

基于无连接多路径路由的负载均衡

徐武平^{1,2}, 晏蒲柳², 夏德麟²

(1. 武汉大学计算机学院, 武汉 430072; 2. 武汉大学电子信息学院, 武汉 430072)

摘要:介绍了一种可以应用于 Internet 网络的无连接多路径路由计算方法, 称为概率无连接多路径路由(probability-Disjoint Multi-paths Routing, p-DMR)。该方法使用概率构造无连接多路径, 降低了在复杂网络环境中计算无连接多路径的复杂度, 并将多路径路由与自适应按比例动态流量分割算法相结合, 使网络性能得到优化, 拥塞得到避免。

关键词:无连接多路径; 路由算法; 流量分割; 动态流量工程

Method of Load Balancing Based on Disjoint Multi-paths Routing

XU Wuping^{1,2}, YAN Puliu², XIA Delin²

(1. School of Computer, Wuhan University, Wuhan 430072; 2. School of Electronic Information, Wuhan University, Wuhan 430072)

【Abstract】A new distributed algorithm for the dynamic computation of multiple disjoint paths is presented, which is called p-DMR (probability-disjoint multi-paths routing), probability is adopted to format disjoint paths. Combined to a traffic balancing algorithm, its average performance is analyzed by simulation and compared against equal cost multi-path(ECMP).

【Key words】Disjoint multiple paths; Routing algorithm; Traffic split; Dynamic traffic engineering

目前Internet中的路由选择由于大多数链路费用值保持静止, 造成即使有另一条更好的路径存在, 同一源点对的网络流量也总是采用相同的路径, 使得网络效率低下。为了解决这些问题, 许多新的方法应运而生。其中一种可能就是给定的流量矩阵进行链路费用全局优化^[1]。这种方法显然可以改善网络的性能, 但不幸的是全局优化是NP难问题。

许多研究表明采用多路由或备份路由可以明显平衡网络负荷, 改善网络性能^[3,4]。其中比较著名的有等费用多路径(ECMP)^[5]和优化多路径(OMP)^[6,7]。ECMP在多条等费用路径之间平均分配流量。然而, 一方面由于链路费用的静止特性, 使得ECMP中的多路径通常也是静止的; 另一方面具有相等费用的路径即使在链路费用粒度很好的网络中出现的机率也非常小; 再者, ECMP没有根据路径性能分配流量, 而是平均分配。OMP的思想和我们的方法类似, 就是寻找近似最优的路径作为辅助路由。但是OMP却需要更复杂的数据结构和算法。

以上提及的多路径路由方案几乎都采用开环多路径(Loop free Multi-paths), 由于开环多路径可能使拥塞向下游节点转移, 导致整个网络的负载均衡难以实现。本文描述了一种实现“近似优化”的负载均衡方案。该方案的核心部分——概率无连接多路径路由(probability-Disjoint Multi-paths Routing, p-DMR)与ECMP不同, 它采用非等费用策略提供更多的有效路径; 同时保持各条路径之间是无连接(或概率无连接)性, 防止或减小流量重新汇聚的影响。最后 p-DMR 和基于路径状态的流量分割算法相结合, 实现有效的负荷均衡。

1 基于 p-DMR 的流量均衡方案

1.1 问题描述

将网络模型看作是一个图 $G = (N, L)$, N 是路由节点, L 是链路集合。让 N^i 表示节点 i 的相邻节点集合。问题归结

为在每个节点 i 为到达节点 j 寻找后继节点。后继节点集合包含两个子集: (1)主后继节点集合: $S_j^i \subseteq N^i$, 提供到达 j 的最优路径; (2)辅助集合: $S_j^{ii} \subseteq N^i$, 提供到达 j 的近似最优路径。当 i 收到发往 j 的负载 v_j^i 时, 将 v_j^i 按比例在相邻后继节点集合 $S_j^i \cup S_j^{ii}$ 之间进行流量分割, 实现负载均衡。通过在每个路由节点重复这一过程, 流量负载可以到达各自的目的节点, 并使网络性能得到优化。

