109 年委託研究報告

物聯網服務與應用研析及網際網路技術標準研析

受委託單位

東海大學資訊管理學系

計畫主持人

林正偉

研究人員

賴囿嘉、陳晏羚、陳示珮、胡詠翔、

吳宜庭、古庭瑋、彭鍾碩、張巧宜、

陳臻、吳光軒、許桓禎、王品力

研究期程：中華民國 109 年 4 月至 109 年 12 月

研究經費：新臺幣 70 萬元

中華民國 109 年 12 月

目　　次

圖　次

第一章　IETF 系列標準

第一節　BGP (Border Gateway Protocol)

一、前言

網際網路(Internet)是由許多大大小小、不同型態的網路彼此互相連網構成的。在網際網路上面傳送的訊息經由多層的網路堆疊處理後，被切割並封裝成一個又一個網路封包(packet)，由底層實體網路傳送。當封包在不同的網路之間傳送時，必須通過路由器(router)的協助將封包轉發(relay)到下一個網路。在不同的網路間，路由器透過查詢路由表(route table)來決定封包轉發的下一個網路。

在小型網路裡面，透過人工的方式來指定每一個路由器的路由表仍是可行的。但在大型網路裡面，網路拓樸(network topology)可能因為很多狀況而隨時發生改變，如果還是採用人工的方式來維護每一個路由器的路由表，會給網路管理人員帶來巨大的負擔。透過路由協定(routing protocol)，路由器可以和網路上的其他路由器或網路設備交換訊息，從而讓路由器可以動態發現網路狀況改變，學習新的路由，並更新路由表。

網際網路上常見的路由協定可以分成兩類，一類稱為內部閘道協定(interior gateway protocol，簡稱 IGP)，適合使用在單一一個自治系

統(Autonomous System，簡稱 AS)內，這個自治系統的網路共用一個 AS Number (簡稱 ASN)，如 RIP (Routing Information Protocol)、IGRP (Interior Gateway Routing Protocol)、EIGRP (Enhanced Interior Gateway Routing Protocol)、OSPF (Open Shortest Path First)、IS-IS (Intermediate system to intermediate system)等等，這些路由協定所維護的路由資訊通常屬於 OSI 網路模型中的網路層，比如 Internet Protocol (簡稱 IP)。

另一類路由協定為外部閘道協定(exterior gateway protocol)，在由多個自治系統構成的複雜網路中使用，如早期使用的 EGP (Exterior Gateway Protocol，RFC 904，已作廢)，以及本文要介紹的邊界閘道器協定(Border Gateway Protocol，簡稱 BGP)。外部閘道協定用來確定在不同的自治系統間網路的可達性，並通過內部閘道協定來解析某個自治系統內部的路由，如圖 1 所示。

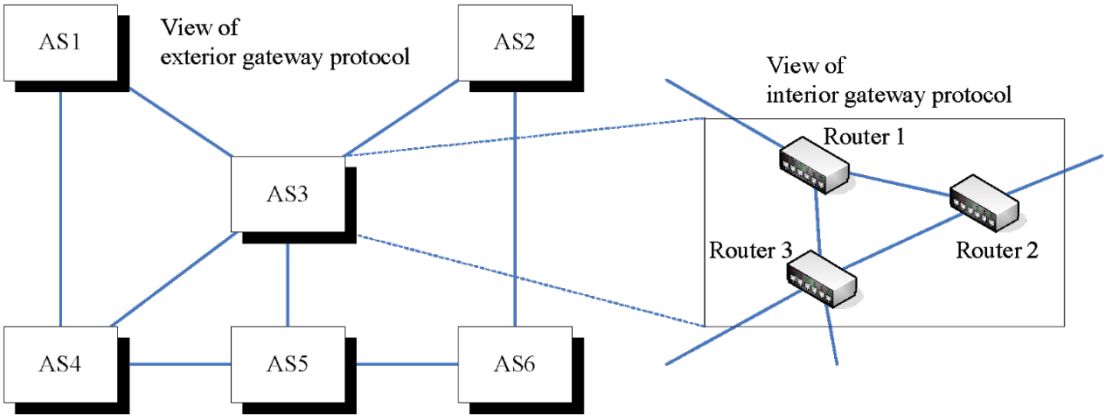

圖 1、View of exterior and interior routing protocols

BGP 是網際網路最重要的協定之一。大多數網際網路服務提供者(Internet Service Provider，簡稱 ISP)使用 BGP 來與其他 ISP 建立路由連接，彼此之間互稱為通信對端或對等實體(peer)，可為單個或多個 ISP 網路提供更好的冗餘網路(RFC 1998)。BGP 取代了早期的 EGP，並透過使用 CIDR (Classless Inter-Domain Routing)和路由聚合(route aggregation)來降低路由表的大小。

BGP 允許網路管理者自定 BGP 路由器(又可稱為 BGP Speaker)的路由政策(routing policy)，並給予較高的優先權，手動選擇最佳路徑、次佳路徑等等，這些作法增加了 BGP 的彈性。然而，自定的路由資訊沒有受到監管，並可能在 BGP 廣播時被其他 AS 信任，使得 BGP 路由器收到錯誤或虛假路由資訊，導致路由意外，甚至如路由外洩、路由劫持等惡意事件的發生。

規模較小的單一自治系統，一般可使用內部閘道協定，如 RIP 或 OSBF。大型複雜的單一自治系統，也可採用 BGP，以面對使用 RIP 或 OSPF 可能存在的效能瓶頸。iBGP (Interior/Internal BGP)指在單一自治網路中使用 BGP，所有的 iBGP 對等實體(peer)之間需要全連接。而在不同的自治網路使用的 BGP 則稱為 eBGP (Exterior BGP)。如圖 2 所示。

