KM : Effective Data Tracing In Distributed Systems

First Name
First Institution

Second Name
Second Institution

Abstract

Modern Internet services often involve large and complex distributed systems with data scattered around. As the number of services grows and data access scenarios get varied, important data might get leaked or stolen with no knowledge of the system administrators. In this paper we propose KM, a system which is used to trace data access and transfer in distributed systems. We present the model for tracing how data are accessed and transferred, showing that this model can succinctly capture users behaviors and thus can be used to replay scenarios where data are leaked or stolen. We also present a efficient implementation of KM and show that with proper sampling strategies our implementation can have negligible impact on existing services.

1 Introduction

Data access control and monitoring in complex distributed systems have been a real world concern [1] but have not yet gathered enough attention. Large internet companies which provide heterogeneous services often have their data stored in different data centers but the technologies they use to process these data, such as Mapreduce [6] and Spark [15], are built with no security in mind, making it hard to precisely grant data access permission to relevant users and, in case of data being stolen, trace how the data are transferred. In this paper we try to mitigate this situation by using a model to capture user program behaviors and design a system called *KM* to efficiently collect information of interest such that we can trace how a piece of data is accessed and transferred within or across machines and then probably stolen (i.e., copy into a USB stick or send to some remote machine).

The goal of KM is to track data access and transfer. The term *event* is used to describe any activity related to a piece of data. An *event* is modeled with a triad *(Subject, Action, Object)*. To succinctly model program behaviors, *event*s are classified into three categories: *file event*s, *network event*s and *process event*s. Examples of *file event*s include opening, reading and writing files; examples of *network event*s are binding a socket, connecting to remote machines and accepting incoming connections; examples of *process event*s are forking subprocesses and executing executables via the *execve()* system call. As will be shown below, these three classes of events are sufficient to model program behaviors.

The whole system is designed to be configurable. Strategies can be add or delete at anytime using a *KMctl* program or with the corresponding API. Therefore one can set up a central configuration service to manipulate strategies used by every single machine. These strategies include turning the tracing system on and off, controlling sampling rate and adding a file to be tracked. Note that our system is *data-oriented*, which mean that instead of collecting all events information, it will only collect information related to the data of interest. If during file transfer some intermediate files are not in our list (*e.g.*, process A read file *a* and write to file *b*. Process B read from file *b*. We have file *a* in our list, but the intermediate file *b* is not in our list), we will add the intermediate files to the global configuration at runtime, which can then be seen by the whole system.

Besides, we have implement a utility called *hotpatcher* which can be used to insert a dynamic shared object (DSO) into a user program, so that one can deploy this tracing system without having to reboot machines and affecting existing services. This utility can also be used to *partially* deploy the tracing system, which is very helpful at the time of testing.

Under the hood we have interposition in libc, using the LD_PRELOAD technique [10, 11, 14] or using the *hotpatcher* utility program described above. Whenever a program try to operate on the data of interest (e.g, reading or writing), a piece of information, which is necessary for data tracing afterward, will be collected by our interposition code and then sent to a message queue in a piece of shared memory. The shared memory is set up by a daemon called *KMagent* when it is started (usually at boot time) and is shared among all processes at that single machine. *KMagent* is the consumer here, continuously polling the shared message queues and collecting data, whereas all the other processes are producers. The message queue in shared memory is wait-free and operations on it are guided by a protocol to make them efficient and safe, affecting no existing services. As will be demonstrated below, with the current implementation, the overhead within a single event is about $1us$ on our[–TODO machine specification –]. With proper sampling strategies, we show that our implementation have only negligible impact on file events, 10% overhead on process event and 10% overhead on network event.

Periodically a central service try to pull from all machines to collect newly generated data. We use a *bulk mode* [5] to make this operation cheap. Besides, our system is *data-oriented*, which makes the volume of collected data small, thus placing only low burden on the system.

