支援接入與核心分離的路由協定擴展機制設計與實現
Design and Implementation of Routing Protocol Extensions Supporting Separation of the Core and Access Network

姚楠 Nan Yao, 楊水根 Shui-Gen Yang, 郭華明 Hua-Ming Guo, 張宏科 Hong-Ke Zhang, 秦雅娟 Ya-Juan Qin
北京交通大學電子資訊工程學院
ynatny@bjtu.edu.cn, ipv6ysg@163.com, guo522@gmail.com, hkzhang@bjtu.edu.cn

摘要

為解決現有互聯網的可擴展性問題，目前有許多研究方案提出將現有互聯網劃分為接入網路和核心網路兩個獨立的演進空間。在分離後的新網路環境當中，現有路由協定需要加入特定處理機制以提供準確的路由可達性資訊。本文在深入研究現有路由協議基礎上，設計和實現在核心與接入網路空間分離下對現有路由協定的擴展機制。

關鍵字：路由協定、接入與核心分離、消息更新、首碼過濾。

Abstract

To resolve the scalability problem in the existing Internet, there are many researches proposing the idea of separating the Internet into two independent evolving spaces, called the access and core network, respectively. In the new network environment, the existing routing protocols require additional mechanisms to provide accurate reachability information. In this paper, based on the in-depth study of the existing routing protocols, we design and implement the extensions to the existing routing protocols to support the separation of the two spaces.

Keywords：Routing Protocol, The Separation of the Access and Core Network, Message Update, Prefix Filtering.

1 緒論

隨著互聯網的飛速發展，互聯網的可擴展性受到越來越多運營商和研究機構的關注。現有網路由成千上萬個自治系統構成，通過在自治系統之間運行邊界閘道協定（Border Gateway Protocol, BGP）、內部運行路由資訊協定（Routing Information Protocol, RIP）、開放最短路徑優先協定（Open Shortest Path First, OSPF）或集成系統到集成系統協定（Integrated System to Integrated System, ISIS）維護路由可達性資訊。從網路結構上看，整個互聯網處於一個平面的路由空間當中。為達到可擴展性的需求，網路的規劃者在構建網路時往往遵循層次化構建的原則，通過將網路的規模和責任範圍劃分到相對較小的區域當中來降低網路複雜性。路由協定也依賴于聚合的路由首碼通告機制將核心網路路由器的路由表維持在一個穩定的規模。然而隨著用戶對於包括移動性、多家鄉和流量負載均衡等應用的需求不斷增加，核心路由表持續擴張，現有互聯網在可擴展性方面的弊端日益顯露出來[1]。

國內外研究機構就這一問題提出了許多對現有網路的改進方案，如[2][3]。其中，許多方案[4][5]都提到了將現有網路劃分為核心網路和接入網路兩個互相分離的路由空間，各自維護自己的路由可達性資訊，在二者之間引入兩個位址空間的映射，完成完整的通信流程。我們將這樣的改進方案統稱為分離映射通信機制（在下文中簡稱為分離網路環境）。在這樣分離的網路環境下，路由資訊也被劃分為兩個互不相干的獨立空間，路由協議只需要在兩個位址空間當中分別維護可達性資訊即可。為達到這樣的目的，現有路由協定需要加入特定機制適應新型網路環境，將兩個路由空間的資訊徹底隔離開來。

需要特別注意的是，路由協議本身也可以通過一定的配置完成這樣的任務。比如對於域內路由協議（RIP、OSPF），如果運行同樣的路由協議，可以通過同時啟動兩個不同的路由進程分別維護分離後兩個路由空間的資訊，來支援分離網路環境；對於域間路由協議（BGP），可以通過特定的配置——指定BGP進程攜帶的網路層可達性資訊（Network Layer Reachability Information, NLRI）——來區分不同的路由資訊。但是，這樣的做法為網路管理員帶來了額外的網路管理負擔，消耗了更多的設備資源；並且由於普遍存在的人為配置錯誤，本應分離的兩個路由空間資訊仍然會不可避免的洩露到另外一個空間當中，這就違背了分離映射機制設計的初衷，而且增加了維護網路的困難度和複雜度。所以，應該從路由協議本身的機制入手，解決這樣的問題。

