计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.200400024

基于区块链的日志安全存储方法研究

刘　静[1]　黄　菊[1]　赖英旭[1]　秦　华[1]　曾　伟[2]
1 北京工业大学信息学部　北京 100124
2 中国联通北京市分公司网络优化中心　北京 101109
(jingliu@bjut.edu.cn)

摘　要　随着计算机科学的高速发展,告警日志的数量呈几何的增长趋势,告警日志记录着攻击行为的相关信息,容易受到数据窃取和恶意篡改,同时告警日志中包含大量的无关告警,导致日志分析的准确性不高。为解决告警日志的安全存储和数据提取两方面的问题,文中提出了一种基于区块链的日志安全存储方法,使用基于区块链的分布式存储架构保存告警日志,采用查询区块索引库的方式代替传统的区块链顺序检索,提高了告警日志的检索速度。通过对攻击源地址的威胁评估,构建密文索引结构存储在区块头中,并根据告警日志之间的相关性分析,实现攻击场景告警日志的关联检索。由实验结果可知,使用基于区块链的日志安全存储方法存储告警日志,存储过程中的区块生成效率并不会由于密文索引构建而大幅度下降,告警日志的检索效率较高并能够检索获得相关攻击场景的告警日志。

关键词:安全存储;区块链;告警关联;攻击场景;索引构建
中图法分类号　TP309

Study on Secure Log Storage Method Based on Blockchain

LIU Jing[1],HUANG Ju[1],LAI Ying-xu[1],QIN Hua[1] and ZENG Wei[2]
1 Faculty of Information Technology,Beijing University of Technology,Beijing 100124,China
2 Beijing Branch of China Unicom,Beijing 101109,China

Abstract　With the rapid development of computer science,the number of alarm logs is increasing geometrically. Alarm logs record the correlation information of attack behavior and are vulnerable to theft and tempering,and the retrieval results contain a lot of irrelevant logs,thus interfering the correctness of log analysis. In order to solve the problems of safe storage and data extraction of alarm logs,this paper proposes a log secure storage method based on blockchain. Alarm logs are stored in distributed storage system based on block chain,which index library records block storage location. The traditional block chain sequential retrieval is replaced by querying the block index library,which improves the retrieval speed of ciphertext log. Through threat assessment of attack source addresses of alarm logs,and build a ciphertext index structure,which is stored in the block header. Alarm logs classified to the same attack scenario are associate retrieved based on correlation analysis. According to the experimental results,using the log secure storage method based on blockchain to store alarm logs,the block generation efficiency will not greatly reduce due to the index construction,and the log retrieval efficiency is high and the attack scenario logs can be obtained.

Keywords　Secure storage,Blockchain,Alarm correlation,Attack scenario,Index construction

1　引言

入侵检测系统(Intrusion Detection System,IDS)负责对网络、系统的运行状况进行实时监控,以尽可能发现所有攻击行为,从而保证网络及系统的正常运行。IDS 告警日志记录检测到攻击行为后,管理员通过日志分析可以及时发现并修补安全漏洞,保证网络及系统安全。然而,在告警日志的管理方面存在两个非常严重的问题:1)告警日志记录着网络中的攻击事件,是分析入侵者攻击行为的关键数据,传统的日志存储模型一般使用云服务器来实现日志的存储和检索,然而该存储方式无法满足告警日志的安全需求,可能会存在日志文件被窃取或篡改的情况,云服务器无法验证告警日志的完整性,会干扰后续日志分析的正确性;2)由于网络环境中的攻击事件发生频繁[1],入侵检测系统在一天内能够产生数以万计

基金项目:北京市自然科学基金(19L2020);信息保障技术重点实验室基金(614211204031117);陕西省网络与系统安全重点实验室开放课题基金资助项目(NSSOF1900105);工业和信息化部 2018 年工业互联网创新发展工程;国防科研试验信息安全实验室基础研究项目(2018XXAQ08)
This work was supported by the Beijing Municipal Natural Science Foundation(19L2020),Foundation of Science and Technology on Information Assurance Laboratory(614211204031117),Foundation of Shaanxi Key Laboratory of Network and System Security(NSSOF1900105),Industrial Internet Innovation and Development Project of the Ministry of Industry and Information Technology of China in 2018 and Basic Research Project of Information Security Laboratory for National Defense Scientific Research and Testing(2018XXAQ08).
通信作者:赖英旭(laiyingxu@bjut.edu.cn)

的告警日志，其中包含着大量的无关告警，这些低级告警在分析过程中极易被当作真实攻击告警，干扰对入侵者攻击意图的正确分析。传统的日志存储模型通常使用关键词匹配的方式来提取告警日志，日志检索结果中包含大量的无关告警，在后续的日志分析过程中将会耗费分析者大量的时间和精力，使得告警日志的分析效率不高。

告警日志的完整性是有效分析的前提，只有安全存储告警日志使其不受到恶意篡改等攻击行为，才能依据告警日志进行正确的后续分析。区块链（Blockchain）是一种按照时间顺序将数据区块以顺序相连的一种链式数据结构[2]，每个区块头中都记录前一区块的哈希值，具有可追溯、不可篡改、不可伪造的特性，将区块链用于存储日志信息可以大大增加日志的安全性。

告警关联（Alarm correlation）通过分析告警日志在时间和空间上的相关性，从大量的日志中获取攻击事件间的逻辑关系[3]，提取出属于同一攻击过程的告警日志，重现攻击场景，从而完善安全设备的防护结构。使用告警关联对日志进行聚类和索引构建，可以通过日志检索获取具有相关性的告警日志，提高后续的日志分析效率。

本文充分考虑了告警日志在安全存储和分析操作上的困难，提出了一种基于区块链的日志安全存储模型。该模型实现了基于区块链的日志存储架构，保证了告警日志的安全存储；进而深入了解来自多类安全设备的告警日志，定义了告警源地址威胁程度的计算方法，将告警日志按照威胁程度进行有效划分，并采用告警关联的方法对日志进行相关性分析，构建安全索引提供告警日志的有效检索，在日志数量庞大的情况下依然能够提取出具有攻击意图的告警日志，同时使用密文检索的方法增加了日志检索过程的安全性。

