Recent BGP Innovations for Operational Challenges

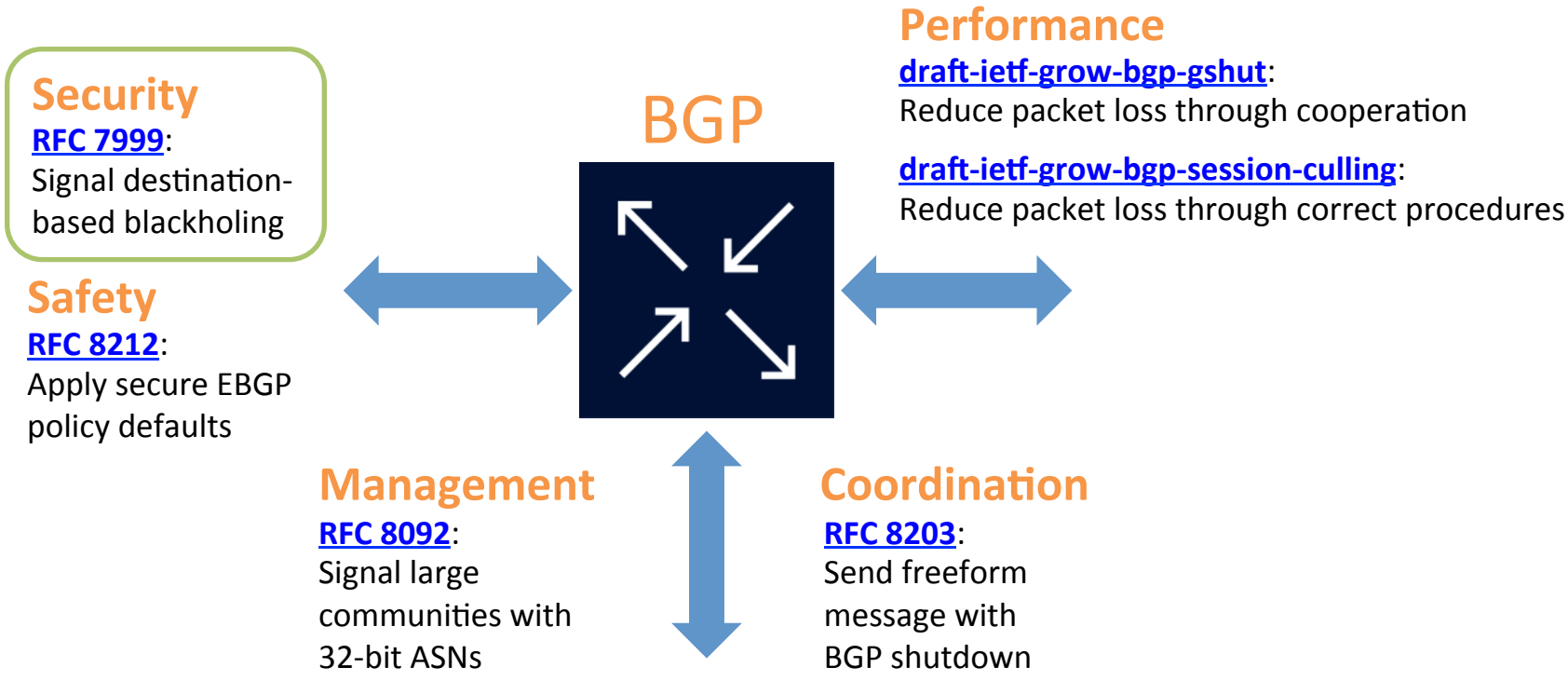
Job Snijders
job@ntt.net

Battle of Operations 2016 - 2017

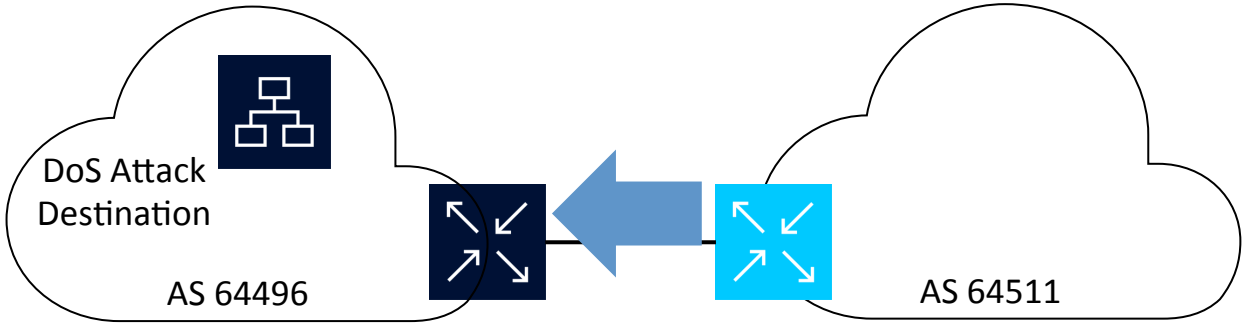
Background

- 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

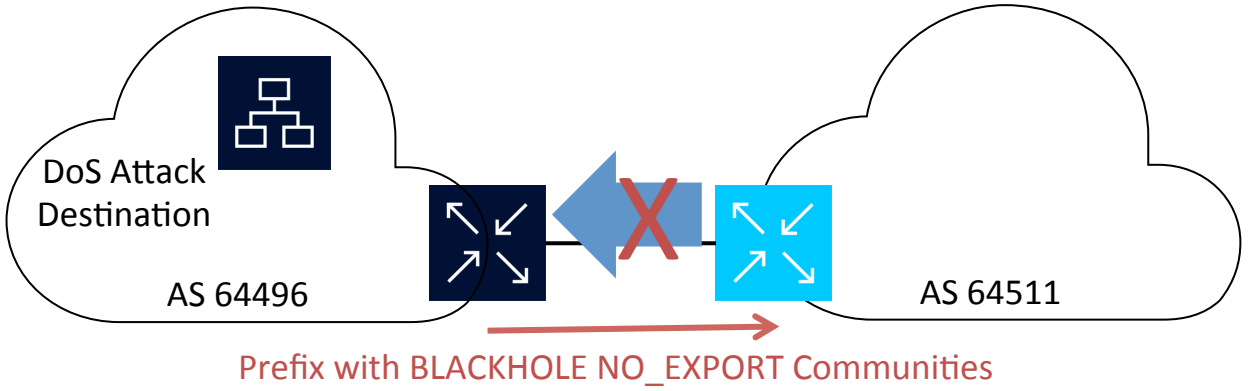

Security
RFC 7999:
Signal destination-based blackholing
Safety
RFC 8212:
Apply secure EBGP policy defaults
BGP
Performance
draft-ietf-grow-bgp-gshut:
Reduce packet loss through cooperation
draft-ietf-grow-bgp-session-culling:
Reduce packet loss through correct procedures
Management
RFC 8092:
Signal large communities with 32-bit ASNs
Coordination
RFC 8203:
Send freeform message with BGP shutdown

RFC 7999 “BLACKHOLE Community”



- Security problem: DoS attacks or unwanted traffic comes into your AS and fills your transit links
- You can block it at your AS borders, but that still wastes transit capacity
- Solution: new optional well-known community to signal destination-based blackholing

RFC 7999 “BLACKHOLE Community”



- Advertise prefix with BLACKHOLE community (65535:666)
- Peer AS honors community and drops traffic to this prefix
- Remove BLACKHOLE community when the attack is over

