KubeRosy:
Dynamic System Call Filtering Framework for Containers

Jin Her

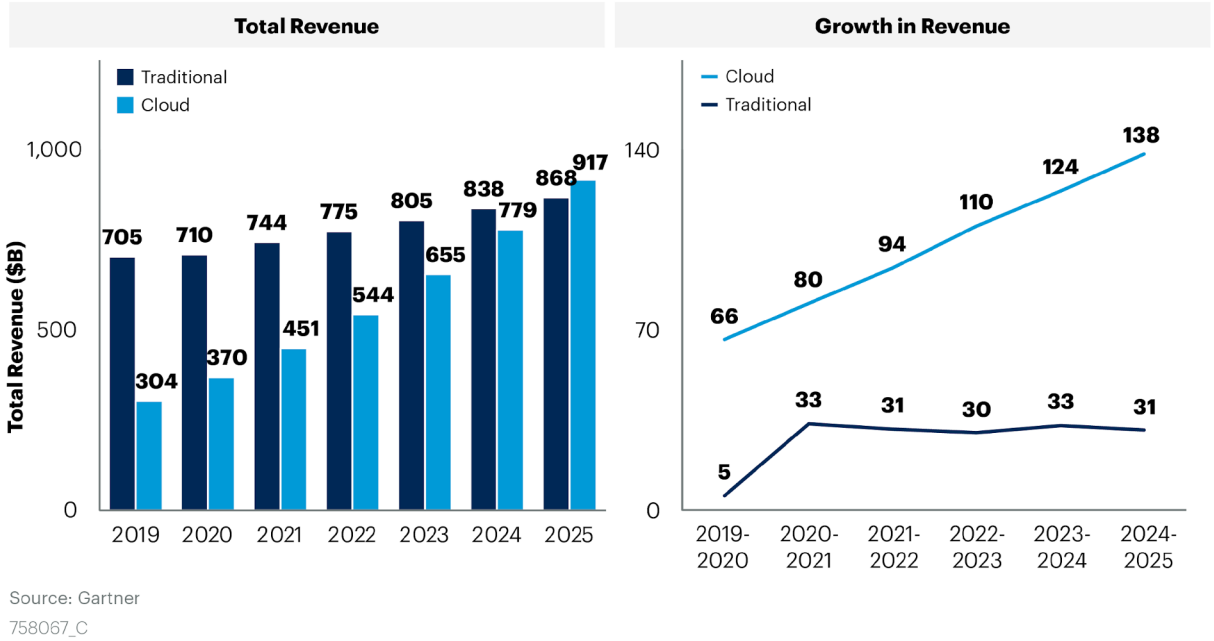
Department of Computer Science & Engineering

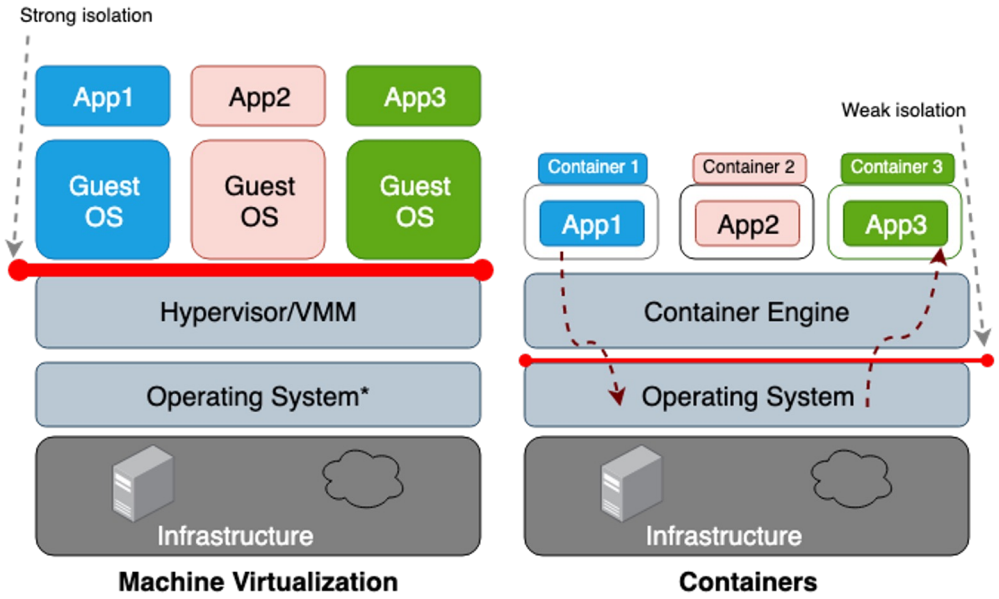
Incheon National University

Index

Introduction

- Rapid migration to cloud-native environments
- Technically, containers share the host's kernel
- Thus, securing system calls in the cloud is **ESSENTIAL** for protecting the kernel