2　相关技术

2.1　区块链技术

区块链[4-5]是一种将数据区块按照时间顺序进行链式相连的数据结构，使用加密技术保护交易记录的安全，采用共识机制达成了各节点处区块结构的一致性，并运用了分布式数据存储、点对点传输等多种计算机技术[6]。区块链的安全特性使得交易记录不被篡改，从而保证了存储数据的完整性，保护了交易记录的存储安全。

2008 年，区块链概念被中本聪在《比特币：一种点对点电子现金系统》[7]中首次提出，2009 年 1 月 3 日，中本聪创造了第一个区块，即创世区块，然后使用比特币与密码学专家哈尔芬尼进行了比特币史上的第一次交易。比特币和以太坊是区块链中非常出色的两个技术，因为有着稳定可靠和容易扩展的特点被人们迅速推广和应用。

如今，区块链技术被广泛应用于医疗数据、电网数据以及一些交易信息的存储，区块链不可篡改的特性保证了数据存储的安全性。Ekblaw[8]基于去中心化的区块链结构实现了一种用于保存医疗记录的分布式存储系统。Wu 等[9]提出了一种基于联盟区块链的智能电网数据的存储系统。Wang 等[10]提出了一种基于区块链的数据安全共享的方式，实现了数据共享可信网络环境的构建。区块链技术还被应用于日志存储[11]、农产品追溯[12]等领域。随着区块链在存储领域上的不断拓展，安全性和效率性也需要进一步的提高。

2.2　告警关联

告警关联是通过对告警日志进行分析，识别出具有攻击意义的日志文件，并了解告警日志间的相关性，从而得到同一攻击过程的告警日志。目前，告警关联研究已经有很多成果，包含了多种告警关联方法。

（1）基于统计分析的告警关联方法。Xin 等[13]提出了基于统计的告警关联方法，该方法引入时间序列模型来计算告警之间的因果指数，如果大于给定的阈值，则将两告警进行关联。该方法不限于关联规则，可以发现新的攻击场景，但是方法中通过计算指数进行告警关联使得计算量较大。

（2）基于因果关系的告警关联方法。Templeton 等[14]通过分析因果关系关联告警日志，提出了一种基于先决条件的告警日志关联方法。Peng 等[15-16]基于该方法深入分析攻击事件之间的先后关系，提出了一种基于因果关系的关联方法。该方法存在一个明显的缺点，如果存在告警漏报的情况，则可能会使得攻击关联图出现断点，从而导致多个攻击事件无法被关联到关联图中，无法产生完整的攻击场景，同时不能识别新的攻击模式。Alserhani 等[17]提出将因果关联方法与统计方法相结合对告警进行关联，经过对测试的结果进行分析可以发现，该方案能够有效地提高检测率，降低告警误报率。

（3）基于攻击图的告警关联方法。Phillips 等[18]提出了使用攻击图表示攻击步骤之间的因果关系，使用结点表示攻击步骤的因果，边表示攻击步骤间的先后关系，模拟出告警事件的攻击流程，从而对告警日志进行关联，但是该方法的效率太低。Zali 等[19]将因果关联图作为知识库从而构建因果关联树，当收到新告警时，在因果关联树中寻找相关的告警，将告警日志与其关联起来。该方案可以有效地提升告警关联效率，但是出现入侵者隐藏攻击事件的情况时，无法与相关的告警进行关联。

（4）基于属性相似度的告警关联方法。Valdes 等[20]提出了基于属性相似度的告警关联方法，通过计算告警属性相似度得出告警相似度，根据条件判断是否属于同一安全事件。Ma 等[21]提出一种基于属性层次树的告警相似度的计算方法，从而进行告警模糊聚类，能有效地重构攻击场景。Mei 等[22]通过计算告警间的相似度来对告警日志进行关联和聚类操作，具体过程为分析告警的相似度，从而得到攻击活动序列集，在序列集中进行比较，将相似的序列进行聚类。Li 等[23]分析现有的告警聚合与关联方法，为了提高告警关联的准确性，提出了基于马尔可夫链模型的告警关联方法。

（5）基于数据挖掘的告警关联方法。Bin 等[24]提出了基于多层感知器和支持向量机的告警关联方法，但该方法需要使用大量的实际数据进行训练，在实际应用中要获取到大量数据的难度较大。Lu 等[25]通过分析当前的告警关联方法，总结了关联结果不足的原因，并且基于二维表改进了 Apriori 方法，能较好地实现告警关联。

（6）基于知识库的告警关联方法。Wang 等[26]基于知识图研究了一个安全事件关联分析系统 KGBIAC，可以整合现有的大量数据信息进行关联，但是需要提前构建一个知识图。

因此，上述告警关联的方法大多依赖于知识库、关联规则、基本条件，在实际应用中依然存在较多的缺点，需要根据

实际场景改进所用的方法。

3 基于区块链的日志安全存储方法

3.1 基于区块链的日志安全存储模型

本文提出了基于区块链的日志安全存储方法，构建一个基于区块链的分布式日志存储架构，如图 1 所示，该系统主要包含 3 个部分：日志上传者、基于区块链的分布式日志存储架构和日志分析者。基于区块链的分布式日志存储架构为分布式存储系统与区块链结合的存储结构，负责存储加密告警日志并支持日志检索结果的完整性验证，日志上传者负责将告警日志上传，将本地的告警日志转移到基于区块链的分布式日志存储架构上进行加密存储，日志分析者提交关键词检索加密告警日志，并对告警日志进行分析操作。

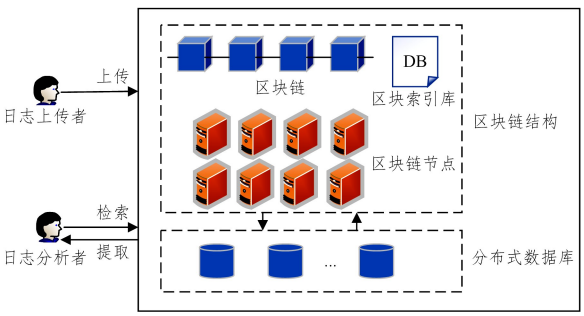

图 1 基于区块链的日志密文存储模型

Fig. 1 Log cipher text storage model based on blockchain

如图 2 所示，基于区块链的分布式日志存储架构中主要包括 3 个存储结构，分别是分布式存储系统、区块链和区块索引库。其中分布式存储系统用于保存告警日志，区块链结构负责进行日志元数据的存储工作和告警日志的检索工作，区块索引库用于提供数据区块的快速查询。基于区块链的分布式日志存储架构主要包含以下几个部分。