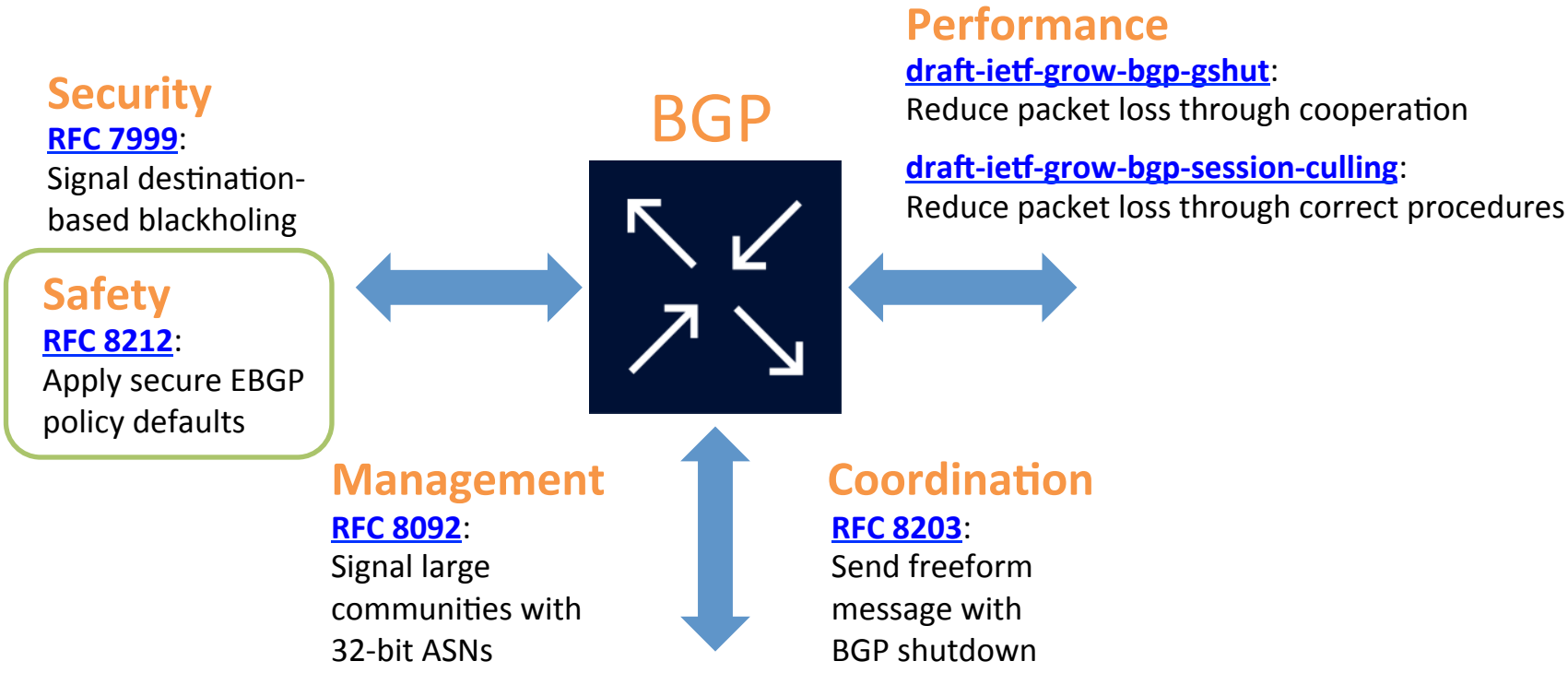
Usage Guidelines

- You have the choice to accept and honor, or to ignore the community
- Usage must be agreed on first between two ASs, and route policies must be configured
- You should strip the community if you aren't using it
- Propagation should be limited to the neighboring AS only, add the NO_ADVERTISE or NO_EXPORT community
- Prefix length is typically as specific as possible, /32 for IPv4 or /128 for IPv6

Security Guidelines

- You should only accept and honor the BLACKHOLE community if
  - The prefix is covered by an equal or shorter prefix that the neighboring AS is authorized to advertise
  - You both agreed to honor the BLACKHOLE community on the particular BGP session
- Route policies must be explicitly configured to drop traffic with the BLACKHOLE community, it does not happen automagically

Agenda


Security
RFC 7999:
Signal destination-based blackholing
Safety
RFC 8212:
Apply secure EBGP policy defaults
BGP
Performance
draft-ietf-grow-bgp-gshut:
Reduce packet loss through cooperation
draft-ietf-grow-bgp-session-culling:
Reduce packet loss through correct procedures
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”

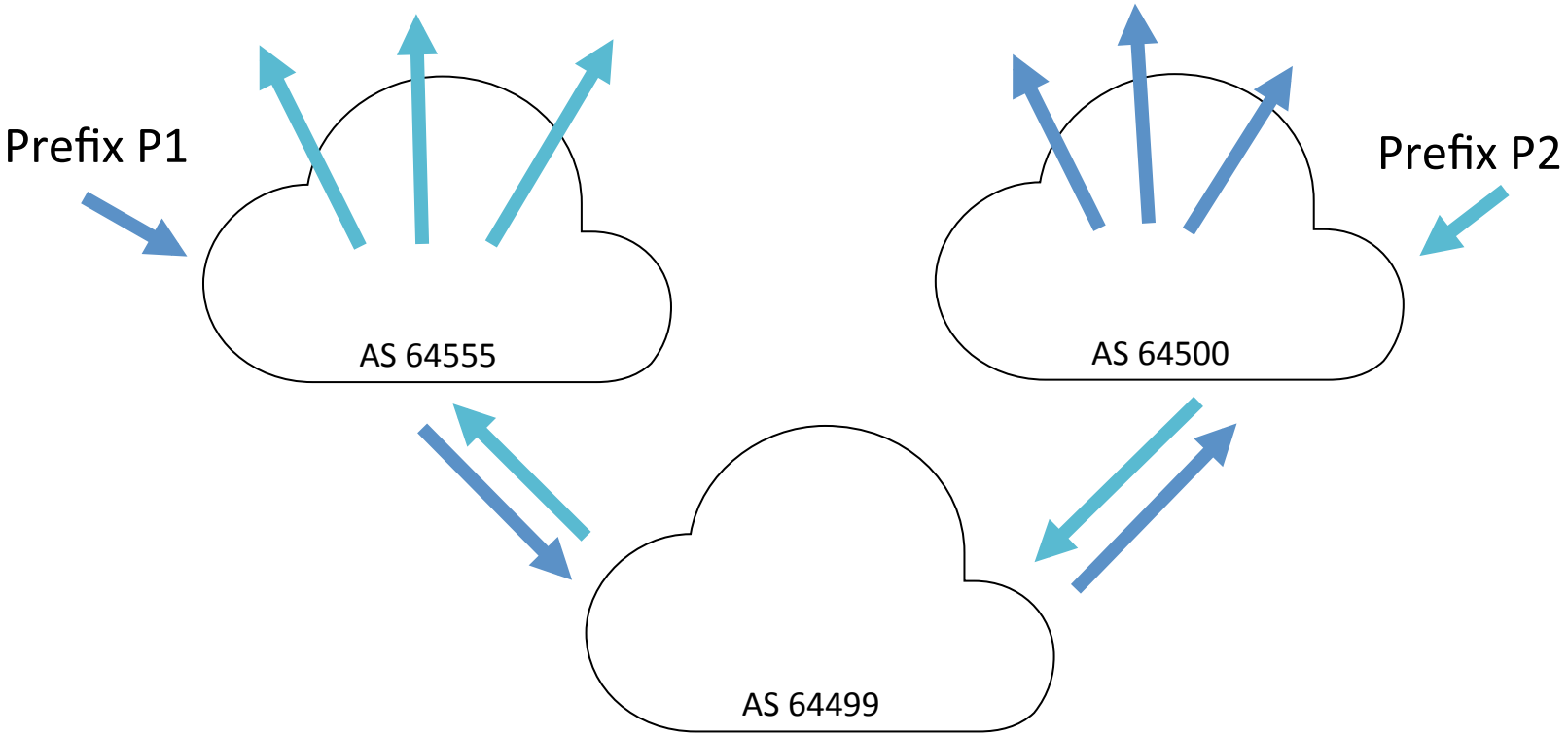
Opponents Argued

- “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 NANOG / NLNOG / RIPE / Community mailing lists
        - » And don’t talk to each other
          - …..

