vkernel 项目开发文档

一、面向轻量级虚拟化的优化技术项目概述

1.1 项目背景

容器作为一种轻量级虚拟化技术，具有启动快速、部署方便、资源占用少、运行效率高等优点，近年来广泛应用于云计算平台和数据中心的资源管理、系统运维和软件部署中，不仅成为云计算领域的业界热点，而且具有改变软件整体生态的潜力。作为业界最具代表性的云计算平台，亚马逊云、谷歌云、微软云和阿里云等都先后在云计算业务中推出各自的容器云服务。最近容器技术开始走出数据中心，除了传统的微服务、DevOps 等领域，已逐步应用于包括移动计算、物联网等在内的新领域，表现出了良好的发展态势和应用前景。

大量的云服务主机采用的是 Linux 操作系统，Linux 原生提供了命名空间（Namespace）和控制组（Cgroup）等技术来实现系统层级的环境虚拟化。命名空间为每个容器提供独立的操作系统视图，让容器中的应用感觉完全在独立的环境中运行。控制组则对每个容器可以获取到的系统共享资源进行分配与管理。虽然命名空间与控制组为容器提供了隔离性，但是由于所有的容器仍然是共享主机系统的内核，容器在安全性方面的防护仍旧较弱。共享内核一方面为容器带来高效性，但是另外一方面也给容器带来了诸多的安全隐患。该问题成为制约容器发展和应用的主要技术瓶颈，目前业界有以下几种主流的方案：

（1）基于主机内核的隔离方案：容器利用 Linux 内核提供的命名空间和控制群组来实现容器的视图隔离和资源隔离，利用 Capability、Seccomp、Apparmor 机制实现容器的安全隔离。

（2）基于 gVisor 的用户态内核：gVisor 为容器实现了一个用户态内核，使用进程虚拟化的方法隔离容器内的恶意代码对 Host Kernel 的访问。它在内核之外实现了一个“内核进程”——Sentry，Sentry 提供了大部分 Linux Kernel 的系统调用。并且通过巧妙的方式将容器内进程的系统调用转化为对这个“内核进程”的访问，实现对系统调用的重定向和重构。

（3）基于轻量级虚拟机的 guest 内核：比较典型的是 Kata 容器，Kata 容器

基于 Hypervisor 技术为每个容器虚拟化出一个经过裁剪的独立内核，让每个容器运行在一个轻量级的虚拟机中，解决了传统容器共享内核的安全和隔离问题。

第一种方案本质上还是多个容器共享服务器操作系统内核，虽然效率高，但隔离性最弱，而后两种方案虽然在隔离性和安全性上有了很大的提升，但它们均以牺牲性能为代价。传统容器隔离性不足的本质在于多个容器共享同一个操作系统内核，内核如果发生漏洞或崩溃会导致所有容器无法正常运行，且统一的内核使用户无法根据容器的特点进行内核服务的定制，这导致多容器场景下，容器安全性差、灵活性差、稳定性差的问题。

针对上述问题，本项目拟进行的工作是：实现既保证隔离性又拥有近似于原生 Docker 容器性能的容器技术。

1.2 预期目标

本项目旨在对现有轻量级虚拟化技术进行优化，改进目前容器技术存在的安全性和灵活性问题。项目的预期实现目标主要有：

低启动延迟：在实现较强的隔离性与安全性的前提下，能够实现接近于原生 Docker 容器的启动延迟，不会为项目部署带来严重的额外开销。

高性能：不同于 gVisor 和 Kata 等方案以牺牲性能为代价换取隔离性与安全性，本项目希望能够实现接近原生 Docker 容器的负载性能，保持较高的资源利用率。

高安全性：本项目相比原生容器具有较高的安全性，能够防护容器中常见的 CVE 提权漏洞，一定程度上保护运行中的容器和宿主机的安全。

性能隔离：原生容器共享主机内核，因此不同容器之间存在一定的性能干扰，会导致容器无法发挥预期性能并产生一定的安全风险。本项目希望设计一种容器性能隔离机制，能够实现容器之间的性能隔离，减少负载之间的性能干扰。

1.3 系统架构

我们的项目名称为 vkernel（virtual kernel），它使用内核可加载模块技术在内核层实现虚拟内核，旨在打破容器场景中共享内核带来的安全限制，具有安全性强、性能高的特点。如图 1-1 所示，vkernel 由以下几个部分组成：