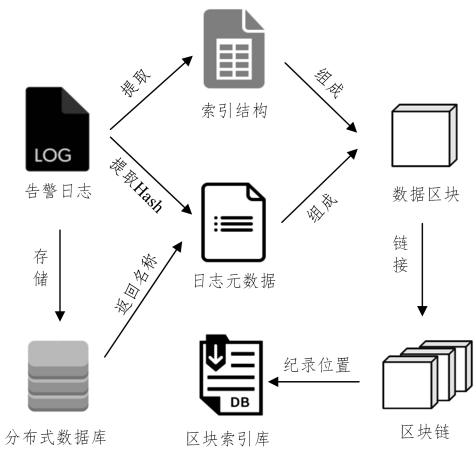

图 2 基于区块链的分布式存储结构

Fig. 2 Distributed storage structure based on block chain

(1)分布式存储系统

分布式存储系统用于存储大量的告警日志，模型中使用的分布式存储系统为 Hadoop 存储系统，使用副本备份的方式增加了告警日志的存储安全性。随着告警日志的快速产生和不断积累，日志数据需要大量的存储空间来保存，本地的存储方式已经无法满足告警日志的存储需要，因此，本文将加密告警日志安全存储在分布式存储系统，通过查询日志文件名称提取日志文件，能够有效地解决告警日志数量太大而不易存储的问题。

(2)日志元数据

日志元数据是告警日志中的关键信息，由告警日志的 Hash 值和日志文件名称组成，其中告警日志的 Hash 值通过使用 MD5 算法计算得到，主要用于验证告警日志的完整性。日志文件名称为分布式存储系统中的文件存储名称，可以根据日志文件名称从分布式存储系统中提取告警日志。由于告警日志数量庞大，而数据区块的存储空间有限，无法直接存储告警日志本身，因此将告警日志加密保存在分布式存储系统中，并提取出告警日志的关键信息作为日志元数据存放在区块体中，从而减少数据区块的空间占用。

(3)数据区块

如图 3 所示，数据区块主要包括区块头和区块体两个部分，本文在区块头中增加了日志对称密钥、密文索引结构和告警时间标识。日志对称密钥在日志检索完成时发送给日志分析者，密文索引结构提供告警日志的有效检索，告警时间标识用于区块查询。区块体中的交易数据为日志元数据。将日志元数据生成默克尔树并存放在区块体中保证日志元数据的安全性，同时在日志检索时提供告警日志的完整性验证。

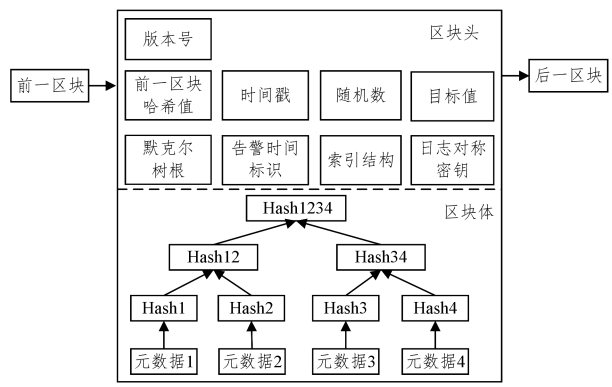

图 3 区块结构示意图

Fig. 3 Schematic diagram of block structure

(4)共识算法

区块链中的区块共识机制采用 PoW 共识算法，区块链节点通过获取随机数计算满足目标难度的区块哈希值作为节点的工作量证明，互相竞争日志元数据的存储权限并且生成数据区块，可以有效地保证各个节点数据区块的一致性。

(5)区块索引库

区块索引库负责存储数据区块的告警时间标识与区块存储地址形成的键值对，提供检索过程中数据区块的快速查询和获取，区块索引库随着新区块的产生进行实时更新，并定时与其他节点验证区块索引库中数据的正确性。本文中的日志密文检索过程分为区块检索过程和告警日志检索过程两个环节，传统的区块检索过程为链上区块查询，通过在区块链上逐个查找上一区块的位置并读取数据区块中的字段进行匹配，从而获得目标区块，当数据区块数量较多时，多次的位置读取和区块查询将会降低数据区块的查找效率。因此，本文通过设置一个区块索引库提供数据区块的快速查询，当日志分析者进行日志检索时，首先区块链节点会通过查询区块索引库快速得到数据区块的存储地址并找到目标数据区块，然后根据数据区块中的索引结构获取告警日志，加快了告警日志的检索过程。

3.2　基于源地址威胁程度的日志索引构建

本文构建一个基于区块链的分布式日志存储架构，用于存储告警日志并支持日志密文检索，然而传统的文件检索方法并不适用于日志文件的检索场景，使用关键词匹配的方式获取告警日志的提取方法导致得到的日志中包含大量的无关告警，这些告警日志并不具有实际意义，使得后续的告警日志的分析效率较低，日志分析的准确率不高。本文将为具有攻击意图的告警日志构建安全索引，可以提高日志分析的效率。因此，本模块设计了一个索引构建方案，根据不同的源地址，针对同一目的地址的攻击行为，计算出源地址对目的主机的威胁程度，从而找到具有攻击意图的源地址并将相关的告警日志提取出来。

定义 1（告警向量，Alarm Vector）　告警向量是从告警日志中提取出告警属性（如源地址、目的地址、告警等级等）组成的 n 元向量，记告警向量为 $\boldsymbol{A}=(a_1,a_2,\cdots,a_n)$，其中 a_k 代表向量中的第 k 个属性。

定义 2（目的告警集合）　目的告警集合是由目的地址相同的告警向量组成的向量集合，记目的告警集合为 $\boldsymbol{DS}=\{\boldsymbol{A}_1,\boldsymbol{A}_2,\cdots,\boldsymbol{A}_n\}$。

定义 3（相似告警集合）　相似告警集合是由源地址和目的地址相同的告警向量组成的向量集合，记相似告警集合为 $\boldsymbol{SS}=\{\boldsymbol{A}_1,\boldsymbol{A}_2,\cdots,\boldsymbol{A}_n\}$。

