强耦合代理模型下的云安全分析算法*

左利云,陈一明
(茂名学院 实验教学部,广东 茂名 525000)

摘 要:针对云计算与终端应用程序的互动瓶颈,提出了一种强耦合网络代理模型,它可以加速云计算与应用程序的交流。还提出了一种云安全分析算法,该算法根据统计学原理和预先登记的网络异常条件,设定其危害指数,对所有经过的网络数据进行过滤,对不安全信息进行识别报警。模拟设置了云计算网络环境,仿真实现了强耦合网络代理模型和云安全分析算法,实验表明,该模型可明显降低云计算与终端应用程序的响应延迟时间,云安全算法识别率高达99.37%,误报率低至0.979%。

关键词:云计算;强耦合;代理网络;危害指数

中图分类号:TP392 **文献标识码**:A **文章编号**:1674-7720(2011)13-0059-04

Cloud computing analysis algorithm based on the model of strong-coupling proxy

Zuo Liyun, Chen Yiming
(Experiment Teaching Center, Maoming College, Maoming 525000, China)

Abstract: Cloud computing for the interaction with the terminal application bottlenecks in the network, this paper presents a strong coupling proxy model that can accelerate cloud computing and communication applications. It also proposes a cloud security content analysis algorithms, which is based on statistical theory and pre-registration abnormal condition of the network, set its hazard index, filter all the network data passing by, identify and alarm the unsafe information. Simulation set up cloud computing network environment, the simulation achieved a strong-coupling proxy model and cloud security network analysis algorithms, experiments show that cloud computing model can be significantly reduced with the terminal application's response delay time, cloud security algorithms recognition rate is as high as 99.37%, false been reported is as low as 0.979%.

Key words: cloud computing; strong-coupling; proxy network; hazard index

随着并行计算、分布式计算和网格计算的发展,一种新型的计算方式——云计算(Cloud Computing)成为IT业内讨论的焦点。云计算指的是Internet上作为服务提供的应用以及部署在数据中心的提供这些服务的软件和硬件。这些服务被称为SaaS(Software as Service),软件即服务。数据中心的软件及硬件就是所谓的"云"。

Sun、IBM、微软、Google、Amazon等信息业巨头都已经参与到云计算的研究和开发中。Sun公司在2006年推出了基于云计算理论的"黑盒子"计划,已进入发售阶段[1];IBM推出的"蓝云"计划[2-3];较为显著的是Google公司专门针对Web应用而设计的AppEngine[4]。同时,学术界也纷纷对云计算进行深层次的研究。例如谷歌与华盛顿大学和清华大学合作,启动云计算学术合作计划,推动云计算的普及,加紧对云计算的研究。我国的计算机研究人员远在"云计算"这个名词提出之前就有透明计算[5-6]的构思。透明计算体现了云计算的特征,即资源池动态的构建、虚拟化、用户透明等。较近的研究有探讨云计算理论及其关键技术等[7-8]。

云计算中的许多应用需即时互动响应终端应用程序的上传、下载数据请求,这会导致严重的性能瓶颈,而且云计算的安全问题亦十分严峻。本文提出一种强耦合网络代理模型以解决云计算固有的互动模式瓶颈,同时提出一种新的基于强耦合网络代理模型的云安全分析算法。

*基金项目:广东省科技计划项目(2007B010400042);广东省自然科学基金(06029274);茂名市科技计划项目(20091009);茂名学院基金项目(203492)

1 强耦合网络代理模型

Google 的实践表明，用廉价服务器组成的大规模集群，在可靠性、稳定性和计算能力上，均能达到大型计算机的标准。也就是说，对于 Google 每天需要处理的海量数据和复杂计算，在保证系统延展性和良好运行效率的基础上，都可以通过架构在廉价集群之上的云计算平台得以实现。在这个由近百万台廉价服务器组成的集群中，就单台机器而言，性能并不出众，但由它们构建起来的整个网络所能提供的数据存储能力和计算能力却大得惊人[9-10]。