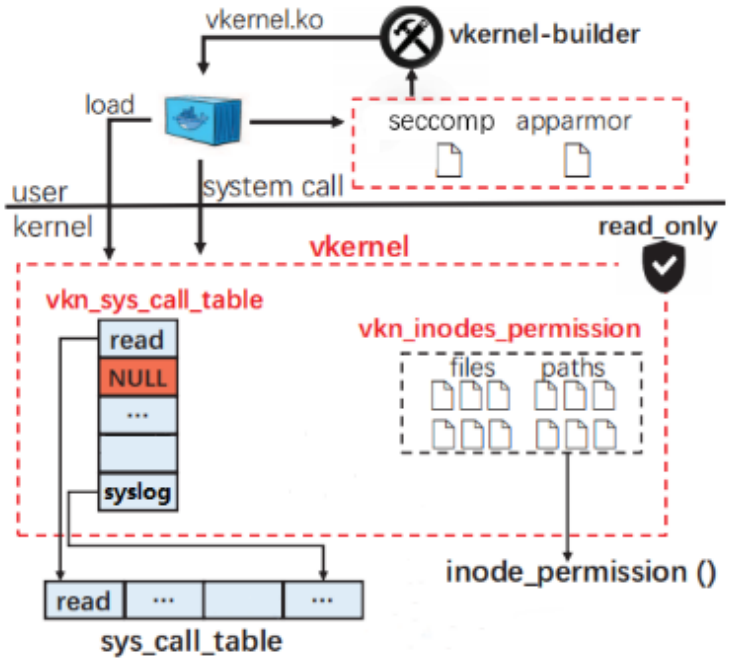

图 1-1 vkernel 系统整体架构图

内核模块隔离系统：模块系统是 vkernel 的关键，实现了容器内核资源的虚拟化和安全。Linux 内核被所有容器共享的，这种共享带来了高性能的同时，也给容器的隔离性带来了巨大的挑战。虽然内核已经面向容器增加了诸多隔离功能，但是内核层的隔离仍然不够完善，需要一步一步进行提升。

系统调用虚拟化系统：系统调用的共享对容器安全性影响十分明显，一方面多个容器竞争调用相同系统调用会降低该系统调用的性能，另一方面恶意容器可以通过篡改系统调用入口进行恶意操作，破坏系统安全性。设计面向容器环境的 Linux 系统调用虚拟化系统对于容器的安全性非常重要。

文件访问控制模块：实现基于 inode 虚拟化的文件保护，在保证内核安全性与强制访问控制能力的同时，提升内核的运行效率，提高系统整体的文件访问速度。

内核定制工具：用于构建 vkernel 模块的工具。它分析容器镜像的特点，并根据 seccomp、apparmor 规则自动构建 vkernel 模块。

二、面向容器环境的 Linux 系统调用虚拟化

2.1 背景简介

随着云计算技术的发展，容器由于其轻量性和高效率等特点，受到越来越广泛的重视，但是由于所有的容器共享主机系统的内核，容器在安全性方面的防护仍旧较弱，具有许多的安全隐患。

在诸多安全隐患中，系统调用的共享对容器安全性影响尤其明显。Linux 内核只提供了统一的系统调用入口以及实现了一致的系统调用。一方面，由于系统调用是共享的，当多个容器竞争调用相同系统调用的时候，产生的剧烈竞争比较容易降低该系统调用的性能。另外一方面，恶意的容器可以通过篡改系统调用入口进行恶意操作，或者通过触发系统调用漏洞产生提权甚至引发宕机。除此之外，某些系统调用申请的资源是有限的（比如内核 futex 锁，内核只在初始化时开辟了有限空间大小的哈希表管理锁），恶意容器可以通过恶意占有系统调用资源来让系统拒绝为其他容器提供服务。

现有的容器虚拟化技术，如 gVisor 和 Kata Containers 等，虽然通过使用用户态内核或者 KVM 虚拟化的方式提高容器的安全性和隔离性，但它们都引入了不容忽视的开销，无法满足现有的容器轻量性和隔离性的需求。