本文工作主要關注三種最為常見並且應用最為廣泛的路由協定——RIP（RIPv2, RIPng）、OSPF（OSPFv2, OSPFv3）、BGP（BGP4, BGP4+）。

2 分離映射路由機制整體設計

2.1 設計思路

RIP是最為簡單的一種距離向量路由協定，它基於Bellman-Ford演算法，通過源和目的之間的跳數來衡量路由優先順序，擁有最短跳數的路由被安裝到路由表

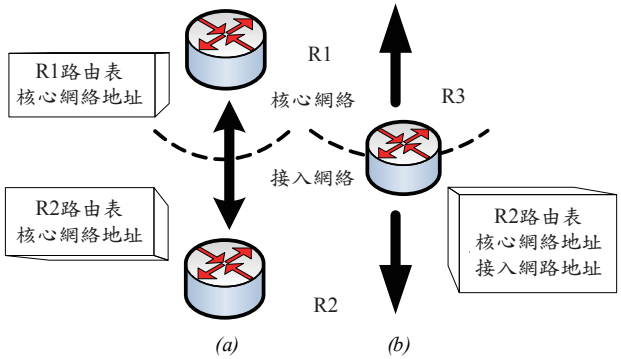
當中，並且路由器週期性和鄰居交互各自的路由表，從而得到全網的路由可達性資訊。OSPF是一種典型的鏈路狀態路由協定，OSPF路由器之間通過交互各自獲取的網路局部狀態資訊（包括網路拓撲結構和鏈路帶寬等資訊）獲取全網的鏈路狀態資訊，每個單獨的路由器再根據Dijkstra演算法計算出到達每一個網路的最佳路由條目，安裝到自己的路由表當中。BGP是當今域間路由協定的事實標準，它是一種策略驅動的路徑向量路由協定，它通過對由內部閘道協議（Interior Gateway Protocol, IGP）重分佈的和自BGP對等體學習到的每條網路層可達性資訊（NLRI）實施豐富的輸入輸出策略來控制路由選擇過程，將符合所有應用策略的最佳路由安裝到路由表當中，並依據當前安裝的路由轉發去往該目的地的流量。

從以上對三種典型路由協議的介紹來看，三種路由協議都屬於分散式計算路由協議。在分散式路由計算當中，路由守護進程依靠與它的鄰居交互路由資訊或鏈路狀態資訊完成對全網狀態的認知。所以，在分離映射的網路環境中實施這三種路由協定，需要在邊界路由器上修改路由協定流程，隔離兩個路由空間的可達性資訊。

在不同網路環境下完成隔離路由可達資訊的任務有不同處理方式。當分離後兩個空間邊界位於路由器之間的鏈路上時，路由協定並不需要做特殊處理，如圖一(a)所示；當分離後空間邊界位於特定路由器上時，邊界路由器自身需要認知其各介面分別屬於哪個路由空間，據此過濾從一個路由空間（如核心網路）收到的路由資訊，抑制該資訊在另一路由空間（如接入網路）中的轉發，實現路由可達性資訊的隔離，如圖一(b)所示。

