



专题:未来网络

MCKite: 一种面向命名数据网络的多生产者多播机制

张 宇, 孙彦斌, 张宏莉

(哈尔滨工业大学计算机科学与技术学院 哈尔滨 150001)

摘 要:命名数据网络(named data networking, NDN)作为一种未来网体系结构,以命名数据代替主机作为网络层的核心对象,直接支持单生产者—多消费者多播应用模式。然而,NDN 对于多生产者应用支持仍然是一个开放问题,为此提出一种多生产者多播支撑机制——MCKite。MCKite 通过集会机制利用挂起兴趣表(pending interest table, PIT)中状态来建立通往生产者的路径,使得生产者拓扑位置对其他生产协作者及消费者透明,进而无需事先建立或更新路由表,降低了协议复杂性。将 MCKite 应用于多生产者数据同步协议并在 ndnSIM 上实现。实验表明,与以往方案相比,延展度更低,流量消耗更少。

关键词:命名数据网络;多播;多生产者;集会

doi: 10.11959/j.issn.1000-0801.2015321

MCKite: A Multi-Producer Multicast Scheme in Named Data Networking

Zhang Yu, Sun Yanbin, Zhang Hongli

(School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China)

Abstract: Named data networking (NDN)-a future internet architecture replacing hosts with named data as the core objects in the network layer and natively supports single producer- multi consumer multicast application pattern. However, the support of multi-producer application in NDN is still an open problem. MCKite-a multi producer multicast scheme in NDN was proposed. MCKite utilized rendezvous mechanism and stated in pending interest table (PIT) to construct paths to producers. MCKite made topological positions of producers transparent to other producers and consumers, so that routing/forwarding information bases didn't need to be created beforehand or be updated which reduced the complexity of protocol. MCKite was applied to an improved multi-producer data synchronization protocol and had been implemented in ndnSIM. Experiments show that MCKite has shorter stretch and less traffic cost compared to previous work.

Key words: named data networking, multicast, multi-producer, rendezvous

收稿日期:2015-11-10;修回日期:2015-12-07

基金项目:国家重点基础研究发展计划(“973”计划)基金资助项目(No.2011CB302605, No.2013CB329602), 国家自然科学基金资助项目(No.61202457, No.61402149)

Foundation Items: The National Key Basic Research Program of China (973 Program) (No.2011CB302605, No.2013CB329602), The National Natural Science Foundation of China (No. 61202457, No. 61402149)

论文引用格式:张宇, 孙彦斌, 张宏莉. MCKite: 一种面向命名数据网络的多生产者多播机制. 电信科学, 2015321

Zhang Y, Sun Y B, Zhang H L. MCKite: a multi-producer multicast scheme in named data networking. Telecommunications Science, 2015321

1 引言

多播对提高数据分发效率具有显著意义,但基于 IP 的多播协议难以在互联网中进行大规模部署和应用,原因之一在于 IP 体系结构的端到端思想并未以数据分发为中心。命名数据网络(named data networking, NDN)^[1,2]作为一种未来网体系结构,秉承以数据/内容为中心的思想,通过对数据的显示命名,以命名数据代替主机作为网络层的核心对象。NDN 直接支持单生产者—多消费者的多播应用模式,这得益于 NDN 的两个基本特性:以数据为中心,即通信直接以数据名为对象,而无需绑定到特定拓扑位置;有状态转发面,即基于挂起兴趣表(pending interest table, PIT)实现的反向路径转发。以数据为中心的特性使得数据生产者与消费者之间无需维护传输层链接,进而降低了多播管理协议的复杂性。有状态转发面的特性使得消费者标识和位置对路由面和生产者透明,直接支持了多消费者的多播应用。

然而,在 NDN 提出之时^[1],虽然指明了单生产者—多消费者的应用模式,但并未给出如何实现对多生产者多播应用的支撑。以当前 NDN 原生机制来实现多生产者多播应用存在以下两点挑战:首先,生产者的可达性依旧要直接依赖于路由面或间接依赖于应用层上映射,例如 CBIS^[3],多生产者的频繁加入和退出将给路由面可扩展性增加压力,或增加应用层映射操作的复杂性和代价;其次,分布式协作数据生产中的组成员发现机制(即多个生产者之间如何发现彼此标识或位置)依旧不明确。这些问题表明 NDN 的多生产者多播支撑机制亟待进一步研究。