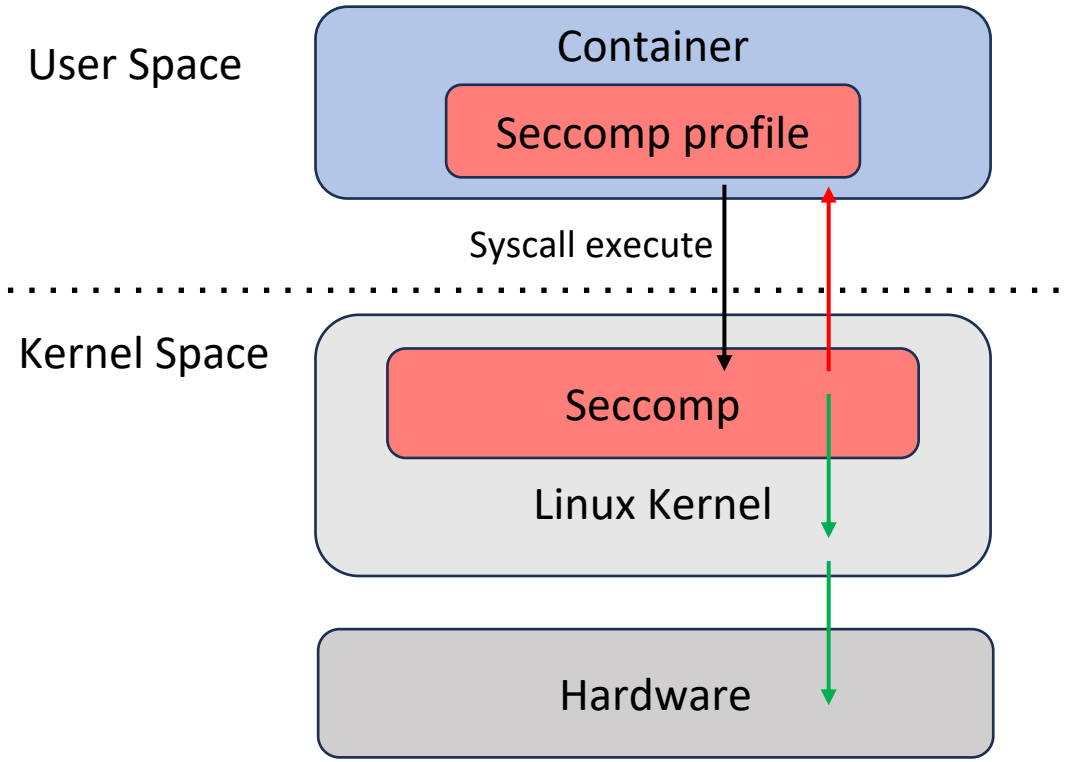




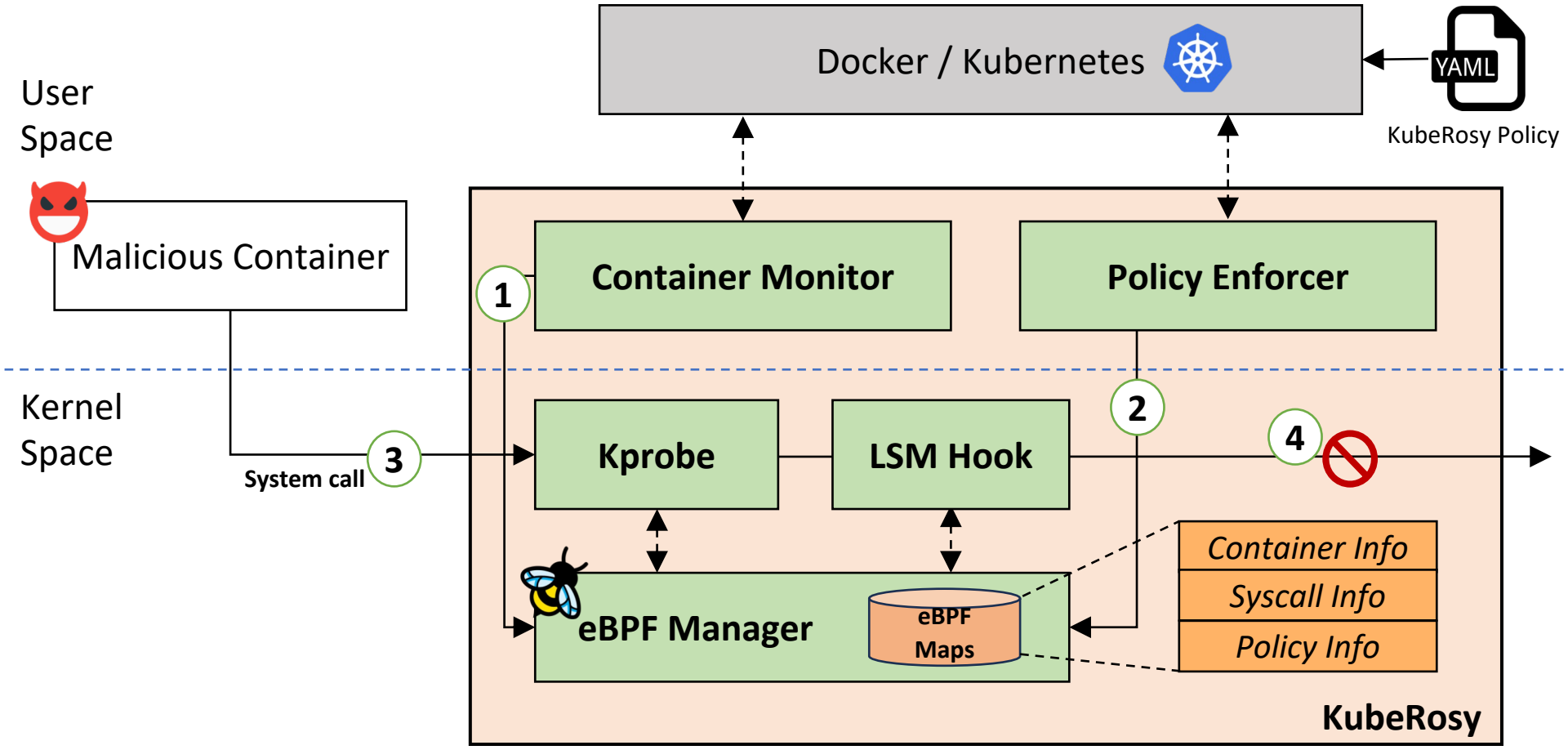
Motivation

- Seccomp(secure computing mode) for containers
  - Attach a seccomp profile to containers when they are deployed
  - **CANNOT** update or delete a seccomp profile at runtime

```
{
    "defaultAction": "SCMP_ACT_ERRNO",
    "architectures": [
        "SCMP_ARCH_X86_64”
    ],
    "syscalls": [
        {.
            "names": [
                "accept4",
                ”execve",
                ”bind",
                ”connect”
            ],
            "action": "SCMP_ACT_ALLOW".
        }
    ]
}
```