2.2 整體設計

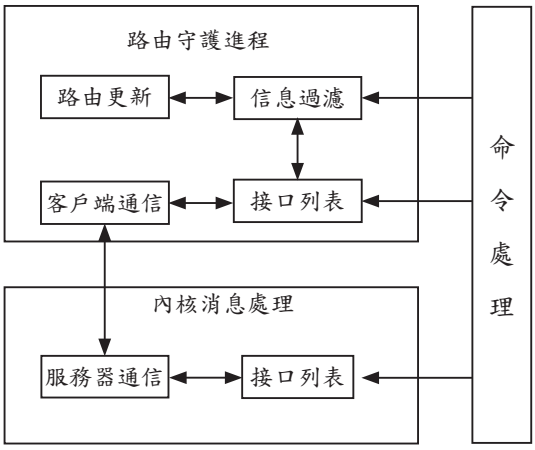
不同的路由協定對需要更新的消息定義有所不同（將在後面的章節具體講述），但作為路由協定的一個整體，協定具體實現的整體框架是基本相同的[6]。本文採用標準Linux上的路由協議實現為基礎。在充分研究原有路由協議處理機制之後，對原有路由協議進行擴展，添加特定更新消息處理、命令處理、介面狀態處理等機制，與原有路由協議構成一個整體。



圖一 網路分離位置示意圖

Fig.1 Location of the separation

分離環境下路由機制擴展的設計總體分為三個功能組成部分（程式啟動時運行三個獨立的Linux用戶層程式）：路由守護進程（Routing Daemon）、內核消息處理和命令處理（即人機交互介面），程式具體細分為七個功能模組：路由更新、資訊過濾、介面列表（路由守護進程和內核消息處理各自維護獨立的列表）、用戶端通信你、伺服器通信和命令處理，如圖二所示。



圖二 分離映射路由協議整體設計

Fig.2 The Design of Routing Protocol in Separation and Mapping Mechanism

每個路由協定擁有獨立的路由守護進程，並處於相互獨立的進程空間當中。但它們都依靠相同的介面——伺服器/用戶端通信模組——和內核消息處理進程通信，從內核消息處理進程中接受關於路由器介面狀態和路由表的資訊。內核處理進程在初始化時可以根據給定資訊或人為配置確定介面所屬路由空間，將其編入每個空間特定介面列表並用標誌位元加以區分，通過伺服器／用戶端通信模組將分類後介面資訊通知路由守護進程。守護進程得到更改介面類型的消息後，更新調整自己進程空間中的介面列表。上述介面操作也可以在程式運行過程中通過命令處理機制即時更改。各路由守護進程在與鄰居路由器交互資料之前應用相應首碼資訊過濾機制，對不同路由空間的路由資訊分類處理，實現路由可達性資訊分離。

3 路由協定更新消息實現

各個路由協議都有不同的路由選擇和消息更新機制，本文在深入研究現有路由協議的路由表更新和維護處理、通信消息更新和處理等機制的基礎上，研究並實現RIP、OSPF和BGP路由協定路由更新消息上的資訊過濾機制，對消息的輸入輸出過程進行跟蹤，在不影響現有路由協議良好運行的前提下，分離核心網路和接入網路兩個路由空間的可達性資訊，使邊界路由器成為分離映射機制下的重要設備。

以下本文將重點介紹不同路由協定更新消息的具體處理機制。認識到所有路由協定整體設計思路的共通性以及不同路由協議具體處理機制的差異性，我們按照消息的不同處理機制將路由協議分為三類：路由表項更新；鏈路狀態更新；輸入輸出策略更新。

2.2 節將詳細介紹不同類型路由協定的處理方式。

3.1 分離後位址空間區分的考慮

在分離映射通信機制當中，大多數方案都將現有網路劃分為兩個完全分離的位址空間，用特定比特區分兩個空間的位址。這就要求全網設備都對新位址格式有一致的認知，現有網路的設備必須要經過特定的轉換機制才能被納入新位址格式體系中。這種做法降低了網路可逐漸部署的特性。