为了评估告警日志中源地址对目的主机的威胁程度，本文提出一个计算源地址威胁程度的方法，如式（1）所示：

$$Danger(src)=\frac{\sum_{i=1}^{5}P_i\times N_i}{\sum_{i=1}^{5}N_i} \tag{1}$$

其中，src 为告警源地址，$Danger(src)$ 为告警源地址 src 的威胁程度，i 为告警等级，P_i 为告警等级 i 的权重，N_i 是告警等级 i 在该时间段的告警日志中出现的次数。在告警等级的权重比例中，告警等级越高对应的权重就越大。由于常规告警通常数量较多且不能体现出源地址的重要程度，因此低级告警等级的权重设置得低一些，可以减小常规告警对源地址威胁程度的评估影响，返回攻击意义更大的源地址和告警日志。将相似告警集合中不同告警等级在日志中出现的次数进行数量统计，并将结果代入公式中，可以得到在该时间段中源地址对目的地址的威胁程度。

基于源地址威胁程度的索引结构如图 4 所示。索引结构的关键词为目的 IP 地址计算的摘要，根据式（1）可以得到源地址的威胁程度，通过将加密的源地址、威胁程度、文件名称存放到一个源地址信息结构中表示源地址的关键信息，记为 Logterm。同一个目的地址告警集合中产生的 Logterm 存放在一个数组中，数组代表了这个目的地址的所有源地址关键信息，并将其作为索引结构的检索值。

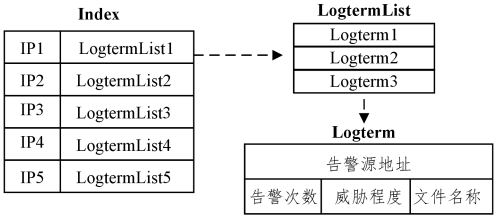

图 4　索引结构图示

Fig. 4　Index structure diagram

密文索引结构的构建算法如算法 1 所示。

算法 1　基于源地址威胁程度的密文索引结构生成算法

输入：同一时段的告警日志 Log

输出：基于源地址威胁程度的索引结构 SrcIndex

1. 初始化集合 SrcIndex 用于存放索引结构，DesSet 存放目的向量集合。
2. 从告警日志中提取出关键属性生成告警向量 $\mathbf{A}=(a_1,a_2,\cdots,a_n)$，将日志文件名称加入到向量 **A** 的尾部，形成告警向量集合 **AS**。
3. 将集合 **AS** 分别按照目的地址进行分类，从而形成多个目的告警集合 **DS**，并放入 DesSet 中。
4. 对于 DesSet 中的每个集合 **DS**：
   4.1. 创建一个新的 SrcSet 存放相似向量集合。
   4.2. 将 **DS** 中的告警向量按照不同的源地址形成相似告警集合 **SS**，并放入 SrcSet 中。
   4.3. 根据 SrcSet 中的集合 **SS** 使用式（1）计算 **SS** 源地址的威胁程度，与加密的告警源地址、告警次数、日志文件名称组成 Logterm 源地址信息结构。
   4.4. 将所有 Logterm 结构放入数组中，并按照威胁程度大小排序，与目的告警集合 **DS** 中的目的地址通过计算得出的摘要值形成键值对，放入索引结构 SrcIndex 中加密生成密文结构。

3.3　基于告警关联的攻击场景日志关联检索

根据告警日志的相关研究可以发现，告警关联可以通过分析关键属性得到告警日志之间是否关联，找出存在相关性的日志，从而重现入侵者的攻击场景。如图 5 所示的攻击场景中，通过对攻击行为的模拟，可以完成入侵路径的追踪，完善对攻击场景的防范工作。本文通过使用 3.2 节的日志索引检索目的地址得到威胁程度较大的源地址和相应的告警日志，同时将包含该目的地址的攻击场景地址序列返回并进行场景选择，通过进一步的攻击场景关联检索可以得到包含该目的地址的攻击场景日志。

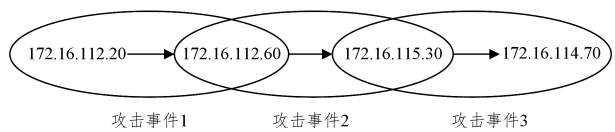

图 5　攻击场景示意图

Fig. 5　Schematic diagram of attack scenario

定义 4（攻击事件，Attack Log）　相同源地址、目的地址和告警类型的告警日志组成的告警集合，称为攻击事件，记攻击事件为 $Al=\{L_1,L_2,\cdots,L_n\}$，其中 $Lj.time-Li.time<T$，$i<j$，T 为攻击事件的时间间隔。

定义 5（攻击序列，Attack Sequence）　将攻击事件按照攻击时间排序的攻击事件序列，称为攻击序列，记攻击序列为 $AS=\langle A_{l1},A_{l2},\cdots,A_{ln}\rangle$。

定义 6（攻击场景，Attack Scenario）　将攻击场景记为 $Atkscene=$（Stime，SIP，Stype，Sgd，Etime，EIP，Etype，Egd，Alset，IPlist），攻击场景中告警时间最早的告警事件称为场景的初始事件，场景中告警时间最晚的告警事件称为场景的结束事件，Stime，SIP，Stype，Sgd 分别为初始事件的告警时间、源地址 IP、告警类型、告警等级，Etime，EIP，Etype，Egd 为结束事件的告警时间、目的地址 IP、告警类型、告警等级，Alset 为攻击场景的告警事件集合，IPlist 为攻击场景的 IP 地址序列。

定义 7（上一攻击事件）　当攻击事件 Al_i 与 Al_j 属于同一攻击场景，$A_j.time>A_i.time$ 且 A_i 的目的地址与 A_j 的源地址相同时，称 A_i 为 A_j 的上一攻击事件，$i<j$。

本文通过告警关联得到攻击场景的告警日志，首先将告警日志预处理生成告警向量集合，并根据告警向量中的时间、源地址、目的地址和告警类型是否完全相同进行分类，其中，同一攻击事件中的告警向量时间间隔应小于 T，间隔超过 T 的告警向量则归类到一个新的攻击事件中，T 为攻击事件的最大时间间隔，用于继续收集属于该攻击事件的告警向量。接着按照告警时间顺序对所有攻击事件排序得到攻击序列，通过对攻击序列中的攻击事件进行告警关联，可以得到攻击场景的告警日志。根据下列规则对攻击序列中的攻击事件按照时间顺序依次进行分析。

当两个攻击事件满足以下任一条件时，认为事件 Al_j 与 Al_i 存在时间相关和 IP 相关：

(1)当事件 Al_j 的源地址与事件 Al_i 的目的地址相同时并且 $Al_j.time-Al_i.time<T_1$ 时，则事件 Al_i 为事件 Al_j 的上一攻击事件。