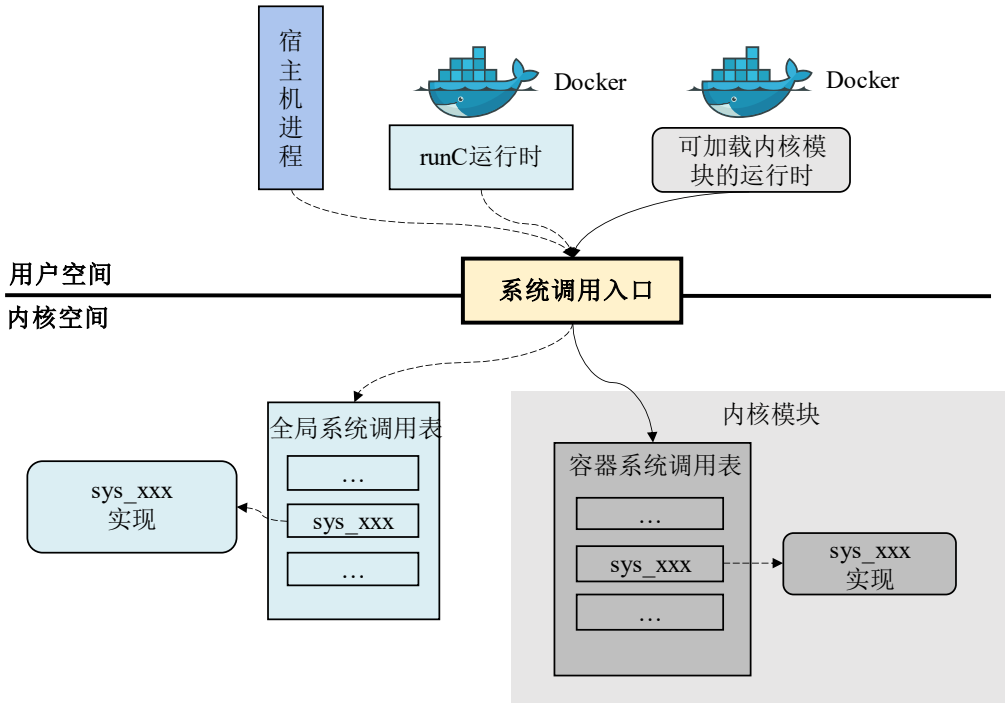

图 2-1 面向容器环境的 Linux 系统调用虚拟化实现架构图

针对以上问题，提出设计一种面向容器环境的 Linux 系统调用虚拟化系统，同时借助内核模块灵活的可插拔性进行实现。

2.2 功能实现

2.2.1 可加载内核模块的容器运行时实现

面向容器环境的 Linux 系统调用虚拟化系统以内核模块的方式实现。为了加载内核模块，需要在容器开始启动后、容器内部应用进程真正运行前，调用内核模块，完成相应功能的初始化以及模块与容器进程的绑定。同时，内核模块要与容器一一对应，伴随着容器的启动停止而动态加载与销毁。

2.2.2 双重 capabilities 保护

容器创建后，只会使用到少量的 capabilities（默认 14 个）。内核内部有一套自己的 capabilities 安全机制。通常情况下，这些有限的 capabilities 集合在内核的 capabilities 机制下可以保证大部分容器的正常运行，同时也能提高容器环境的安全性。但目前公开的一些漏洞显示，在容器环境下，这些漏洞可以通过利用部分内核 bug 进行提权和逃逸，从而摆脱容器的安全限制，进而可能危害宿主机。

vkernel 运行时在容器应用进程创建前会记录容器初始的 capabilities。之后，容器的所有需要进行权限验证的行为会在内核 capabilities 验证的基础上进行 vkernel 的二次验证，保证容器的所有行为都在初始时的 capabilities 约束集合内，避免越权的行为发生。

2.2.3 容器系统调用表虚拟化实现

容器的系统调用表通过内核虚拟化不同的系统调用入口，移除了容器对全局系统调用表的访问权限，代之以每个容器拥有的位于内核空间不同区域的独立系统调用表。内核模块在初始化后，会在内核空间中开辟一块独立的内存区域，用于存放容器的系统调用表。

容器的系统调用表与全局的系统调用表结构相同，但部分内容有所不同。对

于需要进行参数检测的系统调用，会修改系统调用表的表项，对系统调用函数进行二次封装；对于需要屏蔽的系统调用函数，直接返回-1；对于涉及到隔离的子系统资源，会修改重定向该系统调用函数，将其指向内核模块内部的子系统；其余的则默认使用原来的全局系统调用函数。

三、面向容器的内核模块隔离系统

3.1 背景简介

目前，在 Linux 环境下，容器主要采用 Linux 内核提供的命名空间和控制组机制实现隔离和限制。但是本质上，Linux 内核仍然是被所有容器共享的。这种共享带来了高性能的同时，也给容器的隔离性带来了巨大的挑战。虽然内核已经面向容器增加了诸多隔离功能，但是内核层的隔离仍然不够完善，需要一步一步进行提升。