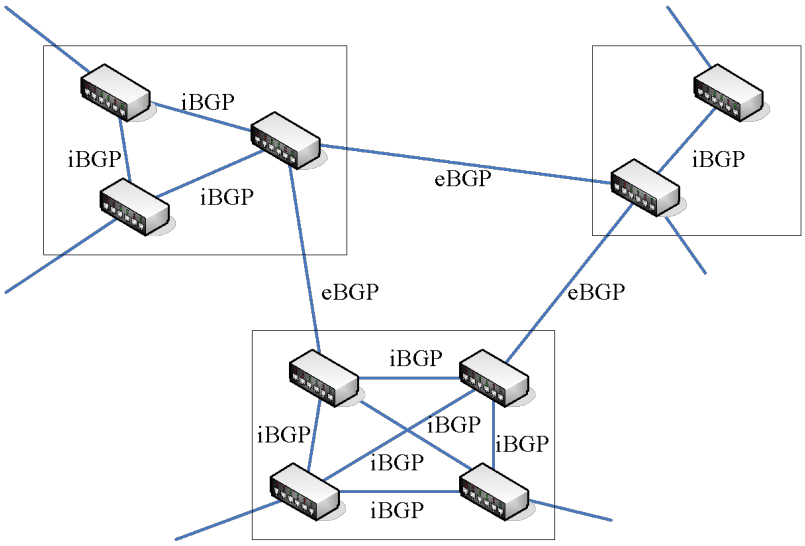

圖 2、iBGP 與 eBGP

二、RFC List

BGP 為目前最常用的路由協定之一，特別是在 ISP 之間使用。IETF 相關的 RFC 標準眾多。本研究列出下列 69 個 BGP 相關的 RFC 標準，如下：

(一) 初期

1. RFC 1105, 1989 (作廢) Border Gateway Protocol (BGP),
2. RFC 1163, 1990 (作廢) Border Gateway Protocol (BGP)
3. RFC 1164, 1990 (作廢) Application of the Border Gateway Protocol in the Internet

(二) BGP3

4. RFC 1267, 1991 Border Gateway Protocol 3 (BGP-3)
5. RFC 1268, 1991 (作廢) Application of the Border Gateway

Protocol in the Internet

6. RFC 1269, 1991 (作廢) Definitions of Managed Objects for the Border Gateway Protocol: Version 3

(三) BGP4

7. RFC 1654, 1994 (作廢) A Border Gateway Protocol 4 (BGP-4)
8. RFC 1655, 1994 (作廢) Application of the Border Gateway Protocol in the Internet
9. RFC 1656, 1994 (作廢) BGP-4 Protocol Document Roadmap and Implementation Experience
10. RFC 1657, 1994 (作廢) Definitions of Managed Objects for the Fourth Version of the Border Gateway
11. RFC 1771, 1995 (作廢) A Border Gateway Protocol 4 (BGP-4)
12. RFC 1772, 1995 Application of the Border Gateway Protocol in the Internet Protocol (BGP-4) using SMIv2
13. RFC 1773, 1995 Experience with the BGP-4 protocol
14. RFC 1774, 1995 BGP-4 Protocol Analysis
15. RFC 4271, 2006 A Border Gateway Protocol 4 (BGP-4)
16. RFC 4272, 2006 BGP Security Vulnerabilities Analysis
17. RFC 4273, 2006 Definitions of Managed Objects for BGP-4

18.RFC 4274, 2006 BGP-4 Protocol Analysis

19.RFC 4275, 2006 BGP-4 MIB Implementation Survey

20.RFC 4276, 2006 BGP-4 Implementation Report

21.RFC 4277, 2006 Experience with the BGP-4 Protocol

22.RFC 4278, 2006 Standards Maturity Variance Regarding the TCP MD5 Signature Option (RFC 2385) and the BGP-4 Specification

(四) 延伸

23.RFC 2283, 1998 (作廢) Multiprotocol Extensions for BGP-4

24.RFC 2858, 2000 (作廢) Multiprotocol Extensions for BGP-4

25.RFC 4760, 2007 Multiprotocol Extensions for BGP-4

26.RFC 8654, 2019 Extended Message Support for BGP

(五) 輔助

27.RFC 2280, 1998 (作廢) Routing Policy Specification Language (RPSL)

28.RFC 2622, 1999 Routing Policy Specification Language (RPSL)

29.RFC 2725, 1999 Routing Policy System Security

30.RFC 4012, 2005 Routing Policy Specification Language next

generation (RPSLng)

31. RFC 7909, 2016 Securing Routing Policy Specification Language (RPSL) Objects with Resource Public Key Infrastructure (RPKI) Signatures
32. RFC 2842, 2000 (作廢) Capabilities Advertisement with BGP-4
33. RFC 3392, 2002 (作廢) Capabilities Advertisement with BGP-4
34. RFC 5492, 2009 Capabilities Advertisement with BGP-4
35. RFC 8810, 2020 Revision to Capability Codes Registration Procedures
36. RFC 2918, 2000 Route Refresh Capability for BGP-4
37. RFC 7313, 2014 Enhanced Route Refresh Capability for BGP-4
38. RFC 1966, 1996 (作廢) BGP Route Reflection An alternative to full mesh IBGP
39. RFC 2796, 2000 (作廢) BGP Route Reflection － An Alternative to Full Mesh iBGP
40. RFC 4456, 2006 BGP Route Reflection – An Alternative to

Full Mesh Internal BGP (iBGP)

41. RFC 6608, 2012 Subcodes for BGP Finite State Machine Error
42. RFC 7606, 2015 Revised Error Handling for BGP UPDATE Messages
43. RFC 7911, 2016 Advertisement of Multiple Paths in BGP

(六) 自治系統

44. RFC 1930, 1996 Guidelines for creation, selection, and registration of an Autonomous System (AS)
45. RFC 6996, 2013 Autonomous System (AS) Reservation for Private Use
46. RFC 7300, 2014 Reservation of Last Autonomous System (AS) Numbers
47. RFC 1965, 1996 (作廢) Autonomous System Confederations for BGP
48. RFC 3065, 2001 (作廢) Autonomous System Confederations for BGP
49. RFC 5065, 2007 Autonomous System Confederations for BGP
50. RFC 4893, 2007 (作廢) BGP Support for Four-octet AS