(2)当事件 Al_j 的源地址与事件 Al_i 的源地址相同时，且 $Al_j.time-Al_i.time<T_2$，认为事件 Al_i 与事件 Al_j 为同一攻击者向不同目的主机发起攻击的两个相似的操作，属于同一攻击步骤。

(3)当事件 Al_j 的目的地址与事件 Al_i 的源地址相同且事件 Al_i 存在上一攻击事件 A_{lk}，$Al_j.time-A_{lk}.time<T_3$ 时，认为事件 Al_j 和事件 A_{lk} 是由攻击者使用不同的源地址向同一目的主机发起的两个相似的攻击事件，属于同一攻击步骤。

(4)当事件 Al_j 的目的地址与事件 Al_i 的源地址相同且 $Al_j.time-Al_i.time<T$ 时，认为事件 Al_j 中存在告警漏报，导致攻击序列的攻击事件的时间顺序出现错误，认为事件 Al_j 是事件 Al_i 的上一攻击事件。

其中，T_1 为两个连续事件的最大时间间隔，T_2 为同时从一个源地址向两个目的地址发起相似攻击的最大时间间隔，T_3 为从两个源地址向一个目的地址发起相似攻击的最大时间间隔，取值由通过分析实际的日志数据得到。当攻击事件 Al_1 与 Al_2 满足以上任一规则时，就认为这两个攻击事件属于同一个攻击场景，不满足上述规则的攻击事件与其他事件不存在关联，被单独划分到一个新的攻击场景中。

在实际的攻击过程中，会出现由于告警漏报而导致攻击场景产生断层的情况，从而使得相关的告警日志无法被关联在一起，为了能将断开的场景聚合到一起，本文对所有攻击场景进行场景相似度计算，计算场景相似度的公式如式(2)—式(4)所示：

$$F_t(x,y)=1-\frac{y.Stime-x.Etime}{120} \tag{2}$$

$$F_d(x,y)=\frac{Samedegree(x,y)}{32} \tag{3}$$

$$F_i(x,y)=\begin{cases}1, & y.Sgd=x.Egd\\ 1-\dfrac{y.Sgd-x.Egd}{5}, & y.Sgd\neq x.Egd\end{cases} \tag{4}$$

其中，式(2)计算场景的时间相似度，式(3)计算场景的 IP 相似度，式(4)计算场景的告警等级相似度，其中，x,y 表示两个不同的攻击场景，$Samedegree(x,y)$ 是计算 $y.EIP$ 与 $x.SIP$ 的二进制数值中相同位数的函数。当攻击场景 x,y 满足 $Ft(x,y)<a$，$Fd(x,y)>b$，$Fi(x,y)>c$(a 为场景时间相似度阈值，b 为 IP 相似度阈值，c 为告警等级相似度阈值)时，认为场景 x 和 y 是由同一攻击场景因漏告警产生的两个相似场景，将两个场景合并生成一个新场景。

用于推送攻击场景告警日志的索引构建描述如算法 2 所示。

算法 2 根据告警日志生成基于告警关联的索引结构

输入：告警事件序列 $AS=\langle Al_1,Al_2,\cdots,Al_n\rangle$

输出：基于告警关联的索引结构 $LogIndex=\{(str_1,Atkscene_1),(str_2,Atkscene_2),\cdots,(str_n,Atkscene_n)\}$

1. 初始化 LogIndex 存放索引键值对，Sceneset 存放攻击场景。
2. 对于 AS 中的每一个 Al：
   2.1. 当 Sceneset 不为空时，遍历 Sceneset 中的场景 scene，若 scene 与 Al 存在时间相关和 IP 相关，则将 Al 加入到场景 scene 的列表中。
   2.2. 若不存在属性相关的 scene，那么将 Al 创建为一个单独的攻击场景。
3. 对于 Sceneset 中的所有场景 scene 进行相似性判断，若存在两个场景的属性相似度和总的相似度达到阈值，则将两个场景融合生成新场景。
4. 对于 Sceneset 中的每一个场景 scene，将其中的 scene.IPlist 进行计算得出摘要值，将该摘要值作为关键词 str，加密的 scene 当作索引值生成键值对(str,scene)加入到 LogIndex 中进行加密生成密文索引。

根据算法得到攻击场景告警日志的密文索引结构，当日志分析者输入 IP 地址得到威胁程度更高的源地址和告警日志，本文将包含 IP 地址的所有攻击场景地址序列返回给日志分析者，日志分析者通过选择地址序列对攻击场景的日志进行关联检索。

本文采用的索引加密方法为：为告警日志构建密文索引结构，索引结构中的关键词与告警日志的数据信息具有较大关联，同时告警日志的关键词只需要通过加密构成密文关键词提供告警日志的检索，无需对关键词进行解密操作，因此对关键词采用散列的方式将关键词转变为密文方式，本文使用多重散列计算的方式增加关键词的安全性。对于索引中的关键词 K，使用 SHA-256 算法进行散列得到 Hash 值 K_h，将 K_h 的前五位字符与后五位字符进行置换，接着使用 SHA-256 算法再次进行散列得到 Hash 值 K_m。告警日志的其他信息在检索时需要返回明文的数据，因此使用能够快速完成加解密的 AES 对称加密算法，提高日志索引结构的安全性。

3.4 告警日志的加密存储和密文检索

(1)告警日志的存储过程如下：

step1 日志上传者上传告警日志到客户端，同时选择 AES 算法从而对称加密生成加密告警日志，将其存储在分布式数据库中，并使用 MD5 算法计算日志 Hash 值与日志文件名称生成日志元数据。

step2 通过对告警日志进行预处理并提取出日志关键属性组成告警向量，根据日志索引构建方案生成密文索引结构，将日志元数据、日志索引结构、日志对称密钥与日志上传者的身份验证信息进行非对称加密一块打包发送到区块链节点。

step3 区块链节点收到加密验证信息后，先对加密信息进行解密获取身份信息，完成对日志上传者的身份验证后，区块链节点将日志元数据存储在区块体中，将索引结构存储到区块头中便于检索告警日志，并将日志对称密钥存储在区块头中在检索完成时返回给日志分析者，区块链节点使用 PoW 共识算法竞争区块生成权限，获得权限的节点将新区块连入自身维护的区块链中，同时更新本地的区块索引库。

