# 应用层组播综述*

李珺晟，余镇危，潘　耘，李　霞，曹建华，武浦军
(中国矿业大学(北京)计算机系，北京 100083)

**摘　要**：为了加速组播的应用，解决现有组播存在的问题，近年来提出了应用层组播。将组播的功能从路由器转移到终端，不需要路由器维护组播组的路由表，且不用改变现有网络设施，方便实现组播功能。论述了现有的应用层组播，并对它们进行比较和评价，最后提出发展前景和进一步的研究方向。

**关键词**：应用层组播；组通信；组播路由；覆盖拓扑
**中图法分类号**：TP393.04　　**文献标识码**：A　　**文章编号**：1001-3695(2004)11-0014-04

# A Survey of the Application-level Multicast

LI Jun-sheng, YU Zhen-wei, PAN Yun, LI Xia, CAO Jian-hua, WU Pu-jun
*(Dept. of Computer, China University of Mining & Technology at Beijing, Beijing 100083, China)*

**Abstract:** In order to accelerate the deployment and solve the existing problems of multicast, there has been proposed application-level multicast. It shifts the multicast ability from routers to the end hosts, therefore routers need not maintain the routing table and does not modify the infrastcture of the current network. The multicast function is easily fulfilled. This paper depicts the existing application-level multicast, evaluates and compares with them. In the end the prospect and the future work is discussed.

**Key words:** Application-level Multicast; Group Communication; Multicast Routing; Overlay Topology

1988 年 Steve Deering 首先在他的博士论文中提出 IP 组播。IP 组播用于一对多、多对多、多对一的组通信。它是一种有效的数据传输应用，发送的同一数据在物理链路中只传输一次，减少了数据包在网络传输中的冗余，节约了带宽，提高了传输效率[1]。组播[2]自提出到现在已经有十多年了，却还没有被广泛地应用到 Internet[3]，主要有以下一些原因：

(1) 技术上的原因。组播还是一个很复杂的研究领域，许多的协议还没有实现，对其监控、管理比较复杂，如对组的管理、组播地址的分配、安全方面的、支持网络管理方面的问题。

(2) 市场的原因。打破了传统的计费模式，当前的组播的服务模式没有支持组播的付费，要在当前的服务模式和协议体系结构下普遍化和商业化，IP 组播会遇到很多困难。

(3) 组播还存在“鸡”和“蛋”的问题。当前组播的研究主要还是受学术的驱动，而客户需求驱动较少。

应用层组播将对组播功能的支持从路由器转移到终端系统，在终端之间运用原来的单播方式进行传输，这样不必改变原有网络中基础设施，也不需要路由器维护组播组的路由表，可以比较容易地实现组播，加速了应用。

## 1　应用层组播介绍

应用层组播的基本模型[4]如图 1 所示。图 1(a) 为 IP 组播数据传输的方式，数据在网络内部的路由器上进行复制；图 1(b) 为应用层组播的数据包在网络的终端系统进行复制。

收稿日期：2003-11-22；修返日期：2003-12-26
基金项目：教育部博士点基金资助项目(20030290003)