本系统主要关注于面向容器的内核模块的隔离。内核模块是一种支持按需加载或卸载的内核代码的机制，可以应用于为容器定制内核功能与访问权限，达到在内核层实现容器强隔离的目的。但是，由于内核共享，内核模块机制本身存在诸多隔离性与安全性的问题。本系统主要面向基于内核模块的容器隔离性解决方案，进行内核模块的隔离以及安全保护。这其中存在的挑战主要在于内核模块机制缺乏容器感知，如何隔离各个内核模块并与相应容器绑定。

为了克服容器场景下内核模块存在的挑战，我们实现了面向容器的内核模块隔离系统。该系统可以在内核中绑定容器与相应模块并克服模块重命名问题。面向容器的内核模块隔离系统充分结合现有内核模块管理机制和容器特性，针对不同的容器加载的内核模块提供安全的内核环境。

3.2 功能实现

3.2.1 内核模块虚拟化

容器通过系统调用与内核交互，在内核空间中，多个容器共享相同功能的内核模块，容器之间会产生竞争，功能模块性能受到严重影响。在这种情况下，需要给相应容器定制一个容器内核模块，提供内核功能和访问权限。一方面，在内核加载模块过程中需进行模块重名检测，如已存在名称相同的内核模块，则该内核模块不能被加载。另一方面，容器内核模块与容器对应，需要在创建容器实例前加载容器内核模块，并与相应的容器 init 进程绑定。如何在内核中绑定容器与

相应模块并克服模块加载重名检测机制是本功能需要解决的关键问题。

内核模块虚拟化是指为内核模块添加容器感知，能够隔离各个内核模块并与相应容器绑定，可以应用于为容器定制内核功能和访问权限。

具体做法是 runC 将容器内核模块文件读取到内存中，根据 ELF 文件格式进行解析，根据模块特定字段名称，定位到内存中容器内核模块名称字符串地址，接着获取容器进程的容器 id，将模块名称修改为模块名加上容器 id，保证容器内核模块名称的不重复。将模块中核心数据添加到 namespace 中，实现内核模块与容器的绑定。

四、基于 inode 虚拟化的文件访问控制模块

4.1 背景简介

本项目是 vkernel 项目的核心组成部分，属于 vkernel 的安全防护模块，主要实现的是基于 inode 虚拟化的文件保护。项目的背景是目前 Linux 的内核强制访问机制如 AppArmor、SELinux 等在为文件访问提供审核管控时，所有文件在被访问时都需要进行相应规则的权限检测，会产生较大的性能开销，资源利用率低。另外，这些文件访问控制也有较为复杂的一套匹配规则，在面临大量的文件访问操作时会带来较大的性能开销，而且在容器内部用户默认拥有 root 权限，可以通过更改 profile 文件来更改 AppArmor 等的相应规则，具有安全隐患。

为了解决上述问题，本项目基于 vkernel 高效高安全性的设计思想，提出了一种基于 inode 虚拟化的文件保护机制，利用内核中已有的权限检测机制，针对 inode 进行虚拟化，只对需要保护的文件进行访问的审核控制，在保证内核安全性与强制访问控制能力的同时，大大地提升了内核的运行效率，提升了系统整体的文件访问速度。

4.2 功能实现

4.2.1 整体架构

为尽量保证 Linux 内部的数据结构不变，选取 inode 中的 i_mode 字段中未被使用的比特位作为标识位，用于标识该文件或目录是否在 vkernel 中有相应的权限限制，而每个容器的 vkernel 在初始化时都会初始化一个哈希表，key 为文件 inode 号，value 为相应文件的访问权限，并根据用户自定义的配置文件，将权限信息存储在该哈希表内。当文件被访问时，先检查相应标志位，当标志位为 0，文件在 vkernel 中并未被限制访问，直接通过，当标志位为 1，文件在 vkernel 中有访问限制，则查询哈希表，进行权限检测。

