
www.kodekloud.com

cks.kodekloud.com

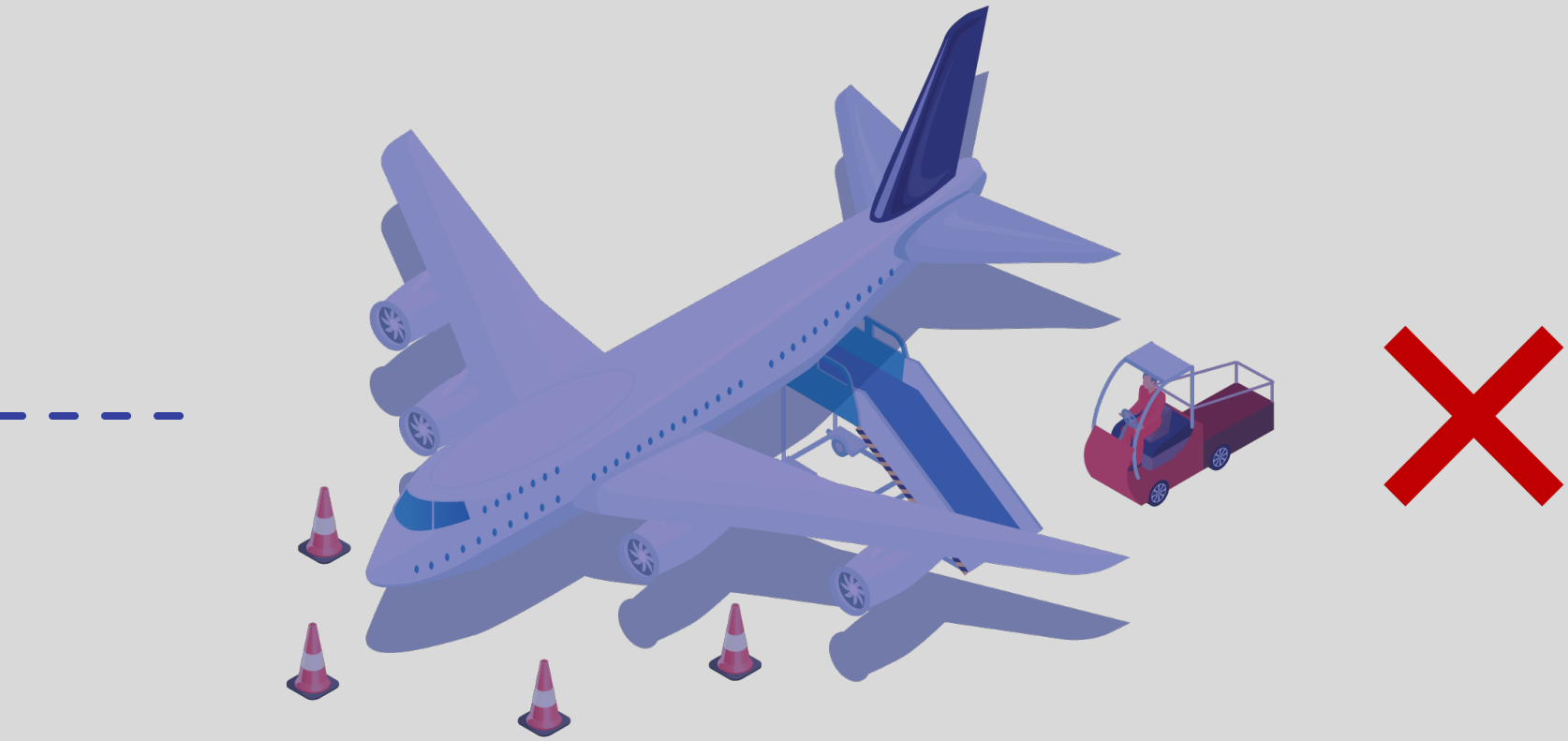
# Disclaimer

# Notice

- This presentation is to refer to course contents only.
- Some of the slides are meant to be animated. So may not be displayed correctly.
- Do not copy and paste command, code or YAML files from this file as it may not be in the right format and may contain hidden characters
- For code refer to the solutions in the lab or the Git repository associated with this course or official Kubernetes documentation pages.
- Some of the code in this deck maybe hidden for brevity

https://github.com/kodekloudhub/certified-kubernetes-security-specialist-cks-course

