



# Multi-Granularity Container Runtime Mandatory Access Control

**Ms. Hrutika Jare, Prof. Chaudhari .V.S, Dr. S.A. Agrawal**

Dept. of Computer Engineering, Vishwabharati Academy's College of Engineering, Ahilyanagar, Maharashtra, India

Dept. of Computer Engineering, Vishwabharati Academy's College of Engineering, Ahilyanagar, Maharashtra, India

Dept. of Computer Engineering, Vishwabharati Academy's College of Engineering, Ahilyanagar, Maharashtra, India

**ABSTRACT:** With the widespread adoption of container-based cloud computing, particularly through Kubernetes, the automation of application deployment and management has advanced significantly. However, this evolution has heightened concerns regarding container runtime security. Traditional mechanisms such as Linux Capabilities, Seccomp, and Linux Security Modules (LSMs) offer limited protection in containerized environments due to their static configurations and coarse-grained control. Emerging eBPF-based approaches provide enhanced observability and flexibility but struggle with real-time policy updates and enforcement when new kernel vulnerabilities arise. This study examines the limitations of existing container security mechanisms and classifies access control into three categories: system call, LSM hook, and kernel function access control. To address these issues, the paper proposes an eBPF-integrated multi-granularity access control technique that leverages Kubernetes metadata. Experimental validation using CVE-2022-0492 demonstrates that the proposed system, BPFGuard, enhances container security with an average performance overhead of only 2.16%

**KEYWORDS:** Container Security, Enforcement, Kubernetes, EBPF.

## I. INTRODUCTION

Container technology has become a cornerstone of modern cloud computing due to its lightweight virtualization, efficient deployment methods, and simplified management. Among various orchestration frameworks, Kubernetes has emerged as the de facto standard for managing and scaling containerized applications across distributed environments. However, as containers continue to proliferate, security challenges have become increasingly prominent, particularly in Kubernetes clusters, where the complexity of multitenant deployments and dynamic workloads introduces new attack surfaces. Early research efforts sought to enhance container security by directly applying traditional Linux security mechanisms such as Capabilities [1], Seccomp [2], and Linux Security Modules (LSMs) [3]. While these mechanisms provide essential kernel-level isolation and access control, they were not specifically designed for containerized ecosystems. Their rigid configurations and limited adaptability often restrict both the effectiveness and practical applicability of these controls within Kubernetes-based infrastructures [4].

To overcome these limitations, researchers have explored the use of extended Berkeley Packet Filter (eBPF) technology for improving runtime observability and control within the Linux kernel. eBPF allows dynamic insertion of monitoring and security logic without modifying the kernel source code, making it an appealing choice for container runtime protection. Recent works [5] have demonstrated eBPF's potential for access control and attack mitigation, particularly by monitoring system calls and kernel functions to detect malicious behaviors post-exploitation.

Nevertheless, existing eBPF-based access control methods face notable constraints. They require predefined hook points and manually implemented callback functions, which limit their adaptability to evolving security requirements. For example, when new vulnerabilities (such as recently disclosed CVEs) emerge, it becomes difficult to update policies in real time for immediate enforcement. Additionally, current solutions often lack a unified framework that can effectively combine system call filtering, LSM hook enforcement, and kernel function interception to achieve comprehensive security coverage. To address these challenges, this paper introduces BPFGuard, an eBPF-based mandatory access control (MAC) system designed specifically for container environments. BPFGuard classifies container monitoring and control mechanisms into three levels

system calls, LSM hooks, and kernel functions. It leverages eBPF to construct a multi-granularity runtime access control framework, tightly integrated with Kubernetes to collect relevant metadata and enforce context-aware security policies.

Furthermore, BPFGuard supports the dynamic interception of kernel functions and enables the execution of tailored actions, such as preventing privilege escalation and managing namespace capabilities, thereby enhancing protection against kernel-level attacks. Through this design, BPFGuard aims to deliver fine-grained, flexible, and real-time security enforcement for containerized environments without sacrificing performance or scalability.

