基于联邦架构的全球网络性能测量

王继龙 孙明敏 张千里
(清华大学信息网络工程研究中心 北京 100084)

摘 要 性能可扩展性是当前互联网面临的主要问题之一,对互联网开展有效的测量和评价,能为寻找提高网络性能的方法提供必要的数据支撑.目前,虽然在大规模分布式测量方面已经开展了大量的出色工作,但是它们无一例外地受到测量点数量和位置的限制.基于此,文中设计并实现了基于联邦架构的全球网络性能测量平台GPERF.该平台实现了大规模异构测量,充分利用了自有资源、伙伴资源和互联网上的开放服务等.文章将从体系结构设计、测量点发展、测量任务调度和应用研究等方面对GPERF展开介绍和讨论.

关键词 互联网;网络测量;分布式测量;联邦
中图法分类号 TP393 **DOI 号**: 10.3724/SP.J.1016.2010.01602

Federation-Based Global Network Performance Measurement

WANG Ji-Long SUN Ming-Min ZHANG Qian-Li
(*Network Research Center of Tsinghua University, Beijing* 100084)

Abstract The scalability of performance is one of the major problems of the current Internet. Accurate measurement and evaluation about the Internet can provide strong support for improving the network performance. To date, many large-scale distributed measurement projects have made great efforts to network performance measurement, but they are all bound to the number and location of measurement points. To solve this problem, the authors propose the global network performance measurement platform called GPERF based on a federation structure in this paper. GPERF is a large-scale heterogeneous measurement platform, whose measurement points include its own resources, partners' resources, and open services on the Internet. This article describes GPERF's architecture design, measurement point development, measurement task scheduling and applications.

Keywords Internet; network measurement; distributed measurement; federation

1 引 言

网络测量已成为开展互联网技术研究的重要支撑,但是由于以下因素,小范围的测量结果往往缺乏普遍意义,难以为设计提高互联网性能的方法提供依据.

(1)网络行为的时空差异性.互联网的行为特性随时间、空间以及应用环境的变化而变化:不同场合,即使网络结构相同,其行为特征也可能差别巨大.因此,测量结果不具备空间普适性;同样的,不同时间段的行为特征也具有较大的差异,测量结果不具备时间普适性.

(2)网络过程的不可控性.有些网络过程是转

收稿日期:2010-04-19;最终修改稿收到日期:2010-08-05.本课题得到国家"九七三"重点基础研究发展规划项目基金(2003CB314807)和国家"八六三"高技术研究发展计划项目基金(2005AA112130)资助.**王继龙**,男,1973 年生,博士,副教授,主要研究兴趣为网络管理和网络测量.E-mail: wjl@cernet.edu.cn. **孙明敏**,男,1985 年生,硕士研究生,主要研究兴趣为网络测量. **张千里**,男,博士,助理研究员,主要研究兴趣为网络安全和流量分析.

瞬即逝的，难以捕捉；有些网络过程则耗时过久. 无论哪种情况，都会使测试验证的效率受到影响，尤其是在需要反复测试的场合.

(3) 网络行为的不可重现性. 因为网络行为是诸多因素复合作用的结果，所以某一时刻观察到的网络行为在未来往往是不可重现的.

因此，全球范围的持续测量将会对研究工作具有重要意义. 但是庞大的互联网规模给开展持续的网络测量带来了众多困难，除了测量方法和工具方面的问题，最突出的困难在于测量基础设施的建设. 一个全球范围的测量基础设施建设意味着巨大的初期投入和持续的运维成本.

针对上述问题，我们设计并实现了基于联邦架构的网络性能测量平台 GPERF. 其它联邦架构如 PerfSONAR，都要求成员主动加入，因此参与者仅仅局限于网络测量领域内的科研工作者，这往往就限制了测量平台的快速发展. GPERF 则不同，它一方面与其它项目开展合作，另一方面还挖掘整合了互联网上大量存在的免费资源. 这类资源数量庞大，分布广泛，但是只有通过我们主动去挖掘它们，整合它们，才能在大范围的网络测量中发挥其价值. 下面我们将从 GPERF 的体系结构设计、测量点发展、测量任务调度及其在 2008 奥运会重要网站性能监控测量中的应用展开讨论.

2　相关工作

2.1　国际知名的大规模网络测量项目

(1) Skitter 和 Ark

Skitter 是 CAIDA 开发的全球测量平台，主要致力于互联网宏观拓扑测量. Skitter 在全球共有 60 个测量点，目前有 20 个可用测量点，主要分布在美国、加拿大，欧洲的英国、荷兰等国以及亚洲的日本、新加坡等国. 目前 Skitter 项目已被 CAIDA 的新测量项目 Archipelago(Ark)取代. Ark 基本上是 Skitter 的延伸.