然而，这样的云计算平台在其面向终端应用程序的互动中所产生的瓶颈会严重影响整个系统的性能。尤其是高峰时段更会给整个系统带来严重的危机。为解决此问题，本文提出了一种强耦合网络代理模型。

1.1 代理网络模型

代理网络模型是基于云计算的服务（泛指大规模的数据、计算或服务资源，提供给最终用户和应用程序），通过强耦合代理网络连接设置，有三个重要组成部分：云服务、代理网络及应用程序启动器，其模型如图 1 所示。例如云服务 S_1 可能是一个大型国际互联网的服务（如 Google 地图等），云服务 S_2 则可能是提供的数据依赖计算机的计算资源（如亚马逊 EC2[11]），云服务 S_n 可能是提供专业化的服务，或被其他应用程序使用的服务。节点 A、B、C、D、E、F、G 是代理节点，而 P 是一个使用云服务的应用程序启动器。

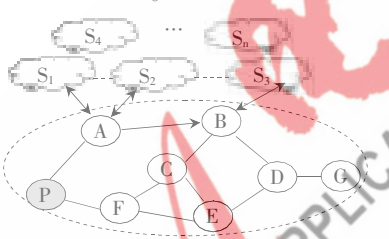

图 1 强耦合网络代理模型

代理网络是由一组逻辑节点相连而成，图 1 中的实线代表代理节点之间及代理与云服务之间的相互作用，虚线表示云服务之间的相互逻辑作用。此代理模型可以提高分布式密集型数据应用程序的性能和可靠性。其表现如下：(1)与云服务的互动。代理可作为客户端访问云服务，这可以使代理以更好的网络连接去访问一个或多个云服务。如一个代理节点能以比客户端高得多的带宽访问云服务。(2)计算。代理可以通过数据处理器对数据进行过滤、压缩、合并、采集和转换等。(3)缓存。代理可以高效存储传送数据到其他节点，也可以缓存可能被云服务重用的即时数据。(4)路由。代理可以作为应用程序流程的一部分传递数据至另一个代理。这很重要，因为应用程序与多个广泛分布的云之间的相互作用需要统一有效的管理安排。而代理的真正优势在于这些角色的联合作用。如一个自瘦客户网络连接到 PDA 上，可利用高带宽的代理与云服务的相互作用获取大量输出数据，并利用其强大的计算能力来处理数据。

应用程序启动器（图 1 中的 P）是代理网络中的一个节点，可代表应用程序，可以是终端用户机，或工作调度等。它决定着应用程序的控制点和最终的资源分配。值得注意的是，终端用户不一定一直作为应用程序启动器，也可以将代理网络中的其他控制点作为启动器。代理网络中也可以有很多应用程序启动节点，并且一个节点此时作为启动节点，在另一时刻也可能会作为普通代理节点。

1.2 强耦合的含义

代理网络模型可加速提高其上应用程序的性能。模型中的每个实体都有其特殊的作用。其中，应用程序启动器最适合保存私有数据，是用于执行应用程序的逻辑控制点；代理网络的特别优势在于，在不同的网络位置能提供最好的网络资源；云服务是提供数据支持和保持资源共享的强大性能保证。而这三者在整个模型中以强耦合方式联系，故强耦合在此的含义是指一个处理单元的输出会受另一个单元的影响。分布式应用程序即是应用程序启动器、代理网络和云服务这三个实体的耦合。如图 1 中应用程序调用两个云服务（S_1 和 S_2），经由数据处理器（如下面的分析过滤器等）实现过滤、压缩、合并、采集和转换等，然后用它们的中间数据作为输入到 S_3 产生终端数据输出。代理 A 和 B 用来加快这一过程。A 平行调用 S_1、S_2 处理输出，并发送数据给调用 S_3 的代理 B。而代理 A 和 B 是由 P 根据它们到 S_1、S_2 和 S_3 的网络性能而选择的。

