基于社区的覆盖网监测与故障推理系统

张涛[1]，程光[1, 2]

（1.东南大学计算机科学与工程学院，南京，211189；
2.江苏省计算机网络技术重点实验室，东南大学，南京，211189）

摘　要：覆盖网为布置网络应用程序提供了强大而灵活的平台，覆盖网的应用越来越多，覆盖网的故障管理对于正确提供覆盖网服务起着关键的作用。故障诊断能够指明故障的根源，因此它也是故障管理系统中最重要的部分，并能解释观测到的网络症状。但是，随着因特网的规模和复杂性越来越大，我们很难清晰的描述它。单点的推理也很难满足大量症状信息的存储与计算，因此本文提出基于社区的覆盖网监测与故障推理系统通过在优势节点上布置代理以解决上述问题。

关键词：覆盖网；监测；社区；故障推理

Community Based Overlay Network Monitoring and Fault Reasoning System

Zhang Tao[1], Cheng Guang[1, 2]

(1. School of Computer Science & Engineering , Southeast University, Nanjing, 211189;
2. Jiangsu Provincial Key Laboratory of Computer Network Technology, Southeast University, Nanjing, 211189)

Abstract: Overlay networks have emerged as a powerful and flexible platform for developing new disruptive network applications. With more and more overlay networks deployed, overlay fault management is playing a crucial role in successfully provisioning overlay services. Fault diagnosis is the most critical component in a fault management system because it identifies the root causes that can best explain observed network disorders. However, as the scale and complexity of internet grow, it is difficult to describe it clearly. Single-point reasoning can not handle the storage and computation of large information, so we provide a community based overlay network monitoring and fault reasoning method to deploy agents on advantage nodes to resolve the problems.

Key words: Overlay network; Community; Monitor; Fault reason

当前在 Internet 上,覆盖网络(overlay network)得到了广泛应用, 包括文件共享和流媒体服务的 P2P 覆盖网络、内容分发网络(content deliver network, 简称 CDN)、应用层组播(application layer multicast, 简称 ALM)、虚拟实验床 EmuLab 和 PlanetLab 等。

覆盖网络是以底层物理网络为基础，在其上构建的虚拟网络系统。在覆盖网络中，节点之间的虚拟链路是逻辑上的，通常对应于底层网络的物理路径。覆盖网络可以根据应用环境和需求定义自己的拓扑结构和路由模式，结构比较灵活，可以用来构建特定于应用(application-specific)的服务，大大扩展了 Internet 的服务。

覆盖网的广泛分布性和用户层灵活性给覆盖网的故障诊断带来了新的挑战，包括底层网络的不可见、网络症状的不完整、不精确以及多层网络的复杂性。

覆盖网故障诊断主要是基于症状-故障关系图（即把症状 S 表示为多个组成部分的集合，图 1 就是症状和各个组成部分的对应关系，S_i 为症状，如时延变大，链路不通等，C_i 为组成部分，可以是一个网络、路由器、链路、自治系统等，图 1 表示症状 S_i 发生时，故障可能出现在直连的组成部分中）。

Yongning Tang 等[1，2]提出建立动态的症状-故障关系图(symptom-fault graph)的思想，推理出一个能解释所有症状的故障（即故障的组成部分，如故障的 AS）集合，并检验该故障集合是否满足一定的置信度，若不满足，则按照算法选择主动测量进行故障诊断。Yongning Tang 等[3，4，5]还提出将每一个症状（覆盖网路径）标识为故障（bad）或

作者简介：张涛，（1988-），男，硕士研究生，E-mail: tzhang@njnet.edu.cn；程光，（1973-）男，教授，博导，E-mail:gcheng@njnet.edu.cn.