1.2 p-DMR 算法

(1)无连接多路径(Disjoint Multi-Paths)

无连接路径强调路径之间没有链路相连, 如图 1 中实线箭头所示。开环主要是防止形成路由回路^[8], 如图 1 中虚线箭头所示。

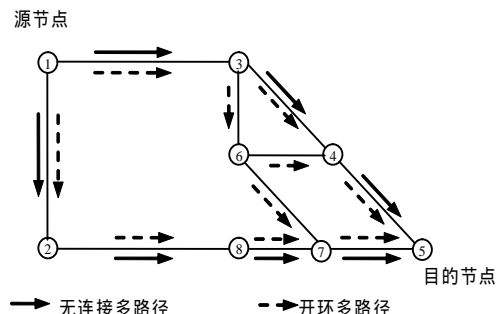


图1 无连接多路径和开环多路径

基金项目:国家自然科学基金资助项目(90204008)

作者简介:徐武平(1970-), 男, 讲师、博士生, 主研方向: 网络通信, 性能管理, 信息系统和数据库; 晏蒲柳, 教授、博导; 夏德麟, 教授

收稿日期:2006-01-22 **E-mail:** whwp@tom.com

在开环多路径中，路由图是一个有向非循环连通图(Directed Acyclic Graph, DAG)，可能会出现拥塞点的转移。以图 2 为例，采用图 2(a)中虚线所示路径分流虽然可以消除链路(1,2)的拥塞，但由于虚线所示路径又汇聚于节点 2，使得节点 1 的负载更快地到达节点 2。如果链路(2,3)没有足够的容量，就会变成新的拥塞点。而在图 2(b)中，分流路径(虚线所示)并没有汇聚于节点 2，所有的路径都是无连接的，拥塞不会向链路(2,3)转移。在无连接多路径中，路由图是一个无连接有向非循环连通图(称为 disjoint-DAG)。

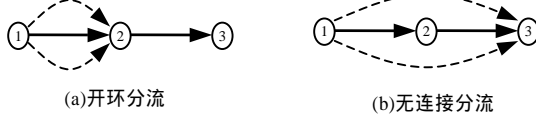


图 2 开环分流和无连接分流

(2) 概率无连接多路径

采用分布式计算模型，网络中的各个节点根据其相邻节点发送的路由信息独立地计算路由表。路由信息包括两个主要部分：1)到达目的节点的多路径长度；2)多路径的形式化表示：路由图 SG_j^i 。多路径长度用来衡量一个 disjoint-DAG 的费用，路由图则用来发现不同邻居节点提供的路由是否相互连接。当收到所有邻居发来的路由信息后，节点首先选择具有最短长度的路径作为主路由，然后选择长度近似最短路径并且和现有路由不相连接的路由作为辅助路由。

如图 1 所示，节点 6 发现到达节点 5 的两条无连接路径 $\{(6,4),(4,5)\}$ 和 $\{(6,7),(7,5)\}$ ，将这个路由表示为路由图 $\{(6,4),(6,7),(4,5),(7,5)\}$ 通知给节点 3。而节点 3 已经发现了到达节点 5 的最短路由 $\{(3,4),(4,5)\}$ ，由于 $\{(3,4),(4,5)\}$ $\{(6,4),(6,7),(4,5),(7,5)\} = \{(4,5)\}$ ，节点 3 没有采用 6 作为后继节点。但当链路(3,4)出现拥塞时，节点 3 非常需要另外的路径来分流负载。假设节点 6 按照比例 $p_{5,4}^6$ 、 $p_{5,7}^6$ 在路径 $\{(6,4),(4,5)\}$ 和 $\{(6,7),(7,5)\}$ 之间分配流量，并且 $p_{5,4}^6 \ll p_{5,7}^6$ ，那么就可以忽略路径 $\{(6,4),(4,5)\}$ 对 $\{(3,4),(4,5)\}$ 的影响。在路由图的表示中增加流量分配比例，即 $SG_j^i = \{(m,n, p_{j,k}^i) | m_{init} = i, n \in S_j^i, k \in S_j^i\}$ 。节点 6 的路由图就变为 $\{(6,4, p_{5,4}^6), (4,5, p_{5,4}^6), (6,7, p_{5,7}^6), (7,5, p_{5,7}^6)\}$ 。流量分配比例看作负载经过某路径的可能性，因此称流量分配比例为路由概率。接下来，在节点 3 上设置一个概率参数 p_5^3 ，并为路由图设计一个过滤操作 FILTER，定义如下：