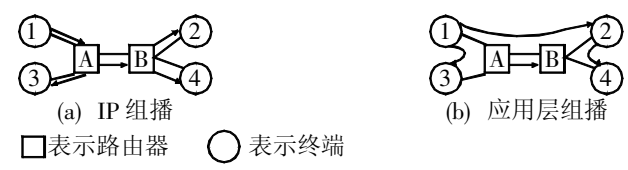


图 1　应用层组播的模型

### 1.1　应用层组播的优点

(1) 应用层组播能够很快就进入应用，不需要改变现有网络路由器。

(2) 接入控制更容易实现。由于单播技术在这方面比较成熟，而应用层组播是通过终端系统之间单播来实现的，所以差错控制、流控制、拥塞控制容易实现。

(3) 地址分配问题也就可以有相应的解决方案。

### 1.2　应用层组播的缺点

(1) 可靠性：终端系统的可靠性比路由器差。

(2) 可扩展性：底层的路由信息对应用层组播来说是隐藏起来的，可扩展性不好。

(3) 延迟比较大：IP 组播主要是链路上的延迟，而在应用层组播中，数据还要经过终端系统，因而延迟相对要大一点。

(4) 传输效率不如 IP 组播：应用层组播在数据传输过程中会产生数据冗余，因此它们比 IP 组播的效率差。

### 1.3　应用层组播的性能参数[5]

评价应用层组播协议一般用以下几种方式：

(1) 数据分发路径的质量

主要有下面三个指标：

①强度(Stress)。在一条物理链路中发送相同数据包的数量。显然 IP 组播进行转发的时候并进行多余的复制，所以是

最优值 1。如图 1( b) 中 1 ～4 的强度为 2。

②伸展度( Stretch) 。就是在覆盖网分发拓扑中从源到成员的延迟与利用单播直接传输的延迟的比例。

③资源利用率( Usage) 。所有参加到数据传输中的成员, 他们的延迟和强度的乘积的总和。这个指标用于评定传输过程中网络资源的利用情况, 假定链路的延迟越高, 花费越大。

( 2) 终端的性能

①失效后包丢失: 单个节点突然失效后, 平均的丢包数量。强调突发事件发生的鲁棒性。

②收到第一个包的时间: 当成员加入到组中, 收到第一个包的时间。

( 3) 控制负荷( Control Head)

为了有效地利用网络资源, 对每个成员的控制负荷必须尽量的小, 这是能否很好扩展重要的指标。

## 2　近年来提出的一些应用层组播

### 2. 1　终端系统组播( Narada ) [6]

#### 2. 1. 1　终端系统组播简介

终端系统组播提出 Narada 协议, 运行完全分发协议, 终端系统以自组织方式形成覆盖网络。终端系统通过适应网络中的动态性和考虑应用层的性能指标优化覆盖网的效率。在进行设计 Narada 提出了以下目标:

( 1) 自组织( Self- organize) 。终端覆盖网的构造要以完全分发的方式, 动态适应组成员变化时, 需具有较强的鲁棒性。

( 2) 有效的覆盖网。覆盖网的构造必须使物理传输链路的冗余性最小化。

( 3) 自身优化的能力。终端系统要能广泛收集网络中的信息, 并借此对 Mesh 网的结构进一步的优化。

#### 2. 1. 2　构造 Mesh 网

基于 Mesh 网的构造, 是因为需要进行多源的服务。Mesh 网允许我们对每个发送源构造最优树, 可以将成员的管理抽象到 Mesh 层。我们在构造分发树时还可以使用已有的路由算法。以下我们介绍 Mesh 网的构造:

( 1) 新成员加入 Mesh 网

当有成员加入组时, Narada 假设该成员可以利用外部引导机制获得组成员的列表。这个列表并不需要非常的完全和准确, 但至少要包含一个当前活动的组成员。

该成员从它的列表中随机的选择一些组成员和给他们发送加入邻居的信息请求。重复这个过程, 直到从列表成员中获得响应。一旦获得加入, 成员开始与它的邻居交换刷新信息, 使成员之间能相互的了解。

( 2) 成员离开和失败

当成员离开组时, 它就通知邻居, 这个信息通过 Mesh 网传送到其他的组成员 。假设 C 成员失效了, Mesh 网中 C 的邻居 A 和 G 接收不到 C 的刷新信息, 他们独立地发送探测消息到 C, 所以每个发送信息( 与 C 的回应) 都丢失的可能性很小。如果是这种情况, A 和 G 就假定 C 失效, 因此把这个信息发送到整个 Mesh 网。如果节点 A 失效, 那就造成 Mesh 网的分离, 就要运行一定的算法进行验证, 一旦确认, 就要加上链路, 将两个分开的 Mesh 网重新连接起来。

每个成员需要保留有失效的组成员, 这样才能区分是新成员的刷新, 还是失效的成员的重申, 失效的成员能在满足一定的时间重新被激活。

( 3) Mesh 网分离的修复

成员的失效可能导致 Mesh 网的分离。例如在图 2 中, 如果成员 A 失效, 到达一定的时间, Mesh 网的两边就不再接收成员之间的刷新信息, Mesh 网就分离了。在此情况下, 成员必须首先检测到 Mesh 网分离的存在, 然后用至少添加一条覆盖链路去重新连接。

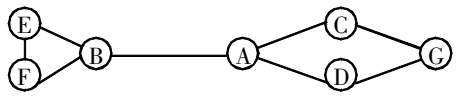


图 2　Mesh 网的失效情况

每个成员保持有 tm 时间, 在这段时间内没有收到成员的更新信息, 并利用一定的算法, 计算出失效点是否是头节点, 是否还有链路连接在它的上面, 并添加链路进行恢复。