Puzzle Time: 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
!
```

Puzzle Answer: Lateral AS-AS-AS Leak



RFC 8212 in a Nutshell

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

- 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)

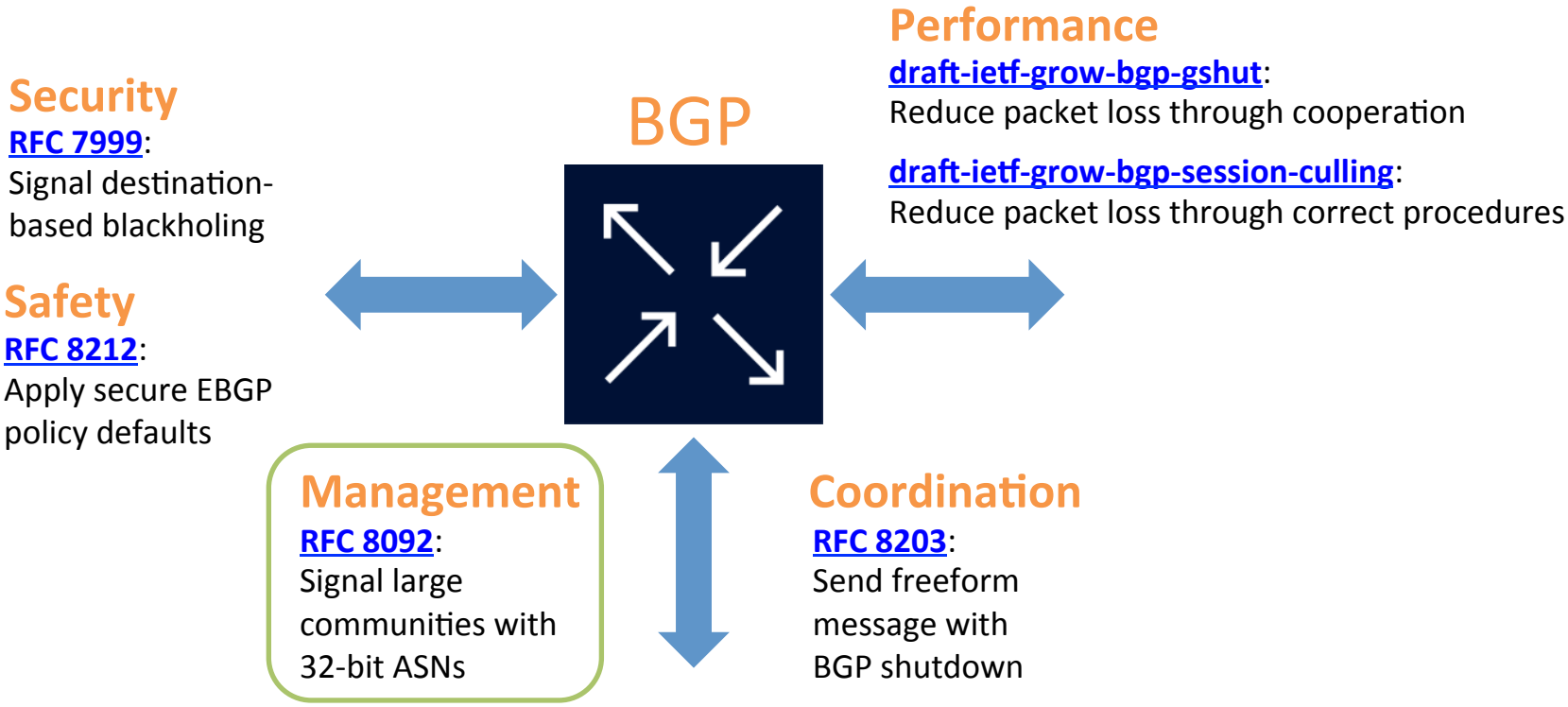
What This Means

- 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**
- 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
  - BIRD
  - OpenBGPD
  - Nokia SR OS
  - Others… (we're keeping track here 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


Security
RFC 7999:
Signal destination-based blackholing
Safety
RFC 8212:
Apply secure EBGP policy defaults
BGP
Performance
draft-ietf-grow-bgp-gshut:
Reduce packet loss through cooperation
draft-ietf-grow-bgp-session-culling:
Reduce packet loss through correct procedures
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

- 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 and fairness 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

BGP Speaker Implementation Status

Implementation	Software	Status	Details
Arista	EOS	Planned	Feature Requested BUG169446
Brocade	IronWare	Planned	First Half of 2018
Brocade	SLX-OS	Planned	First Half of 2018
