# 一种基于最长前缀匹配的分段式 IP 查表方法*)

张文柱　王　炫

(西安电子科技大学综合业务网国家重点实验室、信息科学研究所　西安 710071)

**摘　要**　基于最长前缀匹配,本文提出了一种新的 IP 转发表搜索方法。该方法在实现过程中依赖的主要硬件是一片逻辑控制器以及高速的 DDR II (Double Date Rate Ⅱ)SDRAM (Synchronous Dynamic Random Access Memory)。依据研究 IP 地址前缀所得出的规律,将 IP 地址前缀存储到 DDR II 中。该搜索方法能够将搜索时间限制在两个 DDR II 读周期之内,不超过 4 ns;同时保证转发表更新时间小于 512 ns。

**关键词**　IP 转发表,最长前缀匹配,IP 地址前缀

## A Segmented IP Forwarding Table Searching Scheme Based on Longest Prefix Matching

ZHANG Wen-Zhu　WANG Xuan

(State Key Lab. of ISN and Information Science Institute, Xidian Univ., Xi'an 710071)

**Abstract**　This paper describes a new IP forwarding table searching scheme based on Longest Prefix Matching that can be implemented on a general logic controller together with high-speed DDR II (Double Date Rate II)SDRAM (Synchronous Dynamic Random Access Memory). It uses three pieces of DDR II to store IP forwarding table corresponding to the IP address prefixes whose characteristics have been exploited. The scheme limits the search time in two DDR II's reading periods, which is 4 ns, and provides fast updating speed less than 512 ns.

**Keywords**　IP forwarding table, Longest prefix matching, IP address prefix

## 1　引言

近年来,Internet 得到了迅猛的发展。为应对迅速增长的 IP 业务,网络中的物理链路已经广泛采用了高速率的光纤。今天,密集波分复用技术的应用已使链路速率达到 40Gbps 甚至更高,因此,对骨干路由器分组转发能力的要求也日益提高,要求具有十吉比特转发能力。然而,目前的分组转发的速度还有待于提高。在转发一个 IP 分组时,需要依据分组头提供的 IP 地址搜索 IP 分组转发表,以确定 IP 分组的路由。搜索 IP 分组转发表的速度对分组转发速度起决定作用,是制约分组转发速度提高的瓶颈。传统的在 CIDR (Classless Inter-Domain Routing[1])环境(即允许任意长度的前缀)所采用的最长匹配(Longest Prefix Matching)IP 转发表搜索方法的效率还不尽如人意。我们举例说明一下"最长匹配"原则。假设转发表有 P1=01010110,P2=01011010 以及 P3=010110101011 三个前缀,按照最长匹配搜索方法,则前 12 比特为 010101101011 的 IP 地址与 P1 达到"最长匹配";前 12 比特为 010110101101 与 P3 达到"最长匹配"。

针对"最长匹配"IP 转发表搜索方法的效率比较低这一问题,有些研究人员提出了一些改进方法[2],有些改进方法也取得了比较高的平均搜索吞吐量。然而,这些方法还是基于前缀扩展技术[3],因此会带来转发表更新速度慢这一问题。

本文基于最长匹配,提出了一种新颖的分段式 IP 转发表搜索方法。该方法依据研究 IP 地址前缀所得出的规律,将 IP 地址前缀存储到中 DDR Ⅱ中。该搜索方法能够有效提高搜索速度,将搜索时间限制在两个 DDR Ⅱ读周期之内,同时具有更新时间短的优点。

## 2　分段式 IP 转发表搜索方法

图 1 是依据 MAE-EAST[4]转发表的数据绘制的柱状图。

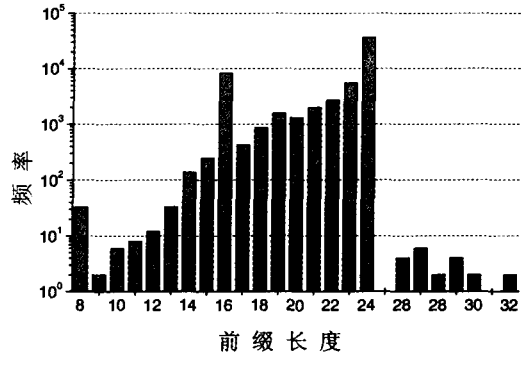



图 1　不同前缀长度分布的频率

从该图我们可以看出:骨干路由器转发的分组其前缀长度的分布是不同的。据此,我们可以在转发表的不同区间存储前缀长度不同的表项。按照我们的设计,对应于不同长度的前缀,转发表由三部分构成,图 2 给出了分段 IP 转发表逻辑框图。