#### 2. 1. 3　Mesh 网性能的优化

以下原因使 Narada 构造的 Mesh 网并不一定是最优的:

( 1) 新节点的加入并没有考虑网络拓扑。

( 2) 分离和修复可能产生多余的链路。

( 3) 组成员之间因加入、离开是不断变化。

( 4) 底层的网络环境、路由、负载情况不尽相同。

所以 Narada 利用一定的策略算法对链路进行评估, 添加和去除链路优化性能。

#### 2. 1. 4　分发树的构造

在 Mesh 网上运行距离向量协议和构造最短路径剪枝树。分发树的质量取决于所构造的 Mesh 网的性能。因为分发树是在 Mesh 网的基础上进行剪枝构造的, 所以 Mesh 网的构造是极为关键的。

### 2. 2　可扩展应用层组播( NICE) [7]

#### 2. 2. 1　NICE 简介

NICE 是一种可扩展的应用层组播, 主要针对大量接收者的低带宽、数据流应用, 它基于分层的结构, 同时具有较小的控制负荷, 有利于它的扩展性。NICE 的数据拓扑隐含在它的控制拓扑中, 因而还可以支持不同源的数据分发树; 并且由于它的分层, 进行错误检测较为迅速。

#### 2. 2. 2　NICE 的控制拓扑和数据拓扑

成员怎么进行分层, NICE 分层将成员赋予不同的层。如图 3 所示, 从 $L_0$ 到 $L_1$ 层在每个层中, 成员被分为多个群集, 每个群集的数量 k ～3k- 1。

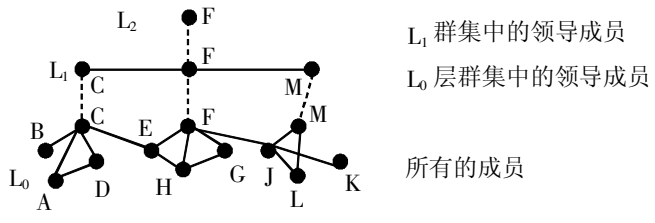


图 3　NICE 成员的分层

控制拓扑的图形, 将每层群集中的领导节点连接起来, 每个节点只包含一定常数量的成员信息, 这样有更好的扩展性。数据拓扑、图形, 可以在控制拓扑上, 单源进行方便的构造, 分发树隐含在控制拓扑中, 可以方便地构造多个分发树。

控制拓扑: 每个层中群集的成员可以很快地清楚其对等成

员的变化信息。如图 4 中 $B_0$ 可以知道 $L_0$ 层上 $A_0$, $A_1$, $A_2$, 还有 $L_1$ 层上 $B_1$, $B_2$, $C_0$ 的信息, 邻居之间信息的交换在控制拓扑上交换。

数据拓扑: 组播数据的分发路径需要防止循环。数据的分发是一个具有固定源的树。如图 5 所示, 从源 $A_0$ 发送数据的情况我们可以看出数据拓扑隐含于控制拓扑中。

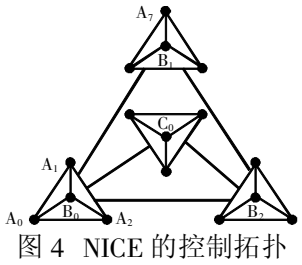

图 4 NICE 的控制拓扑

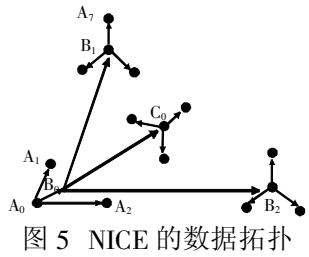

图 5 NICE 的数据拓扑

### 2.2.3 NICE 协议的描述

(1) 成员的加入

当一个成员要加入组播组的时候必须映射到 $L_0$ 层的某个群集上。从高层开始向汇聚点进行提出查询, 寻找最近的成员, 然后顺序查询至 $L_0$ 层。

(2) 群集的维持和精简

①成员的分裂和聚合: 群集中的领导成员定期地检测群集中成员的数目。如果群集的数目超过 3k-1, 那么就将这个群集分裂成两个数量相同的群集; 如果群集的数目少于 k, 那么就将群集进行聚合。