## II. RELATED WORK

Several studies have focused on enhancing container runtime security using Seccomp, Linux Security Modules (LSM), and eBPF.

Seccomp-based approaches include Lei et al. [26], who proposed a phased container execution model to dynamically reduce available system calls across application stages, minimizing attack surfaces. Wan et al. [27] used automated testing to generate Seccomp policies that restrict system calls to those observed during test phases. Similarly, Ghavamnia et al. [28] employed static code analysis to identify necessary system calls and generate optimized Seccomp policies.

LSM-based approaches have also been explored. Mattetti et al. [29] automated AppArmor policy generation during container image building, while Loukidi et al. [30] combined static and dynamic rule generation to strengthen Docker container isolation. Sun et al. [31] addressed LSM's shared nature by introducing Security Namespaces, enabling containers to maintain independent security policies.

Recent work combining LSM and eBPF includes KRSI [32], BPF-LSM [33], and Landlock LSM [34], which attach eBPF programs to LSM hooks for dynamic, high-performance policy enforcement. Findlay et al. [35][36] leveraged eBPF programmability to build runtime detection and restriction frameworks for containers, while CNCF projects such as Tracee and KubeArmor [37] use eBPF for observability and partial system enforcement.

Alternative solutions like gVisor [38] and Kata Containers [39] approach security via process-level and VM-level isolation respectively, trading flexibility for stronger isolation guarantees.

## III. BACKGROUND

### 3.1 Namespaces and Containers

Namespaces are a fundamental Linux kernel feature that provides isolation and management of system resources. They enable the division of global system resources into separate, independent domains, ensuring that processes within one namespace are unaffected by those in others. This isolation prevents interference and resource conflicts, contributing to system stability and security. Within the Linux operating system, namespaces can be used to isolate processes, networks, file systems, users, and other kernel resources, thereby improving resource allocation and containment of faults or malicious behavior.

Containers build upon namespaces to create lightweight, isolated environments at the operating system level. They encapsulate an application along with all its runtime dependencies, ensuring consistent execution across different environments. Docker, one of the most widely adopted container technologies, leverages Linux namespaces and control groups (cgroups) [6] to provide resource isolation and management. Each container runs as an isolated process with its own file system, process space, and network stack, separated from both the host and other containers.

Container technology offers several advantages. It ensures environmental consistency, resolving the common "it works on my machine" problem. Containers also enable rapid deployment, scalability, and efficient resource utilization, as they share the host kernel rather than emulating full hardware environments like virtual machines. By leveraging containers, developers can streamline application development, testing, and deployment processes. Furthermore, integration with orchestration frameworks such as Kubernetes allows for automated management of large-scale

container clusters, ensuring high availability, fault tolerance, and elastic scalability. Namespaces thus serve as the foundation for container isolation, while container orchestration technologies extend these principles to distributed cloud environments, enhancing both security and operational efficiency.

### 3.2 Linux Security Modules (LSM)

The Linux Security Module (LSM) framework provides a standardized method for extending and customizing kernel-level security policies within the Linux operating system. It decouples security decision-making from the kernel's core logic, allowing administrators and developers to implement modular and flexible security mechanisms tailored to specific use cases.

The primary objective of LSM is to enhance the adaptability and extensibility of Linux kernel security by allowing multiple security models to coexist. Administrators can enforce fine-grained access control and define custom security policies for system resources such as files, processes, and sockets by loading specific LSM modules.

LSM operates based on several key principles:

- Hook Mechanism:

LSM inserts hooks at critical points within the kernel, such as during process creation, file access, or system call execution. These hooks serve as interception points where security checks can be performed.

- Hook Invocation:

When a hooked operation occurs, the kernel invokes the corresponding LSM function from the active security module. This function applies the relevant security checks—such as permission validation or policy enforcement—before the operation is completed.

- Security Decisions:

The hooked function determines whether to allow, deny, or modify the requested operation based on predefined rules, process credentials, and object labels. This ensures that only authorized actions are executed according to policy.

