一體化網路中身份與位置映射關係的解析方法研究
A Study on the Resolution Methods of Mapping Relationship Between Identity and Location in the Universal Network

劉曉波 Xiao-Bo Liu, 姚楠 Nan Yao, 董平 Ping Dong, 周華春 Hua-Chun Zhou, 秦雅娟 Ya-Juan Qin
北京交通大學電子資訊工程學院
xbliu2008@gmail.com, ynatny@gmail.com, 03211175@bjtu.edu.cn, hchzhou@bjtu.edu.cn

摘要

一體化網路採用了身份與位置分離映射的機制，為使相關網路節點能夠快速準確地獲取映射資訊，需要在網路中引入一個映射關聯資料庫，以便提供全網統一的映射資訊存儲及解析服務。本文針對映射關聯資料庫在查找效率、可擴展性以及容錯性等方面的需求，提出了基於DNS（Domain Name System）遞迴分級查找、基於結構化DHT（Distributed Hash Table）網路查找和基於BGP（Border Gateway Protocol）集中式查找三種從身份資訊解析歸屬映射伺服器的方案，並進一步分析了三種方案的優缺點。本文的研究工作對於映射關聯資料庫查詢機制的設計以及部署方案的實現具有重要參考價值。

關鍵字：一體化網路、分離映射、映射關聯資料庫、歸屬映射伺服器。

Abstract

Under the mechanism of separating and mapping of identity and location in the universal network, mapping information should be received fast and accurately by related network nodes. Therefore, a database providing storage and resolution services of the unified mapping information needs to be introduced into the network. To meet the demands of the mapping relationship database, such as search efficiency, scalability and fault-tolerance, this paper presents three solutions for storing and resolving the mapping information of identifier and corresponding home identifier mapping server. The schemes based on DNS(Domain Name System), DHT(Distributed Hash Table) and BGP(Border Gateway Protocol), respectively. Furthermore, the paper gives characteristic analysis for each solution. This research work is valuable for the design and implementation of the inquiry mechanism of the mapping relationship database.

Keywords: Universal Network, Separating and Mapping, Mapping Database, Home Identifier Mapping Server.

隨著互聯網技術的快速發展，身份與位置分離成為當前互聯網領域研究的熱點之一。身份與位置分離為互聯網中的路由擴展性、移動性、安全性等問題的解決提供了一個良好的基礎。文獻[1][2]研究和探索新一代資訊網路體系的基礎理論，創建了一體化網路的體系結構模型，並原創性的提出接入標識與交換路由標識分離映射理論，以能夠在一體化網路的基礎上支援多種服務。

在一體化網路中，網路層標識空間劃分為兩部分。一部分是接入標識（Access Identifier, AID），表示終端接入的身份，全球唯一，且不隨著終端的移動而改變；另一部分是交換路由標識（Switching-Routing Identifier, SRID），用於核心網的路由和轉發。接入交換路由器（Access Switching Router, ASR）是完成分離映射的主要實體，它為接入的終端分配一個合法的SRID。在通信連接建立時，ASR通過查詢獲得通信對端的SRID；在通信過程中，對用戶的資料包進行標識替換以使之能在核心網和接入網之間傳輸。這一標識替換的過程就在一體化網路中實現了身份與位置分離映射的機制。