Cisco	IOS XE	Planned	IOS XE 16.9.1 (FCS July 2018) (source)
Cisco	IOS XR	✔ Done!	Beta (perhaps in 6.3.2 for real?)
cz.nic	BIRD	✔ Done!	BIRD 1.6.3 (commit)
ExaBGP	ExaBGP	✔ Done!	PR482
FreeRangeRouting	frr	✔ Done!	Issue 46 (commit)
Juniper	Junos OS	✔ Done!	Junos OS 17.3R1
Nokia	SR OS	Planned	SR OS 16.0.R1
nop.hu	freeRouter	✔ Done!	
OpenBSD	OpenBGPD	✔ Done!	OpenBSD 6.1 (commit)
OSRG	GoBGP	✔ Done!	PR1094
rtbrick	Fullstack	✔ Done!	FullStack 17.1
Quagga	Quagga	✔ Done!	Quagga 1.2.0 (875)
Ubiquiti	EdgeOS	Planned	Internal Enhancement Requested

Tools and Ecosystem Implementation Status

Implementation	Software	Status	Details
DE-CIX	pbgpp	✔ Done!	PR16
FreeBSD	tcpdump	✔ Done!	PR213423
INEX	Bird's Eye	✔ Done!	1.1.0 (commit)
Marco d'Itri	zebra-dump-parser	✔ Done!	PR3
OpenBSD	tcpdump	✔ Done!	OpenBSD 6.1 (patch)
pmacct.net	pmacct	✔ Done!	PR61
RIPE NCC	bgpdump	✔ Done!	Issue 41 (commit)
tcpdump.org	tcpdump	✔ Done!	PR543 (commit)
Yoshiyuki Yamauchi	mrtparse	✔ Done!	PR13
Wireshark	Wireshark	✔ Done!	Wireshark 2.4.0 (patch)

Visit http://largebgpcommunities.net/implementations/ for the Latest Status

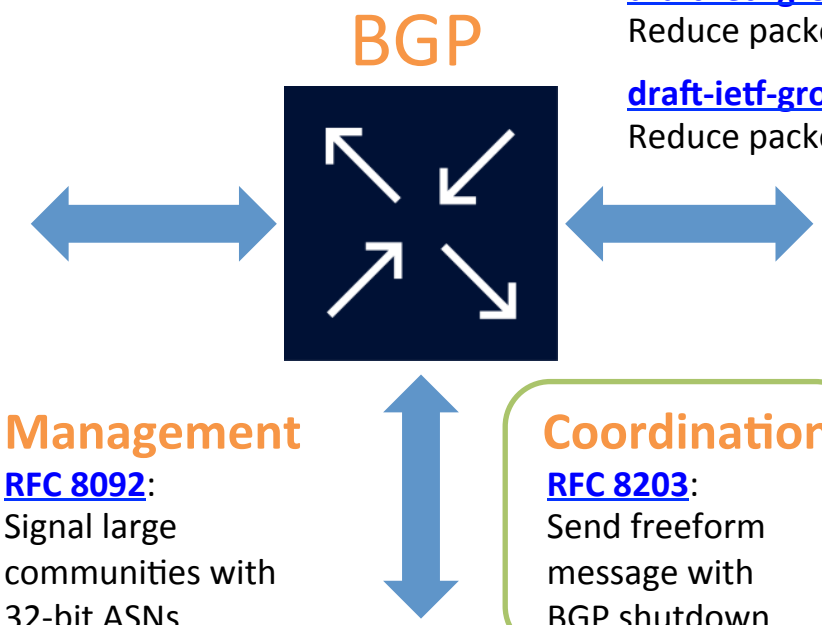
Agenda

Security

RFC 7999:
Signal destination-based blackholing

Safety

RFC 8212:
Apply secure EBGP policy defaults