- Security Policies:

Different LSM modules implement various security models. For instance, SELinux [7] enforces Mandatory Access Control (MAC) using labeled policies, while AppArmor employs profile-based restrictions on process behavior and resource access. Depending on system requirements, administrators can choose the appropriate LSM module and configure custom policies.

By abstracting kernel security through a modular interface, LSM allows for extensible, policy-driven access control, forming the foundation for many modern container and host security solutions.

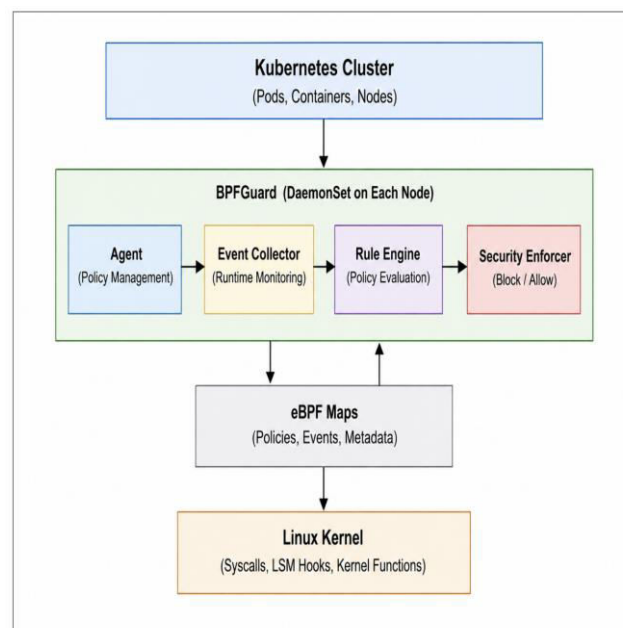


Fig 1: System Architecture

Intruders perform lateral movement within the Kubernetes cluster through the compromised container.

#### IV. GOALS

The primary objective of this work is to develop a comprehensive container runtime security system leveraging eBPF technology, designed for seamless integration with Kubernetes. The proposed system aims to overcome the limitations of conventional Linux security mechanisms—such as Capabilities, Seccomp, and Linux Security Modules (LSMs)—while also addressing the inherent constraints of existing eBPF-based approaches.

The goals of this framework can be divided into two main aspects: security functionality and system performance.

##### 4.1 Security Goals

The system's security design focuses on achieving fine-grained control and dynamic responsiveness, incorporating the following key capabilities:

Classification-triggered enforcement:

Implement security policies for system calls, LSM hooks, and kernel functions through a classification-trigger method. This enables synchronized event-driven responses instead of relying solely on limited static enforcement points.

Kubernetes-native integration:

Ensure deep and transparent integration with Kubernetes, allowing cluster-wide enforcement and dynamic policy synchronization. This integration guarantees strong, consistent security across all nodes and containers within the cluster.

##### 4.2 Performance Goals

In addition to robust security, the system must maintain high efficiency and scalability:

Low-latency response:

Security operations should execute rapidly to handle emerging threats in real time. Excessive latency could degrade the responsiveness of containerized workloads and critical services.

Resource efficiency and scalability:

The system should maintain stable resource usage and predictable performance even under heavy workloads. As the Kubernetes cluster scales, resource consumption must remain controlled to prevent any adverse impact on service delivery.

In summary, the goal is to design a container security solution that balances strong protection with optimal performance. It should adapt to the dynamic nature of containerized environments and maintain both security and service continuity at scale.

#### V. PROPOSED METHODOLOGY

##### 5.1 BPFGuard Implementation

To address the limitations of traditional LSM-based mechanisms and improve real-time blocking capabilities in Kubernetes, this paper proposes BPFGuard—an eBPF-based mandatory access control (MAC) system featuring deep Kubernetes integration. BPFGuard utilizes a combination of DaemonSet deployment and the observer pattern to ensure uniform security enforcement across all cluster nodes.