Number Space

51. RFC 6793, 2012 BGP Support for Four-Octet Autonomous System (AS) Number Space
52. RFC 6286, 2011 Autonomous-System-Wide Unique BGP Identifier for BGP-4
53. RFC 7607, 2015 Codification of AS 0 Processing
54. RFC 7705, 2015 Autonomous System Migration Mechanisms and Their Effects on the BGP AS_PATH Attribute

(七) 其他

55. RFC 4724, 2007 Graceful Restart Mechanism for BGP
56. RFC 8538, 2019 Notification Message Support for BGP Graceful Restart
57. RFC 1997, 1996 BGP Communities Attribute
58. RFC 4360, 2006 BGP Extended Communities Attribute
59. RFC 5543, 2009 BGP Traffic Engineering Attribute
60. RFC 5701, 2009 IPv6 Address Specific BGP Extended Community Attribute
61. RFC 7153, 2014 IANA Registries for BGP Extended Communities

62. RFC 8642, 2019 Policy Behavior for Well-Known BGP Communities
63. RFC 6368, 2011 Internal BGP as the Provider/Customer Edge Protocol for BGP/MPLS IP Virtual Private Networks (VPNs)
64. RFC 8212, 2017 Default External BGP (EBGP) Route Propagation Behavior without Policies
65. RFC 2439, 1998 BGP Route Flap Damping
66. RFC 3765, 2004 NOPEER Community for Border Gateway Protocol (BGP) Route Scope Control
67. RFC 5575, 2009 Dissemination of Flow Specification Rules
68. RFC 7674, 2015 Clarification of the Flowspec Redirect Extended Community
69. RFC 7752, 2016 North-Bound Distribution of Link-State and Traffic Engineering Information Using BGP

圖 3 顯示了這 69 個 RFC 之間的關係。

1989 年，最早的 BGP (已廢除)出爐，經過陸續討論與修正後，1994 年公告了最原始的 BGP 第 4 版(BGPv4)，1995 年的 RFC 1771 進行了一次修正，同年的 RFC 1883 定義了 IPv6 BGP，並在隨後 1998 年 RFC 2208 中修正，使得 BGPv4 能夠支援諸如 IPv4、IPv6 等系列

位址，也因此被稱為 Multiprotocol BGP。

之後，相關的 RFC 標準陸續發表，包含 BGP 對於自治系統的考量、路由政策的規範、實施的經驗與分析等。經過約 10 年的實施後，於 2006 年發表最新的 BGP 標準 RFC 4271，和相關伴隨文件(RFC 4272~4278)，如經驗與分析。之後，也陸續更新既有的 RFC，和發表新功能(如 Graceful Restart)的標準文件。

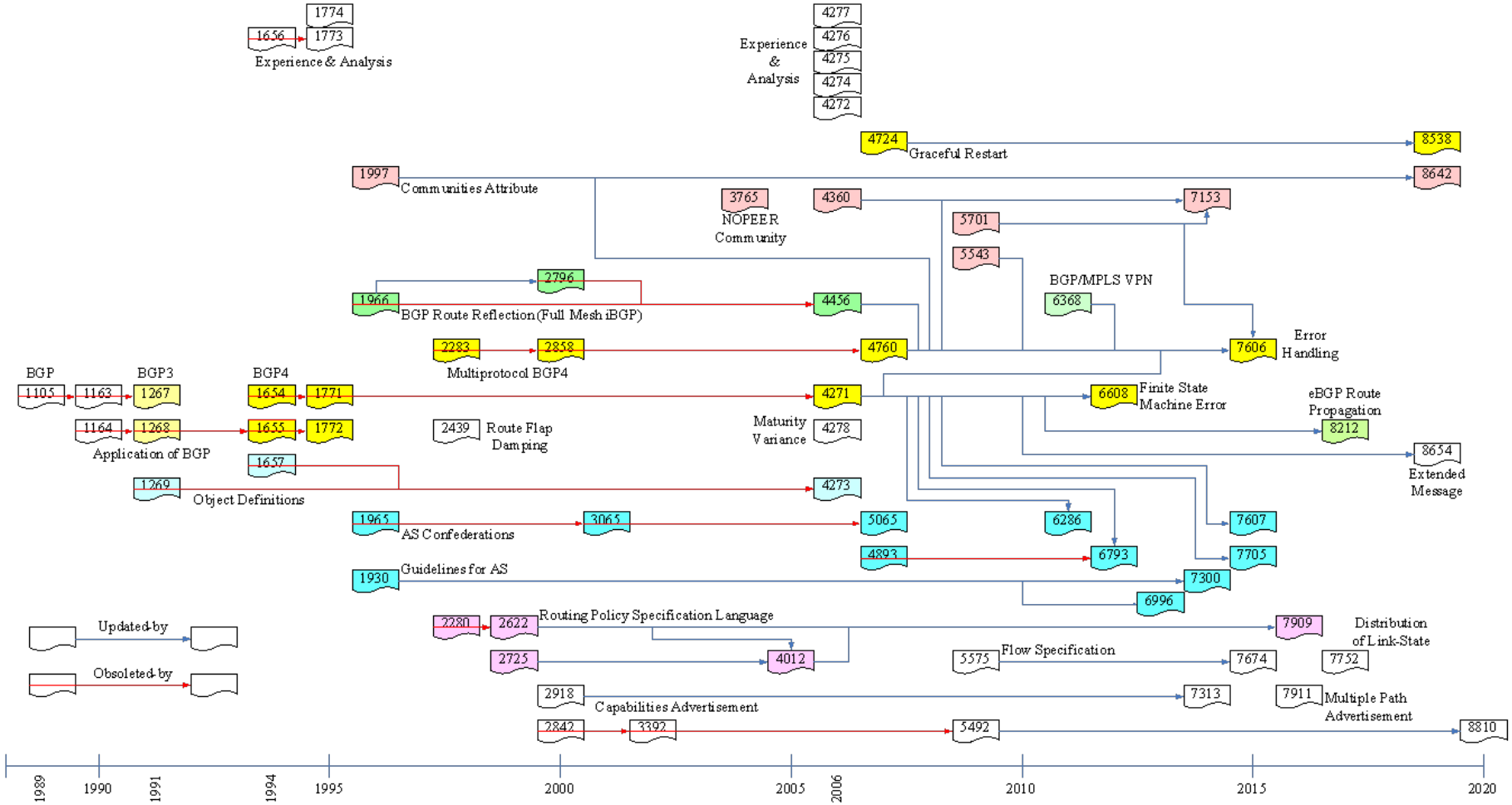

圖 3、BGP 相關 RFC 標準

三、相關知識

(一)　自治系統(Autonomous System，AS)