$$FILTER(SG_j^i, p_j^i) = \{(m,n) | (m,n, p_{j,k}^i) \in SG_j^i \wedge p_{j,k}^i > p_j^i\}$$

如果 $p_{5,4}^6 < p_5^3 \wedge p_{5,7}^6 > p_5^3$ ，则 $FILTER(SG_5^6, p_5^3) = \{(6,7),(7,5)\}$ ，得到 $\{(3,4),(4,5)\} \cap \{(6,7),(7,5)\} = \emptyset$ ，节点 3 采用路径 $\{(3,6),(6,7),(7,5)\}$ 来分流负载，并且不会将拥塞点转移到链路(4,5)。由于这种无连接多路径依赖于概率参数 p_5^3 ，因此称为概率无连接多路径。这里概率参数 p_j^i 相当重要，它是负载的函数，其值随负载的上升而增加，以便节点能够获得更多的路径来平衡负载；负载下降时，其值也相应下降，防止节点占用过多网络资源。为此，引入负载系数 ρ_j^i 来描述负载的轻重程度：

$$\rho_j^i = \frac{V_j^i}{\sum BW_{j,k}^i}, k \in S_j^i \cup S_j^{u_i} \quad (1)$$

V_j^i 表示固定时间间隔内的平均输入负载， $BW_{j,k}^i$ 表示为输出带宽。这样， p_j^i 可以通过下面的函数计算：

$$p_j^i = \begin{cases} 0, \rho_j^i \leq 1 \\ 1 - \frac{1}{\rho_j^i}, \rho_j^i > 1 \end{cases} \quad (2)$$