②群集的优化: 因为成员的加入不一定加到最近的群集中(可能由于查询和响应信息的丢失等), 所以成员定期地检测他的上层, 寻找最近的群集。

(3) 成员的离开和领导节点的选择

正常离开: 发送离开信息到它所属的群集中的所有成员。

突然离开: 群集中的其他主机接收不到离开主机的刷新信息, 经过一定的时间就认为该成员失效了。如果离开的主机是这个群集中的领导成员, 那么就在剩下的成员中选择一个作为新的领导成员。

## 2.3 应用层组播体系结构(ALMI)[8]

### 2.3.1 ALMI 简介

ALMI 采用集中式对成员进行管理, 主要针对成员数量较少的组播应用。在 ALMI 中, 一个组播组由一个会话控制器和多个组播成员组成。利用控制器集中对成员的管理和组播树的构造, 并可以提供基于 Java 的中间组件。会话控制器是一个程序实体, 它运行在所有成员都能访问到的位置, 它可以与某个成员运行在同一台计算机上, 通常是与组的发起成员在一起, 或者位于 ISP 提供的某个组播服务网关上。

组播成员之间形成一棵组播树。组播树中的一条链路代表两个成员之间的一条单播连接。数据信息沿着组播树进行分发, 而控制信息则通过会话控制器与各个成员之间的单播连接进行传输。

### 2.3.2 会话控制器的主要功能

会话控制器(图 6)的主要功能是对加入成员进行注册和维持组播树。它通常放置于成员容易接入的地方。它保证连接性: ①当成员加入、离开会话或网络或主机的失效时保证网络的连接性; ②保证传输效率: 定期从所有成员收集信息计算最小剪枝树。

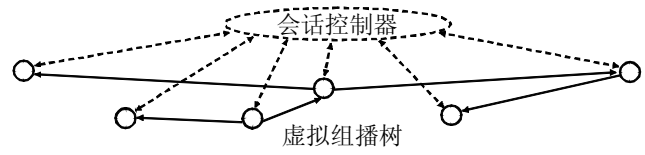

图 6 会话控制器

### 2.3.3 控制协议的介绍

ALMI 利用控制协议在会话控制器和成员之间进行通信。这个协议主要负责成员管理、性能监控、路由等工作。它的包头格式如图 7 所示。

| 0 | 7 | 15 | 31 |
|---|---|---|---|
| 协议版本 | 树的表示 | 标志符 | |
| ALMI 会话 ID | | | |
| 源 ID | | | |
| 序号 | | | |
| 数据负荷长度 | | | |

图 7 包头格式

其中标志位的作用为: ①连接请求和回应; ②性能监测信息; ③分发树信息; ④邻居监测更新信息; ⑤分离信息。

树的表示域指明树的版本数, 可以用来防止组播树的循环和分离。循环可能的原因, 丢失树的更新信息、成员间不同的响应延迟。

### 2.3.4 成员的操作

当有成员要加入组的时候, 首先成员定位到控制器, 在组初始化的时候控制器已经用不同的方式对会话 ID 和控制器地址及端口号进行了声明; 接着成员就向控制器发送加入消息, 并从控制器获得它的 ID 和父节点的地址; 然后该成员就发送嫁接消息给它的父节点, 完成后就可以传输数据了。

### 2.3.5 ALMI 组播树的构造

ALMI 组播树是一棵连接所有成员的虚拟最小剪枝树。它是利用控制器与所有成员用(父, 子)表通信结果计算所得的。可以根据不同的性能指标进行分发树的构造, 如带宽、延迟等。