(2) PingER

PingER[1] (Ping End-to-end Reporting) 是由 SLAC 领导，NUST/NIIT、FNAL 和 GATech ICTP/Trieste 等机构共同参与的大规模主动测量系统. PingER 主要致力于互联网链路的端到端性能测量，PingER 主要使用 ping 作为测量工具. 到目前为止，PingER 在全球 23 个国家部署了 48 个测量点，并从这些测量点对全球 159 个国家 11 个地区的 882 个节点进行测量，所有的测量数据都可以从 PingER 网站上获取.

(3) SAMI (NIMI)

SAMI 美国匹兹堡大学超级计算机中心开发的分布式性能测量系统，其前身是 NIMI[2-3]. 目前 SAMI 的测量点约 35 个，主要分布在欧洲和美国的校园和一些实验网络. 由于 SAMI 过于严格的身份验证和“账户”余额制度，导致 SAMI 的发展并不是非常蓬勃，但是他的一些思想非常值得我们借鉴.

(4) AMP

AMP[4] (Active Measurement Project) 项目由 NLANR 发起，测量的主要内容包括测量点间的丢包、延迟、带宽和网络拓扑. 至今为止 AMP 项目共布设了 150 个测量点. 2006 年 7 月，AMP 项目被 CAIDA 接管，并于 2006 年 9 月停止了大规模测量，只保存了 16 个测量点进行一些示范性的测量.

(5) Surveyor

Surveyor[5] 是由 ANS 发起建设的分布式测量系统，主要测量点到点单向延迟、丢包率和路由信息. Surveyor 在全球布设了约 50 个测量点，主要分布在美国、加拿大、智利、荷兰、新西兰等国. Surveryor 能够精确测量单向延迟，但 Surveyor 项目至今并未公布其测量数据以供公众研究.

(6) RIPE 的 TTM

RIPE 的 TTM[6] (Test Traffic Measurement) 是由 RIPE 发起建设的大规模网络测量项目，主要测量内容包括延迟和路由信息. TTM 测量点使用定制的 Unix 服务器，有 GPS 同步系统，能精确测量单向延迟. TTM 在全球有约 100 个测量点，主要分布在欧洲、美国、以色列等国.

(7) Planet Lab

Planet Lab[7] 本身并非专门的测量系统，而是一个通用的分布式计算平台. 但是由于 Planet Lab 节点数量很多，分布广泛，能够作为大规模分布式测量的支撑平台. PlanetLab 目前有 1086 个节点，分布于 25 个国家，大多数合作伙伴为研究机构.

(8) ScriptRoute

ScriptRoute[8] 是一个依托 PlanetLab 的网络测量系统. ScriptRoute 有超过 200 个测量点，其中绝大多数使用的是 PlanetLab 服务器.

(9) piPEs

piPEs 是 Internet2 上的分布式网络测量项目. piPEs 在全球共有 145 个测量点，所有的测量任务在测量点之间进行. 主要测量包括丢包、带宽和单项

延迟等,也可由用户定制测量任务.

(10) IEPM-BM

IEMP-BM(Internet End-to-end Performance Monitoring-Bandwidth to the World)是由 IEPM 发起建设的全球性能测量项目.其目标是测量 Internet2 到世界其它地区的带宽和性能.IEPM-BM 目前有 10 个测量点,主要分布在欧美等发达国家的高校和研究机构.该项目已经与 perfSONAR 整合.

(11) PerfSONAR

PerfSONAR[9](PERFormance Service-Oriented Network monitoring ARchitecture)是由 ESNET、GEANT2、INTERNET2、RNP 等机构共同建设的全球分布式网络测量项目,在全球范围内发展了大量合作伙伴.

(12) DIMES

DIMES[10]是借助个人电脑资源实现的分布式测量系统.志愿者需下载 1 个客户端程序.DIMES 的客户端通过很低频率的 TRACEROUTE 和 PING 来进行互联网的测绘,最高的带宽占有量是 1KB/SEC.DIMES 目前有 8808 个用户,覆盖全球 119 个国家和 29404 个 AS.用 DIMES 做拓扑发现的效果非常好,这是因为拓扑信息的变化相对较慢.但是在进行性能测量时,会因测量点不固定、测量任务难以计划而表现出一定的局限性.

2.2 相关工作的比较和分析

目前国际上比较知名的大规模网络测量项目主要对网络性能和拓扑进行测量,有以下 3 种形式:

(1) 布设少数测量点,测量其到其它网络的通信性能.如 IEPM-BM 项目.

(2) 布设大量测量点,测量点间互测.比较著名的有 AMP、piPEs 等项目.

(3) 布设大量测量点,对分布在全球的大量目标站点进行测量.如 PingER、SAMI、Skitter 等.

网络测量方面的研究非常活跃而开放,大多数研究机构都愿意把自己的测量工具免费下载和开源.尽管如此,仍然有以下两个挑战制约着网络测量工作的进一步发展.

(1) 测量点发展效率.首先,在全世界不同国家的网络上寻找大量合作伙伴需要长期持续的努力.其次,优质的测量往往对测量点在互联网上的位置有苛刻要求,例如需要与骨干网有直接连接,这使得发展测量点更加困难.