(3) p-DMR 算法描述

假设在图 G 中，到达节点 j 有许多 disjoint-DAG 存在，选择哪些 disjoint-DAG 作为路由图 SG_j 更合适呢？一个很自然的选择是长度最短的路由图。在 p -DMR 中，将具有最短长度路由的后继节点定义为 $S_j^i = \{k | D_j^k < D_j^i, k \in N^i\}$ ， D_j^i 是从 i 到 j 的最短路径长度，它是该路径上所有链路费用的总和。从集合 S_j^i 推导出的路由图 SG_j 是唯一的，称为主多路径(Primary Multi-path)。为了平衡流量负载，需要更多的路由，因此将 $S_j^{u_i} = \{k | D_j^k - dth < D_j^i, k \in N^i\}$ 定义为辅助后继节点集合， dth 是一个近似门限用来衡量辅助路由与最短路由之间的近似程度。由 $S_j^{u_i}$ 推导出的路由图 $S'G_j$ 称为辅助多路径路由(Secondary Multi-paths)，并且必须保证 $SG_j \cup S'G_j$ 也是一个 disjoint-DAG。计算 D_j^i 时， p -DMR 采用 Dijkstra's 的扩散算法(Diffusing computations)^[9]：在一个给定的 disjoint-DAG 中，每个节点使用下游节点报告的长度计算自己的长度，并将计算结果报告给上游节点。节点按固定时间间隔交换路由信息报文，每个报文包含一个或多个路由信息。一个路由信息表示为 $\{j, d, SG\}$ 。 j 是目的节点， d 是发送信息节点到达 j 的长度， SG 是其到达 j 的路由图。在 p -DMR 中，路由节点通常只使用主路由，只有当流量负载系数上升很快，超过某个门限值 Sth (通常介于 0.9~1.1 之间)时，才探测和使用辅助路由。节点收到路由信息报文后调用如图 3 所示的过程 *ProcessEntry* 处理每一条路由信息。

```

00 procedure ProcessEntry( i, m, j, d, SG)
01 {i: thisnode, m: neighbor who send the message, j: destination, d: distance,
   SG: routing graph}
02 begin
03    $D_{j,m}^i = d, SG_{j,m}^i = FILTER(SG, p_j^i)$ ;
04   if  $D_{j,k}^i + q_{j,k}^i < D_j^i$  then  $D_j^i = D_{j,k}^i + q_{j,k}^i, SG_j^i = \{(i,m)\} \cup SG_{j,m}^i, S_j^i = \{m\}$ ;
   endif
05   if (last message is received for j) then
06      $SG_j^u = \emptyset, S_j^u = \emptyset, \rho_j^i = \frac{V_j^i}{\sum BW_{j,k}^i}, k \in S_j^i$ ;
07     do while  $\rho_j^i > Sth$ 
08       foreach  $k \in N^i \wedge k \notin S_j^i \cup S_j^{u_i}$  do
09         if  $D_j^i / D_{j,k}^i \geq dth \wedge SG_j^i \cap SG_{j,k}^i = \emptyset \wedge SG_j^u \cap SG_{j,k}^i = \emptyset$  then
10            $SG_j^u = \{(i,k)\} \cup SG_{j,k}^i \cup SG_j^u, S_j^u = S_j^u \cup \{k\}$ ;
11           break;
12         endif
13       done
14       if  $S_j^u = \emptyset$  then break;
15        $\rho_j^i = \frac{V_j^i}{\sum BW_{j,k}^i}, k \in S_j^i \cup S_j^{u_i}$ ;
16     loop
17   endif
18 end

```

$D_{j,k}^i$: Distance of node k to j as reported by k to i .

p_j^i : Probability parameter at node i for j .

$SG_{j,k}^i$: Routing graph of node k to j as reported by k to i .

图 3 p-DMR 算法描述

1.3 动态负载均衡算法

以文献[10]中的按比例流量分配模型为基础， p -DMR 的流量分割算法如下：

(1) 链路费用(Link Cost)

D_j^i 是从 i 到 j 的最短路径的长度，它是某个路由图中最短路径上所有链路费用的和。链路费用通过一个性能参数——归一化队列长度(normalized queue length)^[8]表示。如果

BW_k^i 代表 i 和邻居节点 k 之间的链路带宽, 归一化队列长度 q_k^i 定义为

$$q_k^i = \frac{Q_k^i}{BW_k^i} \quad (3)$$

其中, Q_k^i 是 i 和 k 之间在固定统计间隔内的平均队列长度。

(2) 路径长度(Past Length)

从 i 经 k 到达 j 的路径长度 D_{ji}^k 等于 k 报告给 i 的 k 到 j 的长度加上 q_k^i , 即

$$D_{ji}^k = D_{jk}^i + q_k^i \quad (4)$$

(3) 负载均衡的流量分割

在多个后继节点之间按比例分配流量, 目标是使各路径的性能尽可能的提高。这个问题可以归结为: 寻找一组优化比例 α_j^k 使得所有多路径的长度之和最小, 即

$$\begin{aligned} \min \quad & \sum D_{ji}^k, k \in S_j^i \cup S_j^i \\ S.T: \quad & \sum \alpha_j^k = 1, k \in S_j^i \cup S_j^i \end{aligned} \quad (5)$$