本文提出一种面向 NDN 的多生产者多播支撑机制——MCKite。MCKite 基于之前提出的一种面向 NDN 的移动支撑机制——Kite^[4],利用 Kite 的 interest 追踪机制来实现生产者间成员的发现以及协作生产者之间、消费者与多生产者之间的 interest 多播。MCKite 的主要思想在于通过集会(rendezvous)机制利用 PIT 转发信息来追踪生产者,使得生产者的拓扑位置对其他生产协作者及消费者透明,进而无需事先建立或更新路由/转发表。对每个多播实例只需建立一个集会点或利用一组公共集会点基础设施,生产者通过发送注册 interest 消息即可加入多播组并建立从集会点到生产者的反向路径,生产者之间通过以集会点为根的双向转发树建立多播信道。作为 MCKite 的一个概念证明和应用示例,在 ndnSIM 上实现了一个改进的 Sync

协议——MCKiteSync。该协议实现了分布式多生产者间数据同步以及数据多播分发。

MSM-IP^[5]指出支持多播与支持移动是类似的,这也表明了 MCKite 与 Kite^[4]的相似性,但与面向移动环境的 Kite 相比,MCKite 的主要特点在于假定生产者拓扑位置相对稳定,进入了一种新的循迹前缀匹配方案,增加了多生产者间的双向转发树(bidirectional forwarding tree, BFT)建立和更新机制,通过硬状态转发来降低不必要的信令开销。参考文献[6,7]均基于位置建立路径,前者根据 interest 中位置信息重定向,后者通过向原位置发出请求建立轨迹,而 MCKite 则是一种与位置无关的方案。MCKite 的双向转发树机制与 BIDIR-PIM^[8]和 CBT^[9]类似,都独立于特定的底层单播路由,在多源情境下具有良好的可扩展性。但 MCKite 与它们分别面向不同的体系结构,得益于 NDN 自身具备的无环路转发特性,MCKite 不需要像 BIDIR-PIM 那样来实现指定转发器的选举或者反向路径转发检查,这一特性极大降低了 MCKite 的复杂性。DCR^[10]提出一种支持任播的可扩展路由方案,或许可以与 MCKite 结合以支持多个集会点。

2 MCKite 原理与方案

2.1 NDN 转发与多播原理

一次典型 NDN 通信过程如下:消费者发送一个包含数据名字的 interest 消息,路由节点以数据名字为目标,根据转发信息库(forwarding information base, FIB)向下一跳转发消息,同时在 PIT 中记录 interest 消息及其入口,直到 interest 消息到达生产者。生产者应答一个包含相同数据名字前缀的 data 消息,路由节点根据之前 PIT 中记录反向转发 data 消息直到消费者,同时路由节点可以将 data 内容缓存在内容库(content store, CS)中。这种“撒面包屑”式的反向路径转发机制使得 interest 和 data 消息中路由/转发信息只有数据名字本身,不包含消费者标识或位置,也不要求包含生产者拓扑位置,尽管通常数据名字前缀作为唯一寻址信息也意味着生产者位置。

当多个消费者向单个生产者请求数据时,形成单生产者—多消费者多播。单生产者的可达性通过单播路由实现,即由单播路由协议以数据名字为目的计算路由并填写 FIB。当消费者在较近时间间隔内发送多个相同 interest 消息,进入同一个路由节点时,只有第一个 interest 消息被转发,后续消息会被抑制,但仍会被记录在 PIT 中。PIT 记录



构成了以生产者根,连接到消费者的一棵多播树,此后 data 消息将沿此棵树上反向路径返回。若 data 消息被中间节点缓存,则此后相应的 interest 消息在到达该中间节点时就将获得 data 消息,而无需从生产者处重新获取数据。可见,NDN 转发机制直接支持单生产者—多消费者多播,无需像 IP 那样需要额外的多播协议来实现。

2.2 多生产者多播方案 MCKite 原理

当多个生产者分布式协作生产数据时,消费者需要获得来自全部生产者的数据,形成多生产者多播。一个典型多生产者应用情景是 Sync 协议^[1],即多生产者间数据同步。由于当前 NDN 体系结构中,interest 转发依赖于路由层构造的 FIB,不同于单生产者多播中利用单播路由协议构造 FIB,多生产者多播路由面临两个问题:可扩展性,即随着生产者规模扩大,由生产者加入和退出所带来的组结构扰动加剧,基于单播路由的机制将增加路由层负担;组成员发现,即生产者之间如何彼此发现,或者说如何避免泛洪实现彼此间消息定向发送。