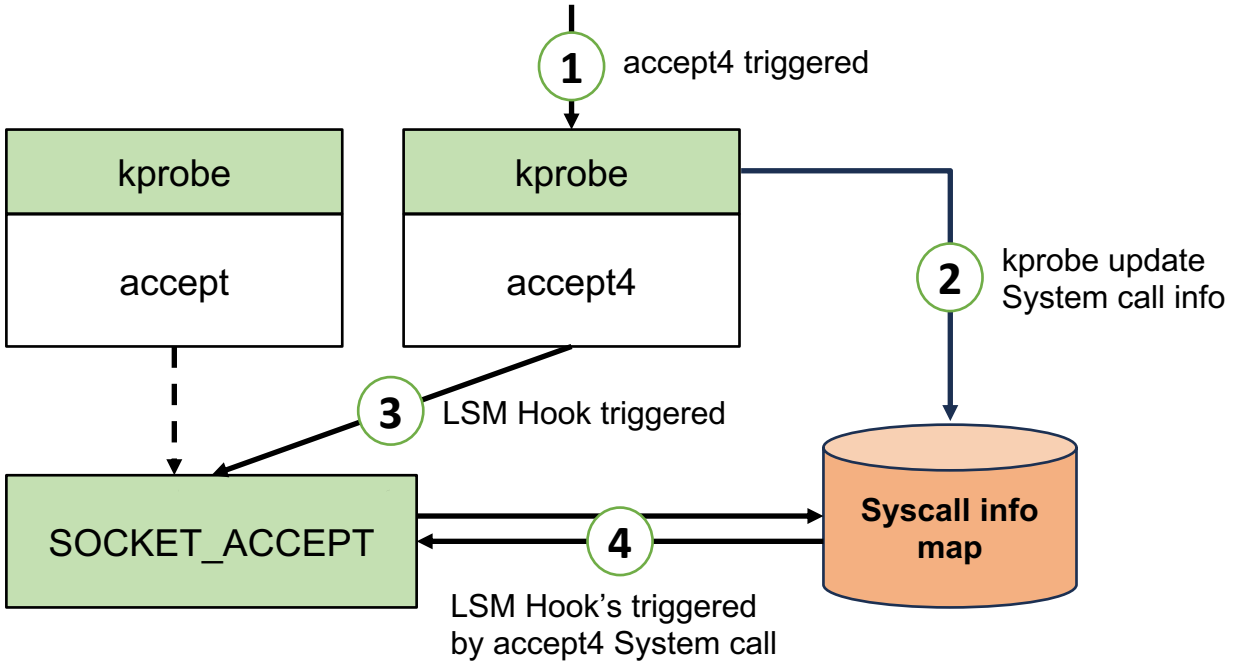
KubeRosy Design: Overall Architecture



KubeRosy Design: Why eBPF and LSM used?

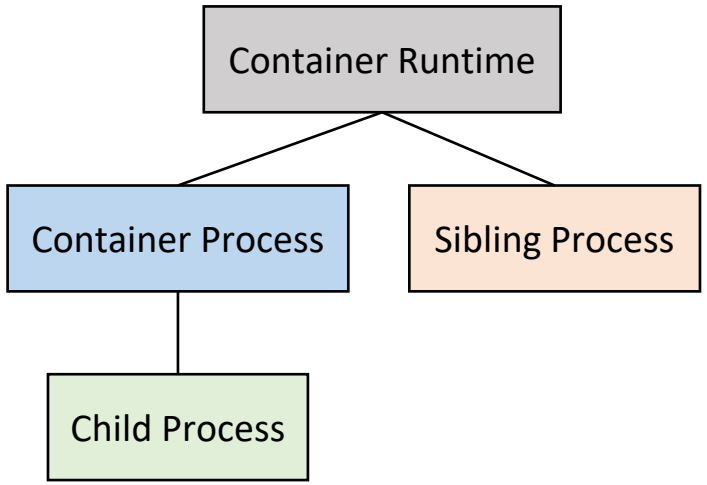
- LSM Hooks can be directly involved in the execution flow of a system call, but triggered by multiple system calls!