Performance

draft-ietf-grow-bgp-gshut:
Reduce packet loss through cooperation

draft-ietf-grow-bgp-session-culling:
Reduce packet loss through correct procedures

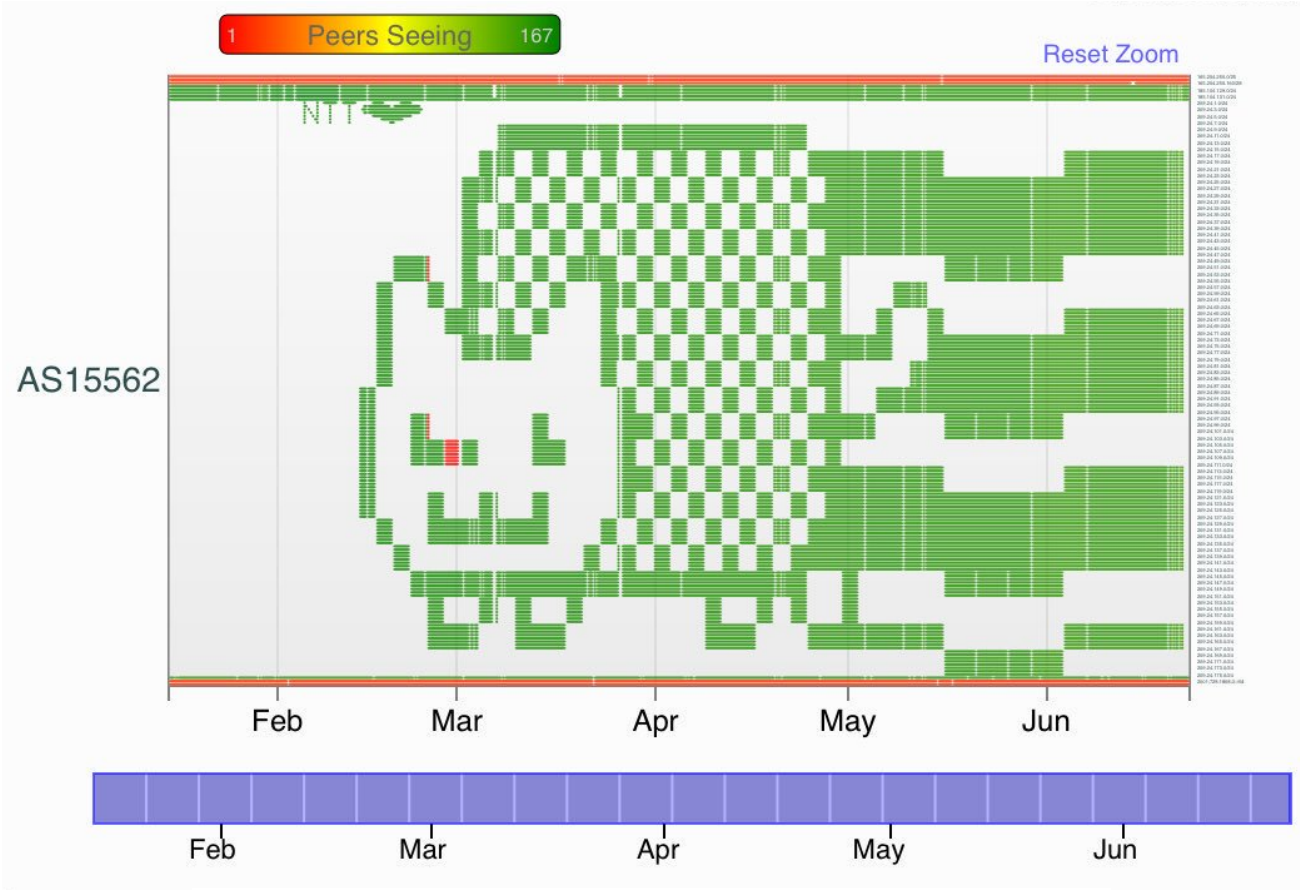
Management

RFC 8092:
Signal large communities with 32-bit ASNs

Coordination

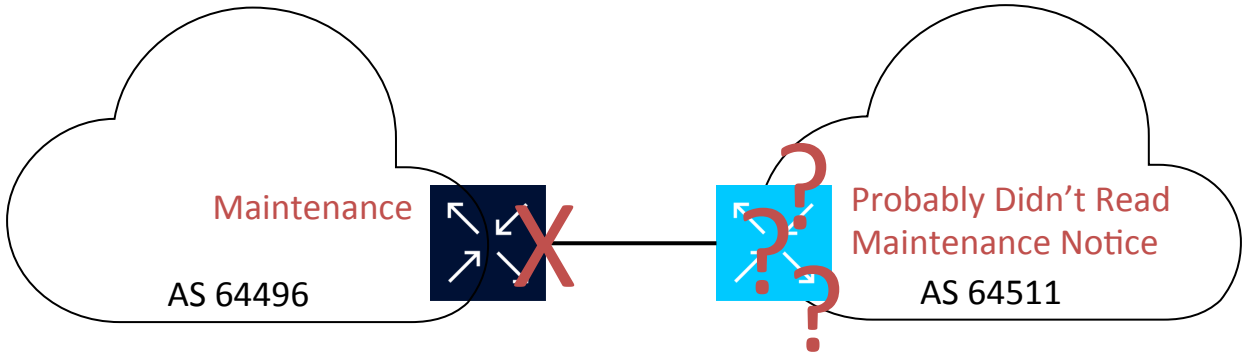
RFC 8203:
Send freeform message with BGP shutdown

Communication can be a Challenge



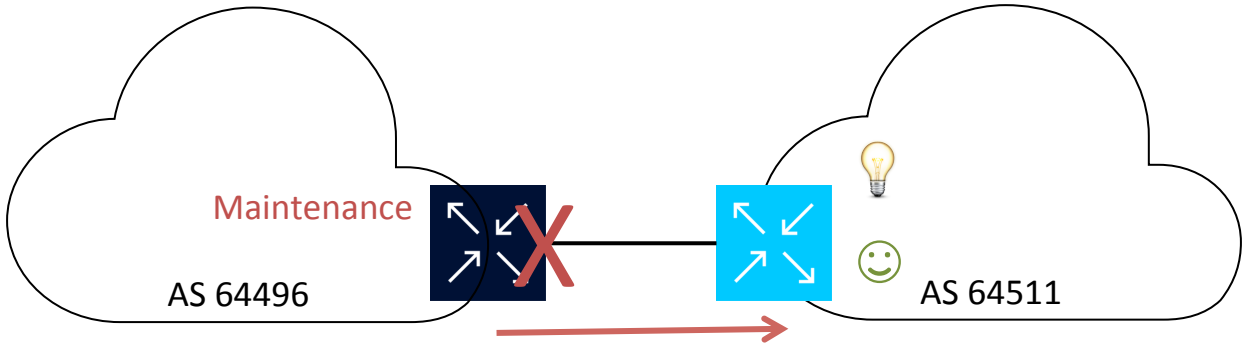
https://labs.ripe.net/Members/cteusche/bgp-meets-cat

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

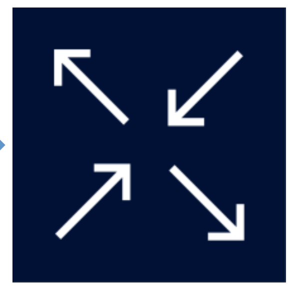
Security

RFC 7999:
Signal destination-based blackholing

Safety

RFC 8212:
Apply secure EBGP policy defaults

BGP



Performance

draft-ietf-grow-bgp-gshut:
Reduce packet loss through cooperation

draft-ietf-grow-bgp-session-culling:
Reduce packet loss through correct procedures

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

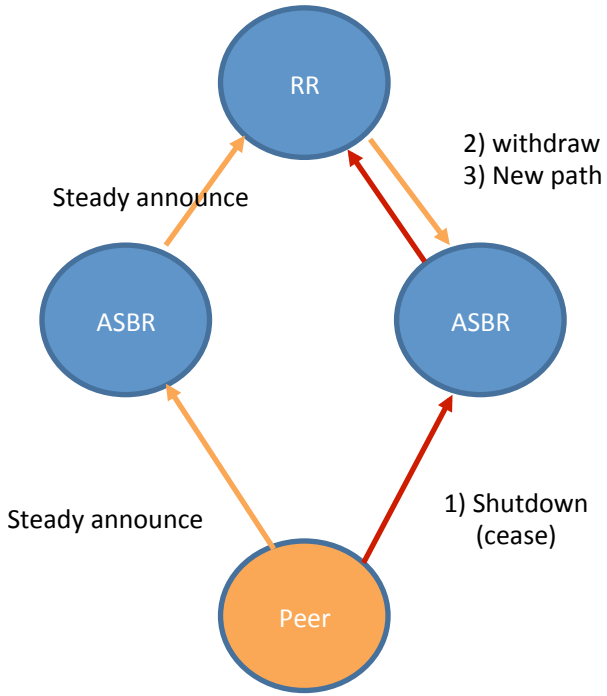
Involuntary Shutdown (other folks)

- 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

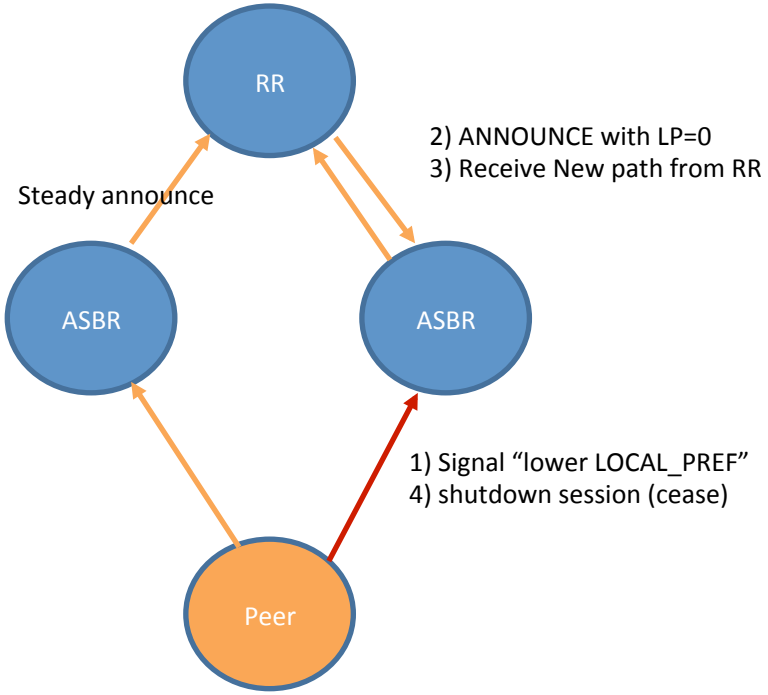
When does blackholing happen with vanilla shutdown?

- 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
```