由图 2 可见,转发表由 SegH、SegM 和 SegL 三部分组成,这三部分都存储在 DDR II 中。对应于前缀长度介于 8 到 16 的目标地址的表项存储在 SegH 中;对应于前缀长度介于 17 到 24 的目标地址的表项存储在 SegM 中;对应于前缀长度介于 25 到 32 的目标地址的表项存储在 SegL 中。将目标地

*)基金项目:国家自然科学基金(No. 60572144)。张文柱　博士,讲师,主要研究领域为分组交换网络、网络的协议设计以及无线 ad hoc 网络性能评估。王　炫　博士生,主要研究领域为网络的协议设计、无线 ad hoc 网络和个人通信系统。

址的前 16 位传送到 SegH，前 24 位送到 SegM。SegM 可能有两个输出：一个是对应于输入的下一跳地址，另一个输出是 12 比特的索引；将目标地址的最后 8 比特和上述 12 比特的索引送到 SegL。SegH、SegM 以及 SegL 三个中的任一个都可能产生输出，该输出与输入相对应，是要将一个到来分组送到其目的地址的下一跳。逻辑控制部分负责决定哪个输出有效，判决规则是：(1)当 SegH、SegM、SegL 都有输出时，SegL 的输出屏蔽来自 SegH 和 SegM 的输出；(2)当 SegL 没有输出而 SegH、SegM 有输出时，SegM 的输出屏蔽 SegH 的输出；(3)如果 SegH 的输出是唯一的输出，则该输出有效。该规则可确保指示下一跳的输出是按照“最长匹配”获得的。图 2 中出现的 FH0、FM0、FM1、FL0 和 C2C1C0 的含义见下一节的相关解释。

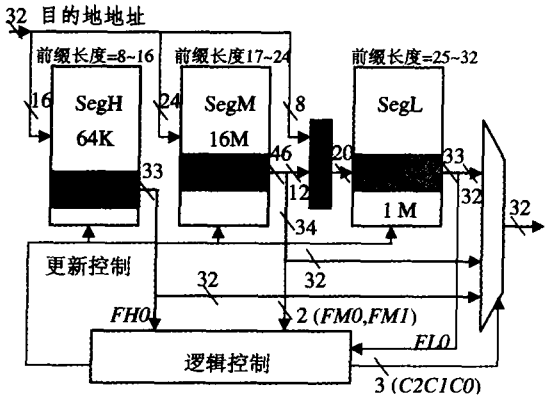
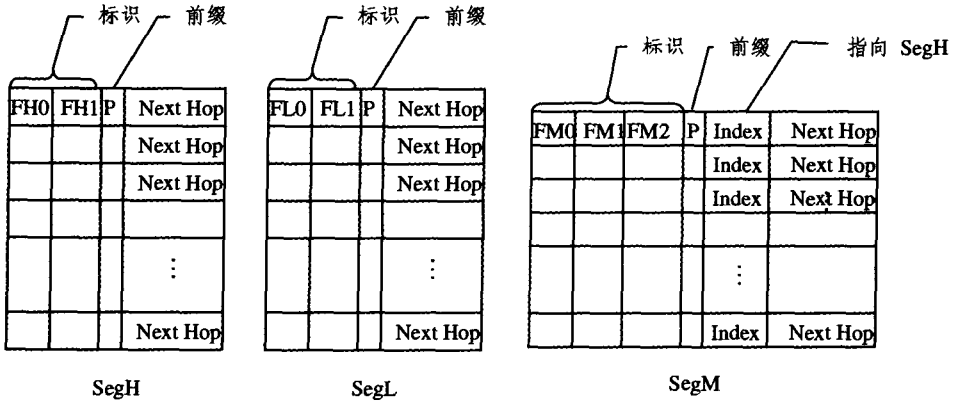