There has already been lots of works in using tracing techniques to gather information about distributed system behaviors [2, 3, 4, 7, 16]. However, those works almost all focus on performance analysis and trouble-shooting. Most of them are targeted to some specified program (*e.g.*, web server) [2, 3] and require help from the operation system kernel. Compared to them, our attention is on tracking how data is accessed and transferred, and then determine whether those operations are legitimate. Our system can affect *every* process on a machine with sufficiently low overhead. With the implementation in userspace, the whole system is also easy to test and customize. We draw some ideas from [8, 9]. Our model can capture user behaviors precisely. Along the way, we design an simple but efficient system which imposes negligible impact on existing services.

To summarize, our contributions are:

1. a model that can precisely capture program behaviors and track how a piece of data are accessed and transferred.

2. a system with efficient implementation which impose negligible impact on existing services.

3. the ability to query and control strategies used in the system and the ability to deploy the whole system while requiring no reboot.

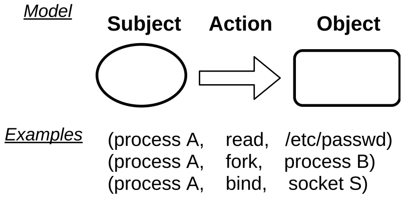
Model
Subject
Action
Object
Examples
(process A, read, /etc/passwd)
(process A, fork, process B)
(process A, bind, socket S)

Figure 1: The Subject-Action-Object Model

The structure of this paper is as follows: in section 2 we present the model we use to capture user behaviors and tracing data transfer. We then show details of our implementation in section 3, along with its strategies. In section 4, we present some experiments carried out on a small cluster and show that it has negligible impact on existing services. We discuss related work and future work in section 5 and section 6 respectively and then conclude in section 7.

2 Model

In this section we present the model used by KM. In KM, any activity related to a piece of data is called an *event*, which is modeled with a triad *(Subject, Action, Object)*, as shown in Figure 1, with examples. All events are classified into three categories: *file event*s, *network event*s and *process event*s. If process A read a file */etc/passwd*, this "read" operation will be recorded by our system which then emits a triad *(process A, read, /etc/passwd)*. Similarly, if process A calls *fork(2)* (and probably communicate with its child process afterward), this "fork" operation will also be recorded and finally a triad *(process A, fork, process B)* will be emitted by the system. Same for the network event when a process A bind to a socket S.

Since most Unix-like systems (*e.g.*, Linux) treat most things as files, if we can record all the relevant file events, we can then capture most of the key operations related to some particular piece of data, such as reading a file and sending it over a socket, thus being confident in knowing how a file is accessed and transferred. If we also have processes events at hand, we can know how processes interact with each other and build more confidence in file transfer within a single machine. Moreover, with network events, we can know how a file is transferred across machines. Therefore, with information of these three types of events, we can capture user behaviors on some piece of protected data and trace its transfer route in a distributed system.

Trace on The Same Host On a single machine, if a file is under monitor, then any access of it by any

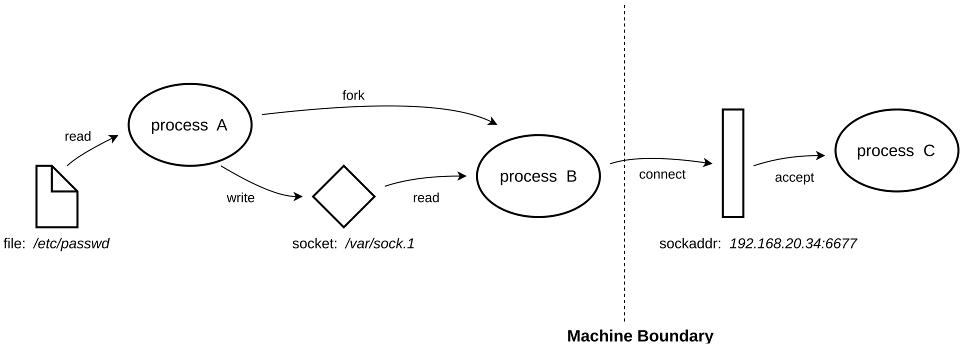

Figure 2: left: file access and transfer in a single machine. right: distributed system

process will be recorded by our system. Subsequent file operations by that particular process (*e.g.*, writing to a file in */tmp* or send it over a socket) will also be recorded. These records are all emitted as file events. Because processes mostly interact with their child processes or parent processes, we also record function calls such as *fork(2)*, *clone(2)* and *execve(2)* and emit those records as process events. Combining file events and process events, we can build a clear picture of how a protected file are accessed and transferred on a single machine, as shown at the left of Figure 2.