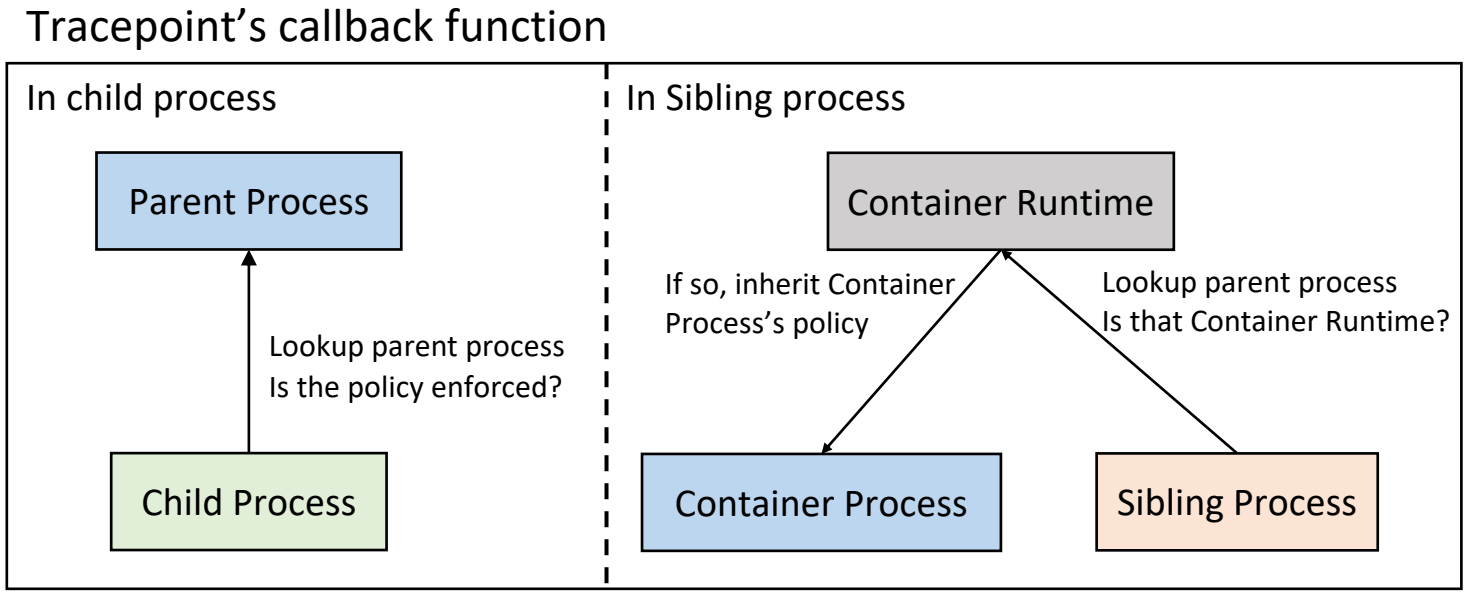
Network	SOCKET_ACCEPT	accept
		accept4
	SOCKET_BIND	bind
	SOCKET_CONNECT	connect
	SOCKET_LISTEN	listen
	SOCKET_RECVMSG	recvfrom
		recvmsg
		recvmmsg
	SOCKET_CREATE	socket
		socketpair



KubeRosy Design: Policy Inheritance

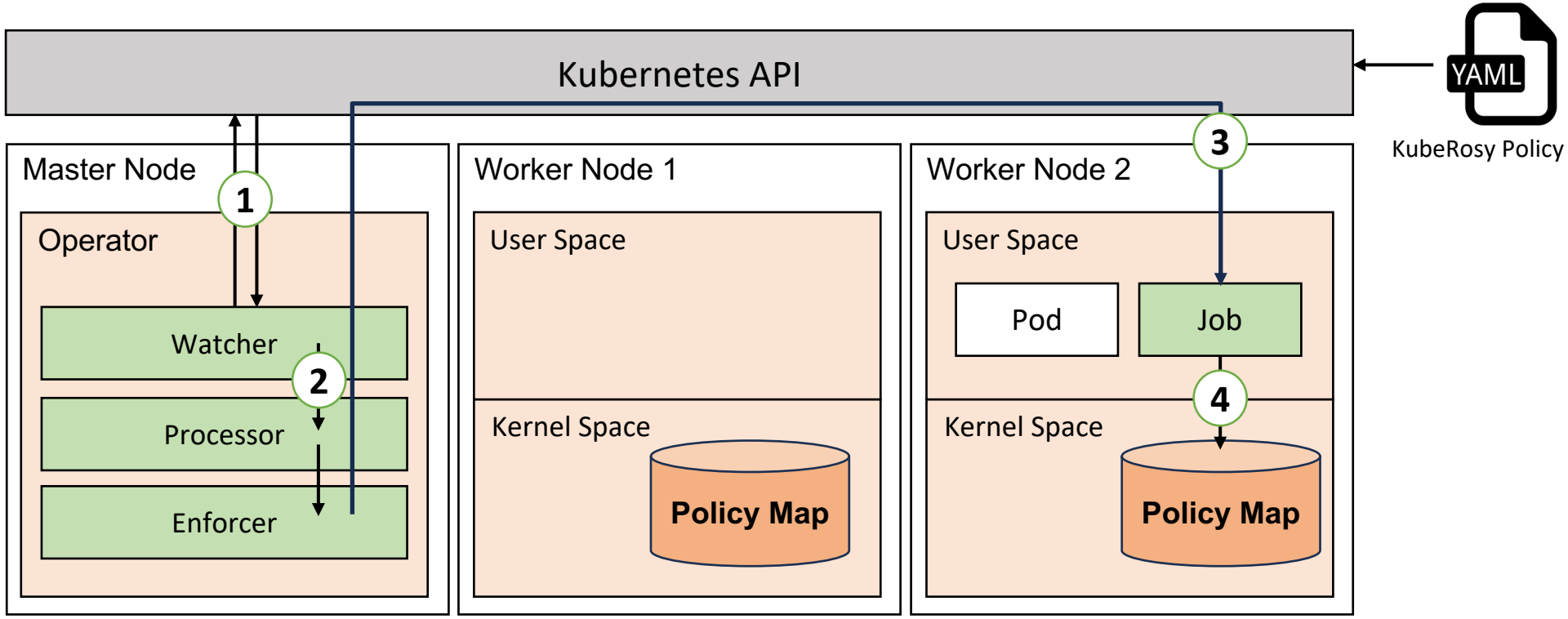
- Attach tracepoints to the `fork`, `vfork`, `clone`, and `clone3` system calls, which creates processes
- Tracepoint's callback function determines if the policy should be applied





