Improving network availability through the graceful shutdown of BGP sessions

PLNOG, Krakow, September 2008

Bruno Decraene  France Telecom  (bruno.decraene@orange-ftgroup.com)
Pierre François  UCL  (pierre.francois@uclouvain.be)
Agenda

Why?  (Problem statement)
What?  (Requirements)
How?   (A solution)
How good? (Test bed evaluation)
Conclusion
Why improving network availability?

- For new applications, customers are requesting Service Providers tighter SLA requirements, especially regarding network availability.
  - e.g., VoIP, online gaming, corporate mission critical applications

- E.g. typical VoIP requirement is a traffic restoration time below 100 or 200 ms after the failure.
How to improve network availability?

- **Failures avoidance** at the IP layer
  - Link: protection below the IP/MPLS layer
  - Node: state of the art hardware & software router, extensive testing.

- **Local concealment**
  - Graceful Restart, Non Stop Routing, In Service Software Upgrade (ISSU).

- **Local reaction**
  - MPLS Fast ReRoute, IP Fast reroute

- **Global reaction**
  - Usual routing convergence: IGP, BGP, IGP+ BGP

- **Mixed of global & local reaction**
  - IGP routing convergence + BGP local reaction
    - BGP protection could be pre-computed & pre-installed in the FIB (e.g. BGP PIC edge)

- **Global anticipation**
  - Make before break
How to improve network availability?

- Most solutions are complementary with different:
  - Applicability: forwarding preservation, type of failure, existence of an alternate path…
  - Cost: states, hardware, implementation, operation…
  - Result: (expected) traffic loss
Network anticipation: make before break

- **Applicability is significant**: every time the BGP session needs to be shutdown
  - Prefix limit reached, session reloaded, unrecognized attribute…
  - Maintenance operations which affect forwarding
    - Most hardware upgrade: router, line card, link
    - Some software upgrades
  - → same applicability than the BGP cease message, but with a different result.

- **Low cost** since speed is not required
  - No need for fast hardware or software, redundant states

- **Possibly very good results**
  - Perfect make before break could achieve 0 packet loss

- **Still does not address all cases** as it requires:
  - a backup path → subset of customers / peers
  - anticipation → subset of events
Graceful shutdown

- Shutdown allowing your peer (router/AS) to gracefully handle the loss.
  - Typically give them some time to accommodate

- Not something new in general:
  - Link state IGP:
    - link max metric, node overload bit (IS-IS)
    - (non urgent) loop free convergence techniques
      - e.g. smart multiple metric increments
  - MPLS, GMPLS: IETF WG draft
    - "Graceful Shutdown in MPLS and Generalized MPLS"
BGP Graceful shutdown

- Currently no agreed procedure for BGP although:
  - BGP is widely used
    - Internet
    - BGP/MPLS VPN services (L3 & L2)
  - BGP routing convergence could be "long"
    - Re-routing of 1 prefix can require multiple steps:
      - Path vector protocol & any router may hide back up paths
      - → multiple messages & best path selections required
    - Hundreds of thousands of routes involved: 280,000 prefixes for Internet
      - → 120,000 BGP updates required to update the RIB
        - With an average of 2.3 prefixes per update
      - → 30 seconds required to update the FIB
        - 300,000 prefixes * 100us/prefix
        - BGP/MPLS VPN usually have bigger scaling numbers
  - Requires bi-lateral / multi-lateral agreements between ASes
    - Cannot be done by an ISP on its own.
Legend:
L : loopback address
L.x: 10.0.1.x/32
Agenda

Why? (Problem statement)
What? (Requirements)
How? (A solution)
How good? (Test bed evaluation)
Conclusion
BGP Graceful shutdown requirements

- In short: minimal / no packet loss when shutting down a BGP session.
  - Providing an alternate path is available in the AS.
  - Otherwise, the path should still be usable until the forwarding failure,
    - Just like today.

- **Should handle common iBGP topologies:**
  - iBGP full mesh, iBGP Route Reflector, hierarchical BGP RR, BGP confederation

- **Regarding eBGP topologies,** the target use case is two ASes directly interconnected through multiple ASBRs
  - Typically a customer dual attached to a provider.
  - Topologies involving more than 2 ASes are out of scope.
    - e.g. a multi-homed AS requiring Internet wide convergence:
BGP Graceful shutdown requirements

- Desired properties (in descending order of importance):
  1. minimize loss of connectivity
  2. applicable to a wide range of networks, BGP topologies and usages
  3. minimize transient forwarding loop
  4. minimize additional BGP load / impact

- More details in: draft-decraene-bgp-graceful-shutdown-requirements-00
Agenda

Why? (Problem statement)