一個自治系統(Autonomous System，簡稱 AS)是一個由彼此互連的網路所構成的集合，其中的每個網路以一個 IP 路由前綴(routing prefix)為代表。一個 AS 被視為一個單一的管理實體(或域)，具有共同而且明確的路由策略。每個 AS 會被分配給唯一的 ASN (AS number)。對 BGP 來說，ASN 是區別整個 Internet 中的各個網路的唯一標識。

一開始的時候，一個 AS 只能被一個實體控制，通常是一個 ISP 或是具有多個網路連接的大型組織(RFC 1771)。現在，一個 AS 可以由一個或多個網路營運商共同控制(RFC 1930)。

2007 年以前，ASN 是 16 位元的，從 1 到 64,511 的 ASN 是公開，從 64,512 到 65,535 的 ASN 是私有的，私有 ASN 只能在一個組織內部的網路中使用。RFC 4893 定義了 32 位元的 ASN，並可用『高 16 位元的 10 進位.低 16 位元的 10 進位』的形式來表達(RFC 5396)，舉例來說，16 進位為 1000800016 的 ASN 可以表達為 4096.32768。從 4,200,000,000 to 4,294,967,294 的 32 位元 ASN 是私有的(RFC 6996)。

網際網路位址分派機構(Internet Assigned Numbers Authority，簡稱 IANA)將一個一個的 ASN 區塊分配給區域網際網路註冊管理機構(Regional Internet Registry，簡稱 RIR)，各地區的 RIR 再進一步從 ASN

區塊中挑選一個 ASN 分配給一個管理實體(RFC 6793)。每個管理實體必須按其所屬地區的 RIR 的規定申請，得到批准後才會分配到一個 ASN。所有 ASN 的分配情況可以在 IANA 的網站找到。截至 2019 年 8 月，已分配的 ASN 數量已超過 92,000。

根據連接性和操作策略，AS 可以分為下列四類。

1. Transit (中轉/過境) AS

末端 AS 只與一個 AS 相連。一般而言，末端 AS 只透過一條網路連線連接到其他的 AS，這種網路是資料流的起點或終點。末端 AS 的路由策略可能其上游 ISP 的路由策略完全相同，這種情形下，末端 AS 其實浪費了一個 ASN。

2. Stub (末端) AS

末端 AS 只與一個 AS 相連。一般而言，末端 AS 只透過一條網路連線連接到其他的 AS，這種網路是資料流的起點或終點。末端 AS 的路由策略可能其上游 ISP 的路由策略完全相同，這種情形下，末端 AS 其實浪費了一個 ASN。

2. Multihomed AS

一種與多個其他 AS 保持連接的 AS。即使其中一個連接完全失效，Multihomed AS 仍可保持與 Internet 連接。與中傳 AS 不同，這種類型的 AS 通常不允許來自其他 AS 的流量通過自身傳遞到另一個

AS。舉例來說，一個 Multihomed AS 網路 A 有二條對外連線連接到兩個不同的 ISP，分別為 ISP B 與 ISP C，以保證網路 A 的可靠性，但不允許 ISP B 到 ISP C 的資料流經過網路 A，反向流量(ISP C 到 ISP B)也是。

3. Internet Exchange Point (IX 或 IXP)

一種可以讓 ISP 或內容傳遞網絡(Content Delivery Networks，簡稱 CDN)之間進行資料交換的實體網路基礎建設。對於一般的網路而言，IXP 通常是透明的。

(二) 路由協定的技術

內部閘道協定可以細分為三類。

1、 Distance-vector routing protocol

包含 RIP、RIPv2、RIPng、IGRP 等。使用這類型路由協定時，路由器通常沒有完整的網路拓樸訊息。每一個路由器在路由表中維護到其他路由器的距離(如 hop count)，並週期性地公告給其他路由器，彼此交換，同時，使用 Bellman-Ford 演算法尋找最短路徑。這類型路由協定的缺點是需要較長的時間收斂，因此，比較適合使用在簡單的小型網路。

2、 Link state routing protocol

包含 OSFP、IS-IS 等。使用這類型路由協定時，路由器知道整個

網路的拓撲訊息，因此，可以獨立的計算目的地址的最佳下一站。路由器之間，僅傳送構造整個網路連通所需要的資訊，收斂較快，但路由器需要較多的記憶體與計算能力來維護網路拓樸與進行路由運算與選擇。

3、 Advanced distance vector routing protocol

如 EIGRP，藉由混合上述兩種技術，來達到較好的效能。

相對於內部閘道協定，BGP 可稱為 path vector (或 distance-path) routing protocol，既不像 link-state routing protocol 那樣去追蹤完整的網路拓樸，也不滿足於像 distance-vector routing protocol 僅僅計算 hop count。在 BGP 中，一條路徑(path)將包含到達目的位址的所有自治系統，並基於路徑及網管人員設定的網路策略和規則集來決定路由。

