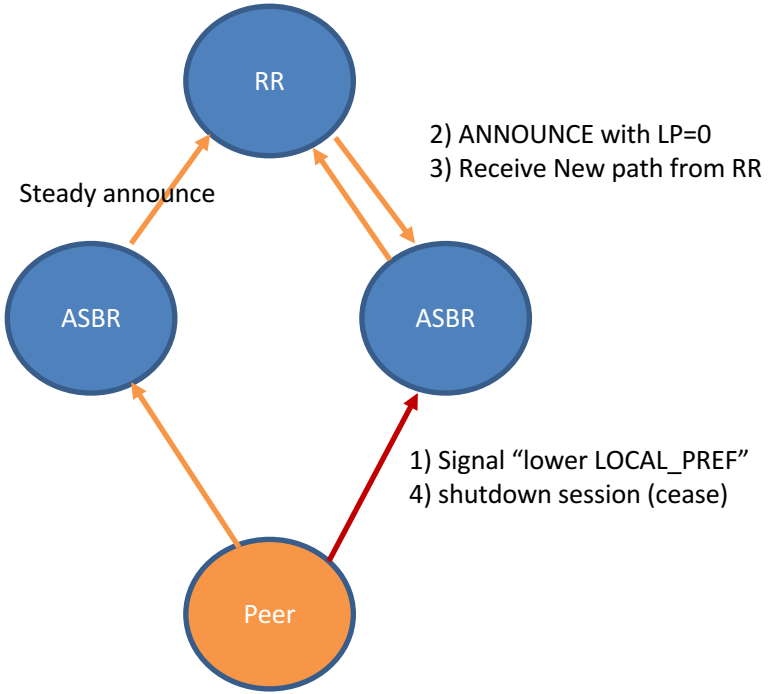
When does blackholing happen with vanilla shutdown?

- Lack of an alternative route on some routers
- Transient routing inconsistency
- A route reflector may only propagate its best path
- The backup ASBR may not advertise the backup path because the nominal path is preferred

Admittedly, the above scenarios usually are short periods of blackholing, but why accept that if they can easily be prevented?



Graceful Shutdown triggers “path hunting”



- Initiated by the operator on the router before maintenance by sending the GRACEFUL_SHUTDOWN well-known community (65535:0 as per IANA)
- Receiving EBGP peer sets LOCAL_PREFERENCE to 0 and selects paths to route traffic away from the initiator, (similar to setting overload in an ISIS)
- When BGP session goes down, minimizes impact to traffic because alternate paths have already been installed

Usage Guidelines

- To support receiving graceful shutdown, update your routing policy to
  - Match the GRACEFUL_SHUTDOWN well-known community (65535:0)
  - Set the LOCAL_PREF attribute to a low value, like 0
- To send graceful shutdown, update your routing policy to
  - Send the GRACEFUL_SHUTDOWN well-known community (65535:0) before you start maintenance
  - When ingress traffic from the peer has stopped, start maintenance and use BGP shutdown communication
  - Remove the GRACEFUL_SHUTDOWN well-known community when you are done

Configuration Example – Simple to Implement

IOS XR

```
route-policy AS64497-ebgp-inbound
  if community matches-any (65535:0) then
    set local-preference 0
  endif
end-policy
!
router bgp 64496
  neighbor 2001:db8:1:2::1
  remote-as 64497
  address-family ipv6 unicast
    send-community-ebgp
    route-policy AS64497-ebgp-inbound in
```

Arista/Brocade/IOS/Quagga/FRR

```
ip community-list standard gshut 65535:0
!
route-map ebgp-in permit 10
  match community gshut
  set local-preference 0
```

Nokia

```
community "gshut" members "65535:0"
policy-statement "ebgp-in"
    entry 10
        from
            community "gshut"
        exit
        action accept
            local-preference 0
        exit
    exit
exit
```