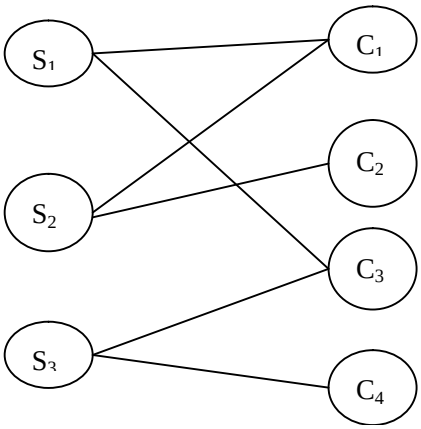

图 1 症状-故障关系图

者没有故障（good）的方法，它根据各个组成部分（component，可能是一个路由器）包含在症状中的情况判断出该部分故障或者不故障的概率。同时，也可以通过合并各个组成部分在诊断精度和粒度间找到平衡。Yongning Tang 等[6]还设计了一个 OF(component)函数，该函数能表示一个组成部分对症状的解释程度，当该函数的结果达到某个阈值，则把它加入症状-故障关系图，找出解释所有症状的最小组成部分集合。George J. Lee 等[7]通过 planetseer[9]获取 codeen 结点上 TCP 连接的情况，包括拓扑、端口、故障概率，来训练贝叶斯网络，根据 TCP 连接端点的 IP 得出端点 AS 号，这样它就能得出 AS 之间连接的故障概率，并用于以后的推断。推断时，我们要给出一条 overlay 路径（如 nodeA à nodeB à nodeC， 其中 nodeA，nodeB 和 nodeC 都是 overlay 节点）。这利用了 planetseer 了解整个覆盖网情况这一特点。监测一般需要给 overlay 节点添加功能模块，研究目的主要在于通过尽量少的监测点来实现对整个覆盖网的监测。Yan Chen 等[8]把覆盖网表示成矩阵，每一行表示一个端到端路径，每一列都表示一个底层链路（矩阵需要列出路径所包含的所有链路），1 表示该路径经过该链路。利用路径之间的链路重叠关系计算出一个 k，选择 k 个 path 进行监测。k 的大小与覆盖网结构有关，若网络是树结构则 k 为 o(n)，若无结构则为 $o(n^2)$。单点的推理很难满足大量症状信息的存储与计算，同时，对某一组件进行观察故障概率在实际中是很难做到的，因此本文提出基于社区的覆盖网监测与故障推理系统通过在优势节点上布置代理以解决上述问题。

1 总体结构

本文所设计的基于社区的覆盖网监测与故障推理系统是基于已有的故障推理算法设计的。系统主要包括两大部分：客户端和代理节点。图 2 是系统工作的逻辑拓扑结构，客户端为运行带有监测模块和推理模块的 P2P 节点，代理节点运行 chord 以及各种服务模块，处理客户端上报和请求的症状信息，客户端监测到故障需要推理时，就向代理索要有用信息，然后推理，将结果以报告的形式输出。

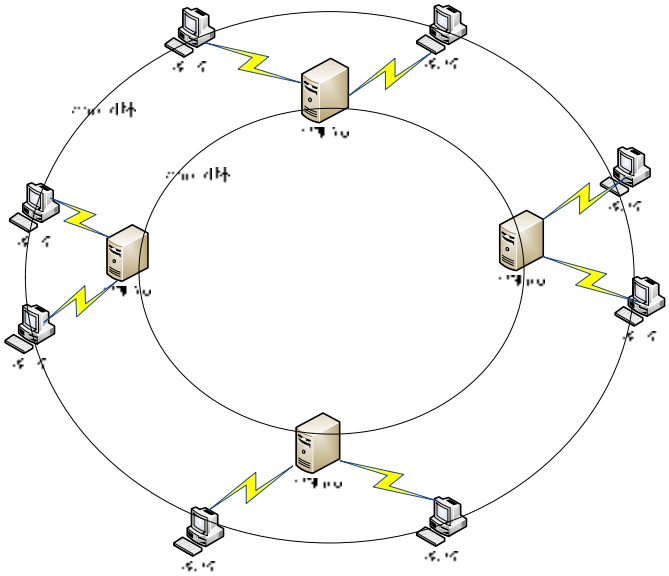

图 2 系统运行示意图

系统的工作流程如下：

（1）监测模块监测客户端覆盖网路径的时延以及联通状况，

（2）获取覆盖网通信的时延，计算每 10 分钟内各覆盖网路径的时延平均值。

（3）根据历史值，判断各路径在该时间段内的时延是否故障

（4）若本地有一个代理地址转（5），若本地没有代理信息转（6）

（5）判断该代理是否在线，若在线则向它索要自己的负责代理地址，转（7），若不在线转（6）

（6）向代理首节点索要自己的负责代理 IP

（7）客户端将该 10 分钟内的症状数据连通是否故障的信息一同上传给负责代理。

（8）代理对接收到的数据按照 IP AS 映射和

AS 拓扑处理后存入 chord，同时，如果故障的话，还要从 chord 中下载需要的数据回送给客户端。

（9）若之前上传了故障数据，则客户端根据接收到的数据进行推理，生成推理报告。结束。

2 功能模块设计

2.1 客户端

从图 3 可以看出，客户端按功能可以划分为 6 个模块：chord 模块、待查询数据生成模块、监测模块、预处理判断模块、获取负责代理 IP 模块以及推理模块。下面对除 chord 模块外的各个模块进行介绍。