现有云服务的互动模式决定了如果终端用户应用程序资源受限，瓶颈问题将更加严重。而该模型可提高云计算中分布式数据密集型应用的性能和可靠性：能更有效地实现云服务的互动，如可以使云服务到客户端拥有更高的带宽；代理网络还兼有缓存、路由和强大的计算功能。以上这些保证了云服务与终端应用程序的交流互动高速而有效。

2 云安全分析算法

2.1 云安全分析处理器

强耦合网络代理模型通过利用低延迟和高带宽连接提供了可扩展资源，有利于云服务与终端用户的互动，提高云服务与终端应用程序的交流速度（这些将通过后面的实验予以验证），同时也不免令人担心置入这样一个代理网络，是否会给本就不令人放心的云计算的数据带来更大的安全隐患？

云服务与终端用户互动的加强、交流速度的加快会给云计算的安全带来很大压力。云计算是以数据为中心的，这意味着数据被转移到了用户不可控制的范围，如何确保这些数据的安全（包括访问控制、入侵防御、反异常部署、防止内部数据泄密和网络内容与行为监控等）

是急需解决的问题(曾有2010年是云计算安全年的说法[12-13])。为此,本文尝试在图1的云服务及强耦合网络代理之间加入一个如图2所示的云安全分析器。期望给云计算的数据安全带来一定保证,以使得代理网络在带来高性能的同时也同样具有安全性。

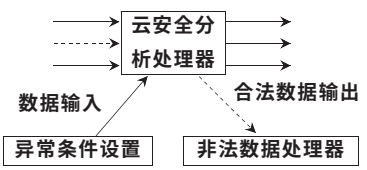

图2 云安全分析处理器

事先需要把大量的非法的需过滤的条件登记在系统中,云安全分析处理器预先处理这些异常条件登记,当源源不断的新数据到达时,处理器即对其进行实时处理。如果满足先前定义的异常条件,则把这个新数据送给非法结果处理器;如果不满足,则将合法的新数据输出。

2.2 异常条件登记

在上面的安全分析器中,异常条件的设定很关键。本分析器的理论基础是根据统计学原理,认为存在这样的关键字集,即在异常数据和正常数据样本中的出现频率有明显差异。关键字集在异常数据样本中的出现频率很高,而在正常数据样本中的出现频率相对很低,具有明显差异。而且出现较多的关键词大多都具有普遍意义,是异常数据行为不可缺少和不可替代的,也就是说,只要准确获取了文本分析的关键字,异常文本的制造者就无法通过替代或者其他方式使异常躲过检测。

异常条件是依据数据中关键字的危害指数而选取的。通过大量的实验积累,根据关键字本身的破坏能力和关键字在异常数据样本中出现的频率给出危害指数的公式:

$$H=\sum_{i=0}^{i_{max}} D(i)W(i)B(i) \tag{1}$$

其中,H表示总的危害指数。$D(i)$表示狄拉克δ分布函数,当$D(i)=0$时,表示未与第i个关键字匹配;$D(i)=1$时,表示匹配。$W(i)$表示关键字与别的文本结合的行为产生的危害指数,简称行为危害系数,由其在异常数据样本和正常数据样本中的出现频率差决定。当选取多个关键字时,实际上这个频率差异的大小标志着它的可信性,不同的关键字出现一次所产生的危害指数与其出现多次的危害指数应当是不同的,根据大量的实验,总结出选择逐次折半的方法描述这个差别,故$W(i)$表示如下:

$$W(i)=\begin{cases}0 & n<1\\ w(i) & n=1\\ w(i)(1+2^{-1}+2^{-2}+\cdots+2^{-n})=2w(i)(1-2^{1-n}) & n>1\end{cases} \tag{2}$$