针对多生产者多播应用中存在的问题,MCKite 提出基于集会与 interest 循迹转发的支撑方案。该方案可以看作之前提出的 NDN 移动支撑方案——Kite^[4]的多播版本。与 Kite 相似,MCKite 中每个多播组需要一个集会点,集会点发出相关数据名字前缀路由声明,通过单播路由实现全网到集会点自身或数据名字前缀的可达性。MCKite 通过集会点使来自不同生产者和消费者的 interest 可到达同一网络位置,这些 interest 消息在 PIT 中留下的踪迹构成了从集会点到多播成员的单向路径树。

interest 循迹转发是 Kite 引入 NDN 转发面的一种转发机制,该机制受 data 消息转发机制启发,使得一个 interest 消息可沿 PIT 中另一个 interest 消息反向路径转发,其中前一个 interest 消息称作“循迹 interest”,后一个 interest 消息称作“被循迹 interest”。一个生产者发送的被循迹 interest 消息经过中间路由节点转发在 PIT 中形成反向路径,其他生产者或消费者发送的循迹 interest 消息利用 interest 循迹转发机制沿该路径到达该生产者。一个 interest 消息可以既是一个循迹 interest,也是一个被循迹 interest。两个相互循迹的 interest 消息在相遇后将沿着彼此的反向路径转发。通过与集会模式的结合,从集会点到多播成员的单向路径树就转变为组成员间的双向路径树,组成员间的 interest 消息会相互循迹,从而实现生产者间多播。

为实现 interest 循迹转发,MCKite 继承了 Kite 在 interest 消息中引入的两个新字段:TraceName 域,即循迹 interest 消息中所包含的被循迹 interest 消息的数据名字;Traceable 标记,即该 interest 消息是否可以被其他 interest 追踪,若置 1,则可被追踪,否则禁止追踪。MCKite 舍弃了 Kite 引入的 TraceOnly 标记。MCKite 也继承了 Kite 在 PIT 中加入的两个新表:循迹转发表(trace forwarding table, TFT),包含所有 Traceable 标记置 1 的被循迹 interest 消息;循迹名字表(trace name table, TNT),包含所有带有 TraceName 域的循迹 interest 消息。

MCKite 引入一种新的 interest 循迹前缀匹配方案,在可区分来自不同生产者的 interest 消息的同时,令循迹 interest 消息可以被多播到所有生产者。该方案把循迹 interest 消息的 TraceName 域 T 作为被匹配前缀,以被循迹 interest 消息的名字 N 来前缀匹配 TraceName 域,即若 T 为 N 的前缀,则匹配。这样,一个循迹 interest 可以匹配多个名字不同的被循迹 interest。

2.3 MCKite 转发方案

以上述新添加的 interest 消息字段、改造的 PIT 数据结构和循迹前缀匹配方案为基础,MCKite 转发算法如下所示:输入一个来自接口 I[IF]的 interest I;输出一个 interest/data 转发操作列表 Oplist, push A、B 操作表示将消息 A 通过接口 B 转发。第(1)~(5)步为标准的 interest 转发流程,即从 CS 中取得 data,否则将 I 加入 PIT;第(6)~(10)步,若 Traceable 置 1,则 I 被循迹,将 TFT 中记录的每个追踪 I 的 interest J 通过 F 转发;第(11)~(16)步,若 Tracename 不为空,则 I 循迹,根据 TNT 将 I 转发向所有被循迹 interest 的入口;(17)~(20)步,根据 FIB 转发 I,算法结束。

算法 1: MCKite interest 转发算法

输入: I 通过接口 I[IF]进入路由器

输出: 一个转发操作列表 oplist

- (1) if (I match data D in CS)
- (2) push D, I[IF] into oplist
- (3) return oplist
- (4) if (I not in PIT)
- (5) insert (I, I[IF]) into PIT
- (6) if (I[Traceable]==1)
- (7) insert (I, I[IF]) into TFT
- (8) for_each interest J (J[Tracename] match I in TNT)

```

(9)      if (has_not_forward J via I[IF])
(10)     push J, I[IF] into oplist
(11)    if (I[Tracename] != NULL)
(12)     insert (I, I[IF]) into TNT
(13)    for_each interest K (I[Tracename] match K in
TFT)
(14)     for_each interface K[IF] (K[IF] != I[IF])
(15)      if (has_not_forward I via L)
(16)       push I, K[IF] into oplist
(17)    if (I match in FIB && next_hop interface U != I
[IF])
(18)     if (has_not_forward I via U)
(19)      push I, U into oplist
(20)    return oplist