这是一个非线性规划问题, 优化比例的计算非常困难, 采用等路径长度策略避免复杂计算。等路径长度的目标是寻找一组优化比例, 使得所有多路径长度都相等。首先计算所有多路径到达目的节点 j 的平均长度 $\overline{D_j^i}$ 。

$$\overline{D_j^i} = \frac{\sum Q_k^i + v_j^i}{\sum BW_k^i}, k \in S_j^i \cup S_j^i \quad (6)$$

v_j^i 是统计间隔内, 经节点 i 到 j 的平均流量负载。然后, 可以计算出在统计间隔内的流量分割比例 α_j^k :

$$\alpha_j^k = \frac{BW_k^i \cdot \overline{D_j^i} - Q_k^i}{v_j^i}, k \in S_j^i \cup S_j^i \quad (7)$$

2 模拟结果与分析

2.1 模拟环境

模拟实验包含两个环境: 一个如图 1 所示, 包含一个源目对; 另一个环境如图 4 所示, 拓扑结构接近实际网络环境。简单起见, 所有链路都是双向连接, 并且具有相同的容量。模拟主要比较 ECMP 和 p -DMR 算法特点, 丢弃了协议实现细节。

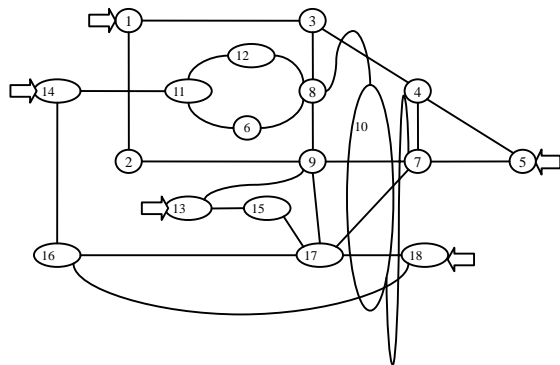


图 4 模拟网络拓扑图

2.2 模拟结果

(1) 多路径选择方式的比较

模拟结果如图 5 所示, 图 1 中的链路(1, 3)是焦点所在。ECMP 模拟中其排队延时持续上升, 而在 p -DMR 模拟中则变得相对平缓。这主要归功于 p -DMR 能够及时采用非等费用辅助多路径 $\{(1,2),(2,8),(8,7),(7,5)\}$ 分担负载。

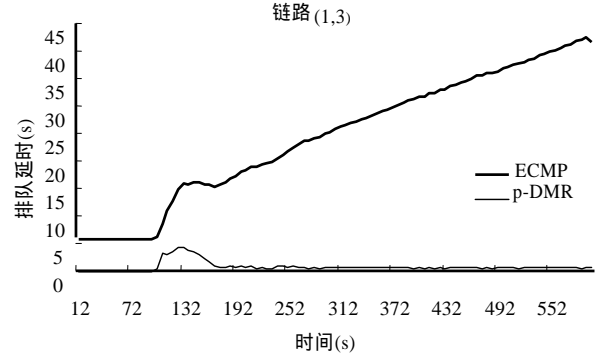


图 5 ECMP 和 p -DMR 中链路(1,3)的排队延时

(2) p -DMR 消除流量重新汇集的影响

使用图 1 的模拟环境来比较 LFMR 和 p -DMR。LFMR 采用的是开环(Loop Free)多路径。模拟结果如图 6 所示, LFMR 中链路(7,5)的排队延时要比在 p -DMR 中略高。这是因为在 p -DMR 中, 节点 3 根据 p (Probability) 没有采用节点 6 作为后继节点, 使得链路(7,5)的性能更好一些。