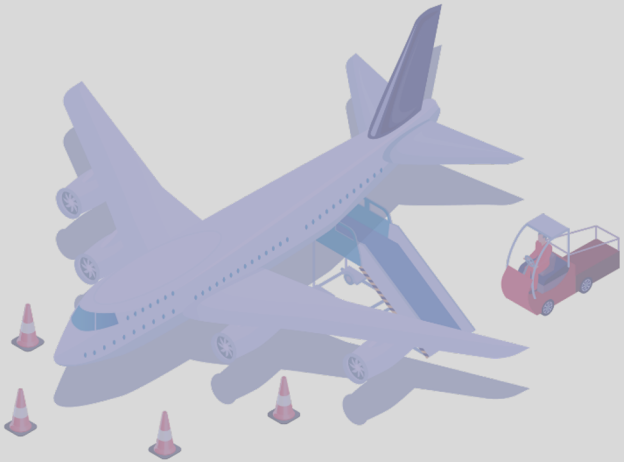
# Least Privilege Principle



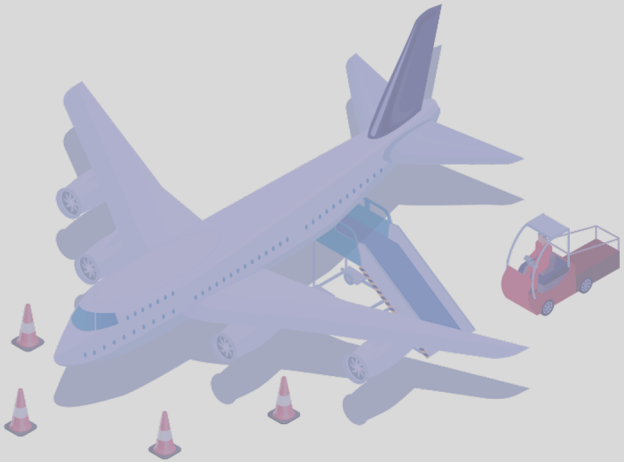
THE TRAVELER


BAGGAGE
COUNTER

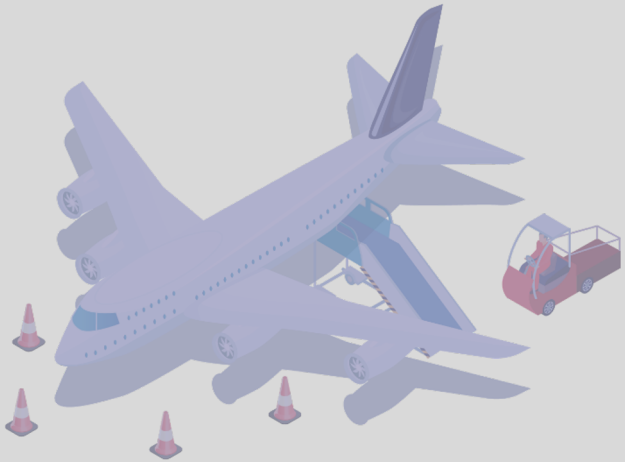

SECURITY
CHECK



STORE EMPLOYEES



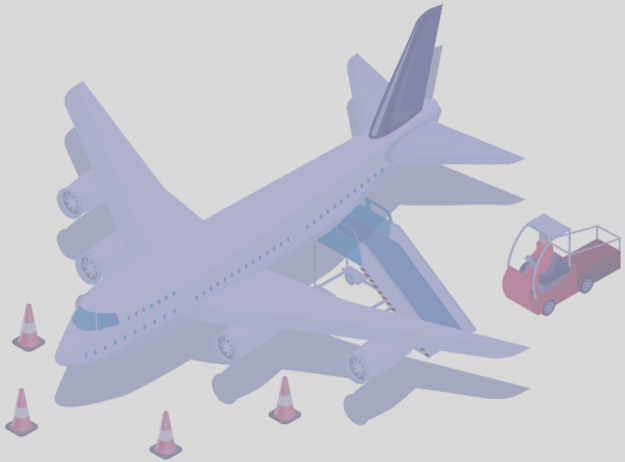
BOARDING GATE



CLEANERS

CARGO LOADERS
& MAINTENANCE
WORKERS

PILOTS STEWARDS
& AIR HOSTESS



BAGGAGE COUNTER