另外一種思路則是通過路由器或者引入其他設備認知兩類位址空間，這樣就需要依靠管理員的配置或者程式自身維護位址空間特徵來區分兩類位址空間。

在具體實現中，本文將兩種位址空間的區分方式同時考慮，同時支援攜帶與不攜帶特定標誌位元區分位址空間的方式，增強了程式的相容性。

3.2 分離映射路由協定消息更新具體流程

本節主要闡述各路由協定的具體實現流程進行。

(1)RIPv2和RIPng屬於路由表項更新類型，RIP是典型的距離向量路由協議，依靠與鄰居交互自己的完整路由表來維護全網的路由可達性。RIP路由器認為在每一個配置並啟動了RIP進程的路由器介面上都有可能存在著RIP的鄰居，於是當路由更新計時器計時到達時，RIP進程遍曆RIP使能介面列表，將自己的路由表發往特定的組播位址。所以，對於RIP路由守護進程的處理較為簡單：

在設備啟動和初始化時依據內核消息處理進程通知，設置各介面類型；

對內核指派和直連路由進行分類，加入標籤後編入路由表；

在發送路由更新之前，依據介面類型判斷需要發送的正確類型路由表項；

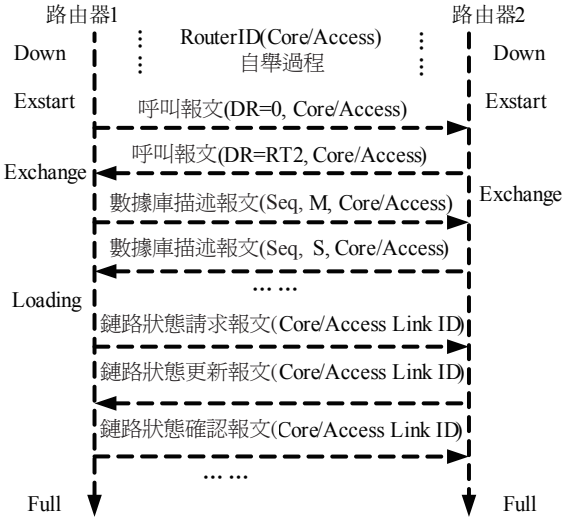
對於接收到的路由表更新消息進行必要的資訊過濾，對相應路由表項執行更新；

程式運行中接收命令處理模組指示，即時更改介面列表中相應介面列表以及路由表項的內容。

(2)OSPFv2和OSPFv3屬於鏈路狀態更新類型，作為鏈路狀態路由協定的代表，OSPF並不直接交互各自路由器的路由表項，而是通過泛洪機制將各個路由器瞭解到的網路局部鏈路狀態特徵資訊傳遞給所有的OSPF路由器，由於沒有在網路中傳遞實際的路由表項，所以分離映射機制下OSPF路由協議的擴展設計應該對可能導致生成錯誤路由表項的鏈路狀態更新加以限制，在構造鏈路狀態更新消息和處理泛洪消息時加入過濾機制。

同時，在選定路由器識別字（Router ID）時，邊界路由器也要同時選定兩個Router ID——分別屬於兩個不同的位址空間——用於在核心網路和接入網路中的DR選舉、網路狀態構成和鏈路狀態更新消息的形成。

OSPF路由器在鄰居交換過程中，交互多種資料包類型，並且各種資料包之間相互關聯，相互依存，完成分離映射OSPF設計需要對整體程式流程的全面考慮。基於對OSPF鄰居消息交互過程的研究，本文提出分離映射環境下OSPF鄰居交換過程，如圖三所示。首先在各個路由器RouterID自舉過程中識別網路空間類型，選擇出適合這個網路類型的路由器識別字，用以維持後續資料庫描述和鏈路狀態資訊的交互。這個路由器識別字也可以由網路管理員通過命令進行配置。之後，在向鄰居發送資料報和接收鄰居的鏈路狀態更新時，進行基於鏈路識別字（Link ID）的首碼資訊過濾，檢查並維護OSPF鄰居鏈路狀態請求列表和重傳列表，清除錯誤的表項。最後，OSPF鄰居完成基於請求和重傳列表的鏈路狀態更新，當鄰居請求和重傳列表為空時，達到全連接狀態，進行計算路由表的操作。