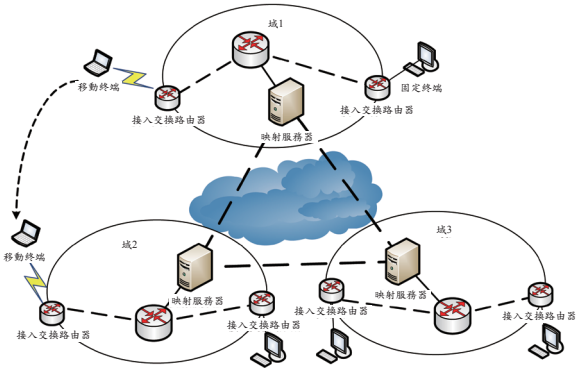
1 一體化網路身份與位置映射資訊管理機制

在一體化網路中，因為表示身份的接入標識是不允許進入核心網的，所以標識的替換過程必須在ASR處一次性完成。在這種嚴格的身份和位置分離的網路中，為了對標識的映射資訊進行更加有效的管理，以及避免由分離映射機制引起的資料包延遲和丟失，一體化網路在核心網內引入一個稱為映射伺服器（Identifier Mapping Server, IDMS）的功能實體，存儲一個區域網路內終端的AID和SRID的映射資訊。IDMS還能夠提供映射資訊的查詢、插入、刪除和修改服務。

一體化網路借鑒了移動通信網路的設計思想，按照地理位置、用戶規模等因素將網路劃分為若干個管理域，並使用「歸屬域」表示終端初始接入的家鄉域。每個域內設置一台或多台映射伺服器，對所有歸屬該域的終端來說，稱為歸屬映射伺服器。為了避免單個映射伺服器存儲過多的映射資訊，採用了分散式的映射資訊存儲結構。每台映射伺服器只保存歸屬本域及當前移動到本域的終端映射資訊。因此，在某些映射查詢或更新過程中，會涉及到域間資訊交互。圖一採用一個簡單的例子說明了一體化網路中多個管理域之間的關係。在圖一

中，當歸屬域1的移動終端（Mobile Node, MN）由歸屬域移動到域2時，域2的映射伺服器在獲得並存儲了MN的接入標識到新交換路由標識的映射關係後，要向域1的映射伺服器（即MN的歸屬映射伺服器）進行映射關係更新。當域3的一台固定終端要與MN進行通信時，固定終端接入的接入交換路由器為了進行標識替換，需要查找MN的交換路由標識，如果本地緩存中沒有所需的映射資訊，則向域3的映射伺服器查詢。域3的映射伺服器還要根據MN的接入標識找到其歸屬域（域1），向域1的映射伺服器查詢。

不同管理域之間的映射伺服器通過一系列的消息交互，能夠有效解決由終端跨域移動帶來的映射關係管理問題。然而，這裏還存在一個問題：映射伺服器之間通信時，要想獲得其他域的映射資訊，必須首先定位相關的映射伺服器，即某個接入標識的歸屬映射伺服器。因為在上述方案中，只有找到該AID的歸屬映射伺服器，才能最終找到其對應的SRID。為了快速準確地解析歸屬映射伺服器，必須引入一種資料庫來提供歸屬域資訊的存儲及解析服務，這樣，映射伺服器之間就構成了一個維護（接入標識，歸屬映射伺服器）映射關係的資料庫網路。需要設計相應的網路結構和通信協定，使這個資料庫網路盡可能高效的維護和擴散映射關係。



圖一　一體化網路拓撲示意圖

2 接入標識與歸屬映射伺服器的解析方案設計

為了實現由接入標識到歸屬映射伺服器的快速而準確的定位，本文提出了資料庫網路的三種部署方案，並對每種方案的特點進行分析。因為接入標識數量巨大，終端的頻繁加入和移動導致AID和SRID的映射資訊不斷變化，對接入標識的定義、分配和管理將直接影響整個網路通信的效率。所以，需要為每個域分配連續的接入標識塊，這些接入標識塊以不同長度或不同數值的接入標識首碼進行區分。這樣，映射伺服器中資料庫的存儲內容就可以是一種按照一定規則（如以接入標識或接入標識首碼為索引）排列的目錄，以達到方便管理和簡化資料庫的優化目的。