Trace Across Hosts In a distributed system, processes in different machines usually communicate with each other through network operation such as *connect(2)* and *accept(2)*. These operations are all recorded by KM which then emits them in the form of network events. These network events bridge the gap of file transfer between different machines, as shown at the right of Figure 2.

Note, however, that while we use only a triad *(Subject, Action, Object)* to model events, the data collected for each event contain sufficient information, such as execution time and process name, to chain up together all relevant events within a host and across hosts. A detail description of this process will be given in section 3.3.

TODO probability model

3 Implementation

3.1 System Overview

KM is implemented in userspace and relies on the LD_PRELOAD mechanism of the dynamic linker. In short, whenever a program starts, the dynamic linker will load the shared libraries listed in the LD_PRELOAD environment variable before any other libraries that the program depends on, including libc.so. After the dynamic linker is done with symbols relocation, symbols exported by shared libraries in LD_PRELOAD will “shadow” those with the same names in libc.so. Because libc is the common runtime which provides system call wrappers and common routines for application to interact with the operating system and most programs depends on libc.so, we can use this feature to add wrappers around libc’s and insert some interposition code between user program code and libc. This is illustrated in Figure 3.

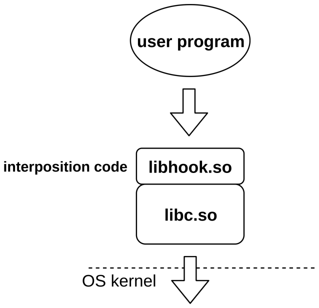

Figure 3: The LD_PRELOAD Mechanism

This schema is common for libc interposition [11]. However, to make the implementation safe and efficient is not easy. To make it safe, one have to adhere strictly to the semantic of every API and every piece of code should be multithread-safe and async-signal-safe. To make it efficient, only a small piece of code should be added, because many system calls (*e.g.*, *accept(2)*) will be invoked very frequently. In order to achieve these, our interposition code in every function wrappers contain only a few memory operations. This requires transforming most operations to memory operations. Therefore, a piece of shared memory is set up by the *KM agent* at the very beginning and attached by every hooked user program at program startup. After that, all operations, such as sending function invocation record, can be done through this shared memory. Also, there are configuration rules in the shared memory, which can be referenced by interposition code in every user program.

At program startup, the constructor of `libhook.so` will be invoked, which then allocates a piece of memory in the process space of the current process. This piece of memory is used to hold relevant information of the current process that may be needed afterward. The relevant information will be gathered by reading the `/proc` filesystem (*e.g.*, to get process starttime) or invoking proper system calls (*e.g.*, getpid()).

In the interposition code, all our wrappers will perform configuration checking, gather information about this particular function invocation and then send those information to the in-shared-memory message queues, which will be described in section 3.2. All these operations involve only some memory operations. Below is a simplified wrapper for *read(2)*:

```
int read_wrapper(int fd, void *buf,
                 size_t count) {

    int fd = read_original(fd, buf, count);
    if(fd < 0)
        return fd;
    if(!read_specified())
        return fd;
    if(!check_conf())
        return fd;
    send_read_info();
    return fd;
}
```

Basically it invokes the original *read(2)* in libc first, and then check configurations in shared memory to see whether or not this file is under monitor and thus operations upon it should be recorded. Because the configuration checking is done through the shared memory, which is attached at program startup, and we employ some hashing techniques (described in section 3.5) for fast searching, it should be fast and safe. After that a traid of file event, as described above in section 2, will be send to message queues in shared memory. Note that we do not have to make explicit effort to obtain information of the current process, because we already have that information stored at some place at the very beginning. In another word, we "defer" the process of information collection to program startup to speed up the interposition code. The `read_specified()` is specified to this particular wrapper. It is used to filter out some useless wrapper for optimization. It will be discussed in section 3.5.