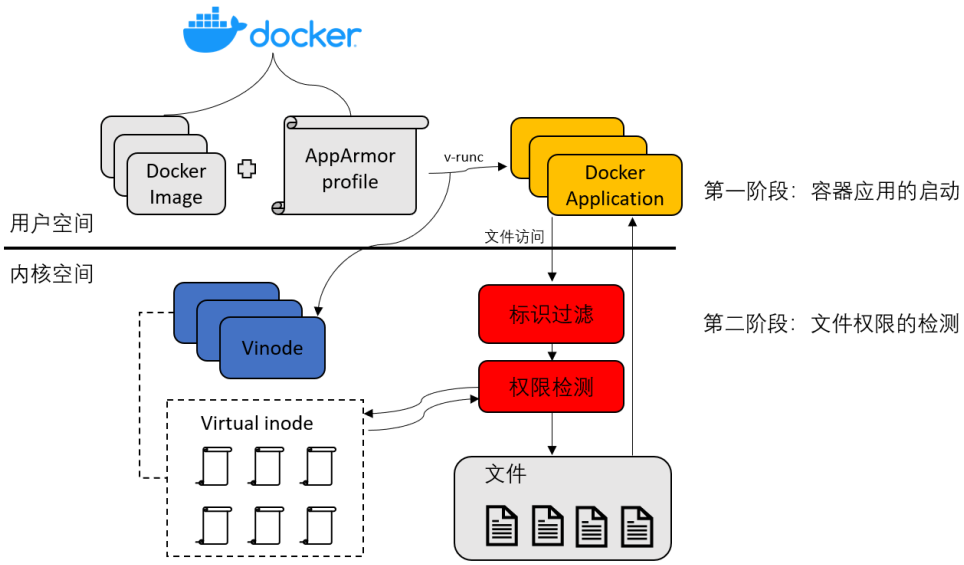

图 3-1 inode 虚拟化整体架构

4.2.2 对单个文件的权限检测

取一位 i_mode 未被使用的比特位，1 表示 vkernel 内有该文件的权限限制，0 表示 vkernel 内无权限限制，在 Linux 内核中的通用权限检测函数 generic_permission 中，添加基于 inode 虚拟化的 vkernel 权限检测，将文件访问申请的权限与 vkernel 内哈希表中的信息进行匹配检测。

4.2.3 对目录的权限检测

在 generic_permission 中对目录的 inode 的权限检测只是针对该目录的权限进行限制，而并非对目录下的文件进行相应的权限限制，结合实际应用中针对目录下文件的权限限制多是针对 procfs、sysfs、cgroupfs，于是取另一位 i_mode 未被使用的比特位用于表示目录在 vkernel 内是否有权限限制，然后在初始化一张新的哈希表用于存储目录的权限控制信息，在 generic_permission 中，针对 procfs、sysfs、cgroupfs 文件系统的文件，会逐级向上查看目录是否有相应的权限进行检测。

五、基于容器镜像的最小化内核定制工具

5.1 背景简介

与运行自己的操作系统的虚拟机相比，用户可以在主机的同一操作系统内核之上启动多个容器，这使得容器更加轻巧，有着更好的资源利用率和更好的性能。但是，容器性能的提升是以隔离度较弱为代价的。由于主机上的容器都共享同一内核，因此容器彼此之间的隔离是进程级别的，只能由底层操作系统内核通过软件机制来保证。因此，如果有攻击者有权通过某个容器来访问主机的内核并对内核的漏洞加以利用，就会对主机和其他容器的安全带来极大的威胁。

尽管操作系统提供了严格的软件隔离机制试图解决这一问题，例如 Linux 中用以进行细粒度进程权限控制的 Capability 和用于资源视图隔离的 namespace 等等，但是，目前的 namespace 仍不支持对系统调用进行视图隔离，因此，恶意的租户仍然可以通过系统调用访问共享内核，并利用内核漏洞绕过这些隔离机制。例如，“waitid”系统调用中的漏洞（CVE-2017-5123）允许恶意用户运行特权升级攻击，并逃脱容器以获得对主机的访问权限。

vkernel 本质上是一个内核模块，旨在替代 Linux 内核中 seccomp 等的安全机制，减少性能开销，增强隔离性。每个容器都有一个自己专用的 vkernel 为自己服务。容器只能通过 vkernel 来访问主机资源，因此可以在 vkernel 中对容器的行为进行一些限制，提高容器的隔离性。

但是，vkernel 是专用的，不同镜像，甚至同一镜像的不同容器，他们所需要的内核功能也是不一样的。如果每次启动容器时都要手动修改代码，编译模块，那用户的使用体验就会大大降低，也不利于实际应用和自身的推广。所以需要一个 vkernel 的自动构建工具，用户只需要提供容器镜像信息及其对应的配置文件，就能分析容器所需要的最小内核功能，并根据分析结果，自动生成容器专属的 vkernel 模块的代码，构建出目标容器专用的 vkernel 模块。