图 2　分段 IP 转发表逻辑框图

## 3　三个存储器中的内容

图 3a、3b 给出了 SegH、SegL、SegM 存储的内容。SegH 中的每个表项包含 4 个域：即“FH0”、“FH1”、“P”以及“Next Hop”。其中“FH0”域的可能取值有两个：“0”和“1”，分别表示与目标地址相匹配的表项是否包含有效的“Next Hop”。数值“0”表示该表项的“Next Hop”有效，而数值“1”表示该表项的“Next Hop”无效；“FH1”域表示前缀的长度，其取值范围是 8～16；“P”域是二进制形式的 IP 地址前缀，如 00000001、0100000000000001，等等，“P”最多可表示 IP 地址的前 16 位，对应的 IP 地址前 16 位范围是从 0.0 到 255.255，因此，SegH 的表项数目是 64k；“Next Hop”表示应该将待转发分组交付的下一跳网络节点的 IP 地址。SegM 的每个表项包含 6 个域：“FM0”、“FM1”、“FM2”、“P”、“Index”以及“Next Hop”。“FM0”域可能取值有“0”和“1”两个值，其含义与“FH0”相同；“FM2”是“P”域的长度，取值范围是 17～24；“FM1”域是全匹配标识，有“0”、“1”两个可能取值：当“FM2”的数值是 24 时，将“FM1”域置“1”，否则将“FM1”域置“0”；“P”域是二进制形式的 IP 地址前缀，如 0000000000000001、11000000000000000000000001，等等，“P”表示 IP 地址的前 24 位，对应的 IP 地址前 24 位范围是从 0.0.0 到 255.255.255，因此，SegM 表项数目是 16M。

当“FM1”的值为“1”时，其含义是仅仅在“SegH”和“SegL”中为到来的分组搜索下一跳地址是不够的，应该继续在“SegL”中搜索。要在 SegL 中继续搜索过程就要用到“Index”，“Index”代表指向“SegL”的指针的高 12 位。在“SegM”中最多有 4096 个有效的“Index”。“SegM”中的“Next Hop”域表示应该将待转发分组交付的下一跳网络节点的 IP 地址。我们将“SegL”的表项分为 4096 个组，每个组有 256 个表项。SegL 中的每个表项有 4 个域，分别为“FL0”、“FL1”、“P”和“Next Hop”。“FL0”域有“0”和“1”两个可能取值，其含义与“FH0”相同；“FL1”域表示前缀的长度，其范围是 25～32；“P”域是二进制形式的 IP 地址前缀。指向一个具体的组的指针具有相同的前 24 位；“Next Hop”表示应该将待转发分组交付的下一跳网络节点的 IP 地址。




a　SegH 和 SegL 的内容　　b　SemM 的内容

图 3

## 4　按照前缀存储路由的规则

总体规则可称为“覆盖规则”，我们以图 5 为例来解释这一规则。

首先指出我们在图 4 ～ 图 7 省略了各列的名称，读者可在图 3a、图 3b 中找出相应列的名称。从图 4 可以看出，已经给出了目标地址前 16 位从 1.0 到 1.255 的 IP 目标地址的下一跳。首先，对于从 1.0.x.x 到 1.255.x.x 的 IP 地址，其前缀为 00000001/8(斜杠后的数值是前缀长度)，相应表项的“Next Hop”已置为“R1”；其次，对于从 1.128.x.x 到 1.255.x.x 的 IP 地址，其前缀为 000000011/9 的表项的下一跳置为“R2”，覆盖了前面的“R1”；最后，对应于从 1.128.x.x 到 1.191.x.x 的 IP 地址，其前缀为 0000000110/10 的表项的下一跳置为“R3”，覆盖了前面的“R2”。对应于 IP 地址前缀长的表项，其“Next Hop”域的值要覆盖前缀短的表项的“Next Hop”域的值。图 5 给出了 SegH、SegM、SegL 中一些表项的内容。