组播树的优化, 成员将它们的监测报告发送给控制器, 控制器就可以根据这些信息对分发树进行优化。

# 3 现有应用层组播的比较与评价

首先我们对现有的应用层组播进行分类。如图 8 所示, 我们可以将应用层组播按分发算法和集中算法分为两大类。其中属于集中算法的有 ALMI, HBM[9], 它们主要对成员进行集中的管理。而分发算法则按其拓扑的建立分为 Mesh 网优先和树优先。利用 Mesh 网优先其组播鲁棒性好, 但控制负荷较大。另外现有的应用层组播依据其发送者的个数, 我们可以将现有应用层组播分成单源和多源。Overcast[10] 属于单源, 它们只有一个发送源, 通常情况网络拓扑比较稳定, 实现起来也比较方便。Narada 属于多源, 它们可以有多个发送源, 因此网络拓扑需要动态的改变, 实现也比较复杂。利用和结合 IP 组播, 有 TAG[11], HMTP[12]。针对少量的接收成员情况, 有 ALMI, Narada。此外, 现有的应用层组播将 P2P 的思想应用到组播路由的设计, 可以实现大量接收者的组播, 如 CAN[13], Bayeux[14], Scribe[15]。还有应用数学上的方法进行网络拓扑的设计, 构造出适于应用层组播的网络拓扑, 如 DT[16]。Scattercast[17] 主要是用于广域网的内容分发。Yoid[18] 提出一系列的协议, 适应普遍的内容分发网络。另外, 还有针对延迟、带宽等

方面的应用层组播实现, 也有针对实际应用的应用层组播。

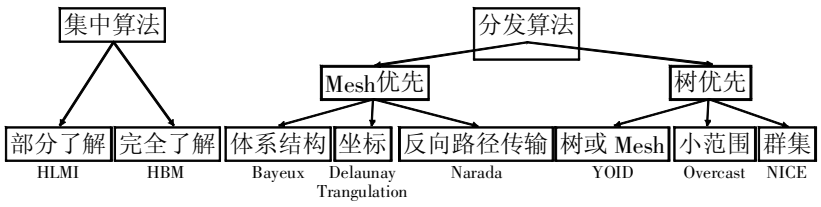


图 8　应用层组播的分类

针对上一部分的应用层组播对它们进行简单的评价[19], 如表 1 所示。

表 1　应用层组播的比较

| 协议 | 类型 | 组播树类型 | 最大路径长度 | 树上度的最大值 | 评价控制负荷 |
|---|---|---|---|---|---|
| Narada | 基于 Mesh 网 | 面向源 | 没有上界 | 有近似上界 | O( N) |
| NICE | 基于树 | 面向源 | O( logN) | O( logN) | 常量 |
| HLMI | 基于集中的算法 | 面向源 | 没有上界 | 有近似上界 | O( N) |

## 4　结论

本文主要谈论了应用层组播出现的原因和近年来提出的一些应用层组播, 并对它们进行了总结。虽然应用层组播有其可以快速应用的优点, 但是也还存在一些问题, 需要进一步的研究和解决。由于应用层组播是在应用层实现的, 所以它的传输效率没有 IP 组播好。为了构造效率的应用层, 首要的问题是在覆盖网上构造一个优化的分发拓扑, 并进行有效的管理。我们提出了进一步研究的工作:

( 1) 可以将 IP 层和应用层组播结合起来, 利用两者各有的优点, 如利用 IP 层的网络拓扑信息, 我们可以建立性能较为优化的应用层组播网络拓扑[20]。

( 2) 可以将应用层组播与 P2P 结合起来, 将对等网络中的路由算法运用到应用层组播中, 这样构造的应用层组播一般可以应用于大量的接收者。

( 3) 可以将主动技术应用到应用层组播, 利用主动技术我们可以很容易地实现接入控制。

( 4) 另外, 我们还需要对应用层组播上的 QoS、拥塞控制、安全机制进行深入的研究。

总之, 应用层网络是一个比较新的研究领域, 还存在许多的问题, 其广阔的发展前景, 值得进一步研究。

参考文献:

[ 1]　Diot, B N Levine, B Liles, et al. Deployment Issues for the IP Multicast Service and Architecture[ J] . IEEE Network, 2000, ( 1) : 78-88.

[ 2]　H Sahasrabuddhe, B Mukherjee. Multicast Routing Algorithms and Protocols: A. Tutorial[ J] . IEEE Network Magazine, 2000, ( 1, 2) : 90-102.

[ 3]　eering S E. Multicast Routing in a Datagram Internetwork[ D] . PhD Thesis, Stanford University, 1991.

[ 4]　uman Banerjee, Bobby Bhattacharjee, Christopher Kommareddy. Scalable Application Layer Multicast[ C] . Proceedings of ACM SIGCOMM, 2002.

[ 5]　l-Sayed, Vincent Roca, Laurent Mathy. A Survey of Proposals for an Alternative Group Communication Service[ R] . Ayman, 2003.