```

2.4 MCKite 双向转发树方案

下面描述基于算法 1 的 MCKite BFT 建立、维护和更新机制。设一次多播会话存在一个单播路由可达的集会点 R, R 声明一个可路由名字前缀, 例如/SyncGroup1。为易于算法描述, 下面给出任意两个组成员 N 和 M 间以 R 为根节点的双向路径, 超过两个成员的双向转发树的建立原理相同。其中, 假设 N 先发送一个用于加入组的 interest 消息, 例如 interest 名字为/SyncGroup1/Join/Member1, TraceName 为/SyncGroup1/Join, Traceable 标记置 1。算法如下所示: 输入集会点 R, 组成员 N 和 M; 输出 N 和 M 间在 PIT 中的双向转发路径, 该路径建立通过 insert INTO PIT 操作实现。第(1)~(6)步, 生产者 N 发送 interest I, 根据单播路由从 N 到达 R, 并在中间路由器上的 PIT 表中留下踪迹, 建立了从 N 到 R 的路径; 第(7)~(15)步, 生产者 M 发送 interest J, 根据单播路由从 M 到达 R, 若在 R 或中间路由器上 I 和 J 相遇, 则根据 interest 循迹转发机制, I 和 J 互为被循迹 interest, 分别通过对方的入口转发, 并最终到达生产者, 形成了从 R 到 M 的路径以及从 M 到 N 的路径。此时, 加上之前从 N 到 R 的路径, 形成了从 M 到 N 的双向路径。类似的, 当生产者规模超过 2 个时, 新生产者发送的注册 interest 将多播转发到集会点和已有成员, 由此形成了以集会点为根的 BFT。不同于 Kite 中 interest 在 PIT 中以软状态保存, MCKite 中注册的 interest 在 PIT 中为硬状态, 无需周期发送新注册消息。

算法 2: MCKite 双向转发树建立算法

输入: 集会点 R, 组成员 N 和 M

输出: 隐含在 PIT 中形成的 M 和 N 间双向转发路径

```

(1)  N send interest I=/SyncGroup1/Join to R
(2)  until I reach R
(3)  for_each intermediate router receives I via
interface I[IF]
(4)      forward I towards R according to FIB
(5)      insert (I, I[IF]) into PIT
(6)  registry I at R and insert (I, I[IF]) into PIT
(7)  M send interest J to R
(8)  until J R reach R
(9)  for_each intermediate router receives J via
interface J[IF]
(10)     if J meet I
(11)      forward J via I[IF] until J reach N
(12)      forward I via J[IF] until I reach M
(13)      forward J towards R according to FIB
(14)      insert (J, J[IF]) into PIT
(15)  registry J at R and insert (J, J[IF]) into PIT
(16)  return

```

当一名生产者成员欲退出时, 该成员先发送退出 interest 消息, 例如/SyncGroup1/Leave/Member1。集合点在收到退出消息后, 通过发送一个退出 data 消息/SyncGroup1/Join/Memeber1/Leave, 将之前的注册 interest 消息/SyncGroup1/Join/Memeber1 沿着从集合点到该生产者路径上的 PIT 中清除, 同时发送一个退出应答 data 消息/SyncGroup1/Leave/Member1 通知该生产者已收到退出请求。此后, 由于从集会点到该生产者的转发路径已经被清除, 即 BFT 上通往该生产者的分支被剪掉, 其他循迹 interest 消息将不再被转发到该生产者, 实现了成员退出和 BFT 更新机制。

3 实验结果

从两方面对 MCKite 性能进行评价: 一方面, 从路径延展度进行度量, 与经典的 PIM 对比; 另一方面, 从流量消耗进行度量, 实现 MCKite。本实验采用 Rocketfuel^[11]测量的不同 ISP 的路由器级网络拓扑。

3.1 延展度

延展度指真实路由路径长度和最短路径长度的比值, 其反映了路径被拉伸程度, 延展度越低说明路径长度越接近最短路径。下面分别讨论 3 种不同情况下多播成员间通



信路路延展度。

(1) 不同网络拓扑

采用 9 个 ISP 网络拓扑进行实验。每次实验随机选择一个节点作为多播集会点,并选择 30% 节点作为多播成员,成员之间相互通信并统计实际路径,同时计算成员间的最短路径。图 1 为不同拓扑下平均路径延展度。虽然各网络拓扑路径延展度略有差异,但整体上 MCKite 路径延展度比 PIM 低近 20.6%,其中 MCKite 的平均路径延展度为 1.46~1.82,而 PIM 平均延展度为 1.76~2.37。其主要原因在于 MCKite 采用循迹转发机制,多播成员间可能无需到达集会点就可通过循迹找到更短的捷径。