(2) 可持续发展能力.有些测量系统的设备要由发起方购置,不仅需要一次性投入,还需要不断更换和维护.有些测量系统的软件是发起方开发的专用系统,部署后还需要长期维护和升级.因此我们经常能够看到一度知名的测量系统活力日下,甚至逐渐淡出.

目前,DIMES 项目测量点分布最广,但测量点稳定性差,无法确定给定时刻可用的测量点;在大规模分布式性能测量方面,以 Planet Lab 的服务器分布最广,数目最多;而近期新兴的性能测量项目在架构设计方面都较多考虑了与同类项目的协作测量和数据交换.但是,目前互联网上还没有能够覆盖全网的完全分布式测量系统.

3 GPERF 联邦测量架构

3.1 GPERF 设计原则

GPERF(全球网络性能测量系统)的设计目标是实现可持续发展的全球互联网性能测量系统.GPERF 不局限于由自己部署的测量点,它不仅通过协商寻找合作伙伴,同时也重视对互联网上离散的开放资源的利用,形成了独特的基于联邦的合作模式.

GPERF 的基本设计原则可以归纳如下:

(1) 基于联邦模型发展测量点.联邦的含义是自治与合作,即联邦中的资源由各成员自主投入,自行维护,自愿合作.GPERF 联邦的发展有 3 种途径:① 与每个可能的合作伙伴协商;② 加入一个已有的合作群体,以少量资源投入一次性获得大量合作伙伴;③ 开发利用互联网上离散的开放资源.通过这样一种联邦模式,GPERF 得以迅速形成规模,并且不会因为规模的增长而显著增加运行成本,具有良好的可持续发展能力.

(2) 采用成熟测量技术和工具.GPERF 没有规定成员必须使用的软硬件规格型号,也没有开发专门的测量方法和工具.GPERF 倡导使用成熟测量技术和工具,并籍此降低部署和后期维护的成本.事实上 GPERF 将尽量适应合作伙伴所可能采用的技术和工具,通过良好的系统结构设计来封装底层异构的测量点.

3.2 GPERF 体系结构设计

GPERF 的体系结构设计如图 1 所示.GPERF 具有 4 层体系结构设计,分别为数据采集层、资源服务层、任务调度层与表示层.其分层架构可以屏蔽底层实现细节,提高可扩展性和开放性.数据采集层为全球分布的测量点,其中既有 GPERF 项目自建的

测量点，也有合作伙伴支持的测量点，还有互联网上的开放测量点. 资源服务层，对所有测量点进行统一管理，可以屏蔽底层测量点的异构性，通过对不同测量点的测量结果进行格式化，实现向上统一的调用接口. 任务调度层实现测量任务的调度和管理，包括任务分发、数据汇总等功能. 顶层为表示层，实现统计分析和性能评价功能.

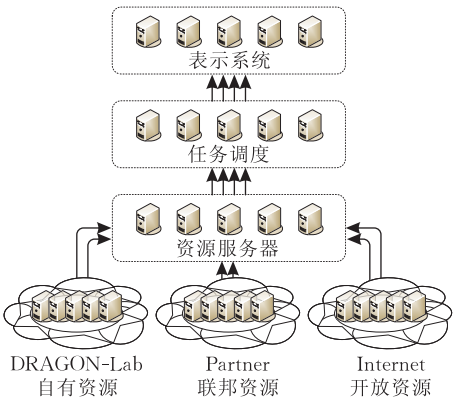

图 1　GPERF 体系结构

3.3　测量任务管理和调度

GPERF 实现了对多种资源的利用，但由于资源的下列特点增加了设计高效任务调度策略的难度. (1) 资源的离散性，表现在资源可以在互联网上广泛分布；(2) 异构性，表现在资源有多种类型，包括自有资源、合作伙伴联邦资源和互联网开放资源，不仅是不同类型资源之间有差异，而且同一类型的资源也有很大的差异；(3) 不稳定性，表现在资源可以选择动态加盟或退出、互联网开放资源可以失效也可以不断增加等方面；(4) 限制性，表现在使用互联网开放资源时，必须考虑到不能给其带来过重的负担.

我们设计 GPERF 的任务调度模型如图 2 所示，调度中心可以根据资源和任务情况作出高效调度. 调度中心可以直接调用底层资源，也可以先把任务分给任务代理，由任务代理完成测量任务，特别是在系统足够庞大的情况下，还能实现调度多层任务代理. 通过这种方式，不仅增加了任务调度的灵活性，而且也能有效降低调度中心的负载.