2.1 基於DNS遞迴分級查找的解析方案

功能變數名稱系統（Domain Name System, DNS）[3]是一種用於TCP/IP應用程式的分散式資料庫。它提供主機名字和IP位址之間的轉換及有關電子郵件的選路資訊。每個站點（校園、公司或部門）保留自己的資訊資料庫，並通過運行伺服器端程式向Internet上的其他用戶端程式提供DNS查詢功能。DNS在整體組織上是個樹形的分散式名稱對應系統。DNS的名字空間也具有層次結構，命名樹上任何一個節點的功能變數名稱就是將從該節點到最高層的功能變數名稱串聯起來。這種嚴格樹形的命名結構使得DNS在查詢時可以按照邏輯拓撲，逐級遞迴地找到任何需要的功能變數名稱與IP位址的對應關係。

與DNS系統對應，一體化網路中可以採用類似DNS這種邏輯樹形的層次結構部署IDMS。

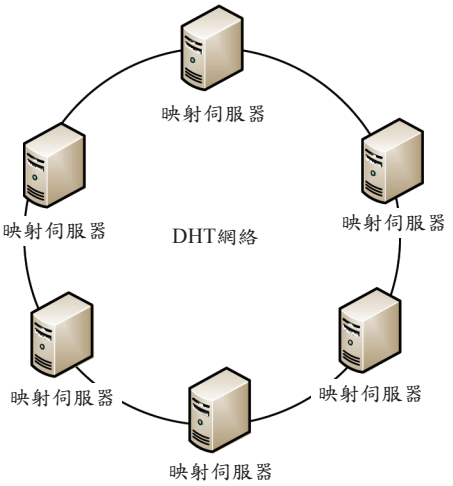
首先，IDMS需要提供接入標識到歸屬映射伺服器的快速解析資訊，即接入標識名稱空間到歸屬映射伺服器位址的轉換,這與DNS系統的設計初衷非常類似。並且，由於整個路由系統和接入網路龐大的標識空間，每個站點並不可能保存全球資訊，採用分散式結構是一種合理的選擇。其次，接入標識與功能變數名稱系統的名字空間同樣具有層次結構和標識站點身份的作用。這種相似性引導我們將接入標識到歸屬映射伺服器的快速定位與DNS查詢機制結合到一起，形成完整的伺服器查詢機制，為網路通信提供保證。

當然，這種設計要求在網路最初設計的時候對整個網路進行比較清晰的分級規劃，進行查找時通常會採用自頂向下的方式，簡單清晰，有比較高的查詢效率。

但是，這種設計也存在一些缺點：第一，DNS系統架構有一些先天無法克服的缺點：DNS使用相對集中的管理方式，配置複雜，極大地限制了服務的靈活性，可擴展性比較差。第二，如果接入標識的定義符合網路的分層結構，必然需要引入一定的位置資訊，這樣安全性不夠高。第三，DNS的樹形結構層次和查詢機製造成資料更新上的巨大延時[4]。映射伺服器查詢的觸發比較頻繁，並且直接關係到網路的連通性，這難以適應即時通信、網路視頻等即時性要求很高的資料流程量。

2.2 基於結構化DHT網路的解析方案

DHT全稱為分散式哈希表（Distributed Hash Table）[5]，是一種分散式存儲方法。DHT各節點並不需要維護整個網路的資訊，僅負責一個小範圍的路由及存儲一小部分資料，從而實現整個網路的定址和存儲。DHT網路模式有效減少了資源定位的開銷，能夠自適應結點的動態加入和退出，有著良好的可擴展性、魯棒性、負載均衡性和自組織能力。而且，由於採用了相容哈希函數獲得識別字並採取特定的機制組成確定的重疊網拓撲結構，DHT可以提供精確的發現。



圖二　基於結構化DHT網路的映射伺服器組織形式

如圖二所示，在這個方案中，由多個IDMS組成一個分散式網路，它基於DHT網路下的Chord演算法來設計，每個IDMS在網路中的地位相等，共同維護整個網路中所有歸屬域的管理資訊。