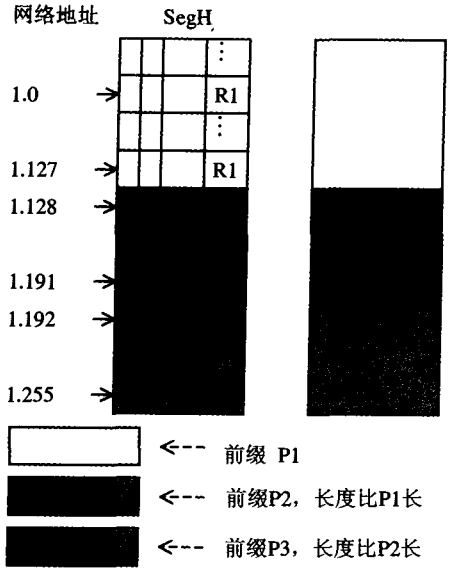



图 4　按照前缀存储路由的规则

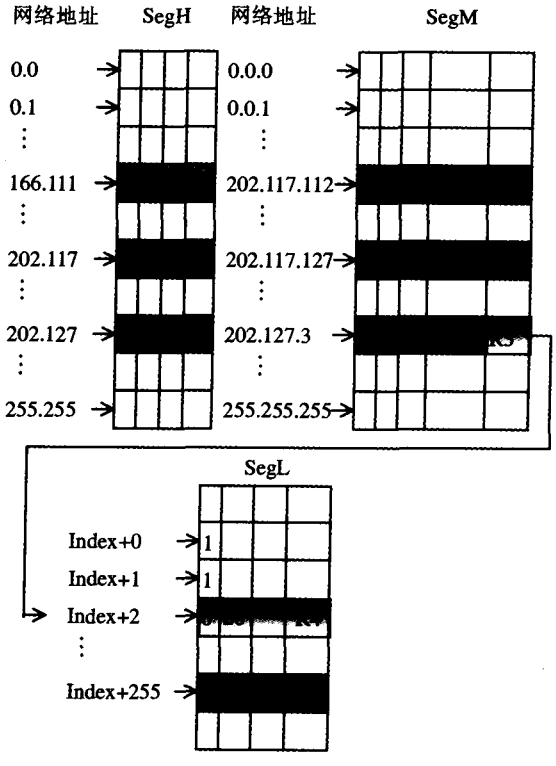



图 5　SegH、SegM、SegL 中内容示例

## 5　查表的过程

我们通过具体的例子来描述查表的过程，见图 6。图 2 中的逻辑控制器中执行的计算公式见下面公式(1)。

$$C2=FH0|\overline{FM0}$$
$$C1=FM0|(FM1\&\overline{FL0} \quad (1)$$
$$C0=FM0|\overline{FM1}|FL0$$

$C2C1C0$ 是逻辑控制器的输出，有三个可能值，分别为 011、101 和 110。当逻辑控制器的输出为 011 时，认为 SegH 的输出的下一跳有效，而 SegM、SegL 的输出无效；当逻辑控制器的输出为 101 时，认为 SegM 输出的下一跳有效，而 SegH、SegL 的输出无效；当逻辑控制器的输出为 110 时，认为 SegL 输出的下一跳有效，而 SegH、SegM 的输出无效。