SECURITY CHECK

STORE EMPLOYEES

THE TRAVELER

CARGO LOADERS & MAINTENANCE WORKERS

BOARDING GATE

PILOTS STEWARDS & AIR HOSTESS

CLEANERS

Limit Access  to Nodes

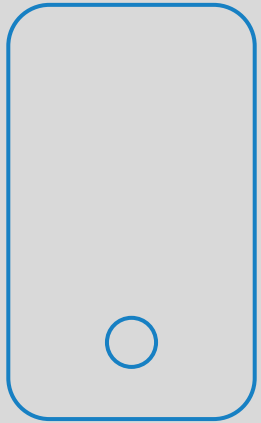
RBAC Access

Remove Obsolete Packages & Services

Restrict Network Access

Restrict Obsolete Kernel Modules

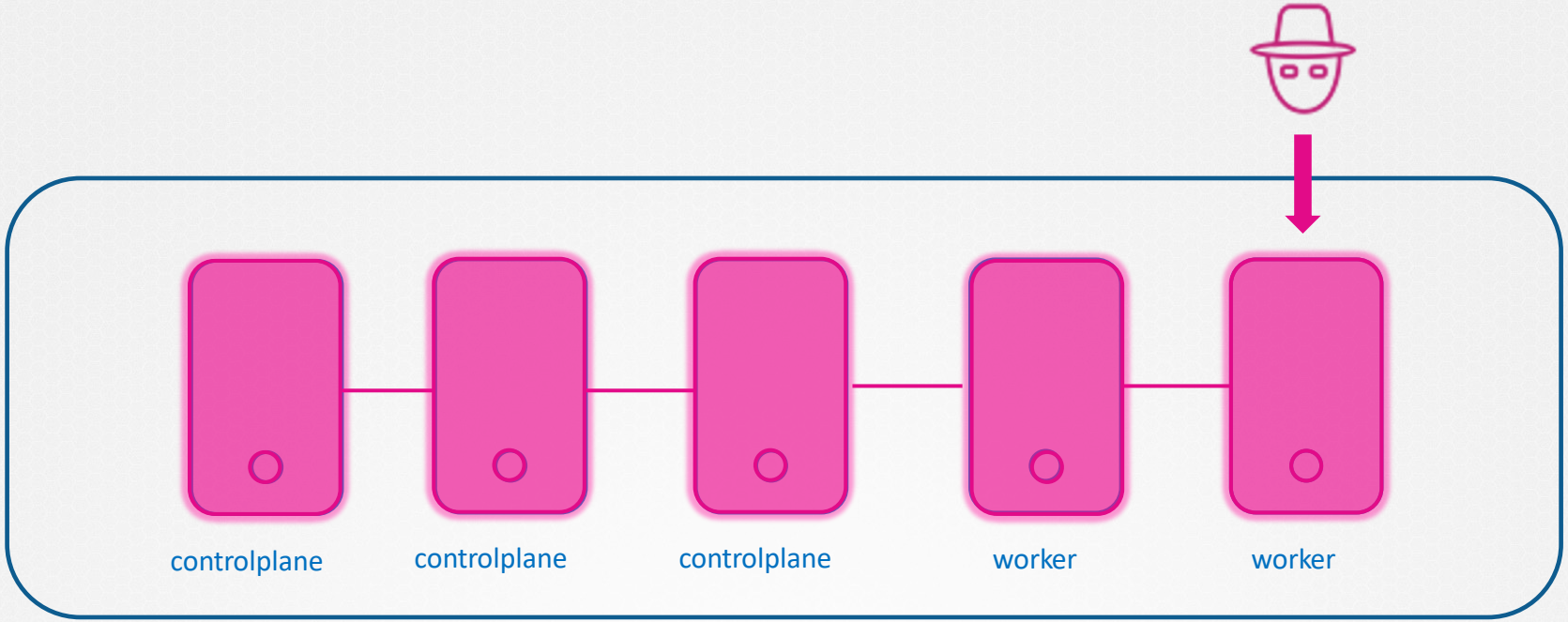
Identify and Fix Open Ports