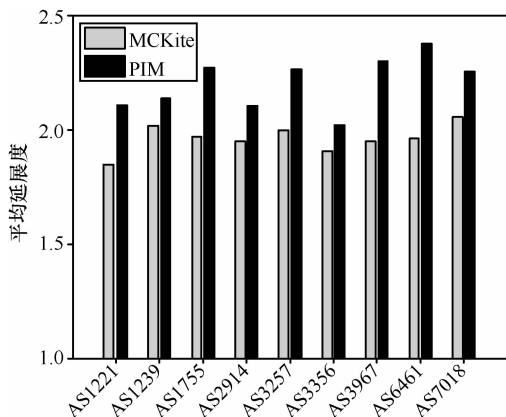


图 1 不同拓扑下平均路径延展度

此外,MCKite 循迹转发机制还可尽可能降低由于距离很近的两个多播成员通过集会点通信产生的高时延。在 AS1221 中,PIM 通过集会点通信造成约 7.99% 的路径延展度大于 3,而通过 MCKite 循迹转发捷径,延展度大于 3 的比例下降到 0.85%。同时 MCKite 最短路径比例比 PIM 高 10% 左右。

(2) 不同多播规模

选定 AS1221 作为网络拓扑,随机选择一个节点作为固定集会点,依次随机选择 10%~50% 的节点作为多播成员。图 2 显示,MCKite 平均路径延展度依然优于 PIM。同时,对于两种多播方案,随着多播成员增加,平均路径延展度略有增长。其主要原因可能在于集会点未处于网络中心,当多播成员增加时,离集会点较远,节点增加较多。其中,PIM 增长幅度为 7%,大于 MCKite 的 5%,可见多播规模对 MCKite 的影响小于对 PIM 的影响。

(3) 集会点选择

以 AS1221 作为网络拓扑,随机选择 30% 节点作为多播成员,根据节点度大小将节点划分为多个区间,每次从不同区间中随机选择一个节点作为集会点。图 3 为不同集

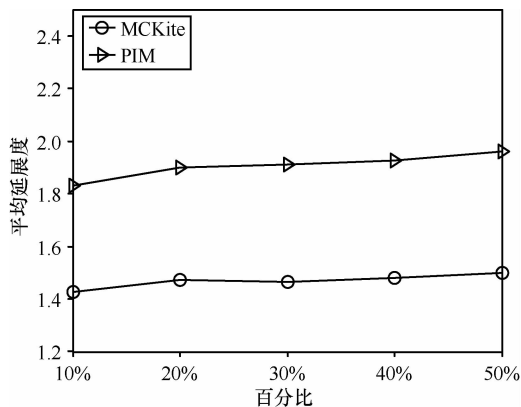


图 2 不同多播规模下路径延展度

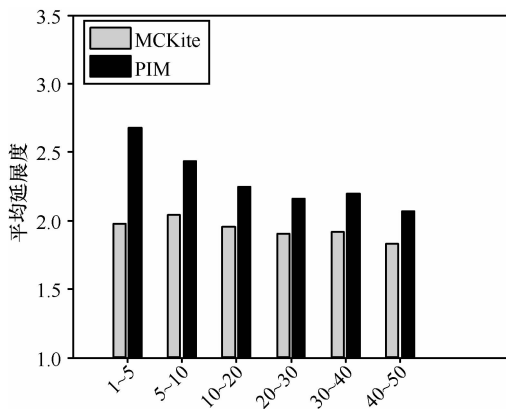


图 3 不同集会点下路径延展度

会点平均路径延展度,MCKite 和 PIM 平均路径延展度整体上都随节点度增大而下降。主要原因在于节点度越小,节点可能越接近网络边缘,使得多播成员到集会点距离大大增加,而节点度越大则越靠近网络拓扑中心,缩短了多播成员到集会点的距离。其中,PIM 平均路径延展度下降幅度为 36.4%,而 MCKite 则能维持一个较低延展度,下降幅度仅为 20%,可见集会点的选择对 MCKite 延展度的影响较小。