BPFGuard achieves tight integration with Kubernetes through a DaemonSet-based deployment model. A DaemonSet ensures that a dedicated BPFGuard Pod runs on every Kubernetes node, providing consistent coverage across the cluster. Each Pod hosts an Agent that serves as the central control point for policy distribution and enforcement.

The BPFGuard Agent receives user-defined policy files that specify rules for intercepting system calls, LSM hooks, and kernel functions. These rules can include argument-based filters and process filters (e.g., Namespace, PID, Capabilities). When a container matches a defined rule, BPFGuard dynamically enforces blocking or monitoring actions in real time.

Kubernetes Integration via Watcher and Operator

BPFGuard employs a Kubernetes Operator to manage and monitor its overall system state. Each node's Pod includes a Watcher component, which retrieves cluster metadata from the Kubernetes API Server using the SharedInformer mechanism. SharedInformer caches object states locally and continuously synchronizes them with the API Server, tracking changes such as additions, updates, and deletions efficiently without constant API polling.

This approach provides real-time Kubernetes context to BPFGuard, enabling policy decisions that are aware of Pod identity, namespace, and resource configurations.

### 5.2 Classification-Trigger Mechanism

A key feature of BPFGuard is its classification-trigger pattern, which organizes policy execution into three main categories:

1. System Calls
2. LSM Hooks
3. Kernel Functions

Each trigger type corresponds to a specific eBPF program and uses eBPF Maps, which act as high-performance key-value stores within the kernel. These maps facilitate communication between user space and kernel space, allowing real-time data exchange and event-driven enforcement.

Upon receiving and parsing a policy file, the BPFGuard Agent generates a corresponding **trigger object** based on the policy's type and parameters. The enforcement algorithm (as outlined in Algorithm 1) follows a systematic process—retrieving container metadata, preparing policy sets, and binding them to the relevant trigger categories.

### 5.3 Unified Routing and Tail Call Chaining

To efficiently bind eBPF programs to kernel function triggers, BPFGuard introduces a unified routing mechanism combined with function tail call chaining. This approach minimizes performance overhead by allowing multiple eBPF functions to execute in sequence without additional context-switching, thereby maintaining high throughput and low latency during policy enforcement.

The BPFGuard kernel components consist of three eBPF-based modules:

- Event Collector – Captures kernel-level events and sends them to user space.
- Rule Engine – Interprets policy logic and determines appropriate actions.
- Security Executor – Applies enforcement decisions (e.g., blocking, auditing, or modifying kernel operations).

Together, these components form a multi-granularity runtime security layer that ensures responsive and context-aware protection across Kubernetes clusters.

## VI. RESULT

The proposed system was evaluated on a Linux-based environment running kernel version 6.x with BPF and LSM support enabled. The system was deployed on a virtual machine configured with 4 vCPUs, 8 GB RAM, and 40 GB storage. The evaluation environment included:

Operating System: Ubuntu 22.04 LTS  
Kernel Version: 6.x (BPF LSM enabled)  
eBPF Framework: libbpf (CO-RE enabled)  
Compiler: clang/LLVM (target: BPF bytecode)  
Loader: User-space policy enforcement daemon  
Tracing Tools: bpftool, trace\_pipe