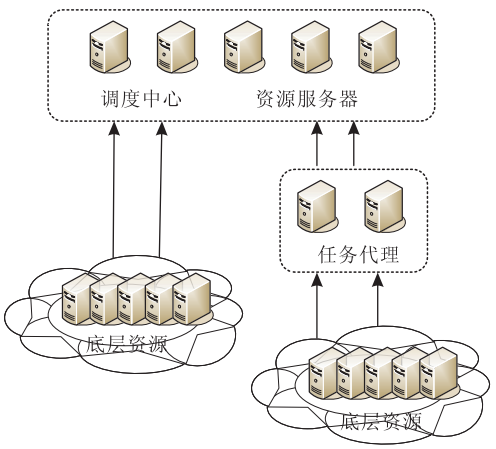

图 2　GPERF 的任务调度模型

那么在上述情况下，系统的调度问题变成一个标准的线性规划问题. 假设 GPERF 的资源节点之间没有交互，一个测量任务可以分解为相互独立的多个子任务. 例如监控欧洲到中国某政府网站的网络性能这样一个任务，可以分解为从欧洲各国分别进行测量. 调度的目标是尽可能地使得每个任务的吞吐率达到最大，即 $f=\max\min_k\left\{\frac{\alpha^k}{w^k}\right\}$（$w^k$ 表示任务 k 的优先级，α^k 表示系统单位时间内处理任务 k 的能力）. 由于问题的复杂性，在实际研究中，并不会去求解该问题的最优解，而是采取一些贪心调度策略，尽可能地去取得好的结果，例如最早结束时间策略（ECT）、最快处理器分配给最重的任务（FPLT）、Round-Robin 策略、动态调整的 FPLT（DFPLT）、结合预测技术的 ECT（ECT-P）、最短任务优先原则等等[11-12]. 对于不同的系统，不同的策略可能会有不同的效果. 对于 GPERF，如何求解该问题，需要详细了解各测量点的网络通信能力、计算处理能力等参数，这将是今后的研究重点.

4　GPERF 的实验系统实现

目前，我们已具有一定的自有资源，但它们仅仅分布在国内. 为了进一步拓展测量点，在系统发展的第一阶段，我们重点将目光投向了互联网上广泛存在的开放资源，主要包括 LookingGlass 资源和 r-DNS 资源. LookingGlass 资源一般是由运营商为了网络诊断而提供的服务器，通常有 ping、traceroute、bgp 路由表查询等功能. r-DNS 资源是提供递归查询功能的 DNS 服务器，可以测量与目标点的双向延迟. 通过这两种免费资源，可以很轻松地实现全球范围的网络测量，这是其它测量项目所达不到的. 下面将具体介绍发展这两类资源的主要方法.

4.1　测量点的发展

4.1.1　LookingGlass 测量资源发展

LookingGlass 资源有 4 个特点：(1) 访问通过 Web 方式进行. (2) 测量点网络位置好，因为 LookingGlass 本身是由网络运营商提供的，所以测量点的部署位置总是靠近主干网或 AS 边界，其测量结

果就比较准确.(3)测量点多分布范围广,这是因为运营商为了能够更充分地了解网络情况.(4)功能丰富,包含多种功能,如 Ping、traceroute、bgp 等.表 1 从多个方面比较了 LookingGlass 测量点与一些知名测量项目测量点分布.

表 1 LookingGlass、r-DNS 资源与知名测量项目测量点的比较

	测量点数	自治系统覆盖数	国家覆盖数	功能	精确性
LookingGlass	1000+	200+	45+	延迟、路由等	好
r-DNS	不计	不计	几乎全部	延迟	差
PingER	48	40−	23	延迟	一般
Skitter	20	20−	20−	延迟、路由	一般
BW	10	10−	10−	带宽、延迟	一般
PlanetLab	1086	不详	20～40	—	—

为了利用好这些资源,需要解决 3 点困难:(1)如何找到在互联网中广泛存在的 LookingGlass 资源;(2)如何将这些资源变成可调度的测量点,因为 LookingGlass 站点提供的网页接口并不适合程序调用;(3)如何定位具体测量的所在地理位置、IP 地址、AS 编号等信息.

对于第(1)点,可以参考由网站[①]整理的 LookingGlass 列表.对于第(2)点,我们首先需要分析 LookingGlass 站点的调用参数,才能进行调用.因为大部分 LookingGlass 站点都是简单通过表单提交查询请求,所以只需对网页源码中的表单部分进行分析就能获得调用参数.我们的 LookingGlass 站点参数分析程序框架如图 3 所示.一些运营商明确提出不允许脚本调用其资源,如著名的 Rogers,对于这些站点,我们将不对其进行收录.最终,程序一共分析出 117 个站点的调用参数,包括 988 个测量点.