3.2 流量消耗

在 ndnSIM^[12]上实现了 MCKite 并将其应用于一个多生产者间数据同步的应用情景,在 AS1221 拓扑下考察不同多播规模下流量消耗,平均每 4 s 产生一个新数据,相同规模下模拟 50 次,每次模拟运行时间 20 s,实验结果如图 4 所示。随着多播规模增长,流量消耗呈线性增长,远远低于泛洪的代价。

4 结束语

针对未来网络体系结构 NDN 中多生产者多播问题,本文提出了一种基于集会模式和寻踪转发的多播方案

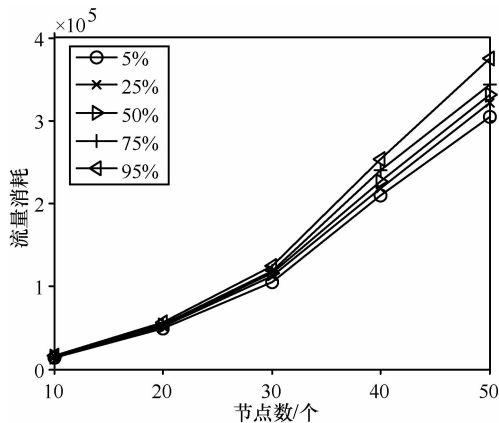


图4 不同多播规模下流量消耗

MCKite。该方案利用集会点作为双向转发树根,利用 PIT 中 interest 踪迹实现通往生产者树枝,并提出一种新的 interest 循迹前缀匹配方案实现树的剪枝。MCKite 方案不仅复杂性低,而且实验表明,MCKite 在路径延展度上优于以往方案,且随多播规模增长,路径延展度和流量消耗的可扩展性都较好。

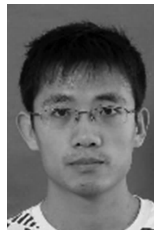
参考文献

- Jacobson V, Smetters D K, Thornton J D, *et al.* Networking named content. Proceedings of International Conference on Emerging Networking Experiment&Technologies, Rome, Italy, 2009: 1~12
- Zhang L, Afanasyev A, Burke J, *et al.* Named data networking. ACM SIGCOMM Computer Communication Review, 2014, 44(3): 66~73
- Jacobson V, Braynard R L, Diebert T, *et al.* Custodian-based information sharing. IEEE Communications Magazine, 2012, 50 (7):38~43
- Zhang Y, Zhang H L, Zhang L X. Kite: a mobility support scheme for NDN. Proceedings of 1st Conference on Information-Centric Networking, Paris, France, 2014: 179~180
- Mysore J, Bharghavan V. A new multicasting-based architecture for internet host mobility. Proceedings of the Annual International Conference on Mobile Computing and Networking, Budapest, Hungary, 1997:161~172
- Hermans F, Ngai E, Gunningberg P. Global source mobility in the content-centric networking architecture. Proceedings of ACM Workshop on Emerging Name-Oriented Mobile Networking Design Architecture, New York, USA, 2012: 13~18
- Kim D, Kim J, Kim Y, *et al.* Mobility support in content centric networks. Proceedings of ACM Conference on Information-Centric Networking, Helsinki, Finland, 2012: 13~18
- Handley M, Kouvelas I, Speakman T, *et al.* Bidirectional protocol independent multicast (BIDIR-PIM). Work in Progress, 2005
- Ballardie A J, Francis P F, Crowcroft J. Core Based Trees. ACM SIGCOMM Computer Communication Review, 1993, 23(4): 85~95
- Garcia-Luna-Aceves J J. Name-based content routing in information centric networks using distance information. Proceedings of 1st Conference on Information-Centric Networking, Paris, France, 2014: 7~16
- Spring N, Mahajan R, Wetherall D. Measuring ISP topologies with rocketfuel. Proceedings of ACM SIGCOMM Conference, Pittsburgh, USA, 2002: 133~145
- Afanasyev A, Moiseenko I, Zhang L X. ndnSIM: NDN Simulator for NS-3, 2012

[作者简介]



张宇,男,博士,哈尔滨工业大学计算机科学与技术学院副教授,主要研究方向为网络安全、网络测量和未来网络。



孙彦斌,男,哈尔滨工业大学计算机科学与技术学院博士生,主要研究方向为可扩展路由和未来网络。



张宏莉,女,博士,哈尔滨工业大学计算机科学与技术学院教授、博士生导师,主要研究方向为网络安全、网络测量和网络计算。