四、BGP 運行

BGP 一個完全分散的路由系統。在 BGP 裡，對等實體之間的鄰居關係是通過人工組態實現的，並通過 TCP 建立會話並交換資料，預設埠(port)是 179。在各種路由協定中，只有 BGP 使用 TCP 作為傳輸層協定。在 AS 網路邊界上與其他 AS 網路交換路由資訊的路由器稱為邊界路由器(border/edge router)，或者簡稱為 eBGP 對等實體。對等實體之間通常直接連接。iBGP 和 eBGP 的主要區別在於轉發路由資訊的行為。例如，從 eBGP 對等實體獲得的路由資訊會分發給所有

iBGP 對等實體和 eBGP 對等實體，但從 iBGP 對等實體獲得的路由資訊通常僅會分發給所有 eBGP 對等實體。這使得在同一個 AS 網路中的 iBGP 對等實體必須是全連接(full-mesh)的。使用 BGP Route Reflector 可以在某種程度上避免 iBGP 對等實體全連接的要求。

BGP 路由器在啟動的時候，會計算本地路由，並發佈給與之相連的 BGP 對等實體。這個過程結束之後，BGP 路由器就是斷斷續續的接收一些更新事件，再將路由重新發佈給它的 BGP 對等實體。BGP 路由器啟動之後，過程可以分為兩個部分，一個是初始化路由並發佈，接下來就是正常工作時的路由發佈。

(一) BGP 訊息

BGP 訊息通過 TCP 傳送，只在完整接收後才能進行處理。最大訊息的長度是 4096 個字節(octet，8 位元)，最小訊息只有 19 個字節，僅有 BGP 表頭(header)。

BGP 表頭包含 16 字節的 Marker，儲存同步訊息和加密訊息，2 字節的 Length，指定包含表頭在內的訊息長度，和 1 字節的 Type，表示當前 BGP 訊息的類型。BGP 表頭如圖 4 所示。

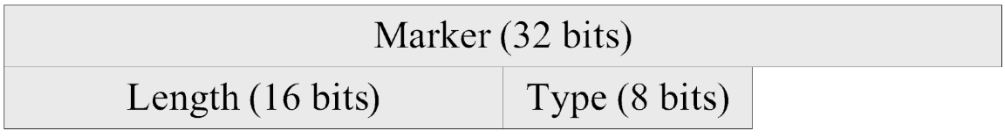

圖 4、BGP 訊息格式

BGP 使用 5 種訊息，分別是 Open (Type = 1)、Update (Type = 2)、Notification (Type = 3)、Keep-alive (Type = 4)與 Route-Refresh (Type = 5) (RFC 2918)。

1、 Open 訊息

在建立 TCP 連線後，每個 BGP 對等實體第一個送出的訊息就是 Open 訊息，接受 Open 訊息後則將回送一個 Keep-alive 訊息作為確認。Open 訊息的格式如圖 5 所示，其中用以表示 BGP 訊息傳送者的一組 32 位元的無號整數，可以是該 BGP 對等實體的 IPv4 位址。

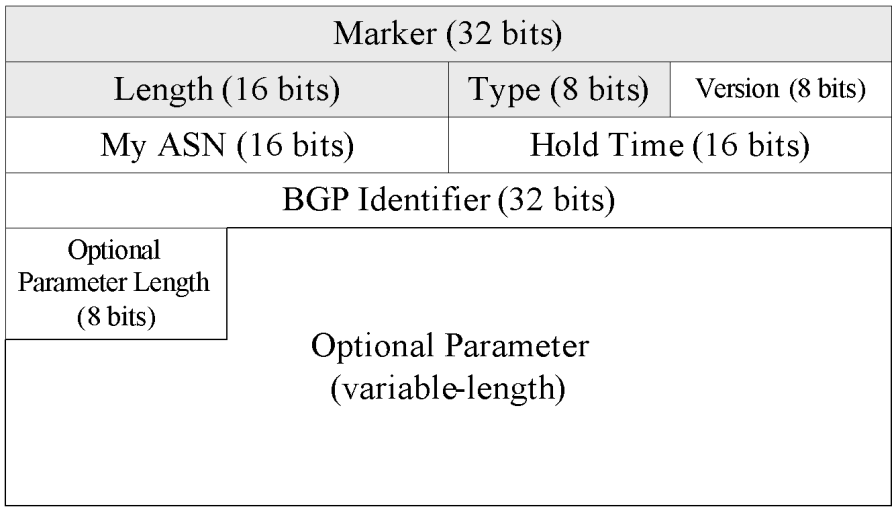

圖 5、BGP Open 訊息

2、 Update 訊息

確認 Open 訊息後，BGP 連接後的首次 Update 訊息會交換整個 BGP 路由表，之後的 Update 訊息只會發送有變化的路由信息。

不包含表頭的 Update 訊息格式如圖 6 所示。

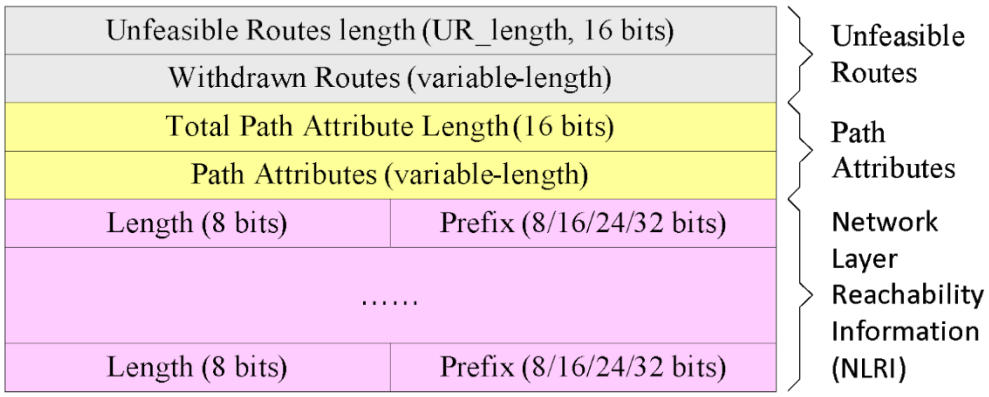

圖 6、Update 訊息

Update 訊息包含撤回和新增的路由，兩者可以同時進行。

每個被撤回的路由都包含一個長度欄位和一個路由前綴欄位。

一個 Update 訊息裡面只會包含一條路徑(path)的路由訊息，因此只有一組路徑屬性(Path Attribute，簡稱 PA)，但是可以包含多條路由，記錄在 Network Layer Reachability Information (簡稱 NLRI)。具體的說，一個 BGP 路由器可能連接了多個 BGP 對等實體路由器，它在發送 BGP Update 訊息時，一次只會發送其中一個 BGP 對等實體路由器的信息。