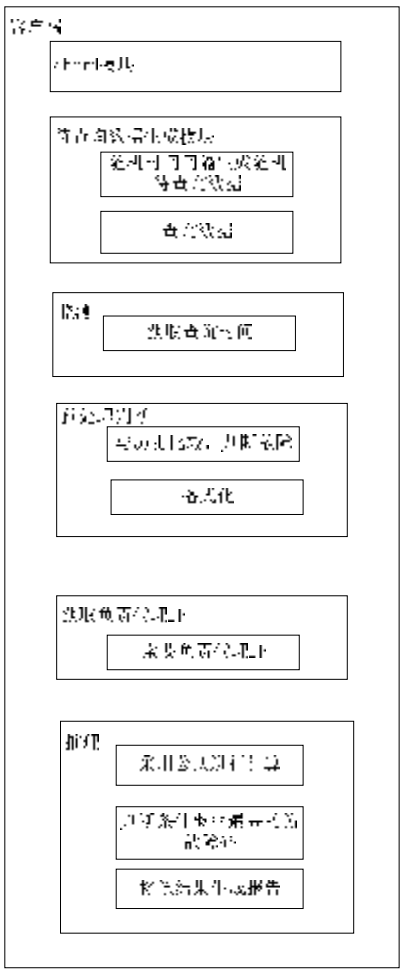

图 3 客户端模块图

1）待查询数据生成模块

功能：生成待查询的数据

设置随机的时间间隔以及生成随机的查询数据，利用 chord 提供的接口进行查询，以产生症状。

2）监测模块

功能：获取覆盖网路径时延。

在客户端覆盖网应用程序中嵌入监测模块，监测模块在每次查询时都设置定时器，记录每次查询的时间，将覆盖网路径和时延记录下来。

3）预处理判断模块

功能：对时延数据进行预处理，得到每个时间段的症状数据上传并判断当前时间段症状的故障情况。

对每个覆盖网路径计算每个时间段内的时延平均值，根据历史值采用平均值标准差的方法判断当前是否故障。

4）获取负责代理 IP 模块

功能：选择该客户端的负责代理并上传症状数据。

先向代理首节点或者本地已有的代理信息索要负责代理的 IP。根据之前的判断结果，将该 10 分钟内的症状数据表示为“AS 路径：是否症状”的形式上传给负责代理。

5）推理模块

功能：根据从负责代理获取的症状数据，推理出故障所在 AS

a) 对每个 AS 的历史症状数据按时间分段，计算每个时间间隔内的故障情况，再综合计算每个 AS 的历史故障概率记为 $p(m_i)$。

b) 选取最近时间段内的症状数据构建症状-AS 关系图。

c) 采用 Yongning Tang 推出的故障概率计算公式。

$$p(m_i|s_i)=p(m_i)\left[\sum_{h=1}^{|M_{s_i}|-1}\left\{\prod_{q=1,q\neq i}^{h}p(m_q)\prod_{k=h+1,k\neq i}^{|M_{s_i}|}(1-p(m_k))\right\}+\underbrace{\prod_{q=1,q\neq i}^{|M_{s_i}|}p(m_q)}_{Item2}+\underbrace{\prod_{q=1,q\neq i}^{|M_{s_i}|}(1-p(m_q))}_{Item3}\right]$$

其中 $p(m_i)$ 表示 m_i 这个组成部分历史的故障概率，也就是一个 AS 的故障概率。M_{s_i} 表示症状 s_i 在症状-组成图中对应的组成部分集合，即症状 s_i 涉及的 AS 集合。借助于这个公