Chord中每個關鍵字和節點都分別擁有一個識別字。所有節點按其節點識別字（NodeID）從小到大沿著順時針方向排列在一個邏輯的標識圓環上（即Chord環）。Chord的映射規則：關鍵字識別字（keyID）為K的（key, value）對存儲在NodeID等於K或者在Chord環上緊跟在K之後的節點上，這個節點被稱為K的後繼節點，表示為successor（K）。為了加快查詢速度，Chord使用擴展的查詢演算法，每個節點維護一個路由表，稱為指標表（finger table）。任何節點收到查詢關鍵字K的請求時，首先檢查K是否落在該NodeID和它的後繼NodeID之間，如果是的話，這個後繼節點就是存儲目標（key, value）對的節點。否則，節點將查找它的指針表，找到表中NodeID最大但不超過K的第一個節點，並將這個查詢請求轉發給該節點。通過重複這個過程，最終可以定位到K的後繼節點。

在Chord中，節點的NodeID是對節點的IP位址或再加上埠號進行哈希運算而得到，keyID是對資源名或其關鍵字進行哈希運算而得到的，所以NodeID和keyID是隨機的哈希值。為了適應一體化網路映射資訊系統的需求，需要對這種隨機性進行修改[6]。本方案用接入標識做Chord環上的keyID，使用和AID對應的歸屬映射伺服器為value。Chord環上的節點使用AID作為NodeID，每個域的IDMS為其管轄的接入標識創建Chord節點。當一個域內的接入標識可以彙聚成一個有相同首碼值的標識塊時，可以為這個標識塊創建一個（key, value）對，將標識塊中的最大值AID（而不是一個隨機的值）作為NodeID，這樣做既確保了一個域的映射伺服器擁有並管理歸屬本域的接入標識在Chord環上對應的部分，又可以減少存儲條目和處理時延。當映射伺服器在本地緩存找不到（接入標識，歸屬域）的對應關係時，就會發送一個映射查詢請求消息。經過在Chord環上的查找，將得到一個正確的回應消息。

Chord網路的一些性能與我們對歸屬映射伺服器資訊系統的期望目標一致，如高的網路工作效率、負載均衡、可擴展性、容錯性，可以在系統設計中充分利用這些優點。但是，該方案還存在以下缺點：如果Chord環上有N個節點，則整個演算法的路由跳數至多為LogN，查詢步驟複雜度為 O（log N）。文獻[7]中提到，傳統的DNS一次完整的查找過程平均需要經過2跳，也就是需要經歷兩次查詢。而在純Chord網路環境下，平均一次完整的查找過程需要5跳，因此在系統查找時延方面會出現很大的問題。而且Chord查詢的開銷隨著節點數目增長而呈對數級增長，查找速度隨著網路規模的擴大將受到越來越大的局限。同時，動態網路的維護需要週期性的網路維護包，當網路規模增大時，網路負載會加重。

為瞭解決上面的問題，可以從下三個方面考慮。第一，在每個節點的指標表上添加更多的路由資訊，從而提高整個系統的查詢效率。這樣做的代價是增加了節點的存儲負擔，它是一個以存儲容量換取查詢效率、以空間換時間的折中問題。隨著存儲設備容量的不斷增加，這一設計可以以較小的犧牲為代價，換來系統性能的明顯提升。第二，採用緩存的方法提高被頻繁訪問資料的查找效率。第三，採用分層的Chord環結構來維護系統路由資訊，多環結構可以有效解決網路規模增大帶來的問題，提高系統的查找效率，但這種方法需要整個網路結構做出較大的改變。

2.3 基於BGP的解析方案

基於遞迴分級查找的方法和基於結構化DHT網路的方法都是基於分散式資料庫的思想，分散式資料查詢所產生的網路傳輸時延是影響其查詢效率的主要因素。為了對存儲內容進行更加有效的組織和管理，提高查詢回應速度，可以讓映射伺服器維護一個全局性的資料庫[8]，保存全網所有接入標識和其歸屬映射伺服器的資訊。為了實現這種全局資料庫的一致和同步，一個IDMS應該能夠自動獲得其他IDMS的資訊，並通過發送更新消息向外域廣播其自身的資料庫資訊。