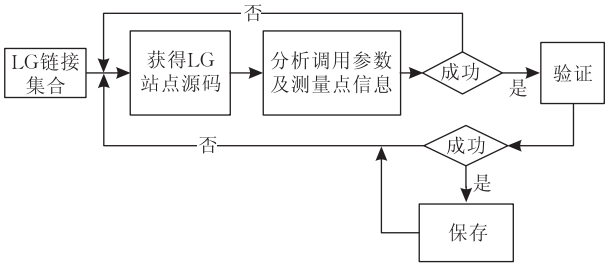

图 3 LookingGlass 站点分析处理流程

对于第(3)点,应用文献[13]的研究成果能查到站点 IP 地址(解析域名得到)对应的 AS 号,同一站点提供的测量点与网站通常都处于同一个自治系统中,如有例外,网站上也会给出说明.要想进一步得到具体测量点的详细信息,如 IP 地址、地理位置等,方法是用该测量点 ping 我们能控制的一台机器,在这台机器上通过抓包的手段就能获得对应测量点的 IP 地址.得到 IP 地址之后,通过网站[②]能查询到 IP 地址对应的具体地理位置.最终,我们确定了 651 个测量点的位置,分布于 109 个不同的自治系统,45 个不同的国家,但是基本都是发达国家和地区.

4.1.2 r-DNS 资源发展

虽然通过 LookingGlass 资源,我们已经找到了上千个测量点,但是这些点过分集中于发达国家和地区,如果想真实地了解全球的网络状况,这显然是不够的.为了进一步扩展测量资源,我们将目光投向在全世界范围广泛存在的 r-DNS 服务器(递归查询 DNS 服务器)上.

文献[14]曾提出估计网络上任意两点之间的双向延迟的方法.它的核心思想在于寻找离需要测量的两点距离最近的两台 DNS 服务器,通过测量这两台 DNS 服务器之间的延迟来近似两目标点之间的延迟,要求是两台 DNS 服务器中至少有一台是提供递归查询服务.基于此,我们可以搜集 r-DNS 服务器作为我们的测量点.

如图 4 所示,如果得出 R 与 A 之间的延迟,可以近似为 R 与 T 之间的延迟,因为 T 与其本地 DNS 服务器往往非常接近.测量发起点首先根据 T 的域名地址构造出一个并不存在的域名,如图中的 xyz.foo.cn,向 R 发送请求解析该地址,则 R 必向 A 提交递归查询请求,R 最终会将 A 返回的结果转发给 S.这样就测得了包通过路径 S-R-A-R-S 所需要的时间 $T1$.对于只知道 T 的 IP 地址(a.b.c.d)的情况,我们需要在 T 的 IN-ADDR.APAR 域名地址前增加一个随机数即可,即向 R 查询 randnumber.d.c.b.a.in-addr.org 即可.另外,很容易测得 S 与 R 之间往返时间 $T2$.这样 $T1-T2$ 便是 R 与 A 之间的双向延迟.

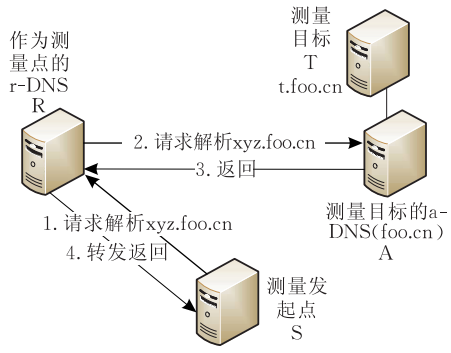

图 4 r-DNS 测量点演示(参考文献[14])

① http://www.traceroute.org
② http://www.ip2location.com/demo.aspx

当然这样得出的延迟只是一个估计值. 在原文中，作者以 Traceroute 得到的延迟数据作为标准对该方法的准确性进行了评估，得出 2/3 的 DNS 测量结果误差在 10%以内，3/4 的误差在 20%以内的结论. 我们以 LookingGlass 的 ping 结果作为标准，与 DNS 测得的数据进行比较，得到的结果与原文中近似，如图 5 所示，60%的 DNS 测量结果误差在 10%以内，78%的 DNS 测量结果误差在 20%以内.

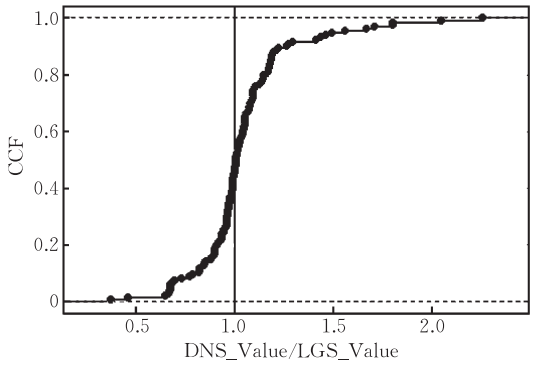

图 5　r-DNS 测量结果与 LG 测量结果比较

尽管有误差，但是它的优点依然明显，即无处不在的测量点，只要具有递归查询的 DNS 服务器都能成为测量点. 而据文献[14]统计，互联网上 75%的 DNS 服务器具有递归查询功能. 正因此，我们决定发展 DNS 测量点，用于某些特定的测量任务.