www.kodekloud.com

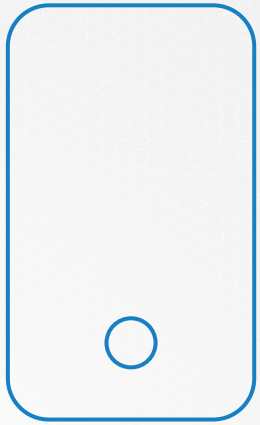
# Reduce the Attack Surface

# Reducing the Attack Surface



# Reducing the Attack Surface

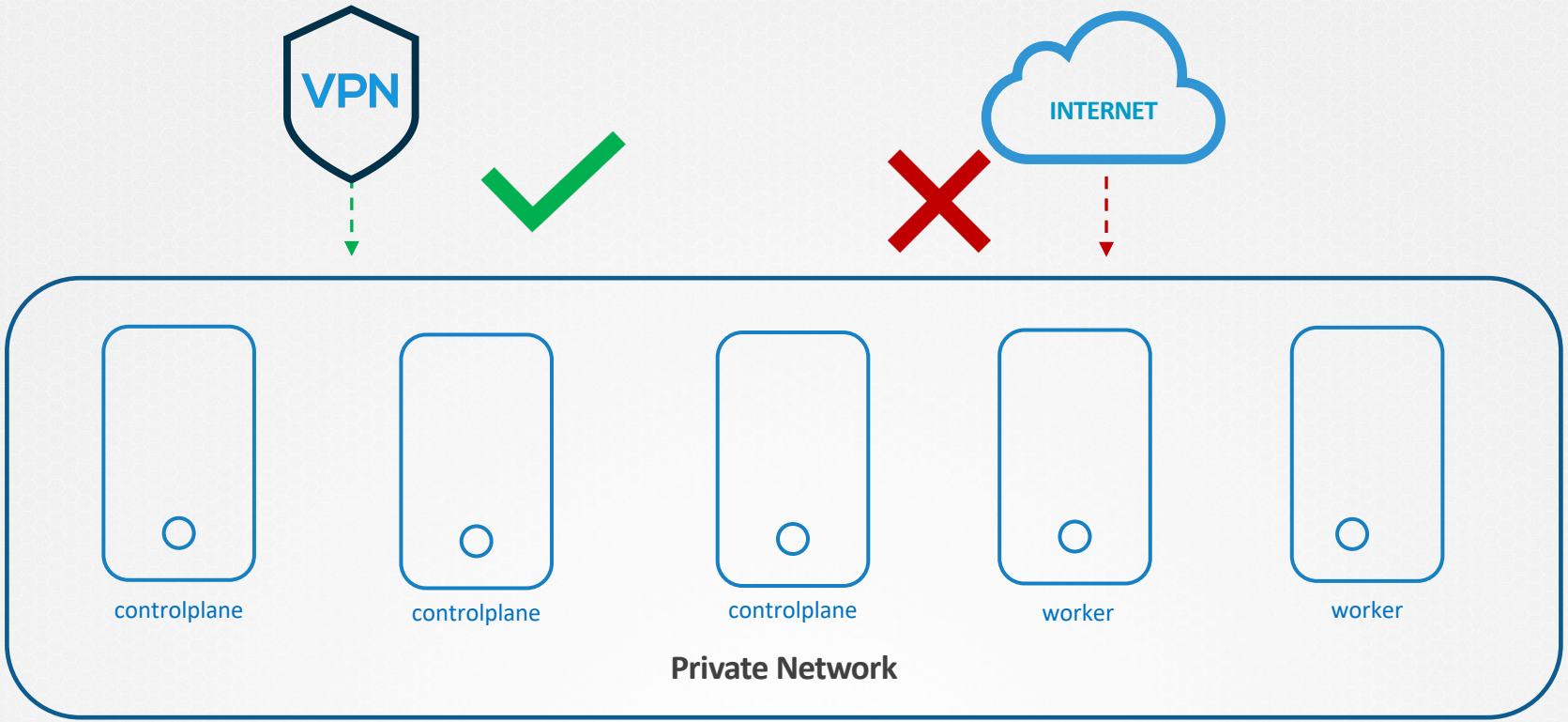
- Use Least Privilege Principle
- Remove Obsolete Software
- Limit Access
- Remove Obsolete Services
- Restrict Obsolete Kernel Modules
- Identify and Fix Open Ports