所有的路徑屬性都包含一個路徑屬性旗標欄位、一個路徑屬性型態欄位，這兩個欄位決定第三個欄位路徑屬性的長度與值。路徑屬性旗標說明路徑是否 well-known、transitive、optional。BGP 路由器必須認得所有的 well-known 路徑屬性。路徑屬性型態有 Origin、AS path、

Next hop、Multi Exit Discriminator、Local Preference、Atomic Aggregate 和 Aggregator。

下面介紹幾種常用的 BGP 路徑屬性：

1. Origin 屬性：用來定義路徑訊息的來源，有以下 IGP、EBP 和 Incomplete 等 3 種類型。IGP 具有最高的優先級，如通過命令列進入到 BGP 路由表的路由，Origin 屬性為 IGP。EGP 優先級次之。Incomplete 優先級最低，通過其他方式學習到的路由信息，其 Origin 屬性為 Incomplete。
2. AS_Path 屬性：按順序記錄了某條路由從本地到達目的地址經過的所有 ASN。
3. Next_Hop 屬性：記錄了路由的下一跳信息。BGP 路由器向 eBGP 對等實體發佈某條路由時，該路由訊息的 Next_Hop 屬性將設為一個本地垃址，BGP 路由器用該位址與 eBGP 對等實體建立 BGP 鄰居關係。iBGP 的情形類似。向 IBGP 對等實體發佈從 eBGP 對等實體學來的路由時，不改變該路由信息的 Next_Hop 屬性。
4. Local_Pref 屬性：表明本地路由的優先屬性，用於判斷流量離開 AS 時的最佳路由。當 BGP 通過不同的 iBGP 對等實體得到目的地址相同但下一跳不同的多條路由時，將優先選擇

Local_Pref 屬性值較高的路由。Local_Pref 屬性僅在 iBGP 對等實體之間有效，不會通告給其他 AS。Local_Pref 屬性可以手動配置。

5. MED 屬性：用於判斷流量進入 AS 時的最佳路由。當 BGP 發現從不同的 eBGP 對等實體學習得到的目的地址相同但下一跳位址不同的多條路由時，在其它條件相同的情況下，將優先選擇 MED 值較小者作為最佳路由。MED 屬性僅在相鄰兩個 AS 之間傳遞，收到這個屬性的 AS 不會再將其通告給任何其他第三方 AS。MED 屬性可以手動配置。

6. 社區(Community)屬性：用於標識具有相同特徵的 BGP 路由，使路由策略的應用更加靈活，同時降低了維護管理的難度。

路由資訊儲存於路由資訊庫(Routing Information Base，簡稱 RIB)。RIB 可說是路由表的同義詞，儲存至少以下三種資訊：目標位址、子網路遮罩和下一跳位址。BGP 使用三種 RIB：Adj-RIBs-In、Loc-RIB 和 Adj-RIBs-Out。

3、 Notification

當偵測到錯誤狀況時，BGP 路由器會送出一個 Notification 訊息，並在送出訊息後立刻中止連線。Notification 訊息包含了一個字節的錯誤碼、一個字節的次錯誤碼和其他相關資料。

4、 Keep-alive

BGP 路由器會周期地傳送 Keep-alive 訊息來維護連接，預設周期為 60 秒。Keep-alive 訊息只有表頭，沒有資料，因此，只有 19 字節，是最小的 BGP 訊息。

5、 Route-Refresh

當 BGP 路由器發現某個路由資訊已經失效，它有三種機制通知鄰居。

1. 發送 Update 訊息，將失效路由的位址前綴寫入 Withdrawn Routes 欄位。
2. 發送 Update 訊息，在 NLRI 中發佈該位址前綴新的路由資訊，取代舊的路由資訊。
3. 關閉連線，來自對方的路由資訊就會自動被刪除。

(二) Multiprotocol BGP

在 BGP 路由器之間建立對話的握手(handshake)期間，交換 Open 訊息時，BGP 路由器彼此可以協商是否使用 BGP 可選功能，包括多協定擴展(multiprotocol extensions，或稱為 Multiprotocol BGP)和各種恢復模式。如果在創建時就決定使用 Multiprotocol BGP，則 BGP 路由器可以為其發布的 NLRI 路由資訊加上位址家族前綴，這些系列包括 IPv4(默認)、IPv6、IPv4/IPv6 VPN 和 Multicast BGP。

(三)　BGP 路由器的狀態

BGP 對等實體使用一個簡單的有限狀態機(Finite State Machine，FSM)，它包含六個狀態：Idle、Connect、Active、OpenSent、OpenConfirm 與 Established。BGP 定義了每個對等實體在某個狀態下應與對方交換的消息，以便將會話從一種狀態更改為另一種狀態。

BGP 路由器的初始狀態是 Idle 狀態。在 Idle 狀態下，BGP 路由器初始化所有資源，拒絕所有從外界來的 BGP 連接嘗試，啟動與對方的 TCP 連接，進入 Connect 狀態。在 Connect 狀態下，BGP 路由器等待 TCP 連接完成，如果成功，則送出 Open 訊息，並進入到第三種狀態 OpenSent。如果不成功，它將啟動 ConnectRetry 計時器，並在時間到時進入為 Active 狀態。在 Active 狀態下，路由器將 ConnectRetry 計時器重置為零，並返回到 Connect 狀態。在 OpenSent 狀態下，路由器已經發送一個 Open 訊息，正在等待 Keep-alive 訊息確認，以便進入 OpenConfirm 狀態。彼此完成交換 Keep-alive 訊息，並在成功接收後，路由器進入 Established 狀態。在 Established 狀態下，路由器可以發送/接收 Keep-alive、Update 與 Notification 訊息。