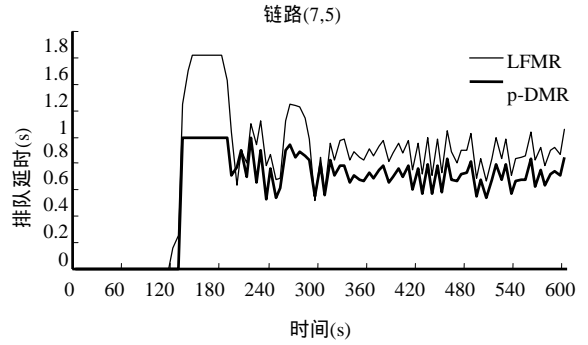


图 6 LFMR 和 p -DMR 中链路(7,5)的排队延时

(3) 负载均衡能力比较

使用图 4 所示的模拟环境。先为所有路由器配置 ECMP, 如图 7 所示, 链路(9,13)的排队延时持续上升, 出现拥塞。原因是从节点 1、节点 5 到达节点 13 的最短路径都经过链路(9,13)。而从节点 14、节点 18 到节点 13 的路径都经过节点 7, 节点 7 有两条费用相等的路径 $\{(17,15),(15,13)\}$ 和 $\{(17,9),(9,13)\}$ 。ECMP 在这两条路径间平均分配流量, 使通过链路(9,13)的流量过大, 如图 8(a)中所示。

但是在 p -DMR 中, 链路费用是动态变化的归一化队列长度, 路径长度则是所有链路费用之和, 也是动态变化的。节点 17 根据路径的性能合理地按比例分割流量, 如图 8(b)所示, 链路(17,15)的流量比链路(17,9)的更大, 使得链路(9,13)的拥塞得以减轻, 如图 7 所示。

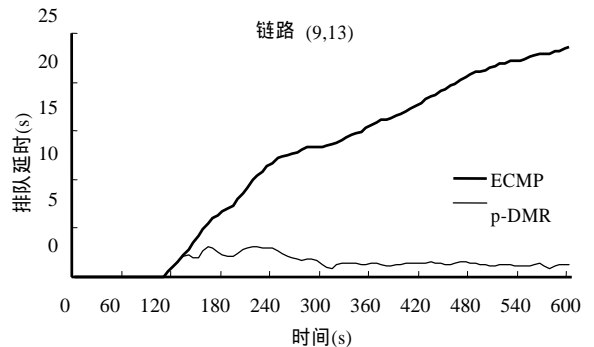
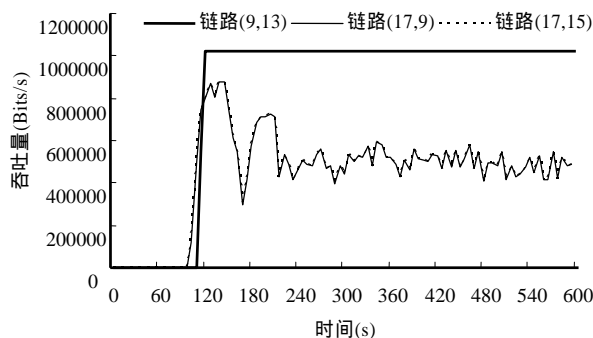
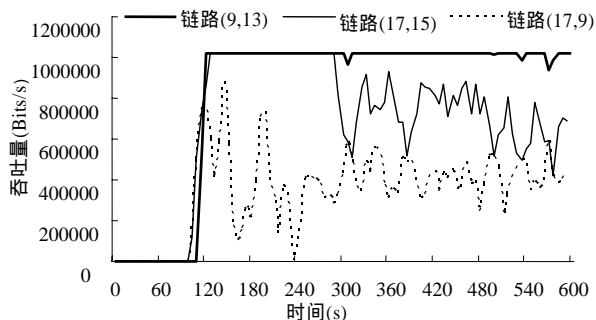


图 7 链路(9,13)的排队延时比较



(a) ECMP 的链路流量分配



(b) p-DMR 的链路流量分配

图 8 ECMP 和 p-DMR 的链路流量分配

3 结束语

本文采用概率无连接多路径路由(p-DMR)实现负载均衡。在无连接路径的计算中引入概率参数,一方面可以忽略分流流量很小的链路,简化高连通度网络的拓扑结构,发现更多有效的无连接多路径;另一方面,具有较大流量的链路被保留,通过连接性的判断可以避免各路由之间流量再次汇聚,防止了拥塞点地转移。另外,概率门限值由负载决定,