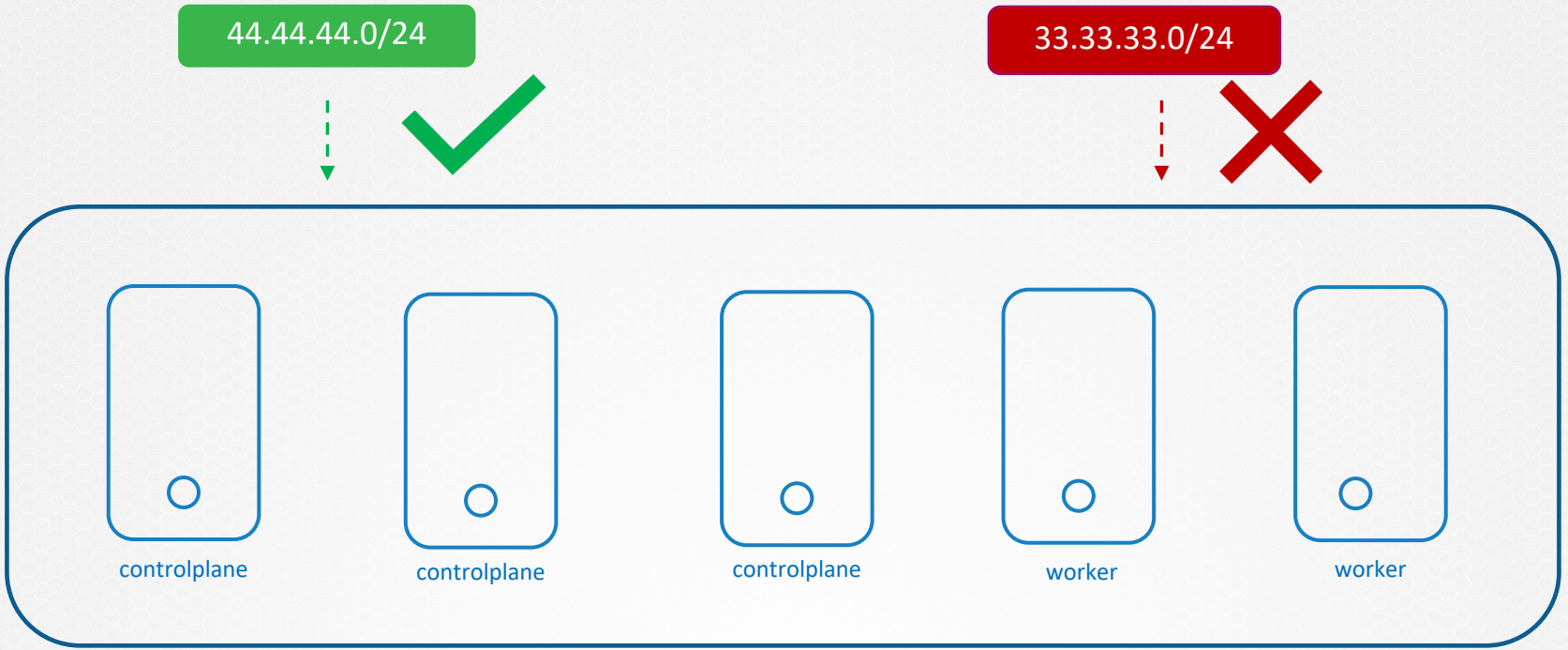

www.kodekloud.com

# Limit Node Access

# Limit Node Access



# Limit Node Access

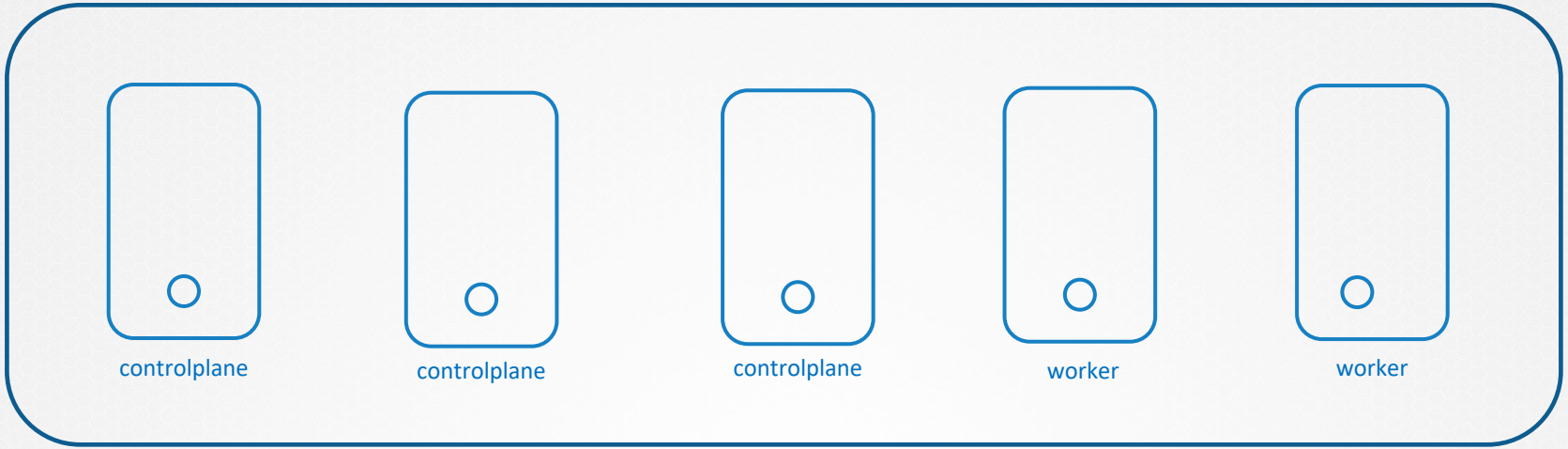


# Limit Node Access

Developers ✗

End Users ✗

Administrators ✓



# Limit Node Access

- bob
- michael
- dave

**01 User Account**

**02 Superuser Account**

UID = 0

- root

- ssh
- mail

**03 System Accounts**

**04 Service Accounts**

- nginx
- http

# Limit Node Access

```
id
uid=1000(michael) gid=1000(michael) groups=1000(michael)1010(admin)
```

```
who
michael      pts/2        Apr 28 06:48 (172.16.238.187)
```

```
last
michael   :1    :1                    Tue May 12 20:00  still logged in
sarah     :1    :1                    Tue May 12 12:00  still running
reboot   system boot 5.3.0-758-gen    Mon May 11 13:00  - 19:00 (06:00)
```

# Limit Node Access

/etc/passwd

```
grep -i  ^michael /etc/passwd
michael:x:1001:1001::/home/michael:/bin/bash
```

/etc/shadow

```
grep -i  ^michael /etc/shadow
michael:$6$0h0utOtO$5JcuRxR7y72LLQk4Kdog7u09LsNFS0yZPkIC8pV9tgD0wXCHut
YcWF/7.eJ3TfGfG0lj4JF63PyuPwKC18tJS.:18188:0:99999:7:::
```

/etc/group

```
grep -i  ^bob /etc/group
developer:x:1001:bob,michael
```

# Limit Node Access

```
usermod –s /bin/nologin michael
```

```
grep –i michael /etc/passwd
michael:x:1001:1001::/home/michael:/bin/nologin
```

```
userdel bob
```

```
grep –i bob /etc/passwd
```

# Limit Node Access

```
id michael
uid=1001(michael) gid=1001(michael) groups=1001(michael),1000(admin)
```

```
deluser michael admin
Removing user `michael` from group `admin` ...
Done.
```

```
id michael
uid=1001(michael) gid=1001(michael) groups=1001(michael)
```