(四)　BGP Route Processing

概念上，一個 BGP 路由器可以接受來自多個鄰居的 Update 訊息，並將 Update 訊息中的 NLRI 通告給相同或不同組的鄰居。

BGP 使用三種 RIB：Adj-RIBs-In、Loc-RIB 和 Adj-RIBs-Out。BGP 並不強制要求 BGP 路由器維護 3 份獨立的資訊，也可以是 1 個指標。

1. Adj-RIBs-In：本地 BGP 路由器從其他 BGP 路由器收到的路由訊息。
2. Loc-RIB：儲存在 Adj-RIBs-In 中未經處理過的路由訊息，經過本地 BGP 路由器的決定程序，依據本地政策所選出的路由資訊。Loc-RIB 記載的每個路由資訊的下一跳都必須是本地 BGP 路由器的路由表可以分辨的。
3. Adj-RIBs-Out：透過本地 BGP 路由器發送的 Update 訊息通告給特定鄰居的路由資訊。

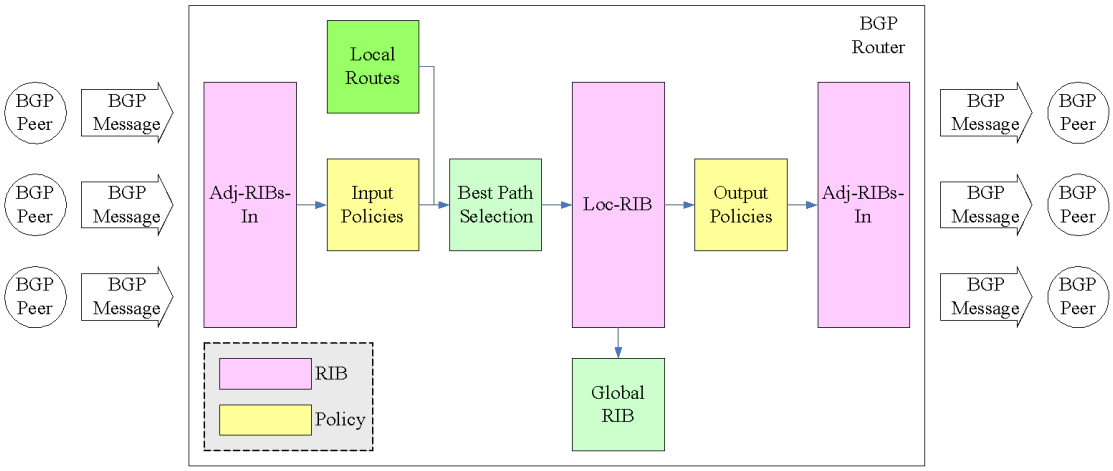

圖 7、BGP 路由選擇

BGP 使用路徑屬性來衡量每一條路徑的成本。BGP 指定了許多決策因素來從 Adj-RIBs-In 中選擇合適的路由進入 Loc-RIB，包含許多一般路由決策過程沒有使用的因素。

評估 Update 訊息中 NLRI 路由是否合適的第一個決策點是其下一跳屬性必須是可到達的或可解析的。接下來，對於每個鄰居，BGP 可以採用不同的標準和實作相關的條件來決定哪些路由可以加入該鄰居對應的 Adj-RIB-in。鄰居可以在 NLRI 中記錄到達某個目的地的多個可能的路由，但在 Adj-RIB-In 中，每個目的地應僅安裝一條路由。在這個過程裡，所有被鄰居撤回的路由也將從 Adj-RIB-In 中刪除。

每當 Adj-RIB-In 發生更改時，BGP 會確定鄰居的任何新路由是否比 Loc-RIB 中已經存在的路由更佳。如果是這樣，它將替換它們。如果給定的路由被鄰居撤回，並且沒有其他路由到達該目的地，則該路由將從 Loc-RIB 中刪除，並且不再由 BGP 發送給主路由表管理器。

驗證下一跳是否可達後，如果路由來自內部(即 iBGP)，則要應用的第一個規則是檢查 LOCAL_PREFERENCE 屬性。如果有來自鄰居的多個 iBGP 路由，選擇 LOCAL_PREFERENCE 最高的路由。若有多個路由擁有相同的最高 LOCAL_PREFERENCE，此時，思科(Cisco)和其他幾家供應商的 BGP 路由器首先會考慮一個稱為 WEIGHT 的決策因素，該因素對於路由器而言是本地的，選擇具有最高 WEIGHT 的路由。LOCAL_PREFERENCE，WEIGHT 和其他條件通常可以通過本地配置和軟件功能進行操作。雖然這種操作超出了 BGP 標準的範圍，

但是很常用，比如使用 COMMUNITY 屬性。其他可能的因素，包含：優先選擇 AS_PATH 最短的路由、優先選擇 ORIGIN 屬性值最低的路由、優先選擇具有最低 MULTI_EXIT_DISC (MED)值的路由等等。

其中，AS_PATH 屬性記錄了某條路由從本地到目的地址經過的所有 ASN。

(五) BGP communities

BGP communities 在概念上可以視為屬性標籤(attribute tags)，可將其應用於(進入或傳出的)位址前綴，以實現某些共同路由目標(RFC 1997)。舉例來說，BGP 允許管理員設置 ISP 如何處理位址前綴的策略，但在實務上，這通常是有困難的或不可能的。例如，BGP 並沒有這樣的一個概念，允許一個 AS 告訴另一個 AS，限制某個位址前綴的廣告(advertisement)僅發送給某個區域的對等實體。然而，ISP 可以發佈一個列表，上面列出公開或專有社區，並為每個社區提供描述。RFC 1997 定義了三個具有全球意義的知名社區：NO_EXPORT 社區、NO_ADVERTISE 社區和 NO_EXPORT_SUBCONFED 社區。RFC 7611 定義了 ACCEPT_OWN 社區。常見的社區範例包括對於本地偏愛調整、地理限制或對等實體類型限制，DoS (Denial of Service)避免(黑洞)和 AS 優先選項。