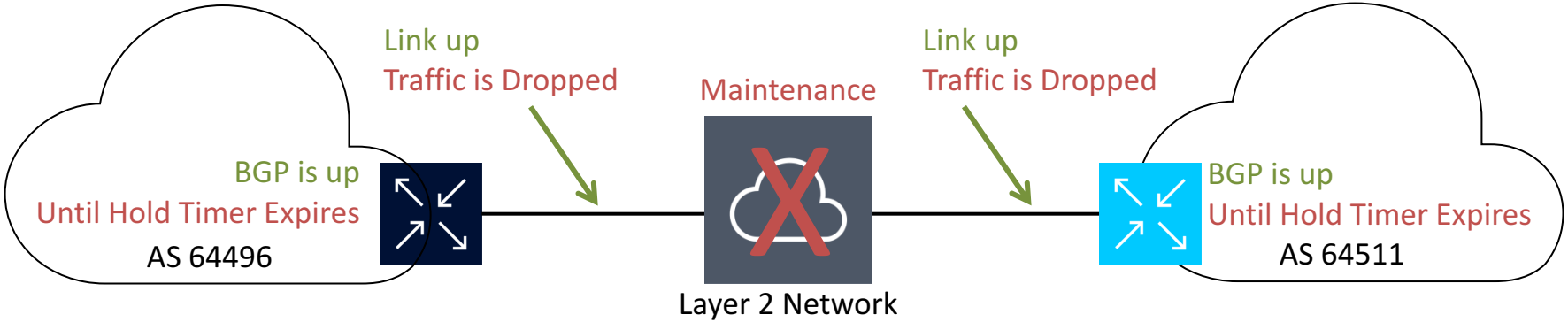
GRACEFUL_SHUTDOWN signals:

“Hello everyone, if you consider this path your ‘best path’, please start considering this path the ’worst path’ and if you find anything better install that into your FIB. This path will disappear within a few minutes.”

Operators known to honor the graceful_shutdown well-known community:

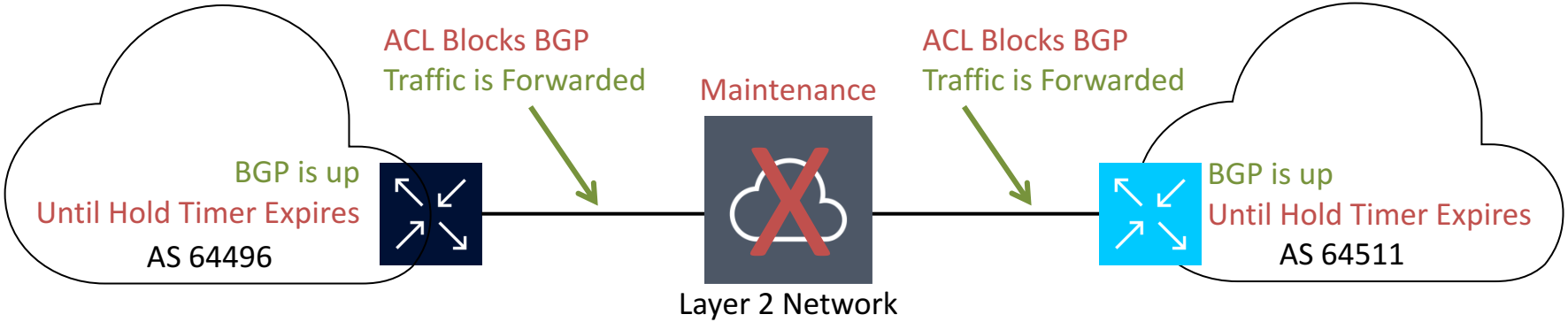
- NTT (AS 2914)
- GTT (AS 3257)
- Github (AS 36459)
- Nordunet (AS 2603)
- Coloclue (AS 8283)
- Amsio (AS 8315)
- BIT (AS 12859)
- Telia (AS 3301/1299)
- Tele2 (AS 1257)
- SVT (AS 201641)
- Netnod (AS 8674)
- Bahnhof (AS 8473)
- DGC Systems (AS 21195)
- ComHem (AS 39651)
- … you? ☺

BCP 214
"Mitigating Negative Impact of Maintenance through BGP Session Culling"



- Performance problem: maintenance on lower layer network breaks path, but link stays up
- Solution: network provider applies Layer 4 ACLs to block BGP control plane traffic while links are up
- Routers continue to forward traffic until hold timer expires, no blackholing

BCP 214
"Mitigating Negative Impact of Maintenance through BGP Session Culling"



- Lower layer network provider applies Layer 4 ACLs to block BGP control plane traffic before maintenance starts
- Data plane continues to forward
- When BGP hold timer expires, BGP chooses a new path
- Then lower layer network starts maintenance, and removes ACLs when maintenance is complete

“Involuntary Teardown” Usage Guidelines

- ACLs are only applied to TCP/179 on directly connected IP addresses
  - Multihop BGP control plane traffic is permitted
  - Data plane traffic is permitted
- ACLs are applied to IPv4 and IPv6 IP addresses
- Maintenance is started when data plane traffic has stopped or dropped significantly
- ACLs are removed after maintenance