Hands-on Labs

cks.kodekloud.com


www.kodekloud.com

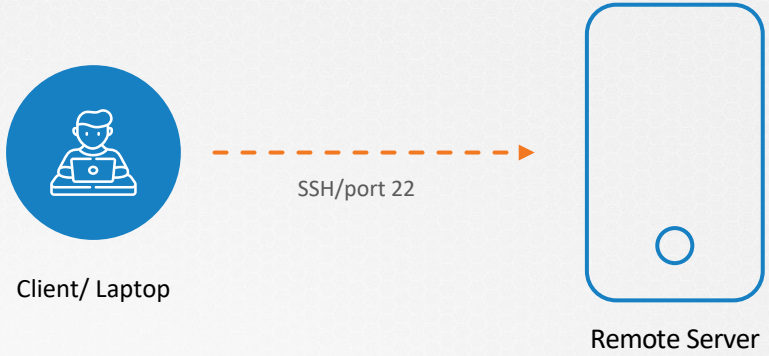
# SSH Hardening

# SSH

```
ssh <hostname OR IP Address>
```

```
ssh <user>@<hostname OR IP Address>
```

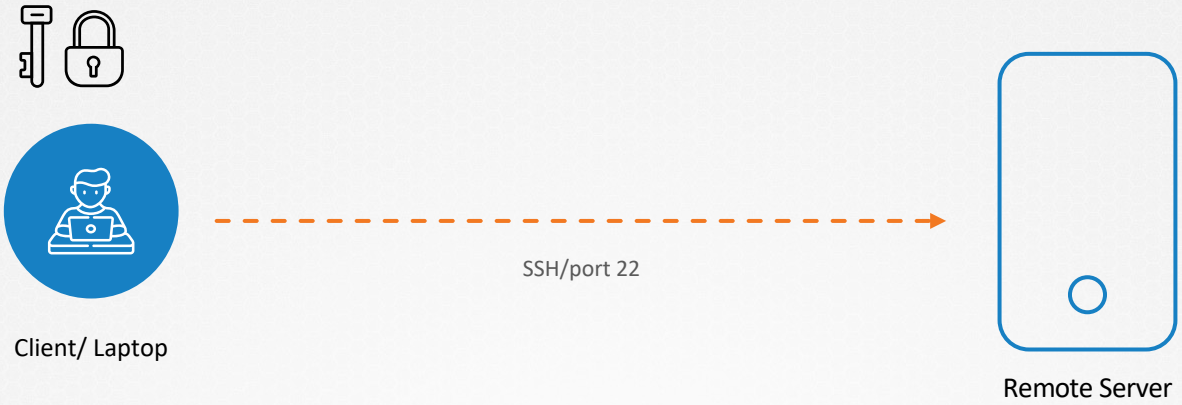
```
ssh -l <user> <hostname OR IP Address>
```




```
[mark@localhost ~]$ ssh node01
mark@node01's password:
Last login: Tue Apr  7 20:08:58 2020 from 192.168.1.109
[mark@node01 ~]$
```

# SSH

```
Key Pair = Private Key + Public Key
```



# SSH

Client/ Laptop

Public Key: /home/mark/.ssh/id_rsa.pub

Private Key: /home/mark/.ssh/id_rsa

```
ssh-keygen -t rsa
Generating public/private rsa key pair.
Enter file in which to save the key (/home/ mark /.ssh/id_rsa):
/home/mark/.ssh/id_rsa already exists.
Overwrite (y/n)? y
Enter passphrase (empty for no passphrase):
Enter same passphrase again:
Your identification has been saved in /home/mark/.ssh/id_rsa.
Your public key has been saved in /home/mark/.ssh/id_rsa.pub.
The key fingerprint is:
SHA256:PCRTdbxxzffzmi8uunjn5V/1LZCG0BvhVJYXBr9gYsE mark@localhost
The key's randomart image is:
+---[RSA 2048]----+
|        .o=o=oo+ |
|       . +E=+oo +|
|      o o * o=. o|
|       = o *.o o.|
|        S o + . +|
|         . . .  =|
|              oo+|
|         .. oo+..|
|        ..o=.oo+o|
+----[SHA256]-----+
```

# SSH

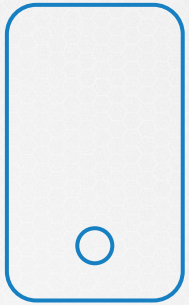
Client/ Laptop

```
ssh-copy-id mark@node01
/usr/bin/ssh-copy-id: INFO: Source of key(s) to be installed:
"/home/mark/.ssh/id_rsa.pub"
/usr/bin/ssh-copy-id: INFO: attempting to log in with the new
key(s), to filter out any that are already installed
/usr/bin/ssh-copy-id: INFO: 1 key(s) remain to be installed --
if you are prompted now it is to install the new keys
mark@node01's password:

Number of key(s) added: 1

Now try logging into the machine, with:   "ssh 'mark@node01'"
and check to make sure that only the key(s) you wanted were
added.
```

```
ssh node01
Last login: Tue Apr  7 20:10:58 2020 from 192.168.1.109
[mark@node01 ~]$
```