On the other side, there is a *KM agent* which keeps periodically polling the message queues in shared memory to collect function invocation records produced by the interposition code in `libhook.so`. There is a protocal between the producers (*i.e.*, the interposition code) and the consumer (*i.e.*, agent) to make this message transfer efficient. We describe that in detail in section 3.2.

All these collected function invocation records are then sent to central service for analysis.

3.2 Efficient Message Passing

One key component in this system is the in-shared-memory message queues which is used to pass message from producers (*i.e.*, the interposition code) to consumers (*i.e.*, *KM agent*)). It is designed as such that message transfer are

1. **Non-blocking**. Because a blocking implementation will not only slow down the interposition code, but also probably change the semantic of some functions that we are going to hook.

2. **Atomic**. By atomic it means message from on process will not get interleaved with messages from any other processes, even though they share the same message queue.

3. **Fast**. In order to reduce overhead, the implementation should be as fast as possible such that it has negligible impact on existing services.

As of this writing our implementation meets all the three requirements above, with a overall overhead of about 1 *us* each function wrapper on our [–machine specification –].

In the shared memory, there are several message queues, the number of which is equal to the number of CPU. Upon program startup, each process will attach to the message queue belonging to the CPU where the process run. Since most processes will spend most of their lifetime running on the same CPU [**?**], assigning each process a "local" message queue will greatly reducer false-sharing, thus boosting performance of the whole system.

Each message queue is essentially a ring buffer [12, 13] with a *index* that can be advanced by both the consumer and producers:

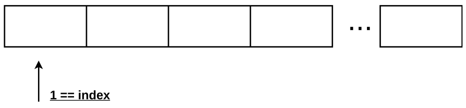

Figure 4a: A single message queue

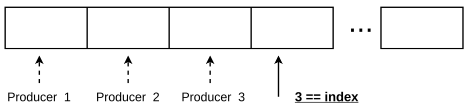

Figure 4b: After three fetch-and-add operations

There are only one consumer (*i.e.*, *KM agent*) but many producers (*i.e.*, the hooked processes). Upon sending message, producers will advance the *index* using the fetch-and-add atomic operation such that every producer get its own slot. This process is illustrated in Figure 4b. Every producer simply take the module of the result of the fetch-and-add operation on the size of the queue to avoid index overflow. To collect messages from producers, the *KM agent* periodically polls all the message queues and read the slots one by one.

Synchronization Between A Consumer and A Producer. In each slot, a flag variable is used to synchronize between a consumer and any one of the producers. Before collecting information in any slot, the consumer will check whether that flag is set to SLOTFULL. If so, it will go ahead reading the slot, after which the flag is set to SLOTEMPTY. Before writing to any slot, the producer will check whether whether that flag is set to SLOTEMPTY. If so, it will go ahead writing information in that slot, after which the flag is set to SLOTFULL. Because there is only one consumer, there is no need to use any atomic operation on that flag. A ordinal read and write will suffices. Therefore, synchronization between the consumer and producers is cheep and fast.

Synchronization Among Producers. Synchronization among producers is more tricky because simply taking the modulo of the result of the fetch-and-add operation on the size of the queue (to avoid index overflow) will probably result in two process collict in the same slot. To avoid this, an internal lock is setup and atomic operations are used to operation on it. This internal lock is a 8 bit integer so operation on it should be fast. The advantages of using this kind of home-brew lock with atomic operation is not only its speed and simplicity, but also that it can be turned into a robust lock so that any process which accidentally dies while holding this lock will not prevent the corresponding slot being used by other process. To achieve this robustness, we use fetch-and-add atomic operation for locking and take advantage of the automatic wrap-around effect of an unsigned integer. When a process exits without unlocking this lock, other processes which try to lock this lock with fetch-and-add operation will eventually overflow the underlying unsigned integer to zero, effectively unlocking it. Because we use a single byte as the underlying lock, 255 times of fetch-and-add would suffice. And because the message queue is long enough, the overflow will not happen too fast. The overall algorithm for this process is shown in Algorithm 1.