Availability Overview

- **Config changes:**
  - Graceful Shutdown
  - BGP Session Culling
- **Code changes:**
  - Large BGP Communities
  - Shutdown Communication
  - EBGP Secure Defaults

Call to Action

Your BGP Software Suppliers

- Ask them to support the following RFCs **now**, even if it's already listed on their roadmap
  - RFC 8092 BGP Large Communities
  - RFC 8203 BGP Administrative Shutdown Communication
  - RFC 8212 Default EBGP Route Propagation Behavior without Policies
- When you write a Request For Proposals (RFP), make sure these three items are on the checklist
- **Vote with your wallet**

Your Peers, Transit Providers & IXPs:

- Ask your transit providers & peers to support
  - RFC 8326 Graceful BGP session shutdown
  - BCP 214 Voluntary Shutdown BCP
- Ask IXPs to apply BGP culling (or equivalent) during maintenance
  - BCP 214 (Involuntary Shutdown BCP) - Mitigating Negative Impact of Maintenance through BGP Session Culling
- When you write a Request For Proposals (RFP), make sure these three items are on the checklist. **PUT THIS IN RFPs!**
- Vote with your wallet

Your Network

- Update your routing policy
  - Assume Secure EBGP defaults
  - GRACEFUL_SHUTDOWN well-known community (65535:0)
  - Large communities
  - Document and publish it
- Add coordination and performance improvements to your maintenance procedures
  - Shutdown communication and BGP graceful shutdown
  - Follow BGP session culling BCP

Movie Credits

(contributors to RFC 8092, 8195, 8203, 8212)

Acee Lindem
Adam Chappell
Adam Davenport
Adam Roach
Adam Simpson
Alexander Azimov
Alvaro Retana
Arjen Zonneveld
Arnold Nipper
Barry O'Donovan
Ben Maddison
Bertrand Duvivier
Bill Fenner
Brad Dreisbach
Brian Dickson
Bruno Decraene
Christoph Dietzel
Christopher Morrow
Dale Worley
David Farmer
David Freedman
Donald Smith
Duncan Lockwood
Eduardo Ascenco Reis
Gaurab Raj Upadhaya
Geoff Huston
Gert Doering
Greg Hankins
Greg Skinner
Grzegorz Janoszka
Gunter van de Velde
Ian Dickinson
Ignas Bagdonas
Jakob Heitz
James Bensley
Jan Baggen
Jared Mauch
Jay Borkenhagen
Jeff Haas
Jeff Tantsura
Jeffrey Haas
Job Snijders
Joe Provo
Joel M. Halpern
John Heasley
John Scudder
Jonathan Stewart
Julian Seifert
Jussi Peltola
Kay Rechthien
Keyur Patel
Kristian Larsson
Linda Dunbar
Lou Berger
Mach Chen
Marco Davids
Marco Marzetti
Mark Schouten
Markus Hauschild
Martijn Schmidt
Martin Millnert
Mikael Abrahamsson
Nabeel Cocker
Nick Hilliard
Niels Bakker
Paul Hoogsteder
Peter Hessler
Peter van Dijk
Pier Carlo Chiodi
Randy Bush
Remco van Mook
Richard Hartmann
Richard Steenbergen
Richard Turkbergen
Rob Shakir
Robert Raszuk
Ruediger Volk
Russ White
Saku Ytti
Sander Steffann
Shane Amante
Shawn Morris
Shyam Sethuram
Sriram Kotikalapudi
Stefan Plug
Stewart Bryant
Susan Hares
Teun Vink
Theodore Baschak
Thomas King
Tom Daly
Tom Petch
Tom Scholl
Warren Kumari
Wesley Steehouwer
Will Hargrave
Wim Henderickx

Presentation created by:

Greg Hankins
Nokia
greg.hankins@nokia.com
@greg_hankins

Job Snijders
NTT Communications
job@ntt.net
@JobSnijders

Reuse of this slide deck is permitted and encouraged!

Bonus slides

The Science Behind Shutting Down BGP Sessions

- Avoiding disruptions during maintenance operations on BGP sessions: https://inl.info.ucl.ac.be/system/files/ucl-ft-bgp-shutdown-inl.pdf (August 2008)
- Requirements for the Graceful Shutdown of BGP Sessions https://tools.ietf.org/html/rfc6198 (April 2011)