圖三　分離映射OSPF鄰居交互過程

Fig.3 The Process of Exchange between OSPF Router Neighbours

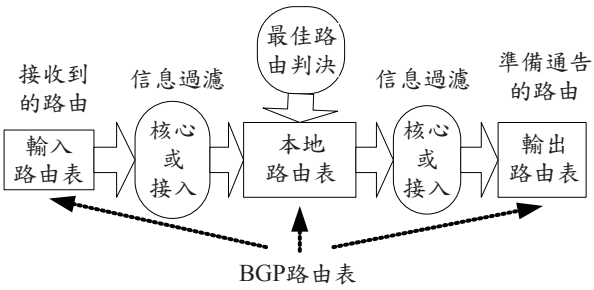
在分離映射環境下，OSPF路由器實際上沒有得到全網的拓撲和鏈路狀態資訊，但是通過對於OSPF鄰居消息交互過程的修改，使得除邊界路由器以外的其他路由器認為得到的資訊已經是全部網路資訊，及時終止對於另外一個路由空間資訊的請求，計算本路由空間的可達資訊。

(3)BGP4和BGP4+屬於輸入輸出策略更新類型。BGP被稱為路徑向量路由協定，是一個基於網路層可達性資訊的輸入輸出策略集[7]。回顧RIP路由守護進程的處理可以發現，對接收和發送路由表項的過濾檢查可以看作對RIP路由表實施輸入輸出策略。因為RIP只維護一個路由表，我們需要對表中具體表項做詳細的過濾處理。而從BGP的路徑選擇

過程來看，它為分離映射網路環境提供了良好的支援。在基於策略的路徑選擇過程中，BGP分別維護輸入、本地和輸出三個路由表，分別應用於不同的過濾及屬性控制機制。

需要提出的是，無論是否過濾BGP網路層可達性資訊通告（NLRI）內容，輸入輸出策略本身都可以幫助管理員完成對於某些特定路由條目的過濾。但對於網路管理員來說，這樣的任務在路由條目繁多的核心網路當中是非常容易出錯的，一旦出現這樣的人為錯誤，將可能造成極為嚴重的損失。

分離映射下的BGP路由協定需要在輸入和輸出策略執行時進行必要的首碼資訊過濾，如圖四所示。這種資訊過濾主要應用在邊界路由器上。特別地，這種BGP路由資訊的分離適用于邊界路由器的兩側空間屬於同一個自治系統的情況。在自治系統（Autonomous System, AS）邊界，外部BGP（exterior BGP, eBGP）對等體之間可以通過重分佈特定路由首碼的方式人工隔離兩類位址空間。



圖四　分離映射BGP路徑選擇過程
Fig.4 The Process of Route Selection of BGP

(4)對分離映射路由協議修改的核心問題在於深入瞭解路由協定交互資訊的原理、流程和具體格式。下面對涉及到的各路由協定交互資料包格式做統一說明（針對IPv4版本路由協議——RIPv2、OSPFv2、BGP4），如表1所示。值得注意的是，在分離映射機制下的路由協議設計中，採用了分離兩個空間路由器介面列表的方法，但並不需要分別構建核心網路和接入網路分離的兩個路由表，用以維護不同路由空間的可達性資訊。因為介面列表是路由器頻繁使用的一項全局資料結構，在消息處理、消息過濾、命令顯示等功能組成部分都要對不同類型的介面進行訪問，所以單獨建立兩個獨立的列表結構有助於程式的優化運行；而在分散式路由計算模型當中，各個路由器分別計算各自可以到達的目的地路由，邊界路由器作為銜接兩個路由空間的重要設備，可以同時到達接入網路和核心網路的目的地。通過介面列表，程式可以判斷路由資訊與該資訊到來的介面的對應關係並作輸入和輸出流程上的必要過濾，而不需要在內核消息處理與路由守護進程模組同時存儲相同的路由表項資訊，從而用最小改動滿足必要功能。