(2)告警日志的检索过程如下：

step1　日志分析者提交关键检索词与时间检索词 Ts，使用多重散列算法对检索关键词提取信息摘要构造查询陷门，使用私钥将查询陷门、时间和日志分析者的身份验证信息进行加密，并发送到区块链节点获取告警日志。

step2　区块链节点收到信息后通过使用公钥进行解密得到检索词信息和验证信息，身份验证完成后根据检索词信息检索告警日志 E，使用时间检索词在区块索引库中找到区块，从区块头中找出索引结构，并在源地址日志索引中找到威胁程度更高的源地址的日志文件名称，在攻击场景日志索引中找到相关攻击场景地址序列。

step3　区块链节点在分布式数据库中根据日志文件名称检索加密告警日志，并计算日志摘要与数据区块的日志元数据中的 Hash 值进行正确性比较，判断日志检索结果的数据完整性。将加密告警日志按照源地址威胁程度进行排序，并根据检索条件选取一定数量的日志结果，与告警日志的对称密钥、攻击场景地址序列一起发送给日志分析者。日志分析者获取日志对称密钥和加密日志后进行快速解密得到明文告警日志。

step4　根据日志分析者输入的目的，推送所有地址序列中包含该目的地址的攻击场景，日志分析者可以通过选择地址序列对推送的攻击场景进行检索，从而得到与该攻击场景相关的告警日志。

4　实验分析

4.1　安全性分析

(1)日志文件的数据安全性。告警日志 C 使用算法 $Encrypt(C)$ 生成加密日志 E 存储在分布式数据库中，攻击者从数据库中窃取的数据是加密告警日志 E，无法直接从日志数据中获取关键信息。日志元数据通过算法 $Getmetadata(Hash(C),Filename(C))$ 生成，并存放在区块链中，若是告警日志发生变化生成 C'，在完整性验证中哈希计算得到的 $Hash(C')\neq Hash(C)$，能够及时地发现日志数据出现错误，从而进行数据恢复操作，不会将篡改的告警日志返回给日志分析者。

(2)区块链的安全性。区块链中每个区块记录上一区块的哈希值 $Hash_{last}$，当攻击者对其中的区块 $Block$ 中的日志元数据进行篡改时，由于下一区块 $Block_{next}$ 中记录着当前区块的 $Hash$，攻击者需要对下一区块 $Block_{next}$ 同样进行篡改，并依次往下处理，篡改区块链的耗费是非常巨大的，因此区块链具有较高的安全性。

(3)索引结构的安全性。在两个索引结构中，关键词 K 通过算法 $Encindex(K)$ 得到关键词 K_m，算法的加密过程中包含两次 SHA-256 算法，分别为 $K_h=Hashkey_1(K)$ 和 $K_m=Hashkey_2(K_t)$，两次的哈希计算不可逆，并在两次哈希计算间进行了数字置换 $Changedig_1(K_h)$。假设攻击者能够采用暴力破解的方式从 K_m 得到 K_t，若要将得到的 K_t 变换为原始的 K，则需要将 K_t 进行一次数字置换 $Changedig_2(K_t)$ 得到 K_h，然后进行一次暴力破解得到 k，然而无法得知 $Changedig_2(K_t)$ 的具体过程，因此无法从散列的关键词 K_t 中得到原有的关键词 K。攻击者从区块中获取的索引结构为密文索引，无法从关键词 K_m 获取 K 的相关信息和数据，因此无法获得明文日志数据。

4.2　性能分析

本文通过收集多个安全设备的告警日志进行存储模型性能实验，实验中使用的开发语言为 Java 语言，版本为 JDK1.8，分布式存储系统为 Hadoop 存储系统，区块链结构共包含了 8 个节点。为了评估数据区块的生成效率，本文通过记录有无索引区块生成耗时随着告警日志数量增长而变化的数据，来分析在日志存储过程中数据区块的生成效率，有无索引区块生成耗时的对比如图 6 所示。

由图 6 可知，随着告警日志的数量增长，有无索引的区块构建耗时都在不断增长，其中有索引的区块构建耗时增长速度要比无索引的区块构建耗时更多一些。这是因为索引的构建耗时与日志的数量有直接关系，索引构建过程中需要对日志进行预处理和相关性分析，告警日志数量越多，分析过程所花费的时间也越多，数据区块的生成与日志元数据的数量十分相关，数据区块中的默克尔树根据日志元数据的哈希值计算生成，因此区块中存储的日志元数据应该保持在一个比较合适的数量以提高区块的生成效率。

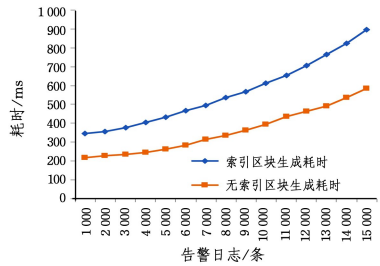

图 6　有无索引区块生成耗时的对比

Fig. 6　Comparison of time-consuming generation of blocks with or without index

本文将基于告警关联的攻击场景日志索引进行明文索引和密文索引的构建效率对比实验，实验记录在告警日志数量不断增长的情况下明文索引和密文索引的构建耗时变化，两种索引分别进行 10 次索引构建实验，对比数据取构建耗时的平均耗时，索引构建耗时的对比结果如图 7 所示。

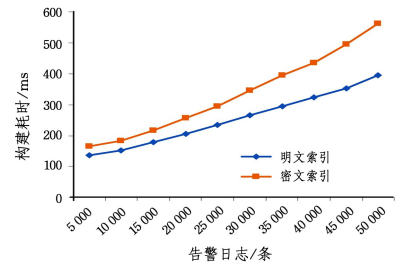

图 7　明文密文索引构建耗时对比

Fig. 7　Time-consuming comparison of plaintext ciphertext index construction

由图 7 可知，密文索引的构建耗时比明文索引的构建耗时更多一些，但是密文索引构建耗时的增长幅度并不迅速。这是因为告警日志中的属性数量不多，构建索引使用的关键词并不复杂，加密索引的耗时增长速度不高，日志索引中的告警数量较多并不会导致密文索引的构建效率明显下降，与明文索引的增长趋势较为接近，因此模型中索引加密方法能够保证日志索引的构建效率。