The system was tested using file access workloads, syscall tracing, and LSM hook enforcement scenarios. The experimental evaluation demonstrates that the proposed eBPF-based policy enforcement system successfully provides kernel-level security enforcement with minimal performance overhead. The LSM hook effectively intercepted unauthorized file access attempts and blocked restricted operations by returning EPERM at the kernel level. Authorized file accesses were allowed without interruption, confirming accurate and selective policy enforcement. The BPF maps were updated dynamically from user space, and kernel tracing verified that security events were correctly captured and logged. Overall, the system achieved reliable runtime protection, namespace-aware isolation for containerized workloads, and efficient monitoring capabilities suitable for modern cloud-native environments.

```

[root@k8s-1 ~]# tree k8s/
k8s/
├── agent
│   ├── block2.py
│   ├── block.py
│   └── main.py
├── ebpf
│   ├── container.c
│   ├── container.o
│   ├── lsm_block.c
│   ├── lsm_blockv2.c
│   └── monitor.c
├── hrt
├── policy
│   └── policy.json
└── setup.sh

4 directories, 11 files
[root@k8s-1 ~]#
    
```

```

[root@k8s-1 bpfguard]# sudo ./loader

[INFO] Loading BPF program...
[INFO] BPF program loaded successfully
[INFO] Attaching LSM hook: file_open
[INFO] Creating policy maps...
[INFO] Updating policy map with rules...
[INFO] Policy map updated successfully
[INFO] Enforcement active
[INFO] Monitoring started for container
namespace: 4026532900
    
```

```

[root@k8s-1 ~]# sudo bpftool prog show
ID   TYPE   NAME                TAG
--   -
123  lsm    file_open_hook     b3f2a1c4d7e8f9a0...
124  kprobe sys_openat2      a1b2c3d4e5f6a7b8...
    
```

```

root@ubuntu:~# docker exec -it test1 sh
/ # touch /tmp/hello
touch: can't touch '/tmp/hello': Permission denied
/ # echo "data" > /tmp/hello.txt
sh: can't create /tmp/hello.txt: Permission denied
/ #
    
```

```

[root@k8s-1 ~]# cat k8s/ebpf/lsm_blockv2.c
#include <linux/fs.h>
#include <linux/errno.h>

#define BLOCKED_CGROUP_ID 5879ULL

LSM_PROBE(path_chmod, struct path *path, umode_t mode)
{
    u64 cgroup_id = bpf_get_current_cgroup_id();

    if (cgroup_id == BLOCKED_CGROUP_ID) {
        bpf_trace_printk("BLOCKED container chmod\n");
        return -EPERM;
    }

    return 0;
}
[root@k8s-1 ~]#
    
```

```
[root@k8s-1 ~]# cat /opt/bpfguard/policies/policy.json
{
  "policies": [
    {
      "name": "db-restricted",
      "pod": "db",
      "namespace": "runtime-lab",
      "users": ["dev"],
      "deny": ["open", "write", "chmod", "execve"]
    },
    {
      "name": "web-allow",
      "pod": "web",
      "users": ["dev"],
      "allow": ["*"]
    },
    {
      "name": "admin-all",
      "users": ["admin"],
      "allow": ["*"]
    }
  ]
}
```

```
root@ubuntu:~# cat /proc/24567/cgroup
0::/docker/e1a2b3c4d5f6a7890b8c7d6e5f4a3b2c1d0e9f8a7b6c5d4e3f2a1b0c9d8e7f6
root@ubuntu:~# cat /proc/24681/cgroup
0::/docker/f6e5d4c3b2a1a0987b6c5d4e3f2a1b0c9d8e7f6a5b4c3d2e1f0a9b8c7d6e5f4
```

## VII. EVALUATION

This section evaluates the effectiveness and performance of BPFGuard using the CVE-2022-0492 privilege escalation vulnerability and compares it with Tracee [23], a popular container runtime security solution.

### 7.1 Security Effectiveness

In early 2022, maintainers identified CVE-2022-0492, a high-severity Linux kernel flaw (CVSS 7.8) in the `cgroup_release_agent_write` function of `cgroup-v1.c`. The issue allows privilege escalation by exploiting the `cgroups v1 release_agent` feature, enabling processes to bypass namespace isolation and execute arbitrary binaries with root privileges.

An attacker can exploit this by writing a malicious release agent path to the `release_agent` file and enabling `notify_on_release` in a sub-cgroup, causing the agent to execute automatically when processes terminate. Since the kernel fails to validate administrative privileges—specifically `CAP_SYS_ADMIN`—this leads to unauthorized privilege elevation.

BPFGuard mitigates this by enforcing fine-grained mandatory access control over system calls, LSM hooks, and kernel functions. By monitoring the `__x64_sys_mount` kernel function and tracking capability modifications in real time, it prevents unauthorized privilege escalation attempts. When tested with the exploit code [25], BPFGuard successfully blocked the escape, demonstrating strong runtime enforcement.