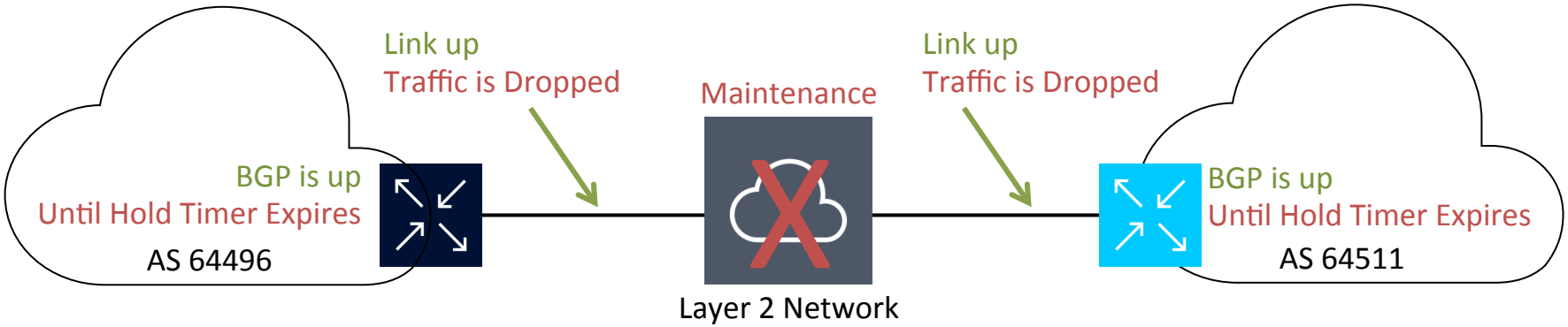
GRACEFUL_SHUTDOWN signals:

"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:

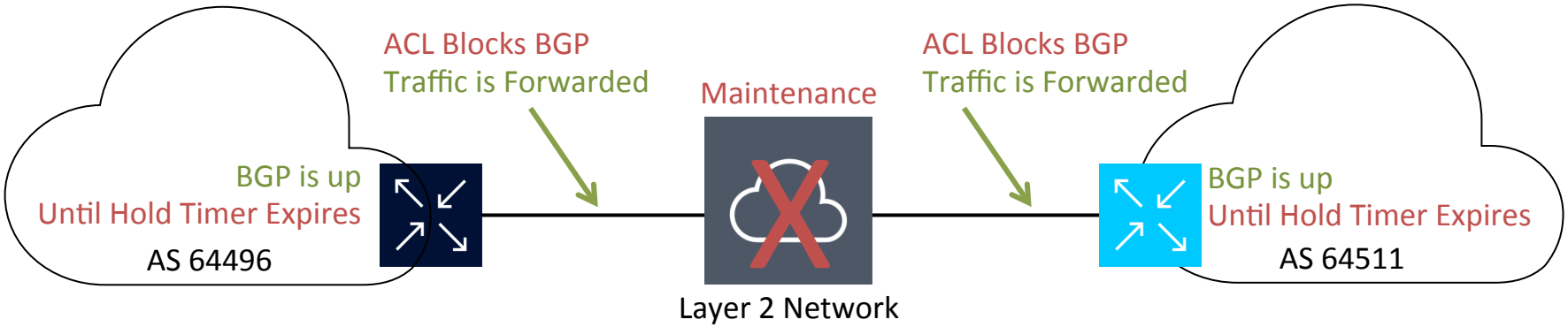
- NTT Communications (AS 2914)
- GTT (AS 3257)
- Github (AS 36459)
- Nordunet (AS 2603)
- Coloclue (AS 8283)
- Amsio (AS 8315)
- BIT (AS 12859)
- … you? ☺