2006 年擴展了 Extended Community Attribute，目的是擴展屬性的

範圍，RFC 4360 中記錄了此 Extended Community Attribute 的定義。IANA 管理 BGP Extended Communities Types。

(六) BGP Route Reflector

大型 AS 網路內，也可以選擇使用 iBGP。比如說，由於 BGP 使用 TCP 傳送路由訊息，當網路規模足夠大的時候，iBGP 比 OSPF 能更加穩定可靠的傳輸路由訊息。對於路由器來說，eBGP 和 iBGP 只是對應的參數不一樣，但都是通過同一個 BGP 行程(process)來運行，不會增加路由器的負擔。此時，只需要建立好 eBGP 和 iBGP 之間的連接，相應的路由就可以會通過 BGP 傳送。同時，AS 內部的中介路由器(intermediate router)的路由表可以比較小，因為它們並不需要關心 AS 以外的路由信息。

從 iBGP 對等實體獲得的路由資訊通常僅會分發給所有 eBGP 對等實體，不會傳遞給其他的 iBGP 對等實體，這是為了防止產生循環(loop)。不像 eBGP 對等實體通過可以通過 BGP 裡面的 AS_PATH 和其他元素過濾一個來自於自身 AS 的路由，但是 iBGP 運行在一個 AS 內部，沒有 AS_PATH。這也要求，同一個 AS 網路中的 iBGP 對等實體必須形成一個全連接(full-mesh)網路。這會使得 iBGP 對等實體之間的連接數隨著節點數量成二次方成長，對大型複雜的 AS 網路來說，是一個配置與管理上的困難。

BGP Route Reflector (RFC 4456)是一種特殊的 iBGP 路由器，它會將學習到的 iBGP 路由，傳遞給所有連接的 BGP Route Reflector 客戶端，可以避免 iBGP 對等實體全連接的要求，如圖 8 所示。

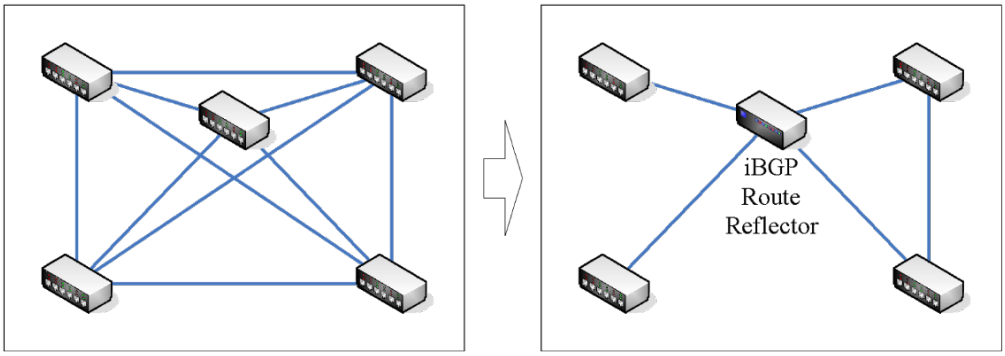

圖 8、iBGP Route Reflector

BGP Confederation (RFC 5065)是另一種處理方式，如將一個大型的 AS 裡面的 iBGP router，劃分到多個小的 Sub-AS，如圖 9 所示。

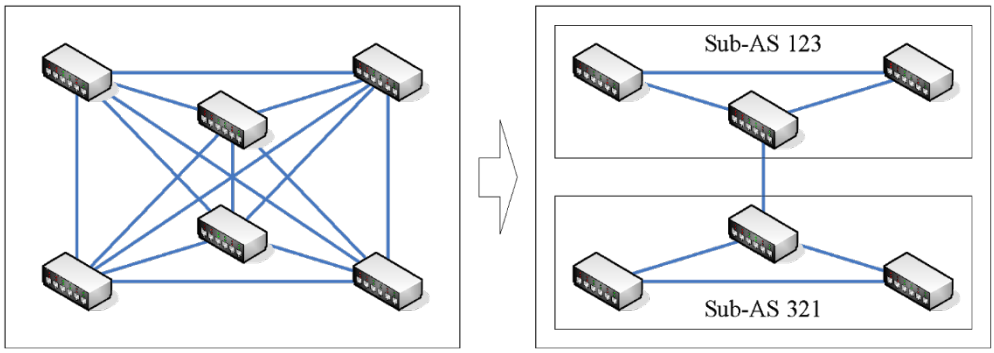

圖 9、BGP Confederation

(七) Graceful Restart

當一個 BGP 路由器與一個對等實體的連接中斷時，它會認為這個對等實體已經不能工作，它會刪除之前從這個對等實體學習到的路由訊息，並希望能以最快的速度收斂到整個網路最新的狀態。但與對

等實體連接中斷並不一定代表對路由無法進行了，可能只是與對等實體之間的 TCP 連接有問題，或是 Keep-alive 訊息丟失了。對等實體的路由功能仍有可能是正常的，具備正常封包轉發的能力。也有可能因為某些因素，這個對等實體正在重新啟動中，很快就能恢復到原來的狀態。這些狀況將造成該 BGP 路由器必須重新學習在之前一段相對短的時間內才剛剛被刪除掉的路由訊息。

如果對等實體之間的連接才剛剛斷開，BGP 路由器就開始刪除相關的路由，一會之後又很快地重新建立連接、新增路由，可能會造成 BGP 路由器上路由反轉，產生短時間內的路由環路或者路由黑洞。這樣的路由反轉甚至可能會傳遞到整個數據中心，不僅消耗了路由器控制平面的計算能力，甚至會引起整個網路的抖動。特別是現在許多 BGP 路由器是採取軟體實現的方式，軟體的維護與更新經常需要重新啟動 BGP 路由器。

RFC 4724 為 BGPv4 和 Multiprotocol BGP 新增了 Graceful Restart。使用 Graceful Restart，在 BGP 對等實體之間發生連接中斷的時候，BGP 路由器不會在很短時間內就刪除相對的路由信息，從而確保網路的穩定性。