Accelerated Routing Convergence for BGP Graceful Restart

draft-keyur-idr-enhanced-gr-00

Keyur Patel, Enke Chen, Rex Fernando, John Scudder

IETF 81, July 2011, Quebec City, Canada

Motivation

- Full Table re-announcements and cleanups across session resets are becoming expensive in BGP
  - Newer AFs added to BGP adds to number of tables that BGP stores and announces
  - BGP AF table size is growing as well. VPN AF table sizes already in excess of 1.4M routes
- Would like to perform incremental updates within BGP to speed up convergence

Advantages of Incremental Table Exchanges

- Avoid table/prefix cleanups upon session resets
  - Stale path timer cleans up table/routes if session does NOT come up within *restart* time period
- Avoid exchanging Full BGP tables upon successive session restarts
  - Results in faster Convergence
- Highly beneficial in terms of CPU and transient memory usage

Requirements for BGP Incremental Updates

- Need to preserve ADJ-RIB-IN and ADJ-RIB-OUT during session resets
- Need an ability to exchange incremental updates – Aka versioning of prefixes and routing updates
- Need to signal if outbound and inbound RIBs have been preserved during the session reset or not
  - Crucial in generation of incremental updates
- Seems like a natural extension to an existing BGP Graceful Restart mechanism

Enhanced GR aka Incremental updates

- Augmented BGP GR to support Incremental route updates
- Reuse GR to preserve
  - BGP ADJ-RIB-IN and BGP ADJ-RIB-OUT during BGP session resets
- Introduced new BGP GR Capability AF Flags
  - (R) Flag used to indicate if the received routing state of ADJ-RIB-IN has been preserved or not
  - (T) Flag used to indicate if the send routing state of ADJ-RIB-OUT has been preserved or not

Enhanced GR

- Introduced a new BGP Capability known as Enhanced GR Capability
  - Used to indicate the support for a newly defined BGP Update Version message
  - Used to indicate support for new AF level GR Capability flags
- Introduced a new BGP message known as a BGP Update Version message
  - Has a message subtype indicating if the message is a 1) send version number message, 2) Ack version number message, 3) Req version number message
  - 8 byte Version number

Enhanced GR Operation

- Enhanced GR Capability needs to be exchanged for enabling Incremental Updates
- Every BGP speaker uses version number (per AF per peer) to track
  - routing updates and other states announced
  - routing updates and other states received
- Received version number is an opaque value from receiving BGP speaker's perspective

Enhanced GR Operation

- BGP speakers supporting Enhanced GR needs to exchange Update version messages
  - Send version messages (subtype = 1) after batch of update messages
  - Ack version messages (subtype = 2) for every version message with (subtype = 1) received
  - Optionally request a peer to send update message from a certain version number (subtype = 3)
- Upon session restarts BGP speakers explicitly exchange their ADJ-RIB-IN and ADJ-RIB-OUT state since the session reset
  - If ADJ-RIB-OUT is not preserved then full table needs to be announced. Otherwise incremental updates are good enough
  - If ADJ-RIB-IN is not preserved then full table is requested. Otherwise incremental updates are good enough

Current GR Scenario for Session Restart

Session disconnects, Marks prefixes as stale
Starts the restart timer

Mark prefixes as stale, start restart timer

OPEN, Adjacency Up

Stop restart timer. Start stalepath timer

Stop the restart timer. Start stale path timer

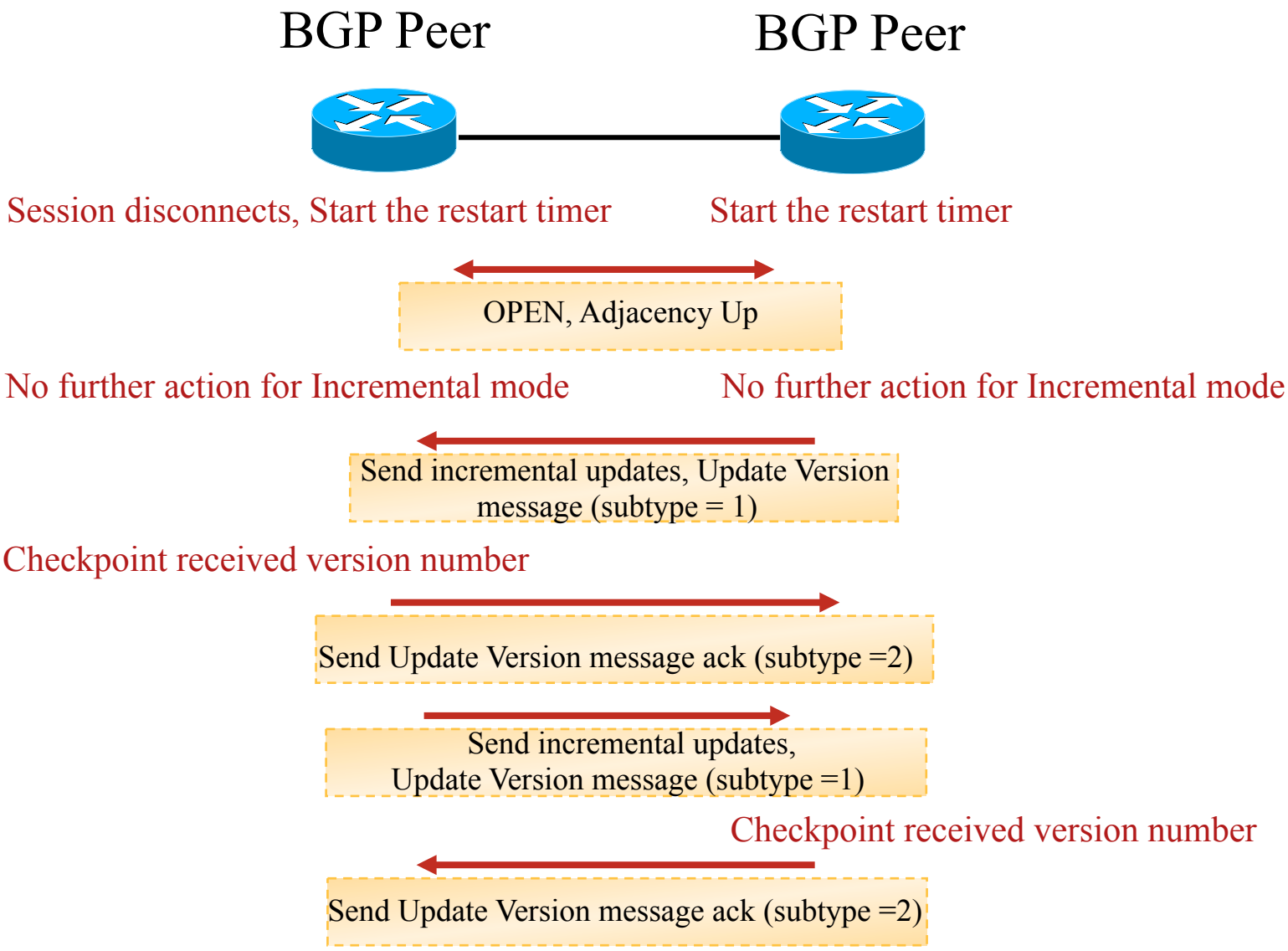
Send complete table updates, EOR

Send complete table updates, EoR

Stop stalepath timer. Delete stale prefixes needed

Stop stalepath timer. Delete stale prefixes if needed

Enhanced GR Scenario for Session Restart



Questions?