維護各個IDMS中資料庫的同步和一致需要比較大的通信開銷，需要有一種可靠的、即時的機制將資料庫內容的變化同步到所有其他IDMS上。為此，本方案提出了通過修改邊界閘道協議（Border Gateway Protocol, BGP）[9]，並使其獨立運行在IDMS之間來完成歸屬域資訊的更新和傳遞。BGP是一種用來在自治系統之間交換網路層可達性資訊的域間路由選擇協定。由於對大型和複雜網路強大的支援能力，使得BGP在現有互聯網中獲得了大規模的部署。不難發現，BGP為了維護路由可達性資訊而具有的一些特性，同樣符合我們對接入標識到歸屬映射伺服器對應資訊維護管理的需求，兩者在本質上存在著許多相似點。具體如下：

2.3.1 可靠性

BGP應用傳輸控制協議（Transmission Control Protocol, TCP）提供可靠性傳輸，解決BGP資料包分

段、重傳、確認和前後順序等問題。同時，為了維護路由選擇資訊的精確性，當某些路由變得不可達時，BGP會迅速地從它的對等體中撤回這些不可達路由。

映射伺服器需要快速準確的定位AID的歸屬域，歸屬域查詢的失敗會直接影響標識映射關係的查詢，進而導致通信的失敗。並且，當某些AID的歸屬域發生轉移或變得不可達時，維護協定應該迅速在伺服器網路中更新這些不可達資訊。

2.3.2 穩定性

在大規模網路中，大量路由的不穩定震盪將會對網路產生災難性的影響。通過豐富的計時器以及路由衰減、溫和重配置等機制，BGP可以很好的抑制網路中出現的路由起伏。

接入標識對應的歸屬映射伺服器資訊需要在網路中長時間的維持，甚至少量的震盪都會直接影響網路為終端提供服務的品質。

2.3.3 靈活有效的更新機制

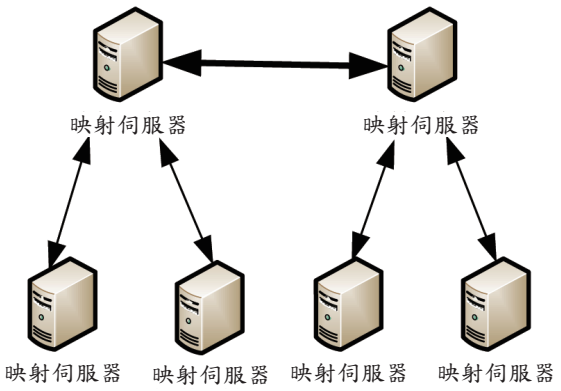
BGP使用「遞增式」的更新機制，這種機制在最初建立好連接、交換完整的路由資訊以後，在對等體之間就只交換關於那條路由的更新資訊，大大減少了鏈路開銷和存儲開銷。

由於域的結構相對比較穩定，引起資料庫更新消息的因素主要有接入標識的新分配和作廢，域的增加和減少，域的細分和合併，所以映射伺服器之間在最開始交換完整的資料庫資訊以後，只需要交互少量的消息進行更新維護，以減少查詢時延及網路傳輸負擔。

2.3.4 分層彙聚的思想

分層彙聚本身是維持BGP可擴展性的重要原則，BGP有很多諸如路由聚合的路由策略，從而關於一組網路的資訊可以表示為一個路由,具有了很好的伸縮性和穩定性。

本方案要求映射伺服器有足夠的存儲容量並提供快速準確的檢索能力，這需要有複雜的存取結構和有效存取資料的技術。為了增強方案設計的可行性，在實際映射伺服器部署中可以考慮引入分層彙聚的網路構建思想，以減少其他伺服器的存儲容量和協定消息數量。