其中,$w(i)$为关键字自身原始具有的行为危害指数,会因与其他关键字结合而影响;n为匹配次数。H的收敛区域为$2\times w(i)$,即任意一个关键字在检测时所带来的行为危害指数不会小于其自身原始的行为危害指数,同时也不会大于其本身原始的行为危害指数的两倍。

$B(i)$为关键字本身先天所具有的危害指数,不会受与其他关键字结合的影响:

$$B(i)=\begin{cases}9(\text{可用于创建异常对象的关键字})\\ 8(\text{本身不具有危害的关键字})\\ 10(\text{可能具有危害的关键字})\end{cases} \tag{3}$$

2.3 云安全分析算法实现及分析

算法描述如下:已知X个正常数据样本,Y个异常数据样本,通过对所有数据样本的有效关键字进行统计学习,从而得到这些关键字的统计数值,找出在正常数据样本与异常数据样本中出现频率差异较大的i个关键字,记录其频率和危害指数等属性的一般规律后,对未知的待检测数据样本,统计这i个关键字及其属性,将统计结果与事先得到的基于样本的统计学习结果进行对比,由此来判别待测数据样本是否为异常数据样本。

假设有x个正常样本,y个异常样本。$w_x(i)$表示第i个关键字在正常数据样本中的出现次数,$w_y(i)$表示第i个关键字在异常数据样本中的出现次数,$w(i)$为第i个关键字的行为危害指数,$B(i)$为第i个关键字本身具有的危害指数,H为检验中的危害指数总和,ρ为阈值调整参数,t为中间变量。算法实现过程如图3所示。

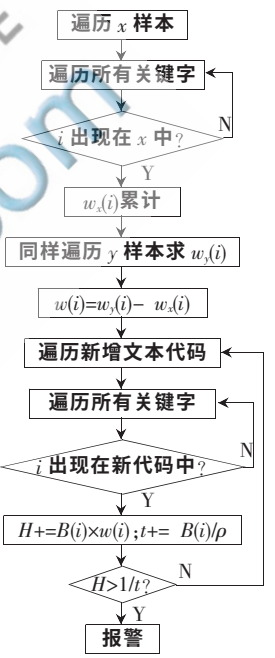

图3 云安全分析算法

而算法中的报警阈值和匹配的关键字个数与匹配的危害组有关。考虑到整体选定的关键字个数和其所在的组可以得到它的归一化系数:

$$\lambda=\frac{\rho}{\sum_{i=0}^{i_{max}} B(i)} \tag{4}$$

其中ρ为阈值调整系数,用它来确定正常数据样本与异常的边界范围。

本算法是基于不同的关键字在正常数据样本和异常数据样本中出现的频率差异,通过计算这些关键字在待测数据样本中的出现频率及其他统计信息,判断待测数据样本是否为异常数据样本,反映了异常数据样本与正常数据样本在关键字出现的频率和行为特性上的特征。该特征不仅存在于已知的异常数据样本中,也存在于未知数据样本中。而且该算法也可单独用于网络数据安全监控,只是由于其统计信息的特点,更适合具有大量密集型分布应用程序的云计算系统,只有在云计算中方更能显示其优势。

同类的基于统计学的关于数据安全方向研究有RABEK J C.等对可执行程序中的系统调用进行统计分

析，来检测动态生成的变种恶意代码[14]；BHATTACHARYYA M 等则针对通过 Email 传播的恶意代码构建了一个实验系统，其主要思想也是统计分析[15]；基于统计的理论在入侵检测中也有过一些应用[16]。但它们的初始应用均是普通网络，适用于云计算的报道较少见，而国内同类相关研究成果尚不多见。