本文将基于区块链的日志安全存储方法与一些日志存储方法进行方案对比，方法对比结果如表 1 所列。

表 1 本文方法与其他日志存储方法的对比

Table 1 Comparison between proposed method and other log storage methods

方法	区块链	加密	完整性验证	链上链下分离存储	链下检索
文献[27]	×	×	×	—	—
文献[28]	×	√	√	—	—
文献[11]	√	√	×	√	×
文献[29]	√	√	√	√	×
本文方法	√	√	√	√	√

通过进行多组攻击场景日志检索的实验得到告警日志，分析计算本文中存储模型的日志检索准确率和召回率。为了客观地评估日志检索的效率，实验中共采用了 5 个时段的 snort 告警日志，实验中取数值 T 为 60 s，T_1 为 120 s，T_2 为 60 s，T_3 为 80 s，通过实验分析可知当阈值 a 为 0.4、b 为 0.875、c 为 0.8 时告警日志的检索效率最高。通过整理和预处理后进行日志检索，将检索结果进行统计分析，能够得到召回率和准确率。在 5 个时段内分别对告警日志构建索引并检索文件，通过将检索得到的相关攻击场景告警日志进行数量统计，通过计算可以获取检索结果的召回率和准确率，并与实际相关攻击场景告警日志进行数据对比，得到的日志结果如表 2 所列。

表 2 基于告警关联的攻击场景日志检索方法的结果分析

Table 2 Result statistics of attack scenario log retrieval method based on alarm correlation

实验组数	告警数量	检索结果数量	原有场景日志数量	检索场景日志数量	召回率/%	准确率/%
1	27356	7280	7352	6973	94.8	95.8
2	35260	10500	10683	10095	94.5	96.1
3	52365	15371	15635	15197	97.2	98.9
4	53126	6532	6798	6098	89.7	93.4
5	53283	8237	7953	7579	95.3	92.0

由表 2 可知，基于告警关联的攻击场景日志检索方法中攻击场景告警日志的检索召回率都在 85%以上，准确率都在 90%以上，5 个时间段的平均检索召回率为 94.3%，平均检索准确率为 95.2%。总体来说，本文中的检索方案召回率较高，将目的 IP 地址的相关攻击场景的告警日志返回，可以满足告警日志的检索需求。

根据本文中攻击场景日志检索方法中的分析结果，与文献[30]的日志检索结果进行准确率数值对比，对比结果如图 8 所示。

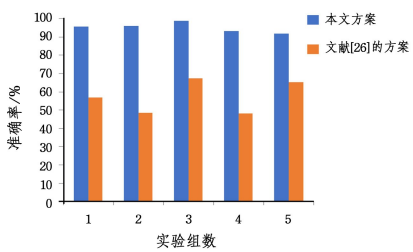

图 8 有无攻击场景日志索引的检索结构准确率对比

Fig. 8 Comparison of retrieval results with or without attack scene log index

由图 8 可知，本文中存储模型的检索结果准确率明显较高，文献[30]没有构建攻击场景日志索引，日志检索结果的准确率一般较低，因为检索结果中包含大量的无关告警，从而使得攻击场景告警日志的占比较低，使用基于告警关联的攻击场景日志索引能够保证日志检索结果中包含大量的攻击场景日志，提高告警日志检索结果的准确率，同时可以获取攻击场景中其他目的地址的告警日志，有效地提取具有攻击意图的告警日志，基于告警关联的攻击场景日志检索方法具有较高的可用性。

为了更好地评估数据提取的时间性能，本文将使用区块索引库的日志检索方案与传统的区块链上日志检索方案进行检索时间对比，数据区块中存储的日志元数据数量保持不变，随着告警日志的持续增加，记录日志检索的时间变化。告警日志数量从 10000 增加到 300000，平均每个区块中存储日志元数据的数量保持在 1000 条左右，随着告警数量的不断增加，记录检索方案耗费的时间，其中每组实验进行 20 次，采取检索耗时平均值作为实验的最终结果，检索方案耗时对比如图 9 所示。

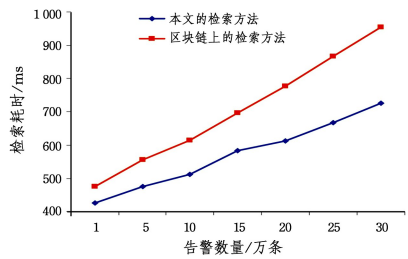

图 9 告警日志检索时间

Fig. 9 Alarm log retrieval time

为了更好地评估数据提取的时间性能，本文将使用区块索引库的日志检索方案与传统的区块链上日志检索方案进行检索时间对比，数据区块中存储的日志元数据数量保持不变，随着告警日志的持续增加，记录日志检索的时间变化。告警日志数量从 10000 增加到 300000，平均每个区块中存储日志元数据的数量保持在 1000 条左右，随着告警数量的不断增加，记录检索方案耗费的时间，其中每组实验进行 20 次，采取检索耗时平均值作为实验的最终结果，检索方案耗时对比如图 9 所示。

由图 9 可知，随着告警数量的不断增长，本文的检索方案和区块链上检索的方法的耗时都有所增加。与本文的检索方案相比，区块链上检索的方法的耗时增长更加快速。随着告警数量的增长，区块链上检索的方法耗时的增长趋势基本呈线性增长，本文检索方案的耗时以较慢的趋势增加，在增长速度上明显较好。由于区块链上检索的方法是通过在区块链中顺序查找数据区块来检索目标告警日志，数据区块的数量越多，日志检索耗时就越多，检索速度与日志存储时生成的区块数量有直接关系，因此耗时会随着区块数量的增加较为快速地增加。本文中的检索方案基于区块索引库查询目的区块，通过将告警时间标识与区块位置建立索引，日志检索时能够快速的根据告警时间标识直接得到目标区块，从而在区块头的索引中检索日志，因此随着告警数量的增加，检索耗时的增加并不快速，总体来说，本文中检索方案的效率性能较好。

结束语 本文针对日志管理上出现的日志存储困难、分析效率低下等问题，提出了基于区块链的日志密文存储模型。该模型充分了解区块链的安全特性，将区块链应用于告警日志的分布式存储，将日志元数据安全存储在区块链结构中，以保证告警日志的完整性。为了提高后续的分析效率，采用告警关联的方法从日志中提取出具有攻击意图的告警日志并构建安全索引，使用区块索引库提供日志的快速检索，并采取漏

告警分析的方式对产生断层的攻击场景进行补充和完善，尽可能地提取所有存在关联的告警日志。该模型不仅有效地实现了日志的安全存储，而且可以去除大量无关告警，快速并且全面地获取存在逻辑相关性的告警日志，能够为高效的日志分析提供依据和支持。未来的工作中，将会继续增强告警日志的存储安全性，并提高告警日志的检索效率。

参 考 文 献

[1] LAI Y X, CHEN Y N, ZOU Q C, et al. Design and analysis on trusted network equipment access authentication protocol[J]. Simulation Modelling Practice and Theory, 2015, 51: 157-169.