为了在尽可能多的国家和地区收集 DNS 测量点，我们的方法是通过网站①的功能查询某一国家对应的 IP 地址段，并选择其中一个 IP，利用 in-addr.arpa 查询该 IP 的 a-DNS，验证该 DNS 的有效性之后，进一步确认该 DNS 的地理位置和所属自治系统(方法同 LookingGlass). 最终，我们收集到 476 个 DNS 服务器，跨越 156 个国家，240 个 AS.

4.2　任务调度的高效性与礼貌性的折中

在 2.3 节中，GPERF 的任务调度系统是基于任务代理的方式实现的. 在实验系统的实现中，我们用两个进程分别模拟了所有关于 LookingGlass 和 DNS 资源任务的代理，不过没有采用多层次的代理形式. 为了达到 2.3 节中提到的性能目标，我们必须对测量点的通信能力、计算能力作进一步的测定. 然而在测试这些数据的时候，一些网络管理员抱怨我们过度频繁地使用这些资源.

因此，我们提出了资源访问的礼貌性这个概念，参考了网络爬虫对网页进行更新访问的礼貌性概念，也就是要有节制地使用测量资源，特别是互联网开放资源. 显然，礼貌与高效是相互矛盾的，在我们的实验系统调度中，只能对这两者进行折中. 在一位网络管理员的回应中提到，他们的 Looking Glass 服务器能够忍受的最高使用频率为每 2min 一次[15]. 考虑到不同站点之间可能存在的差异，在实验系统中我们将互联网资源的使用间隔统一提高到 4min. 这样降低了系统的效率，但也降低了系统关于互联网资源的调度复杂性. 期待将来有更多的伙伴资源加入，届时我们将更一步研究任务调度策略以提高系统性能. 在现阶段，我们通过搜集更多的互联网开放测量点，来弥补因为礼貌性而丧失的效率.

5　实验系统的应用：奥运和世博相关网站的全球访问性能测度

2008 奥运会及 2010 世博会都受到全球瞩目，全世界的用户能否快捷访问相关网站是一个值得关注的问题. 基于 GPERF 实验系统，我们已经完成了奥运相关重要网站的全球访问性能监控，对世界各地的用户体验进行测量. 对于世博相关网站的监控正在进行. 下面以奥运监控结果为例.

图 6 是目标站点的全球访问性能排序和按地理位置的性能分布图. 其中，对访问性能较好的国家和地区用浅色表示，随着访问性能的下降，颜色逐渐变深，对于无法连通，或者性能很差的国家用深色表示.

Ranking Today

Rank	Hostname	Score
1	www.google.com	92.566
2	www.olympic.org	92.406
3	fr.beijing2008.cn	88.0579
4	en.beijing2008.cn	87.9783
5	www.beijing2008.cn	87.6868
6	esi.volunteers.beijing2008.cn	69.6441
7	2008.baidu.com	66.1864
8	www.2008eshop.cn	65.7207
9	2008.163.com	65.2165
10	2008.qq.com	64.32
11	www.olympic.cn	63.7587
12	2008.sina.com.cn	60.787
13	2008.sohu.com	49.2292
14	www.tickets.beijing2008.cn	35.2708

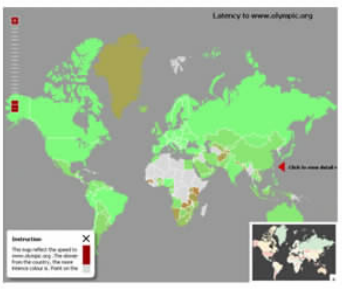

图 6　相关网站全球访问性能排名及按地理位置性能分布图(右图目标站点为 www.olympic.org)

① http://ip.wisa.com.cn/worldip.php

另外,在监控过程中,我们也发现了一些网络特征,例如访问延迟中值的稳定性、访问性能随时间的稳定性、访问性能较差网络分布的稳定程度等.

5.1 使用均值和中值评估性能的对比

在计算监控站点的全球访问性能评分时,我们使用两个参数:平均值和中值.平均值指的是全球各个测量点的平均访问延迟;中值指的是各个测量点访问延迟的中值.实际上,我们发现平均值受到外界的影响(例如由于网络异常导致延迟迅速增加)非常大,因此即使在一小时内平均值的波动也非常大.而中值则相对稳定,按统计学的观点,中值不受异常值影响.图 7、图 8 分别用平均值和中值作为指标比较了两个网站 www.beijing2008.cn 和 2008.qq.com 的全球访问性能.由图可见,平均值的波动性确实要比中值大得多.从这个意义上来说,中值是一种更好的评估参数.