# SSH

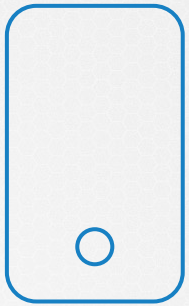
```
cat /home/mark/.ssh/authorized_keys
ssh-rsa
AAAAB3NzaC1yc2EAAAADAQABAAABAQCgVV5wgH37kNwjnEIxgeX4j6LASNckjKi4bRpjPGecyxEi
EeJhIU4x31XPEFzUFp/1xX2rjeiM2Ko3oPmTGCCTEQMpQogerR7NS+bA9eXs34jWIg+xoSQjeQu1
+lXgrRippJn2YhWYVAY3sKWIiiklowuMXmxjmBBr48L52di1J+8EASwnM4ILX/YL72Czq3uFFhVW
1fNUKBPUbW58h4QSAd2r9abzZfrHH48ThPJW4/5i8LOHEo3W0BXl3foEV0c6pk3TgxcjTuZQOimd
48mM2pxWJh9WxA0xcXwbD3+JrcnZeMJq4TbrKjaXQ0pBGenglxurxnRT2og9DeTIqGN3
mark@localhost
```



# HARDEN SSH SERVICE

```
vi /etc/ssh/sshd_config
PermitRootLogin no
PasswordAuthentication no
```

```
systemctl restart sshd
```



Remote Server

# CIS Benchmark Reference

## *5.2 SSH Server Configuration*

SSH is a secure, encrypted replacement for common login services such as `telnet`, `ftp`, `rlogin`, `rsh`, and `rcp`. It is strongly recommended that sites abandon older clear-text login protocols and use SSH to prevent session hijacking and sniffing of sensitive data off the network.

**Note:** The recommendations in this section only apply if the SSH daemon is installed on the system, if remote access is not required the SSH daemon can be removed and this section skipped.

**Note:** Once all configuration changes have been made to `/etc/ssh/sshd_config`, the sshd configuration must be reloaded:

```
# systemctl reload sshd
```

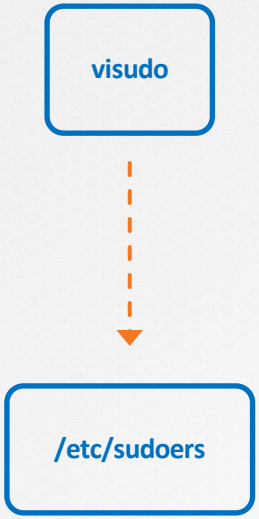
Hands-on Labs

cks.kodekloud.com


www.kodekloud.com

# User Privilege Escalation

# SUDO




```
apt install nginx
E: Could not open lock file /var/lib/dpkg/lock-frontend -
open (13: Permission denied)
E: Unable to acquire the dpkg frontend lock
(/var/lib/dpkg/lock-frontend), are you root?
```

```
sudo apt install nginx
[sudo] password for michael:
```

```
cat /etc/sudoers
User privilege specification
root    ALL=(ALL:ALL) ALL
# Members of the admin group may gain root privileges
%admin ALL=(ALL) ALL
# Allow members of group sudo to execute any command
%sudo    ALL=(ALL:ALL) ALL
# Allow Bob to run any command
mark  ALL=(ALL:ALL) ALL
# Allow Sarah to reboot the system
sarah localhost=/usr/bin/shutdown -r now
# See sudoers(5) for more information on "#include"
directives:
#includedir /etc/sudoers.d
```

```
grep -i  ^root /etc/passwd
root:x:0:0:root:/root:/usr/sbin/nologin
```

# SUDO

```
cat /etc/sudoers
User privilege specification
root    ALL=(ALL:ALL) ALL
# Members of the admin group may gain root privileges
%admin ALL=(ALL) ALL
# Allow members of group sudo to execute any command
%sudo    ALL=(ALL:ALL) ALL
# Allow Bob to run any command
mark  ALL=(ALL:ALL) ALL
# Allow Sarah to reboot the system
sarah localhost=/usr/bin/shutdown -r now
# See sudoers(5) for more information on "#include"
directives:
#includedir /etc/sudoers.d
```

| Field | Description | Example |
|---|---|---|
| 1 | User or Group | bob, %sudo (group) |
| 2 | Hosts | localhost, ALL(default) |
| 3 | User | ALL(default) |
| 4 | Command | /bin/ls, ALL(unrestricted) |