[2] YUAN Y, WANG F Y. Development status and prospect of block chain technology [J]. Acta Automatica Sinica, 2016, 42(4): 481-494.

[3] DOU B L, ZHU Y C, SHANG L B. Research on alarm correlation method[J]. Computer Applications and Software, 2006, 23(1): 74-76.

[4] ZHANG Y H, SHU J G, YANG K, et al. TKSE: trustworthy keyword search over encrypted data with two-side verifiability via blockchain [J]. IEEE Access, 2018, 6: 31077-31087.

[5] HUCKLE S, BHATTACHARYA R, WHITE M, et al. Internet of things, blockchain and shared economy applications[J]. Procedia Computer Science, 2016, 98: 461-466.

[6] SWAN M. Blockchain: blueprint for a new economy[M]. USA: O'Reilly Media inc., 2015.

[7] NAKAMOTO S. Bitcoin: a peer-to-peer electronic cash system [EB/OL]. [2019-05-06]. https://bitcoin.org/bitcoin.pdf.

[8] EKBLAW A, AZARIA A, HALAMKA J D, et al. A case study for blockchain in healthcare: "medrec" prototype for electronic health records and medical research data[C]// Proceedings of 2nd IEEE Open & Big Data Conference. Piscataway, NJ: IEEE, 2016: 25-30.

[9] WU Z Q, LIANG Y H, KANG J W. Smart grid data security storage and sharing system based on alliance block chain[J]. Journal of Computer Applications, 2017, 37(10): 2742-2747.

[10] WANG J Y, GAO L C, DONG A Q. Research on data security sharing network system based on block chain[J]. Journal of Computer Research and Development, 2017, 54(4): 742-749.

[11] FEI Y, NING J, HU Q. Log storage system based on blockchain [J]. Cyberspace Security, 2018, 9(6): 80-85.

[12] TIAN F. An agri-food supply chain traceability system for China based on RFID & blockchain technology[C]// 2016 13th International Conference on Service Systems and Service Management (ICSSSM). IEEE, 2016: 1-6.

[13] QIN X, LEE W. Attack plan recognition and prediction using causal networks[C]// Proceedings of the 20th Annual Computer Security Applications Conference. Piscataway, NJ: IEEE, 2004.

[14] TEMPLETON S J, LEVITT K. A requires/provides model for computer attacks[C]// Proceedings of the 2000 New Security Paradigms Workshop. New York, ACM, 2000: 31-38.

[15] NING P, CUI Y, REEVES D S. Constructing attack scenarios through correlation of intrusion alerts[C]// Proceedings of the 9th ACM Conference on Computer and Communications Security. New York, ACM, 2002: 245-254.

[16] NING P, XU D. Learning attack strategies from intrusion alerts [C]// Proceedings of the 10th ACM Conference on Computer and Communications Security. New York, ACM, 2003: 200-209.

[17] ALSERHANI F, AKHLAQ M, AWAN I U. MARS: multi-stage attack recognition system[C]// Proceedings of the 2010 24th IEEE International Conference on Advanced Information Networking and Applications. Piscataway, NJ, IEEE, 2010: 753-759.

[18] PHILLIPS C, SWILER L P. A graph-based system for network-vulnerability analysis[C]// Proceedings of 1998 Workshop on New Security Paradigms. New York, ACM, 1998: 71-79.

[19] ZALI Z, HASHEMI M R, SAIDI H. Real-time intrusion detection alert correlation and attack scenario extraction based on the prerequisite-consequence approach [J]. The ISC International Journal of Information Security, 2013, 4(2): 126-136.

[20] VALDES A, SKINNER K. Probabilistic alert correlation[C]// Proceedings of the 4th International Symposium on Recent Advances in Intrusion Detection. Berlin: Springer, 2001: 54-68.

[21] MA L R, YANG L, WANG J X. Reconstruction of intrusion detection alarm association diagram by fuzzy clustering[J]. Journal on Communications, 2006, 27(9): 47-52.

[22] MEI H B, GONG J, ZHANG M H. Research on multi-step attack pattern discovery based on alarm sequence clustering[J]. Journal on Communications, 2011, 32(5): 63-69.

[23] LI H C, WU X P. Alarm multi-level aggregation and association method based on self-expanding time window[J]. Advanced Engineering Sciences, 2017, 49(1): 206-212.

[24] ZHU B, GHORBANI A A. Alert correlation for extracting attack strategies[J]. International Journal of Network Security, 2006, 3(3): 244-258.

[25] LU X G, DU X H, WANG W J. Alarm correlation algorithm based on improved FP growth[J]. Computer Science, 2019(8): 64-70.

[26] WANG W, JIANG R, JIA Y, et al. KGBIAC: knowledge graph based intelligent alert correlation framework[C]// International Symposium on Cyberspace Safety and Security Springer. Berlin, Springer: 2017: 523-530.

[27] WU G J, WANG S P, CHEN M, et al. Massive structed data oriented storage and retrieve system[J]. Journal of Computer Research and Development, 2012(S1): 1-5.

[28] CHENG M C, XU K Y. Audit log secure storage system based on trusted computing platform [J]. Computer Science, 2016, 43(6): 146-151.

[29] LU J F, LAI Y X, LIU J. Log Security Storage and Retrieval Based on Combination of On-chain and Off-chain[J]. Computer Science, 2020, 47(3): 298-303.

[30] WANG R D, JING Y N, WANG H G, et al. Research on parallel retrieval technology of log files based on timestamp index[J]. Computer Applications and Software, 2011, 28(2): 145-147.

LIU Jing, born in 1978, Ph.D, lecturer, is a member of China Computer Federation. Her main research interests include network security and trusted computing.

LAI Ying-xu, born in 1973, Ph.D, professor, Ph.D supervisor. Her main research interests include cloud computing, information network security and trusted computing.