Algorithm 1 Synchronization Among Producers

```
procedure SYNCPRODUCER
    if $flag = SLOTFULL$ then
        return false
    if $0 \neq fetch-and-add(lock)$ then
        return false
    copy message to slot
    unset flag and lock altogether
    return true
```

3.3 Forward Tracing

Once we have data of all relevant events, we can chain up events to get a clear picture of how those events relate to each other, and more importantly, how a piece of data is transfered with hosts and across hosts. This procedure starts by the data under monitor (which we call the *tracingpoint*), and populates by a breath first search.

An example of this is given at Figure 2. The event order of it is:

1. Process A read file */etc/passwd* and then open a socket */var/sock.1* at time *T1*.

2. Process A fork a process B and write data to */var/sock.1* at time *T2*.

3. Process B read data from */var/sock.1* at time *T3*.

4. Process B connect a remote host at address *192.168.20.344:6677* at time *T4*.

5. Process C accept the connection from B at time *T5*.

where time *T1-5* are in ascending order.

Clearly these events are suspicious and should be put under administrators' attention.

To perform forward tracing, it is required that file */etc/passwd* is added to the monitor list at the very beginning. If it is added, then process A's read operation will be intercepted and its subsequent operations of opening a socket, forking a subprocess and writing data to that socket will all be under interception. Because the socket */var/sock.1* was opened by process A, which is under interception, the path */var/sock.1* will be added to the monitor list automatically by our interception code, such that process B's read operation on it will be intercepted. The connect and accept operations are monitored all the time so we can chain together process B and process C's events.

As described in the example above, once a read or write to a file in the monitor list is intercepted, we can start from that *tracingpoint* and perform a breath first search in our database to get a clear picture of how a piece of data is manipulated and transfered. Moreover, with a proper login control system, we can have information of which user starts which process at which time, and therefore can which users are involved in this series of events.

3.4 DSO Runtime Injection

Since the whole system works by libc interception, processes which are already running at the time of deployment will not be affected. To address this we can either reboot the machines or inject a library (in ELF format) into the memory space of the running process and change relevant function pointers to have them point to the corresponding pointers in the injected library. Since rebooting machines is not a universal solution and can do harm to many time-sensitive services, we develop the DSO runtie injection technique. Besides, being able to dynamically inject a dynamic shared object at runtime allows the system adminstrators to *partially* deploy the system. For example, some adminstrators might not want to intercept a long-running *logind* process. With the DSO runtime injection technique, they are free to do so.

The technique of DSO runtime injection consists of 1) injecting the shared object into the memory of a running process and 2) change relevant function pointers (in libc) to point to that in the injected object. Injecting the shared object can be done by using the *ptrace(2)* system call [**?**], as in most debuggers. Changinng function pointers can be done by analysing the current process's ELF structure, finding the correct pointers and changing them. The procedure of doing this is straightforward, but it is not easy to doing it in a safe way that does not crash a running program, especially when the program is compiled as PIE with RELRO [**?**]. Also, we have to guarantee that all our operations are safe multithread-safe and async-signal-safe such that the behavior of a running process will not be affected.

TODO more detail here.

3.5 Optimization

Describe the optimization techniques we use.

4 Experiments and Evaluation

1. the data output (how things are traced)
2. the performance (overhead)

5 Future work

6 Acknowledgments

A polite author always includes acknowledgments. Thank everyone, especially those who funded the work.

7 Availability

It's great when this section says that MyWonderfulApp is free software, available via anonymous FTP from

```
ftp.site.dom/pub/myname/Wonderful
```

Also, it's even greater when you can write that information is also available on the Wonderful homepage at

```
http://www.site.dom/~myname/SWIG
```

Now we get serious and fill in those references. Remember you will have to run latex twice on the document in order to resolve those cite tags you met earlier. This is where they get resolved. We've preserved some real ones in addition to the template-speak. After the bibliography you are DONE.