5.2 功能实现

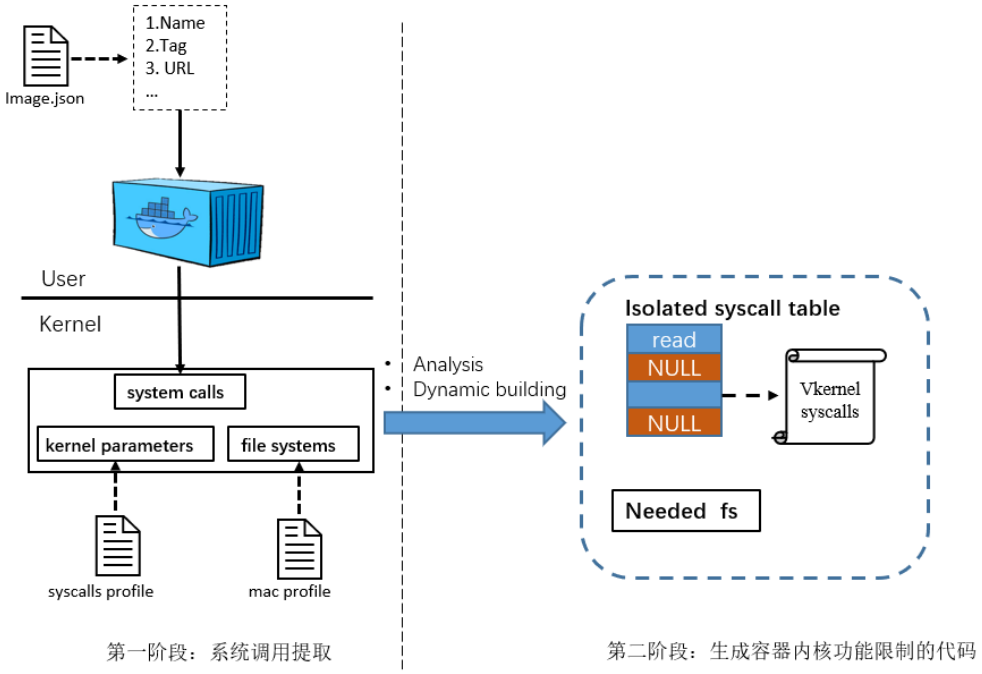

图 5-1 基于容器镜像的最小化内核定制工具架构图

5.2.1 容器镜像系统调用提取

该模块用以对用户提供的容器镜像进行动态分析，以提取其所需要用到的系统调用，供后面生成限制性策略的模块使用。

具体实现上，首先需要对用户提供的包含容器镜像信息的 json 文件进行解析，获取容器镜像的名称，版本号，URL 等信息，随后将主机上现存的容器全部停止并删除，再启动目标容器，保证主机上只有目标容器在运行。然后启动 sysdig 监控程序，设定运行时间，并将这段时间内监控的结果以文本的形式保存在指定目录下。最后对监控结果进行分析，提取与目标容器相关的条目，截取系统调用相关的字段并保存。

5.2.2 自动构建 vkernel 模块

工具会对第一阶段的容器镜像分析结果进行解析，结合用户提供的容器的系统调用和文件访存的配置文件，并生成对应的限制性策略的代码。

具体来说，针对系统调用配置文件，由于文件定义了容器允许访问哪些系统调用，因此只需要对配置文件进行解析，并结合第一阶段提取的系统调用，两者取并集，就能得到允许容器访问的系统调用的结合。其次，配置文件还定义了哪些系统调用在访问时需要进行参数检测，因此系统还会对这些系统调用进行重新封装，插入参数检测的代码，用重新封装好的系统调用去替换系统调用表上的默认的系统调用，实现限制系统调用的策略。而针对文件访存的配置文件，文件中的条目定义了容器对于某个路径下的文件是否具有各类权限（读、写、链接权限等），因此工具会对配置文件进行解析。由于文件路径是 glob 正则表达式，因此首先需要对路径进行解析，得到 Linux 下的标准路径，并且根据路径去读取出该路径下的所有文件或文件夹，随后使用位图来表示这些文件对应的权限。在得到这些信息之后，将文件及对应的权限以键值对的形式存入哈希表中，并自动生成哈希表初始化的代码，容器在访问这些文件时，需要先去哈希表中进行查询，确认自己持有相关权限才可访问，进而达到限制容器对文件进行访存的目的。当系统调用限制和文件访存限制两部分代码生成之后，工具会自动执行编译指令，对 vkernel 源码进行编译，生成.ko 模块文件，供容器运行使用。