可以防止各源一目对竞争网络资源,做到按需分配。该算法采用分布式计算模型,具有诸多优点,如较低的通信开销、简单易于实现、高效等。

参考文献

- 1 Nelakuditi S, Zhang Zhili, Tsang R P. Adaptive Proportional Routing: A Localized QoS Routing Approach[C]. Proc. of INFOCOM, 2000: 1566-1575.
- 2 Xue Guoliang. Optimal Multi-path End-to-end Data Transmission in Networks[C]. Proc. of ISCC, Antibes, France, 2000.
- 3 Moy J. Request for Comments 2328, Internet Engineering Task Force[Z]. Network Working Group, 1998.
- 4 Villamizar C. MPLS Optimized Multipath MPLS-OMP[Z]. Internet Draft, 1999.
- 5 Villamizar C. OSPF Optimized Multipath (OSPF-OMP)[Z]. Internet Draft, 1999.
- 6 Gojmerac I, Ziegler T, Ricciato F, et al. Adaptive Multipath Routing for Dynamic Traffic Engineering[C]. Proc. of IEEE GLOBECOM 2003: 3058-3062.
- 7 Garcia L A, Behrens J. Distributed, Scalable Routing Based on Vectors of Link States[J]. IEEE Journal on Selected Areas in Communications, 1995, 13(8): 1383-1395.
- 8 Basu A, Lin A, Ramanathan S. Routing Using Potentials: A Dynamic Traffic-aware Routing Algorithm[C]. Proc. of ACM SIGCOMM, Karlsruhe, Germany, 2003: 37-48.
- 9 Dijkstra E W, Scholten C S. Termination Detection for Diffusing Computations[J]. Information Processing Letters, 1980, 11(1): 1-4.
- 10 XU Wuping, Yan Puli, Wu Ming. Multi-Path Routing and Resource Allocation in Active Network[J]. Wuhan University Journal of Natural Sciences, 2005, 10(2): 398-404.

(上接第 35 页)

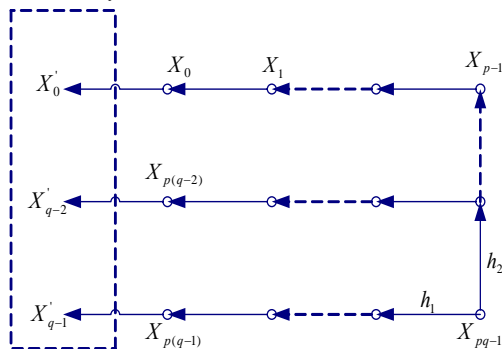


图 4 推广的二维 hash 链

5 总结

本文分析了 Quan 的多维 hash 链结构的不可行之处,对 Quan 方案进行了改进,给出了相应 Payword 方案。总体上,二维 hash 链的效率比一维 hash 链有了大幅提升,使最大 hash 运算次数由 $O(N)$ 降为 $O(\sqrt{n})$ 。在一维扩展成二维的基础上,可以很容易地将其更进一步扩展为 n 维。

参考文献

- 1 Rivest R, Shamir A. Payword and Micromint: Two Simple Micropayment Schemes[C]. Proceedings of the 4th Security Protocols International Workshop, 1996: 69-87.
- 2 Sunhyoung K, Wonjun L. A Payword-based Micropayment Protocol Supporting Multiple Payments[C]. Proceedings of the 12th IEEE International Conference on Computer Communications and Networks, 2003.
- 3 Yan Zongkai, Lang Weimin, Tan Yunmeng. A New Fair Micropayment System Based on Hash Chain[C]. Proceedings of the IEEE International Conference on E-technology, E-commerce and E-service, 2004.
- 4 Yang Chingnung, Lin Tzong, Chen Tse-Shih. Enhanced Fair Micropayment Scheme Based on Hash Chain to Avoid Merchant Collusion[C]. Proceedings of the IEEE International Symposium on Consumer Electronics, 2005.
- 5 Quan S N. Multi-dimensional Hash Chains and Application to Micropayment Schemes[C]. Proc. of International Workshop on Coding and Cryptography, Bergen, Norway, 2005.