3 仿真实验及结果分析

为验证强耦合代理模型及云安全分析算法的性能，本文利用云仿真软件 CloudSim 设计了仿真实验系统，其步骤如图 4 所示。

图 4 云计算仿真实验过程

首先评估网络差异(如延迟、带宽等)在多大程度上影响代理节点与现有云服务的部署，这将决定着云服务与客户端之间的代理网络模型是否有价值。

在具体操作时，先广泛收集 25 个不同商业网络服务加入云任务列表，每次代理网络通过检测获取一个非常小的文件的时间来记录每一个服务的响应延迟时间。随机抽样 6 个节点，并且每个节点取 5 次实验的平均值，作出其最高响应时间、最低响应时间及平均响应时间曲线；针对同样的 25 个云服务，获取同样的文件来记录未使用代理网络响应的延迟时间，作出响应时间曲线，如图 5 所示。

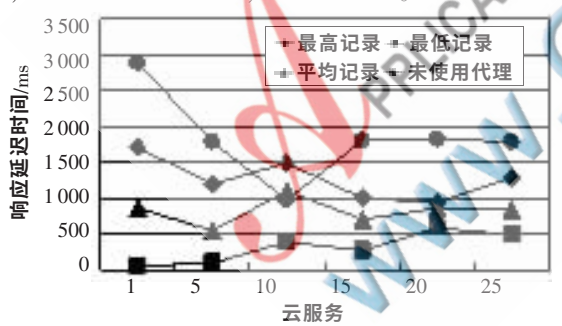

图 5 云服务与应用程序响应延迟时间

实验结果表明，不管网络核心带宽多么丰富，网络节点仍会随不同的云服务而显示出差异，而具有较好连接的代理节点其响应时间有更大差异，但其响应时间的均值相对其他节点偏低。同时，使用强耦合代理网络比不使用代理网络有明显较低的响应延迟时间。

通过观察关键字个数、识别率、误报率和阈值等参数来考察云安全分析算法的性能。实验选取了 40 个病毒样本作为异常样本，这些样本经诺顿和 360 等病毒识别软件识别，并经实际的病毒执行测试确定。正常样本同样采用之前 25 个商业网站的客户端脚本，同样经病毒识别软件识别和实际执行测试。从以上异常、正常样本中选取可靠特征作为关键字。

识别率通过以下公式获得：

$$识别率=1-漏报率=1-\frac{漏报样本数}{异常样本总数} \quad (5)$$

误报率情况与此类似，不再赘述。

而参数中最能影响算法性能的是关键字个数和报警阈值 ρ。因为关键字越多，每一匹配到的关键字的权重就越小，造成匹配率增高，同时增加了冗余信息，并使重要的权重信息无法凸显，会造成一定漏报；相反，较少关键字可能会丢失一些关键字信息，会造成一定漏报和误报。为此，经反复实验验证得出图 6 中不同关键字下的漏报率和误报率。

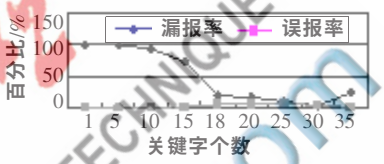

图 6 不同关键字下漏报率和误报率

由图 6 可以看出，在关键字个数为 30 时，漏报率和误报率都相对较小，因此选择 30 作为算法最终关键字个数。

在选定不同关键字及其危险指数条件下，阈值的改变无疑也会对检测结果造成相当大的影响。图 7 为不同阈值下漏报率和误报率的检测情况。

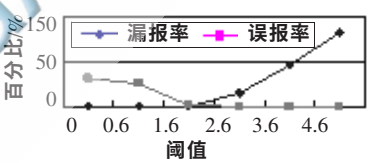

图 7 不同阈值下漏报率和误报率

由图 7 分析得出，在随机选取的测试样本中，正常脚本的误报集中在阈值的 0.6 倍和 1.4 倍左右。特别是当阈值取 0 的时候，得到了 32%的误报率，这说明抽取的大多数样本中并不存在这些关键字。另外，算法中的阈值最终取 1.6 倍，因为此时误报与漏报都处于最小，此时漏报率为 1.63%，即识别率高达 98.37%。但在某些网络环境中，为提高可信指数可略微增大阈值系数。