[ 6]　ang-Hua Chu, Sanjay G Rao, Hui Zhang. A Case for End System Multicast[ C] . ACM SIGMETRICS, 2000. 1-12.

[ 7]　uman Banerjee, Bobby Bhattacharjee, Christopher Kommareddy. Scalable Application Layer Multicast[ C] . Proc. of ACM SIGCOMM '02, 2002.

[ 8]　Pendarakis, S Shi, D Verma, et al. ALMI: An Application Level Multicast Infrastructure[ C] . Proceedings of 3rd USENIX Symposium on Internet Technologies and Systems ( USITS) , 2001. 49-60.

[ 9]　incent Roca, Ayman EL-Sayed. A Host-Based Multicast ( HBM) Solution for Group Communications[ C] . IEEE International Conference on Networking ( ICN '01) , Colmar, France, 2001. 610-619.

[ 10]　Jannotti, D Gifford, K Johnson, et al. Overcast: Reliable Multicasting with an Overlay Network[ C] . Proceedings of 4th Symposium on Operating Systems Design and Implementation, 2000.

[ 11]　inseok Kwon, Sonia Fahmy. Topology-aware Overlay Networks for Group Communication[ C] . Proceedings of International Workshop on Network and Operating Systems Support for Digital Audio and Video, 2002.

[ 12]　eichuan Zhang, Sugih Jamin, Lixia Zhang. Host Multicast: A Framework for Delivering Multicast to End Users[ C] . Proceeding of IEEE INFOCOM, 2002.

[ 13]　Ratnasamy, M Handley, R Karp, et al. Application-level Multicast Using Content-addressable Networks[ C] . Proceedings of 3rd International Workshop on Networked Group Communication, 2001.

[ 14]　Q Zhuang, B Y Zhao, A D Joseph. Bayeux: An Architecture for Scalable and Fault-tolerant Wide-area Data Dissemination[ C] . 11th ACM/IEEE NOSSDAV, New York, 2001.

[ 15]　iguel Castro, Peter Druschel, Anne-Marie Kermarrec, et al. Scribe: A Large-scale and Decentralized Application-level Multicast Infrastructure[ J] . IEEE Journal on Selected Areas in Communications, 2002, 20( 8) .

[ 16]　Liebeherr, M Nahas, W Si. Application-layer Multicast with Delaunay Triangulations[ R] . University of Virginia, Department of Computer Science, Technical Report, CS-2001-26, 2001.

[ 17]　atin Chawathe. Scattercast: An Architecture for Internet Broadcast Distribution as an Infrastructure Service[ D] . Ph. D. Thesis, University of California, Berkeley, 2000.

[ 18]　Francis. Yoid Extending the Multicast Internet Architecture[ EB/OL] . http: //www. aciri. org/yoid/, 1999.

[ 19]　uman Banerjee, Bobby Bhattacharjee. A Comparative Study of Application-layer Multicast Protocols [ M] . Submitted for Publication, 2002.

[ 20]　akeshi Sano, Taku Noguchi, Miki Yamamoto. Improving Efficiency of Application-level Multicast with Network Support [ J] . Networked Group Communication, 2003, 13-22.

作者简介:

李珺晟, 男, 福建泉州人, 硕士, 主要研究方向为覆盖网络和组播路由算法; 余镇危, 教授, 博士生导师, 主要研究方向为计算机网络、下一代网络体系结构; 潘耘, 博士生, 主要研究方向为下一代网络体系结构、主动覆盖网络。

( 上接第 13 页)

[ 17]　M D 'Souza, X Fan, et al. Modifications to the pGVT Algorithm to Eliminate Acknowledgement Messages and Improve the GVT Broadcast Frequency[ J] . World Congress on Systems Simulation: Conference on Parallel & Distributed Simulation, 1997: 288-292.

[ 18]　Mattern. Efficient Algorithms for Distributed Snapshots and Global Virtual Time Approximation[ J] . Journal of Parallel and Distributed Computing, 1993, 18( 4) : 423-434.

作者简介:

周喜明( 1978-) , 男, 硕士研究生, 研究方向为数字系统的并行逻辑模拟; 吴悦( 1960-) , 女, 副教授, 从事计算机辅助设计和 EDA 技术方面的科研与教学; 杨洪斌( 1962-) , 男, 高工, 从事计算机系统结构方面的科研与教学。