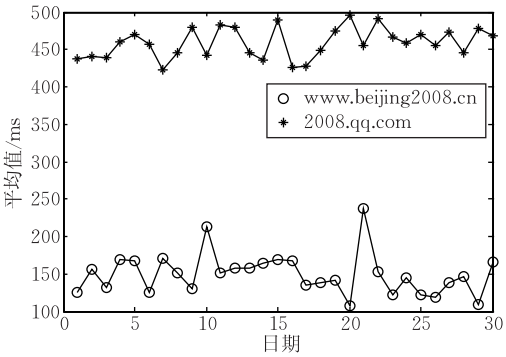

图 7 每日平均值曲线

(数据采集时间:2008.4.1～2008.4.30)

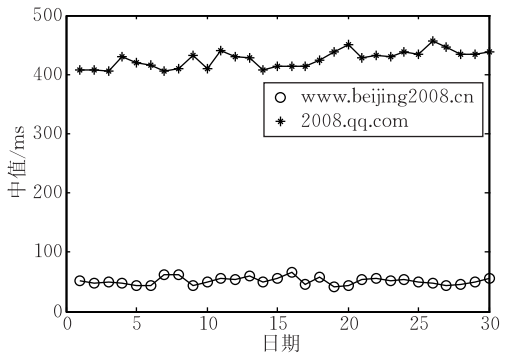

图 8 每日中值曲线

(数据采集时间:2008.4.1～2008.4.30)

5.2 网络性能的时间稳定性

为了解站点的访问性能随时间的变化情况,我们测量了一天内各个小时段相应站点的平均值和中值变化,如图 9～图 11 所示.由图可见,在一天的各个小时之间,不管是平均值还是中值,都相差不大.也就是说,一天内时间的差异对于测量结果影响不大.这是由于测量点分布在全球各地,处于各个不同的时区.虽然当中国进入白天之后,由于网络的繁忙,会导致延迟的增长,但与此同时,美国加拿大等地则进入黑夜,网络变得相对空闲,从而导致延迟的降低.因此,从全球范围看,网络性能随时间波动相对较小.

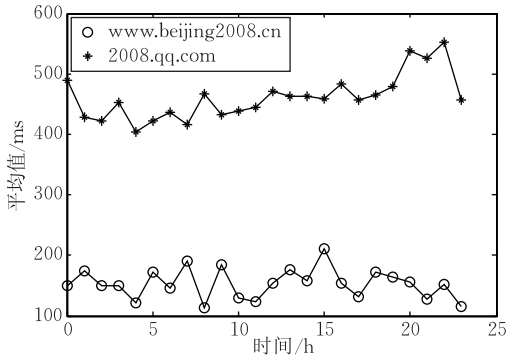

图 9 每小时平均值

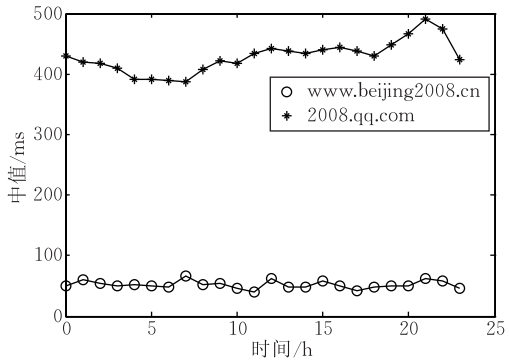

图 10 每小时中值

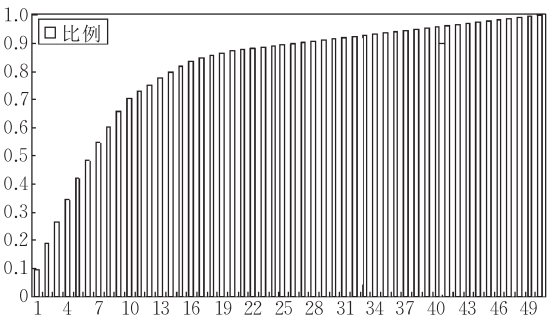

图 11 2008-08-11～2008-08-17 Top10 最差网络所占比例图

5.3 访问性能较差的网络分布的稳定性

在 2008-08-11～2008-08-17 这 7 天中,我们每天选出 10 个到测量目标表现最差的 AS.当然,一些 AS 会每天重复当选.最终,共有 50 个 AS 曾经出现在表现最差的 TOP10 AS 中.如图 11 所示,我们先将 50 个 AS 按照出现次数排序,出现次数多的排在前面(7 天最多出现 7 次),然后统计前 X 个 AS 占总出现次数的百分比.由图可见,在这 50 个 AS 中,出现次数相对集中,20% 的 AS 占了 70% 的出现次

数.这说明对于设定的各目标站点来说,表现最差的AS相对来说比较稳定,这些AS或者是由于接入带宽低,或者是由于用户过多而拥塞,从而导致其到其它站点的速度都很慢.

6 结　论

通过分析网络测量领域中主要存在的问题,我们发现最大的困难在于测量点的覆盖率不足.基于此,我们设计并实现了GPERF联邦测量平台,通过采用联邦架构,我们以较小的投入整合了大量资源,通过开发互联网开放资源以及与其它测量项目的合作,GPERF完全有能力解决困扰着网络测量研究者的测量点覆盖率难题.目前,在全球范围内已发展的测量点超过一千,覆盖150多个国家和200多个AS,成功地对2008奥运重要网站进行了全球访问性能监控,并正在对世博相关网站进行监控.当然,目前的GPERF的实验系统还远远不够,为了继续改进GPERF,我们将在主干测量点的部署及调度算法优化方面重点开展研究,同时希望更多有兴趣的学者能参与进来.