What? (Requirements)

How? (A solution)
  - LoC during planned maintenance
  - G-shut for outbound & inbound traffic
  - Deployment consideration
  - Further options

How good? (Test bed evaluation)

Conclusion
LoC during planned maintenance

Potentially large number of p/P
G-Shut: Outbound traffic

\[
p/P\text{ in : } lp = 100
\]

\[
p/P\text{ out : } lp = 0
\]

\[
p/P : lp = 90
\]
G-Shut: Inbound traffic

- Need to trigger outbound g-shut at the other side of the peering link
  - Use of a specific BGP community

z/Z:
tag community: GSHUT

iBGP out-policy: match GSHUT remove community set local pref 0

No need for action from the customer at maintenance time
Preconfigured policies on customer's & SP ASBRs

- Preconfigured on ASBRs
  - Outbound policy on iBGP sessions:
    - G-shut community \(\rightarrow\) set local\_pref = 0
    - Remove g-shut community

```
[edit protocols bgp group ibgp]
JM7B@p-jm7b# show type internal;
local-address 10.0.1.2;
export allow-BGP-gshut;
neighbor 10.0.1.6;

[edit policy-options]
JM7B@p-jm7b# show policy-statement allow-BGP-gshut {
  term 1 {
    from {
      protocol bgp;
      community gshut;
    }
    then {
      local-preference 0;
      community delete gshut;
    }
  }
  community gshut members 3215:6666;
```
SP policies at maintenance time

1. Apply an outbound & inbound policy on the eBGP session to be shutdown.
   - Add G-shut community
2. Wait for BGP convergence
3. Shutdown the BGP session (as usual)

```
[edit protocols bgp group customer-65511]
JM7B@p-jm7b# show type external;
export set-BGP-gshut;import set-BGP-gshut;
peer-as 65511;
neighbor 10.0.20.6;

[edit policy-options]
JM7B@p-jm7b# show policy-statement set-BGP-gshut {
   term 1 {
      then {
         community add gshut;
      }
   }
   community gshut members 3215:6666;
```
Deployment considerations

- One g-shut community value per customer/peer/provider is
  - difficult to manage
  - error prone
- **G-shut community should be standardized**

- Good deployment properties:
  - Incremental deployment possible
    - per eBGP session
    - Incremental gain

- Can be implemented now by ISPs through configuration
  - As detailed in draft-francois-bgp-gshut-00.txt
  - Vendors could help make it simpler
    - by automating this before sending the BGP cease message
Is g-shut enough? – micro-forwarding loop

- "Micro" forwarding loops are still possible during iBGP convergence.
- Caused by transient inconsistent FIBs between routers along the forwarding path
  - No atomic change at the network scope

Possible solutions:
- Simultaneous RIB/FIB update
- Order RIB/FIB update across routers of an AS
- Tunnels between ASBR
  - MPLS LSP, GRE, L2TP…
  - Available now
**Is g-shut enough? – diversity in iBGP signaling paths**

- BGP graceful shutdown tries to avoid abrupt route withdrawal

- But even a BGP update can initiate loss of connectivity
  - A route update can be translated into a withdrawal along the iBGP signaling path
  - Both messages (update & withdraw) can use two different iBGP signaling paths and the withdraw can possibly be quicker.

- Example:
  - Full mesh iBGP
  - Primary path (F) selected on the local pref
  - eBGP session brings UP on router F

- A solution: BGP external best
  - draft-marques-idr-best-external-00.txt
  - Instead of withdrawing a route, a router advertises its best external route
    - Can be different from its overall best.
  - Available now in some implementation.
Is g-shut enough? – convergence concealment

p/P : \( lp = 100 \) nh : self

p/P : \( lp = 100 \) nh : ASBR1 (cust)

p/P : \( lp = 50 \) nh : self (peer)

Abrupt Withdraw sent due to policies

SHOULD use loc-pref value that conceals convergence within customer paths
Agenda

Why? (Problem statement)
How? (A solution)
What? (Requirements)
How good? (Test bed evaluation)
Conclusion
Tests goals

- **Evaluate** the BGP graceful shutdown solution
  - Check correctness
  - Evaluate the gain

- **Focus on the gain brought by BGP g-shut**, everything else being equal.
  - Absolute convergence times will be very hardware and software dependant.
Legend
10.0.3.x/32 Loopback L.x
- eBGP
- iBGP client
- iBGP

ISP AS
Customer AS
Cluster-id: 10.0.3.8
Cluster-id: 10.0.3.6
Cluster-id: 10.0.3.4
Cluster-id: 10.0.3.13
AS 2222
AS 100
2.2.1.0/24 to 2.2.10.0/24
1.1.1.0/24 to 1.1.10.0/24
BGP sessions
P-linux
iBGP client
AS 200
RT 104.1
RT 104.2
Testbed