4 結論

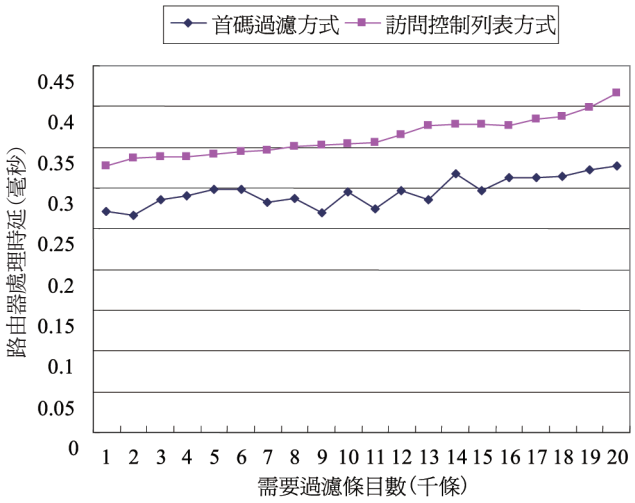
本文提出了在分離映射機制下的三類路由協議設計實現。針對目前網路發展遇到的可擴展性問題，研究機構提出許多方案解決這一問題。由於在可部署性、網路升級開銷方面的優勢，分離映射成為受到強烈關注的解決方案之一。針對於分離映射後的網路環境，本文提出適當的路由協議修改，著重體現以下幾個方面的目標：

高效性：對於現有路由協議做儘量少的修改，簡化對於首碼資訊的過濾處理操作，引入較少的時延。傳統訪問控制列表需要對要進行過濾的首碼資訊精確匹配，每過濾一個首碼就需要在路由器配置當中添加一個控制列表條目；改進後的機制通過明確定義分離後兩個空間當中的位址格式，用特定比特位來區分位址，所以路由器配置中不需要繁瑣的首碼條目過濾，而是通過判定首碼中的特定比特來進行過濾，加大了處理效率並降低了演算法的複雜度。

表1　各路由協定需要過濾的交互消息
Table 1 Interactive Messages Filtered of Routing Protocols

路由協議	消息類型	需要過濾欄位	消息功能描述
RIPv2	RIPv2回應消息	IP地址域	定時產生或對鄰節點更新請求的回應，包含網路位址的路由資訊
OSPFv2	呼叫報文	路由器ID，DR，BDR	屬於OSPF Hello呼叫協議，用於建立和維護鄰接節點關係
	資料庫描述報文	路由器ID，LSA鏈路狀態ID	描述OSPF路由器鏈結狀態資料庫內容，用於交換前初始化鄰接關係
	鏈路狀態請求報文	路由器ID，鏈路狀態ID	請求鄰居路由器鏈路狀態資料庫的特定部分
	路由器LSA	通告路由器，鏈結ID	描述路由器鏈路到某區域的狀態和代價
	網路LSA	通告路由器，相連路由器	描述路由器直連網路中所有鏈路狀態和代價資訊的集合
	Type - 3聚合LSA	通告路由器，鏈路狀態ID	向OSPF網路中鄰居區域通告本區域聚合的路由資訊
	Type - 4聚合LSA	通告路由器，鏈路狀態ID	向OSPF網路中自治系統邊界路由器通告網路外部路由
	AS部LSA	鏈路狀態ID，轉發位址	描述OSPF自治系統的外部目的地可達資訊
BGP4	UPDATE報文	網路層可達資訊	列出將要通告給遠端對等體的路由目的地

我們在實驗環境中選取兩種方式進行對比：(1)在核心路由器上配置首碼訪問控制列表，由終端持續發送資料包，記錄資料包時延；(2)在核心路由器上配置分離網路環境，依據比特位元過濾資訊，分別向核心路由器加入等量的路由表條目，記錄路由更新消息接收時間。試驗結果如圖五所示。可以看出，傳統訪問控制列表方式處理時延隨需過濾條目數增長呈明顯上升趨勢，而首碼過濾方式則基本維持在同樣的水準。