式，我们就可以计算出在某个症状 s_i 发生时，m_i 出现故障的概率，也就是当症状 s_i 发生时，各个 AS 为故障的概率。

d) 其中概率值最大的 AS 即为故障 AS。

2.2 代理节点

代理部分有负责节点选择模块、症状数据预处理模块、存储模块、数据获取模块以及 chord 模块。

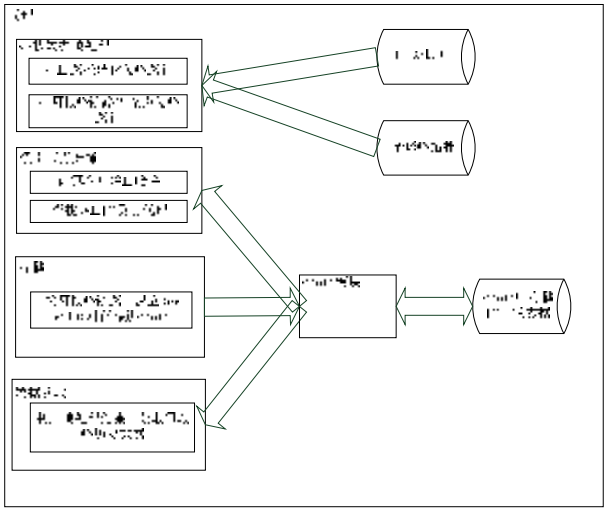

图 4 代理节点模块图

1） 负责节点选择模块

功能：为客户端选择它的负责代理

借助于 chord 模块，对代理节点和客户端节点的 IP 分别进行哈希，并将哈希结果列在环上，把客户端节点在环上顺时针方向第一个代理节点称作该客户端节点的后继，客户端在想任意一个代理节点发送请求后，该代理节点把该客户端节点的后继代理节点 IP 告诉客户端，让后继负责该客户端。这样可以实现把客户请求均匀分配给各个代理节点。

2） 症状数据预处理模块

功能：对接收的症状数据进行存储前的预处理

客户端把观测到的症状（可能为故障）表示为 ip1->ip2->ip3:症状结果。代理节点将这三个点分别查询 IP AS 映射关系，将其转化为 AS1->AS2->AS3:故障，然后查询全球 AS 拓扑，比如 AS1->AS2 的底层路径为 AS1->ASa->ASb->ASc->AS2，再将两段 AS 路径表示为一个 AS 的集合{故障：AS1，ASa，ASb ……}。

3） 存储模块

功能：设计索引，将症状存入 chord

将索引设计成 key_时间值的格式，先确定一个有意义的时间跨度。然后，根据该跨度寻找接近的时间粒度单位（如日，周，双周，月等）。对一个有 3 个 AS 的症状集和，我们存储的内容为 AS1:3:故障、AS2:3:故障、AS3:3:故障。意思是有一个表现为故障的症状涉及 3 个 AS，AS1 是其中的一个。

4） 数据获取模块

功能：按索引获取所需的 chord 数据

按需要的时间范围和 AS 链路，从 chord 中获取需要的症状数据。比如我们症状的为 ip1->ip2->ip3:症状结果，转化为 AS 路径为 AS1->AS2->AS3:故障，然后查询全球 AS 拓扑，比如 AS1->AS2 的底层路径为 AS1->ASa->ASb->ASc->AS2，再将两段 AS 路径表示为一个 AS 的集合{故障：AS1，ASa，ASb ……}，这样对其中的每个 AS 都获取比如一个月内的数据，根据索引 AS1_2011_11，就得到 AS1 在 2011 年 11 月内的症状情况，最后将这些数据发送给客户端去处理。

3 实验结果

实验环境为华东北网络中心服务器上的虚拟机网络，采用 zebra 路由软件，利用 linux netem 组件设置时延。其中四台客户端检测结果如下：75.00%，66.67%，77.78%，74.51%。平均正确率为 73.49%。通过实验，我们可以看到该系统故障诊断的正确率在可允许的范围内。

4 结束语

该系统能有效的解决上述提到的单点处理所带来的问题以及有些症状数据难以获取的问题，但是，实际环境中的网络状况非常复杂，如何让该系统应用到实际中，是我在今后的工作中要深入的问题。

参考文献

[1] Yongning Tang, Ehab Al-Shaer, Raouf Boutaba, Efficient fault diagnosis using incremental alarm correlation and active investigation for internet and overlay networks[J], Network and Service Management, 2008

[2] Yongning Tang, Guang Cheng, Zhiwei Xu, Probabilistic and Reactive Fault Diagnosis for Dynamic Overlay Networks[J], Peer-to-Peer Networking and Applications, 2010

[3] Yongning Tang, Ehab Al-Shaer, Reasoning about Uncertainty for Overlay Fault Diagnosis Based on End-User Observations[C], INFOCOM, 2009

[4] Yongning Tang, Ehab Al-Shaer,Overlay Fault Diagnosis Based on Evidential Reasoning[C], INFOCOM, 2009

[5] Yongning Tang, Ehab Al-Shaer, Sharing End-user Negative Symptoms for Improving Overlay Network Dependability[C], DSN, 2009

[6] Yongning Tang, Ehab Al-Shaer, Towards Collaborative User-Level Overlay Fault Diagnosis[C], INFOCOM, 2008

[7] George J. Lee, Lindsy Poole, Diagnosis of TCP Overlay Connection Failures using Bayesian Networks[C], SIGCOMM, 2006

[8] Yan Chen, David Bindel, Randy H. Katz, Tomography-based overlay network monitoring[C], SIGCOMM, 2003

[9] M. Zhang, C. Zhang, V. Pai, L. Peterson, and R. Wang, PlanetSeer: Internet Path Failure Monitoring and Characterization in Wide-Area Services[J], Proc. Sixth Symposium on Operating Systems Design and Implementation., 2004