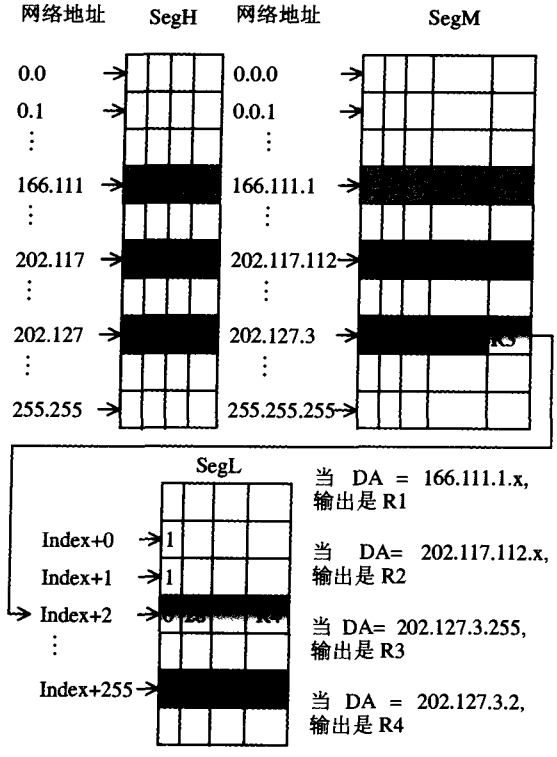



图 6　搜索转发表的过程

按照上述规则，当目标地址 IP 为 166.111.1.x 时，SegH 输出的 R1 被选为有效输出。相似地，当目标地址为 202.117.112.x 时，R2 被选为有效输出；当目标地址为 202.127.3.255 时，R3 别选为有效输出；当目标地址为 202.117.3.2 时，R4 被选为有效输出。详细图解见图 6。实际应用中，我们可以将转发表的表项存储在读写速度快的 DDR II SDRAM 中。随着半导体技术的发展，现在的 DDR II 读写周期可在 2ns 之内。从前面的讨论可以知道：在转发表中搜索一个 IP 地址最多只需读取 DDR II 两次，因此可以在 4ns 之内完成，即 1 秒钟之内可以完成 $2.5\times10^8$ 次的 IP 地址搜索，这个速度是文[2]中获得的搜索速度的 2～2.5 倍。

## 6　更新表项

骨干路由器的转发表的表项需要经常更新[5]。在更新转发表的同时必须停止查表过程，因此，更新表项的时间过长，必将降低路由器的吞吐量，导致性能下降。这就要求在路由器的设计过程中，尽量减小搜索引擎更新转发表的时间，以提高路由器的吞吐量，避免性能下降。利用本文中设计的转发表就可以获得快速的转发表更新。下面我们就描述一下更新过程。当运行在骨干路由器上的 BGP 协议收到一个指示更新的消息时，该协议就把这个消息转送给图 2 中的逻辑控制器，逻辑控制器立即检查目标 IP 地址和相应的前缀长度，以确定更新哪一个表项：当前缀长度介于 8 到 16 时，可确定表项位于 SegH 中；当前缀长度介于 17 到 24 时，可确定表项位于 SegM 中；当前缀长度介于 25 到 32 时，可确定表项位于 SegL 中。这样就首先确定了表项位于哪个 Seg＊。再通过分析目标 IP 地址，从中提取出网络地址，从而可以在确定表项在 Seg＊中的准确位置。我们来看图 7。对于 IP 地址从 166.114.128.x/17 到 166.144.225.x/17 的 IP，SegM 中相应的表项更新其“Next Hop”。这个例子中有将近 128 个表项被更

新。准确地说，更新的表项数是 $2^{24-FM2}$ ($17 \leqslant FM2 \leqslant 24$)，可见更新的表项数最大值是 128。相似地，SegH 中一次性更新的表项数是 $2^{16-FH1}$ ($8 \leqslant FH1 \leqslant 16$)，更新的表项数最大值是 256；SegL 中一次性更新的表项数是 $2^{32-FL1}$ ($25 \leqslant FL1 \leqslant 32$)，更新的表项数最大值是 128。大多数情况下，需要更新的是对应于前缀长度介于 16 到 24 的表项，因此需要的时间只是一个 DDR II 读周期。最差的情况是一次需要更新 256 个表项，假设我们转发表存储在是读写周期为 2ns 的 DDRII 中，此时更新操作所需要的时间也仅仅是 512ns，这个值大约是文[3]中获得的平均更新时间的 25%。

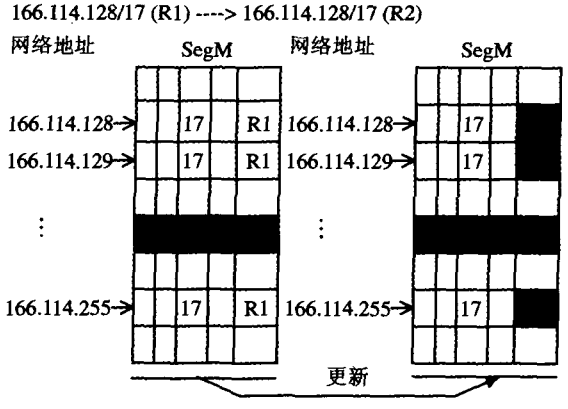



图 7　更新 SegM 中的表项

## 7　性能

转发表的逻辑框图见图 2，假设该图中的 SegH、SegM 以及 SegL 采用的是读写周期为 2 ns 的 DDR II。根据前面的分析，当 IP 地址的前缀长度介于 8 到 16 时，相应的表项位于 SegH 中，此时搜索下一跳只需要一个 DDR II 读写周期，即 2ns；当当前缀长度介于 17 到 24 时，可确定表项位于 SegM 中，此时搜索下一跳仍然只需要一个 DDR II 读写周期，2ns；当前缀长度介于 25 到 32 时，可确定表项位于 SegL 中，这时需要分别读取 SegM 和 SegL，即搜索过程包括两次 DDR II 读周期，需要 2×2 ns = 4ns。将上述结果用搜索时间～前缀长度的曲线来表示，参见图 8。