圖三 基於BGP的層次化映射伺服器組織形式

如圖三所示，在一定範圍內的相鄰幾個域中選擇一個域的映射伺服器，存儲這幾個相鄰域內接入標識到其歸屬映射伺服器的對應資訊。這樣，逐級向上彙聚，最後由少數幾個映射伺服器集中管理全局的資訊。如果某個IDMS在本地資料庫內找不到某個域的位址，則可以向上層IDMS查詢。同級伺服器之間的同步，以及上下級伺服器之間的更新、查詢等資訊都可以通過維護協定傳遞。

當映射資訊採用分散式方式存儲時，每個映射伺服器僅存儲所管轄網路範圍內的接入標識的歸屬域資訊，所以映射資訊表並不會對映射伺服器形成過大的存儲壓力。而本節所述的歸屬資訊表採用的是集中式的存儲方式，所以下面對歸屬資訊表的容量進行分析，以判定集中式的歸屬資訊表存儲方式是否合理。為了分析歸屬資訊表的存儲容量，定義如下參數：

nAID：歸屬資訊表中條目的總數量

nIDMS：歸屬資訊表中每個條目所平均包含的歸屬IDMS數量（考慮到支援基於網路的多家鄉技術，每個AID條目都可以有多個歸屬映射伺服器）

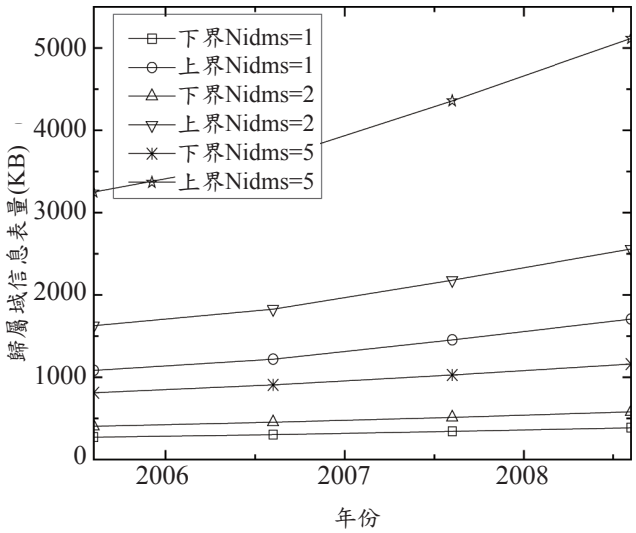
SAID：AID條目的位元元組數

SIDMS：IDMS條目的位元元組數

歸屬資訊表的存儲容量可以描述為：

Stotal = nAID×（SAID+nIDMS ×SIDMS）

假設每個用戶網路的接入標識都可以聚合為歸屬資訊表中的一個條目，那麼歸屬資訊表中條目的總數量nAID就是用戶網路的數量。每個用戶網路均聚合為一個AID條目時得到的nAID可以看做是nAID的下界。由於現有網路端自治系統的數量可以近似的看做用戶網路數量，可以採用端自治系統數量來估計nAID的下界；如果用戶網路直接按照現有互聯網的位元元址分配方式分配接入標識，那麼可以使用現有網路端自治系統中的位址首碼數量來估計nAID，這時的nAID可以看做nAID的上界，可以採用端自治系統中的首碼數量來估計nAID的上界。



圖四 歸屬域資訊表容量分析

圖四描述了根據2005-2008年端自治系統數量和位址首碼數量，取AID為32bit，SAID為20位元組，SIDMS為8位元組，並且在nIDMS分別取值1、2和5的情況下，對歸屬資訊表存儲容量的上界和下界進行了計算。可以看出，在2008年6月當AID為32比特時，nIDMS取值為5的情況下，歸屬資訊表的容量上界僅為5117.232KB，所以集中式存儲的歸屬資訊表不會在存儲容量上給映射伺服器帶來大的壓力。