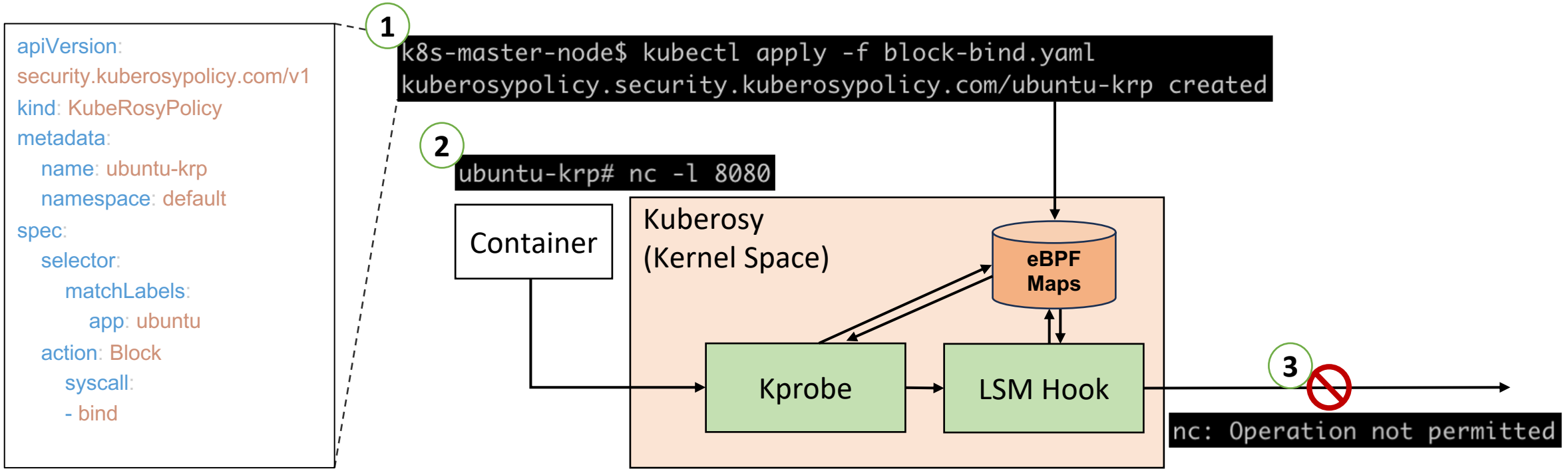
KubeRosy Design: Operator

- Watcher monitor the Kubernetes API
- Enforcer creates a job on the node where the Pod with the policy is deployed



Evaluation

- Running on Ubuntu 22.04, Kernel v5.15
- 28 system calls supported by KubeRosy



Conclusion

- In this paper, we propose a system call security framework to secure the limitations of seccomp using LSM and eBPF.
- Limited by the small number of supported system calls relative to the total number of system calls

- Future work
  - Adding supported system calls
  - Implement more detailed security policies with filtering based on argument values for system calls