**结论**　本文提出了一种新颖的基于最长前缀匹配的分段式 IP 查表方法。利用这种方法可以获得高达 $2.5 \times 10^8$/秒的吞吐量，可满足具有 40Gbps 高速光纤链路的骨干路由器的要求。此外，利用这种方法更新转发表，可以使得更新时间最多不大于 512 ns，能极大程度满足骨干路由对更新时间的要求。

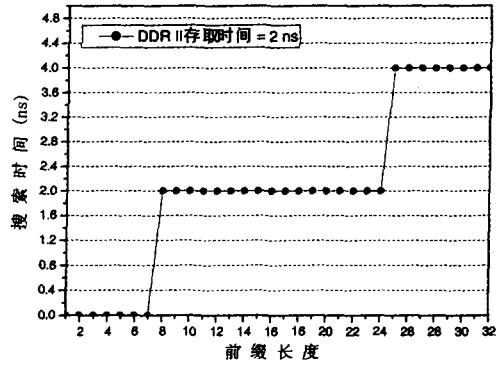



图 8　不同前缀长度需要的搜索时间

## 参考文献


1　Al-Khaffaf B A, karuppiah E K, Abdulah R. Efficient partition based IPv6 lookup algorithm for packet forwarding. In: Proc. of the 9th Asia-Pacific Conference on Communications, Penang, Malaysia, Vol. 1, Sep. 2003. 238～242

2　Berger M. IP lookup with low memory requirement and fast update. In: Workshop on High Performance Switching and Routing, Torino, Italy, Jul. 2003. 287～291

3　Jean S, Chung S H, et al. Scalable IP lookup scheme with small forwarding table for gigabit routers. Electronics Letters, 2002, 38 (6): 298～230

4　MAE-West routing database. The Internet performance Measurement and Analysis (IPMA) Project. http://www.merit.edu/impa/routing_table, Oct. 2002

5　Sundstron M, Larzon L A. High-performance longest prefix matching supporting high-speed incremental updates and guaranteed compression. In: Proc. of 24th Annual Joint Conference of the IEEE Computer and Communications Societies, Miami, FL, USA, 2005, 3: 1641～1652


(上接第 71 页)


22　Battiti R, Villani A, Le Nhat T. Neural network model for intelligent networks: deriving the location from signal patterns. In: Proceedings of The First Annual Symposium on Autonomous Intelligent Networks and Systems UCLA, May, 2002

23　Brunato M, Battiti R. Statistical Learning Theory for Location Fingerprinting in Wireless LANs. Computer Networks, 2005, 47 (6): 825～845

24　Roos T, Myllymäki P, Tirri H, et al. A Probabilistic Approach to WLAN User Location Estimation. Int Journal of Wireless Information Networks, 2002, 9(3): 155～164

25　Youssef M, Agrawala A K. Handling Samples Correlation in the Horus System. IEEE Infocom, Hong Kong, March 2004

26　Youssef N, Agrawala A. On the Optimality of WLAN Location Determination Systems. In : Communication Networks and Distributed Systems Modeling and Simulation Conference, San Diego, California, January 2004

27　Xiang Z, Song S, Chen J, et al. A Wireless LAN-based Indoor Positioning Technology. IBM Journal of Research and Development, 2004, 48(5-6): 617～626

28　Castro P, Chiu P, Kremenek T, et al. A Probabilistic Location Service for Wireless Network Environments. In: Proceedings of Ubicomp 2001, Springer Verlag, September 2001. 18～24

29　Madigan D, Elnahrawy E, Martin R P, et al. Bayesian Indoor Positioning Systems. In: Proceedings of the 24 th Joint Conference of the IEEE Computer and Communication Societies (INFOCOM 2005), Miami, FL, March 2005

30　Tekinay S, Chao E, Richton B. Performance Benchmarking for Wireless Location Systems. IEEE Communications Magazine, April 1998

31　Krishnan P, Krishnakumar A S, Ju Wen-Hua, et al. A System for LEASE: Location Estimation Assisted by Stationary Emitters for Indoor RF Wireless Networks. INFOCOM, 2004