基於以上的分析可知，利用BGP協議在映射伺服器之間傳送歸屬映射伺服器資訊並獲得一些預期的性能是合理可行的。

3 結論

本文在分析一體化網路身份與位置分離映射資訊管理系統的基礎上，設計了基於DNS遞迴分級查找、基於結構化DHT網路查找和基於BGP集中式查找的三種根據身份資訊解析歸屬映射伺服器的方案，並分析比較了各個方案的特點，如表1所示。每種方案有各自的優勢和缺陷，具體的選擇要視實際部署時的網路規模和性能要求等因素而定。下一步工作，將對各個方案做詳細的性能分析，並進一步考慮可行性和實施細節。

表 1　三種歸屬映射伺服器解析方案的特點比較

特點／方案	基於DNS遞迴分級查找	基於結構化DHT網路查找	基於BGP的查找
網路邏輯拓撲	樹形	環形	平面或分層
資料庫組織形式	分散式	分散式	集中式
查找方式	分級遞迴	Chord演算法	本地直接查找
查找效率	較高	較低	高
節點負載均衡	否	支持	否
更新速度	慢	較快	慢
可擴展性	差	好	好
可靠性／容錯性	低	較高	高
資料庫存儲開銷	較小	小	大

基金專案

國家863計畫（2007AA01Z202）。

參考文獻

[1] 張宏科、蘇偉，*新網路體系基礎研究——一體化網路與普適服務[J]*，電子學報，Vol. 35，No. 4，2007，pp.593-598。

[2] 董平、秦雅娟、張宏科，*支援普適服務的一體化網路研究[J]*，電子學報，Vol. 35，No. 4，2007，pp.599-606。

[3] 範建華，胥光輝，張濤等（譯），W. Stevens著，*TCP/IP詳解卷1：協議[M]*，北京：機械工業出版社，2005，pp.142-157。

[4] 李丹、吳建平、崔勇等，*互聯網名字空間結構及其解析服務研究[J]*，軟體學報，2005，pp.1445-1455。

[5] I. Stoica, R. Morris, and D. Liben-Nowell, et al., *"Chord: A Scalable Peer-to-peer Lookup Protocol for Internet Applications,"* IEEE/ACM Transactions on Networking, Vol. 11, No. 1, 2003.

[6] L. Mathy, L. Iannone, and O. Bonaventure, "*LISP-DHT: Towards a DHT to Map Identifiers onto Locators,"* Internet Draft, draft-mathy-lisp-dht-00, 2008.

[7] R. Cox, A. Muthitacharoen, and R. Morris. "*Serving DNS Using a Peer-to-peer Lookup Service [C],*" Proceedings of the Int'l Workshop on Peer-To-Peer Systems 2002 (IPTPS 2002), 2002, pp. 155-165.

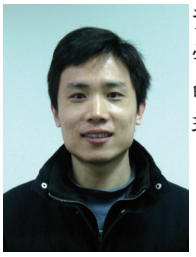
[8] D. Jen, M. Meisel, D. Massey, et al., *"LISP-APT: A PracTical Transit Mapping Service,"* Internet Draft, draft-jen-apt-01, 2007.

[9] 黃博、葛建立（譯），R. Zhang, M. Bartell著，*BGP設計與實現*，北京：人民郵電出版社，2005。

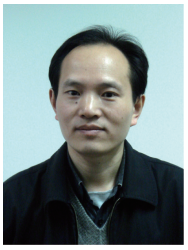
作者簡歷

劉曉波（Xiao-Bo Liu），現為北京交通大學碩士研究生。研究方向主要為IP網路路由理論、下一代網路位置管理、移動性管理。

姚楠（Nan Yao），現為北京交通大學在讀博士研究生，主要研究興趣：IP網路的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。



董平（Ping Dong），現為北京交通大學在讀博士生，主要研究方向：IP網路的路由理論與組播技術、下一代網路理論。



周華春（Hua-Chun Zhou），北京交通大學副教授，主要研究方向為IPv6路由器、移動互聯網路、網路與資訊安全、通信軟體等。

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。