云计算日趋发展壮大，但其中存在的问题亦不容忽视，在云服务与终端应用程序的互动中所产生的瓶颈会严重影响整个系统的性能，尤其是高峰时段更会给整个系统带来严重的危机。本文提出的强耦合网络代理模型二者之间做了很好的缓存、调度、计算、路由等工作，仿真实验证明了模型的多功能多角色的作用大大降低了网络响应的延迟时间。而基于统计学原理的云安全分析算法根据正常数据中文本代码的危害指数与异常代码出现的频率，从而过滤不安全信息，它不但能对经过数

据进行检测，而且可以根据统计规律对未来数据中可能的出现的新的异常进行识别。实验对异常数据的识别率高达98.37%，而误报率虽然随报警阈值系数ρ增大而增加，但均在合理可接受范围内。该算法的缺点是主要针对网络内容监控，如病毒入侵等，而对于访问控制及防止内部数据泄密的效果不明显，这有待进一步的研究。

参考文献

[1] Sun 的黑盒子—移动数据中心.http://www.biia.org.cn/a/jishushebei/20091222/821.html，2007.

[2] SIMS K. IBM introduces ready-to-use cloud computing collaboration services get clients started with cloud computing. http://www-03.ibm.com/press/us/en/pressrelease/22613.wss，2007.

[3] IBM. IBM virtualization. http://www.ibm.com/virtualization，2009.

[4] 蔡键，王树梅.基于Google的云计算实例分析[J].网络通讯及安全，2009(9)：7093-7095.

[5] Zhang Yaoxue，Zhou Yuezhi. 4VP+：a novel meta OS approach for streaming programs in ubiquitous computing[C]. Proceedings of IEEE the 21st Int'l Conference. on Advanced Information Networking and Applications (AINA 2007)，Los Alamitos：IEEE Computer Society，2007：394-403.

[6] Zhang Yaoxue，Zhou Yuezhi. Transparent computing：a new paradigm for pervasive computing [C]. Proceedings of the 3rd Int'l Conference. on Ubiquitous Intelligence and Computing (UIC 2006).Berlin，Heidelberg：Springer-Verlag，2006，4159：1-11.

[7] 陈全，邓倩妮.云计算及其关键技术[J].计算机应用，2009(9)：2562-2567.

[8] 陈涛.云计算理论及技术研究[J].重庆交通大学学报(社科版)，2009(8)：104-106.

[9] 孙健，贾晓菁.Google云计算平台的技术架构及对其成本的影响研究[J].电信科学，2010(1)：38-43.

[10] 陈康，郑纬民.云计算：系统实例与研究现状[J].软件学报，2009(5)：1338-1347.

[11] Ec2/s3. http://aws.amazon.com.

[12] 陈怡桦.2010年云端安全年.http://domynews.blog.ithome.com.tw/post/1252/67523，2010.

[13] 乐天.云计算还须迈过安全关[J].计算机世界，2008(7)：39.

[14] RABEK J C，KHAZAN R I，LEWANDOWSKI S M，et al. Detection of injected，dynamically generated and obfuscated malieiouso code [J]. Proceedings of the 2003 ACM Workshop on Rapid Malcode，Washington，DC，USA. 2003：76-82.

[15] BHATTAEHARYYA M. An experimental system for malieious email tracking. http://www.linkedin.com/pub/manasi bhattacharyya/4/968/A60，2002.

[16] WAENER D，DEAN R. Intrusion detection via static analysis [C]. Proceedings of the IEEE Symposiums on Security and Privacy，2001：156-168.

（收稿日期：2011-03-11）

作者简介：

左利云，女，1980年生，硕士，讲师，主要研究方向：云计算，网络数据库应用。

陈一明，男，1964年生，硕士，副教授，主要研究方向：云计算，FreeBSD应用。