圖五　首碼過濾與訪問控制列表方式處理時延比較

Fig.5 The Comparison between the Method of Prefix Filtering and Access Control List

安全性：在路由守護進程輸入輸出方向同時過濾，防止第三方錯誤資訊干擾分離映射下路由協定工作進程；

可部署性：通過豐富的介面、調試配置命令，使網路管理員通過簡單配置操作就可以使路由器正常工作，並且能夠通過少量配置更改適應網路拓撲的變化。

基金專案

國家973重點基礎研究發展規劃（2007CB307101）；教育部高等學校科技創新工程重大專案培育資金（706005）；國家863計畫（2007AA01Z202）。

參考文獻

[1] T. Li, *"Design Goals for Scalable Internet Routing, Internet draft （work in progress）,"* draft-irtf-rrg-design-goals-01, July, 2007.

[2] G. Tsirtsis, P. Srisuresh, *"Network Address Translation-- Protocol Translation （NAT-PT）,"* IETF RFC 2766, February, 2000.

[3] B. Zhang, D. Massey, D. Pei, L. Wang, L. Zhang, R. Oliveira, and V. Kambhampati, *"A Secure and Scalable Internet Routing Architecture （SIRA）,"* Technical Report TR06-01, University of Arizona, April, 2006.

[4] D. Farinacci, V. Fuller, and D. Oran, D. Meyer, *"Locator/ID Separation Protocol （LISP）, Internet draft （work in progress）,"* draft-farinacci-lisp-06, November, 2007.

[5] C. Vogt, *"A Scalable and Backwards-Compatible Solution for Provider-Independent Addressing and IPv4/IPv6 Interworking,"* On-line http://users.piuha.net/chvogt/pub/ 2008 /vogt-2008-six-one-router-spec.pdf, February, 2008.

[6] 張宏科、張思東、蘇偉，*路由器原理與技術*，北京：國防工業出版社，2005。

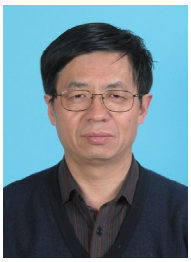
[7] R. Zhang, M. Bartell，黃博、葛建立譯，*BGP設計與實現*，北京：人民郵電出版社，2005。
R. Zhang, M. Bartell, *"BGP Design and Implementation,"* Cisco Press, 2005.

作者簡歷

姚楠（Nan Yao），男，1985年2月出生，北京朝陽區人，現為北京交通大學在讀博士研究生，主要研究興趣：IP網路路由協定及相關技術、下一代網路路由理論、互聯網路由體系結構、域間流量工程及路由優化等。

郭華明（Hua-Ming Guo），男，1982年9月出生，吉林四平人，2005年9月獲得北京交通大學（原北方交通大學）學士學位。目前為電子學院下一代互聯網研究中心博士研究生。研究方向主要為普適網路、網路路由協定和網路處理器。

楊水根（Shui-Gen Yang），男，1981年8月出生，福建省三明市清流縣人，2004年6月在北京交通大學獲得學士學位，目前正在北京交通大學攻讀博士學位。研究興趣包括：移動IP、移動互聯網、下一代網路的移動性管理等。



張宏科（Hong-Ke Zhang），男，1957年9月出生，山西大同人，北京交通大學教授，博士生導師。目前主要從事下一代資訊網路關鍵理論與技術的研究工作，並作為首席科學家主持國家973專案「一體化網路與普適服務體系基礎研究」的研究工作。

秦雅娟（Ya-Juan Qin），女，山西晉城人，博士，博士生導師。2003年獲北京郵電大學工學博士學位。近年來主要從事互聯網體系結構、移動互聯網路由與交換、寬頻無線通信等領域的技術研究，主持或主研完成多項國家863、國家自然科學基金及國家發改委專案。