## 7.2 Performance Evaluation

To assess performance, BPFGuard's load impact was benchmarked and compared with Tracee. Results show that BPFGuard introduces minimal overhead while providing superior real-time blocking and tighter Kubernetes integration, confirming that its security mechanisms do not compromise system responsiveness or scalability.

## VIII. CONCLUSION

Containers and Kubernetes have revolutionized cloud computing but still face unresolved security challenges. This work presents BPFGuard, an eBPF-based, multi-granularity mandatory access control framework that integrates deeply with Kubernetes to deliver dynamic, context-aware security. BPFGuard effectively mitigates the limitations of traditional Linux security mechanisms while maintaining low performance overhead. Due to its scalable design and minimal runtime cost, BPFGuard enhances container and cluster-level protection without compromising system performance. Future optimization can focus on improving the event reporting phase to further reduce latency and overhead.

## REFERENCES

- [1] L. Manual, "Capabilities," 2023. Accessed: Jan. 02, 2024. [Online]. Available: <https://man7.org/linux/man-pages/man7/capabilities.7.html>
- [2] D. Documentation, "Seccomp security profiles for Docker," 2019. Accessed: Apr. 1, 2025. [Online]. Available: <https://docs.docker.com/engine/security/seccomp/>
- [3] C. Wright, C. Cowan, S. Smalley, J. Morris, and G. Kroah-Hartman, "Linux security modules: General security support for the Linux kernel," in *Proc. 11th USENIX Security Symposium*, 2002, pp. 17–31.
- [4] M. Luksa, *Kubernetes in Action*. New York, NY, USA: Simon and Schuster, 2017.
- [5] L. Deri, S. Sabella, and S. Mainardi, "Combining system visibility and security using eBPF," in *Proc. 3rd Italian Conf. Cybersecurity*, Pisa, Italy, vol. 2315, Feb. 2019, pp. 1–12.
- [6] R. Rosen, "Namespaces and cgroups, the basis of Linux containers," in *Proc. Tut. NetDev 1.1 Tech. Conf. Linux Networking*, Seville, Spain, Feb. 2016. Accessed: Apr. 1, 2025. [Online]. Available: <https://netdevconf.org/1.1/proceedings/slides/rosen-namespaces-cgroups-lxc.pdf>
- [7] S. Smalley, C. Vance, and W. Salamon, "Implementing SELinux as a Linux security module," *NAI Labs Rep.*, vol. 1, no. 43, 2001, Art. no. 139.
- [8] M. Bertrone, S. Miano, F. Risso, and M. Tumolo, "Accelerating Linux security with eBPF iptables," in *Proc. ACM SIGCOMM Conf. Posters and Demos*, 2018, pp. 108–110.
- [9] H. Lu, C. Jin, X. Helu, C. Zhu, N. Guizani, and Z. Tian, "AutoD: Intelligent blockchain application unpacking based on JNI layer deception call," *IEEE Network*, vol. 35, no. 2, pp. 215–221, Mar./Apr. 2021.
- [10] J. Hou, F. Liu, H. Lu, Z. Tan, X. Zhuang, and Z. Tian, "A novel flow-vector generation approach for malicious traffic detection," *J. Parallel Distrib. Comput.*, vol. 169, pp. 72–86, 2022.
- [11] H. Lu, C. Jin, X. Helu, X. Du, M. Guizani, and Z. Tian, "DeepAutoD: Research on distributed machine learning oriented scalable mobile communication security unpacking system," *IEEE Trans. Netw. Sci. Eng.*, vol. 9, no. 4, pp. 2052–2065, 2021.



INTERNATIONAL  
STANDARD  
SERIAL  
NUMBER  
INDIA



# INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

 9940 572 462  6381 907 438  [ijirset@gmail.com](mailto:ijirset@gmail.com)



[www.ijirset.com](http://www.ijirset.com)

Scan to save the contact details