References

[1] Armbrust, M., Fox, A., Griffith, R., Joseph, A. D., Katz, R., Konwinski, A., Lee, G., Patterson, D., Rabkin, A., Stoica, I., and Zaharia, M. A view of cloud computing. *Commun. ACM 53*, 4 (Apr. 2010), 50–58.

[2] Barham, P., Donnelly, A., Isaacs, R., and Mortier, R. Using magpie for request extraction and workload modelling. In *Proceedings of the 6th Conference on Symposium on Opearting Systems Design & Implementation - Volume 6* (Berkeley, CA, USA, 2004), OSDI'04, USENIX Association, pp. 18–18.

[3] Barham, P., Isaacs, R., and Narayanan, D. Magpie: online modelling and performance-aware systems. In *9th Workshop on Hot Topics in Operating Systems (HotOS-IX)* (Lihue, Hawaii, May 2003), USENIX, p. 8590.

[4] Chen, M., Kiciman, E., Fratkin, E., Fox, A., and Brewer, E. Pinpoint: problem determination in large, dynamic internet services. In *Proceedings of the 6th Conference on Symposium on Opearting Systems Design & Implementation - Volume 6* (2002), DSN '02, IEEE, pp. 595–604.

[5] Dean, J., and Barroso, L. A. The tail at scale. *Commun. ACM 56*, 2 (Feb. 2013), 74–80.

[6] Dean, J., and Ghemawat, S. Mapreduce: Simplified data processing on large clusters. *Commun. ACM 51*, 1 (Jan. 2008), 107–113.

[7] Fonseca, R., Porter, G., Katz, R. H., Shenker, S., and Stoica, I. X-trace: A pervasive network tracing framework. In *Proceedings of the 4th USENIX Conference on Networked Systems Design &#38; Implementation* (Berkeley, CA, USA, 2007), NSDI'07, USENIX Association, pp. 20–20.

[8] King, S. T., and Chen, P. M. Backtracking intrusions. In *Proceedings of the Nineteenth ACM Symposium on Operating Systems Principles* (New York, NY, USA, 2003), SOSP '03, ACM, pp. 223–236.

[9] King, S. T., Mao, Z. M., Lucchetti, D. G., and Chen, P. M. Enriching intrusion alerts through multi-host causality. In *Network and Distributed System Security Symposium* (2005), NDSS '05.

[10] Krentel, M. W. Libmonitor: A tool for first-party monitoring. *Parallel Comput. 39*, 3 (Mar. 2013), 114–119.

[11] Lee, H.-c., Kim, C. H., and Yi, J. H. Experimenting with system and libc call interception attacks on arm-based linux kernel. In *Proceedings of the 2011 ACM Symposium on Applied Computing* (New York, NY, USA, 2011), SAC '11, ACM, pp. 631–632.

[12] Lee, P. P. C., Bu, T., and Chandranmenon, G. A lock-free, cache-efficient shared ring buffer for multi-core architectures. In *Proceedings of the 5th ACM/IEEE Symposium on Architectures for Networking and Communications Systems* (New York, NY, USA, 2009), ANCS '09, ACM, pp. 78–79.

[13] Mitropoulou, K., Porpodas, V., Zhang, X., and Jones, T. M. Lynx: Using os and hardware support for fast fine-grained inter-core communication. In *Proceedings of the 2016 International Conference on Supercomputing* (New York, NY, USA, 2016), ICS '16, ACM, pp. 18:1–18:12.

[14] Saito, Y. Jockey: A user-space library for record-replay debugging. In *Proceedings of the Sixth International Symposium on Automated Analysis-driven Debugging* (New York, NY, USA, 2005), AADEBUG'05, ACM, pp. 69–76.

[15] Shanahan, J. G., and Dai, L. Large scale distributed data science using apache spark. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (New York, NY, USA, 2015), KDD '15, ACM, pp. 2323–2324.

[16] Sigelman, B. H., Barroso, L. A., Burrows, M., Stephenson, P., Plakal, M., Beaver, D., Jaspan, S., and Shanbhag, A. Dapper, a large-scale distributed systems tracing infrastructure.