draft-ietf-grow-bgp-session-culling
"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

draft-ietf-grow-bgp-session-culling
"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

- **Shipping now:**
  - Graceful Shutdown
  - BGP Session Culling
  - BLACKHOLE Community
- **Partially available:**
  - 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 to support
  - RFC 7999 BLACKHOLE Community (destination-based blackholing)
- Ask your transit providers & peers to support
  - draft-ietf-grow-bgp-gshut Graceful BGP session shutdown
  - draft-ietf-grow-bgp-session-culling Voluntary Shutdown BCP
- Ask IXPs to apply BGP culling (or equivalent) during maintenance
  - draft-ietf-grow-bgp-session-culling (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
  - BLACKHOLE well-known community (65535:666)
  - 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 7999, 8092, 8195, 8203, 8212)

Acee Lindem
Adam Simpson
Arnold Nipper
Bill Fenner
Christian Seitz
David Farmer
Eduardo Ascenco Reis
Greg Hankins
Ian Dickinson
Jan Baggen
Jeff Tantsura
Joel M. Halpern
Julian Seifert
Kristian Larsson
Marco Davids
Martijn Schmidt
Nick Hilliard
Peter van Dijk
Remco van Mook
Robert Raszuk
Sander Steffann
Sriram Kotikalapudi
Terry Manderson
Thomas Mangin
Warren Kumari
Yordan Kritski

Adam Chappell
Alexander Azimov
Barry O'Donovan
Brad Dreisbach
Christoph Dietzel
David Freedman
Gaurab Raj Upadhaya
Greg Skinner
Ignas Bagdonas
Jared Mauch
Jeffrey Haas
John Heasley
Jussi Peltola
Linda Dunbar
Marco Marzetti
Martin Millnert
Niels Bakker
Petr Jiran
Richard Hartmann
Ruediger Volk
Shane Amante
Stefan Plug
Teun Vink
Tom Daly
Wesley Steehouwer
Richard Turkbergen

Adam Davenport
Alvaro Retana
Ben Maddison
Brian Dickson
Christopher Morrow
Donald Smith
Geoff Huston
Grzegorz Janoszka
Jakob Heitz
Jay Borkenhagen
Joe Provo
John Scudder
Kay Rechthien
Lou Berger
Mark Schouten
Mikael Abrahamsson
Paul Hoogsteder
Pier Carlo Chiodi
Richard Steenbergen
Russ White
Shawn Morris
Stewart Bryant
Theodore Baschak
Tom Petch
Will Hargrave
Job Snijders

Adam Roach
Arjen Zonneveld
Bertrand Duvivier
Bruno Decraene
Dale Worley
Duncan Lockwood
Gert Doering
Gunter van de Velde
James Bensley
Jeff Haas
Joel Jaeggli
Jonathan Stewart
Keyur Patel
Mach Chen
Markus Hauschild
Nabeel Cocker
Peter Hessler
Randy Bush
Rob Shakir
Saku Ytti
Shyam Sethuram
Susan Hares
Thomas King
Tom Scholl
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)