Houdini’s Escape: Breaking the Resource Rein of Linux Control Groups

至彤

2020 年 7 月 18 日

1 摘要

LCG，也叫cgroups是操作系统级容器化的基础，cgroups机制将进程划分成有层次的进程组，并且应用不同的控制器来管理系统资源（CPU，内存，块IO），子进程会自动复制父进程的cgroups属性进行资源控制。但是这种情况下，子进程的cgroups限制和资源记账（resource accounting）在特殊情况下会失效。

2 基础

2.1 容器资源隔离

容器资源隔离和管理基础是Linux Namspaces和Linux Control Groups，容器部署安全的机制包括Capabilities，SELinux，AppArmor，Seccomp和Security Namespace。

2.2 cgroups层次和controller

cgroups是树状结构，线程只能关联至每个层次结构的一个cgroups，但是课关联至多个层次结构，每个层次结构有多个子系统关联。所以cgroups机制能限制进程组的总资源。

2.2.1 cpu controller

2.2.2 cpusets controller

2.2.3 blkio controller

2.2.4 pid controller

2.3 cgroups继承性

Fork或Clone的子进程和父进程的cgroups相同

3 Exploiting策略生成额外工作负载

3.1 内核Upcalls

利用内核线程启动新进程，避免cgroups资源限制。在用户空间的进程启动内核方法（在内核空间），比如用户空间的系统调用例子：usemode helper API

3.2 内核线程

例子：kthreadd，kworker，ksoftirqd，migration，kswapd

3.3 服务进程

服务进程：进程管理，系统信息日志，debug信息

3.4 中断上下文

硬中断，软中断softirqs（ksoftirqd）

4 案例分析

5个案例分析现实情况下Docker 容器破坏cgroups的资源限制，进一步在多租户的容器环境下放大资源消耗从而攻击其他容器。

4.1 异常处理

调用usermode helper API，通过异常触发用户空间进程，消耗200倍CPU资源，降低同宿主机容器性能85

4.2 数据同步

磁盘数据同步的writeback机制，系统调用：sync，syncfs，fsync，其中sync能降低系统级别的IO性能，造成Resource-Freeing Attack (RFA)和隐藏信道

4.3 系统进程-Journald

内核日志，系统日志与审计记录有系统进程journald记录，su（切换用户命令），添加用户或用户组，exception

4.4 容器引擎

增加内核线程（kworker）和容器引擎的工作负载，资源消耗可高达三倍

4.5 softirq处理

4.5.1 NET softirq—NIC在包传输后触发，可被iptables放大

4.5.2 Block softirq–块设备IO中断

5 实验

在Amazon EC2云环境下进行实验，容器可以消耗200倍以上的资源，减少百分之九十五co-resident容器的计算和IO资源。

6 相关工作-容器安全

* 系统调用限制 [*Speaker: Split-Phase Execution of Application Containers*]
* 基于内存的伪文件系统攻击 [*A Study on the Security Implications of Information Leakages in Container Clouds*]
* Namespaces安全策略冲突处理 [*Security Namespace: Making Linux Security Frameworks Available to Containers*]
* 利用Intel SGX 构建secure Linux containers [*SCONE: Secure Linux Containers with Intel SGX][SCONE: Secure Linux Containers with Intel SGX*]
* 容器内的隐藏信道攻击 [*Whispers between the Containers: High-Capacity Covert Channel Attacks in Docker*]

7 相关工作-云安全

* Co-residence
  - 侧信道攻击 [*Exploring Information Leakage in Third-Party Compute Clouds*]
  - 隐藏信道攻击 [*4k-Aliasing Covert Channel and Multi-Tenant Detection in IaaS Clouds*]
  - 资源计费问题 [*Peeking behind the curtains of serverless platforms*]
* DOS攻击
  - 内存，IO攻击 [*An Experimental Study of Cascading Performance Interference in a Virtualized Environment*]
  - resource-freeing攻击 [*Resource-Freeing Attacks: Improve Your Cloud Performance*]