- Real / commercial routers used
  - Packets forwarding done in hardware
    - not impact on control plane CPU / BGP convergence

- All routers of the test bed will be emulated on a single box by using Virtual Router
  - Not possible / too costly to have 15 (identical) commercials routers
  - PRO: Perfect time synchronization.
  - CON: Virtual Router shares hardware resources (CPU/RAM)
    - Care taken to avoid overloading the router.
    - Can affect absolute times
      - however tests were hardware dependant,
Testbed (very) specifics

- **BGP load**
  - BGP loaded with 6,000 routes external to the testbed
    - extracted from the Internet full routing table
    - advertised by the ISP to the customer.
  - No route flapping.

- **Router**
  - Juniper M7i, RE-5.0, Junos 7.1B2.2

- **Customer traffic**
  - Agilent Router Tester N2X version 6.5, build 4.10B
  - 5 bidirectional flows of 1000 packets per second \(\rightarrow\) +/- 1ms accuracy
  - Low TTL (25) to avoid forwarding loops, delayed packets, overloaded interfaces.

- **G-shut BGP policies only**
  - No BGP external best, no convergence concealment.
Tests plan

- Multiple topologies tested because results are expected to be topology dependant.

- 2 eBGP topology tested
  - “V”: 1 CE – 2 PE
  - “U”: 2 CE – 2 PE

- 4 iBGP topology tested:
  - full mesh
  - Route Reflectors (RR)
    - Hierarchical RR
      - With different cluster-id
      - With identical cluster-id

- 2 BGP best path decision criteria
  - IGP cost (hot potato routing)
  - Local Pref (policy routing)

- 3 forwarding types:
  - IP hop by hop (pervasive BGP)
  - MPLS (BGP free core)
  - VPN (L3 BGP/MPLS)
Tests plan

- Each test repeated 5 times ➔ keeping mean value

- 2 Events:
  - eBGP down (beginning of maintenance)
  - eBGP up (end of maintenance)

- Each topology is tested twice:
  - Vanilla BGP
  - BGP graceful shutdown

- 270 tests performed: 5 times * 27 topologies * 2 (UP/DOWN)
Summary of tests results

- Average gain is very significant
  - 100% for MPLS and VPN forwarding: 0 packet loss
  - 89% for IP forwarding
Agenda

Why? (Problem statement)
What? (Requirements)
How? (A solution)
How good? (Test bed evaluation)

Conclusion
Conclusion

- High availability is a strong requirement for IP/MPLS networks but standard BGP convergence does not meet such requirement.

- Multiple solutions to improve availability
  - Mostly complementary solutions

- BGP graceful shutdown can improve network availability:

- Applicable to a subset of cases: alternate path, anticipation

- 0 packet loss is achievable:
  - BGP graceful shutdown procedures
  - Tunnels between ASBR (e.g., MPLS LSP)
  - BGP external best

- Applicable now by ISP but vendors could help automating it.
Thank you!

Questions & feedback welcomed
References

- [http://bgp.potaroo.net](http://bgp.potaroo.net)

- Projecting Future IPv4 Router Requirements from Trends in Dynamic BGP Behaviour
  - Geoff Huston, Grenville Armitage
  - Australian Telecommunication Networks & Applications Conference (ATNAC), Australia, December 2006

- Graceful Shutdown in MPLS and Generalized MPLS Traffic Engineering Networks
  - [draft-ietf-ccamp-mpls-graceful-shutdown-06.txt](https://datatracker.ietf.org/doc/draft-ietf-ccamp-mpls-graceful-shutdown-06.txt)

- Disruption-free topology reconfiguration in OSPF Networks.
  - Pierre François, Mike Shand and Olivier Bonaventure

- Avoiding transient loops during the convergence of link-state routing protocols.
  - Pierre Francois, Olivier Bonaventure.
  - IEEE/ACM Transactions on Networking, December 2007
References

- Requirements for the graceful shutdown of BGP sessions
  - draft-decraene-bgp-graceful-shutdown-requirements-00.txt

- Graceful BGP session shutdown
  - draft-francois-bgp-gshut-00.txt

- Avoiding disruptions during maintenance operations on BGP sessions
  - Pierre Francois, Pierre-Alain Coste, Bruno Decraene and Olivier Bonaventure.

- Advertisement of the best-external route to IBGP
  - draft-marques-idr-best-external-00.txt

- Intermediate System to Intermediate System (IS-IS) Transient Blackhole Avoidance
  - RFC 3277