致　谢　感谢清华大学计算机系网络所博士孟坤在此期间提出的建议,与您的讨论让我们受益匪浅!

参　考　文　献

[1] Matthews Warren, Cottrell Les. The PingER project: Active Internet performance monitoring for the HENP community. IEEE Communications Magazine, 2000, 38(5): 130-136

[2] Paxson V, Adams A, Mathis M. Experiences with NIMI// Proceedings of the SAINT'02. Nara, Japan, 2002: 108-118

[3] Paxson V, Mahdavi J, Adams A, Mathis M. An architecture for large-scale Internet measurement. IEEE Communications Magazine, 1998, 36(8): 48-54

[4] McGregor A J. NLANR's active measurement program: Network knowledge leads to practical payoffs. SDSC/NPACI Online Biweekly Newsletter, 2003, 7(3)

[5] Kalidindi Sunil, Zekauskas Matthew J. Surveyor: An infrastructure for Internet performance measurements//Proceedings of the INET99. San Jose, California, 1999

[6] Georgatos Fotis, Gruber Florian, Karrenberg Daniel et al. Providing active measurements as a regular service for ISPs// Proceedings of the PAM 2001. Amsterdam, 2001

[7] Chun Brent, Culler David, Roscoe Timothy et al. PlanetLab: An overlay testbed for broad-coverage services. ACM SIGCOMM Computer Communications Review, 2003, 33(3): 3-12

[8] Spring Neil, Wetherall David, Anderson Tom. Scriptroute: A public Internet measurement facility//Proceedings of the USENIX Symposium on Internet Technologies and Systems (USITS). Seattle, WA, 2003: 17

[9] Hanemann A, Boote J W, Boyd E L, Durand J, Kudarimoti L, Lapacz R, Swany D M, Zurawski J, Trocha S. PerfSONAR: A service oriented architecture for multi-domain network monitoring//Proceedings of the 3rd International Conference on Service Oriented Computing. Amsterdam, The Netherlands, 2005: 241-254

[10] Carmi Shai, Havlin Shlomo, Kirkpatrick Scott et al. A model of Internet topology using k-shell decomposition. Proceedings of the National Academy of Sciences, 2007, 104(27): 11150-11154

[11] Beaumont O et al. Centralized versus distributed schedulers for bag-of-tasks applications. IEEE Transactions on Parallel and Distributed Systems, 2008, 19(5): 698-709

[12] Iosup A et al. The performance of bags-of-tasks in large-scale distributed systems//Proceedings of the 17th International Symposium on High Performance Distributed Computing. New York, USA, 2008: 97-108

[13] Chang H, Jamin S, Willinger W. Inferring AS-level Internet topology from router-level path traces//Proceedings of the SPIE ITCom 2001. 2001, 8: 19-24

[14] Gummadi K, Saroiu S, Gribble S. King: Estimating latency between arbitrary Internet end hosts//Proceedings of the SIGCOMM Internet Measurement Workshop (IMW 2002). Marseille, France, 2002: 5-18

[15] Spring N, Mahajan R, Wetherall D, Anderson T. Measuring ISP topologies with rocketfuel//Proceedings of the ACM SIGCOMM Conference. Pittsburgh, PA, 2002: 133-146

WANG Ji-Long, born in 1973, Ph. D., associate professor. His research interests include network management and network measurement.

SUN Ming-Min, born in 1985, M. S. candidate. His research interests focus on network measurement.

ZHANG Qian-Li, Ph. D., assistant researcher. His research interests include network security and traffic analysis.

Background

Network measurement has been an important subject in network research community. But it is always not valuable to do the measurement at a small scale because network behavior is different in time and space, unpredictable, and not reproducible. The network always changes a lot according to the position, time and other factors. Like the flowing river, we cannot measure the same network twice.

So how do we carry out the network measurement study? On the one hand, we should develop measurement tools to do the specific measurements; On the other hand, we should build large distributed system to support the measurements. However, it is so expensive to build the infrastructure and keep it running that no one is able to build a large enough platform alone for network measurement. To overcome the difficulty, the authors propose this global network performance platform structure called GPERF based on federation policy. The authors didn't try to invest a lot of money to build a large distributed system. They just want to unite everyone's resource to achieve this goal. In GPERF, the infrastructure resource may belong to us or our partners. And they also integrate large number of open network measurement resources including Looking Glass resource, traceroute server and so on.

This work was supported by the National Basic Research Program (973 Program) of China under grant No. 2009CB320505, National Natural Science Foundation of China under grant No. 60973144 and National High Technology Research and Development Program (863 Program) of China under grant No. 2008AA01A303.