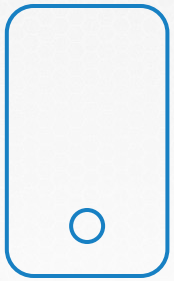
Hands-on Labs

cks.kodekloud.com


www.kodekloud.com

# Remove Unwanted Packages and Services

# Install only the Required Packages



kubelet

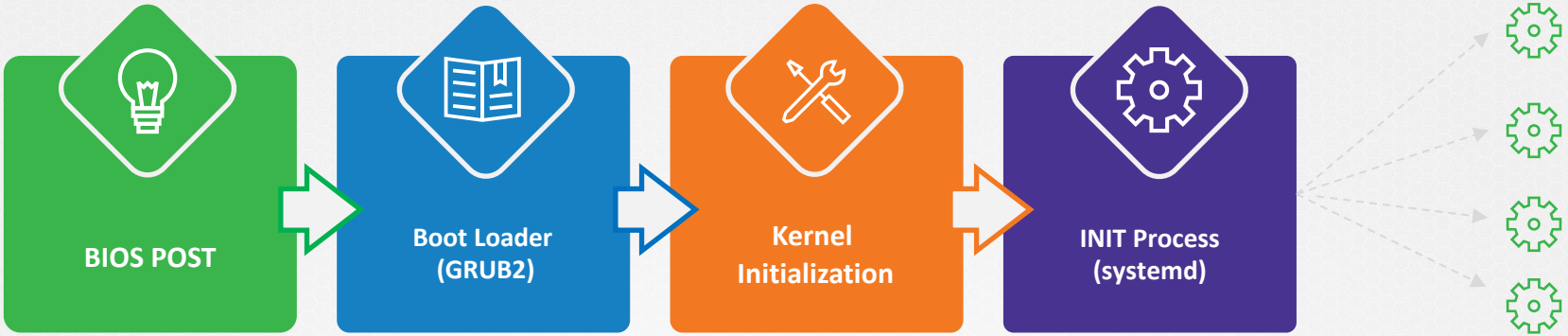
kubeadm

Container runtime

kubectl

apache2

# Remove Unwanted Services




```
systemctl status apache2
```

```
Loaded: loaded (/lib/systemd/system/apache2.service; enabled; vendor preset: enabled)
  Drop-In: /lib/systemd/system/apache2.service.d
           └─apache2-systemd.conf
   Active: active (running) since Mon 2021-03-29 18:01:14 UTC; 1s ago
  Process: 19026 ExecStart=/usr/sbin/apachectl start (code=exited, status=0/SUCCESS)
 Main PID: 19037 (apache2)
    Tasks: 55 (limit: 7372)
   CGroup: /system.slice/apache2.service
           ├─19037 /usr/sbin/apache2 -k start
           ├─19038 /usr/sbin/apache2 -k start
           └─19039 /usr/sbin/apache2 -k start
```

# Remove Unwanted Services

```
systemctl list-units --type service
```

```
Apache2.service                    loaded active running The Apache HTTP Server
apparmor.service                   loaded active exited  AppArmor initialization
containerd.service                 loaded active running containerd container runtime
dbus.service                       loaded active running D-Bus System Message Bus
docker.service                     loaded active running Docker Application Container Engine
ebtables.service                   loaded active exited  ebtables ruleset management
kmod-static-nodes.service          loaded active exited  Create list of required static device n
kubelet.service                    loaded active running kubelet: The Kubernetes Node Agent
proxy.service                      loaded active running kubectl proxy 8888
systemd-journal-flush.service      loaded active exited  Flush Journal to Persistent Storage
```

```
systemctl stop apache2
```

```
systemctl disable apache2
```

```
Synchronizing state of apache2.service with SysV service script with /lib/systemd/systemd-sysv-
install.
Executing: /lib/systemd/systemd-sysv-install disable apache2
```

# Install only the Required Packages

```
apt remove apache2
Reading package lists... Done
Building dependency tree
Reading state information... Done
The following packages were automatically installed and are no longer required:
  apache2-bin apache2-data apache2-utils libapr1 libaprutil1 libaprutil1-dbd-sqlite3
  libaprutil1-ldap liblua5.2-0 ssl-cert
Use 'apt autoremove' to remove them.
The following packages will be REMOVED:
  apache2
0 upgraded, 0 newly installed, 1 to remove and 23 not upgraded.
After this operation, 536 kB disk space will be freed.
Do you want to continue? [Y/n] Y
(Reading database ... 15908 files and directories currently installed.)
Removing apache2 (2.4.29-1ubuntu4.14) ...
invoke-rc.d: policy-rc.d denied execution of stop.
invoke-rc.d: policy-rc.d denied execution of stop.
```

# CIS Benchmark Reference

## *2 Services*

While applying system updates and patches helps correct known vulnerabilities, one of the best ways to protect the system against as yet unreported vulnerabilities is to disable all services that are not required for normal system operation. This prevents the exploitation of vulnerabilities discovered at a later date. If a service is not enabled, it cannot be exploited. The actions in this section of the document provide guidance on some services which can be safely disabled and under which circumstances, greatly reducing the number of possible threats to the resulting system. Additionally some services which should remain enabled but with secure configuration are covered as well as insecure service clients.

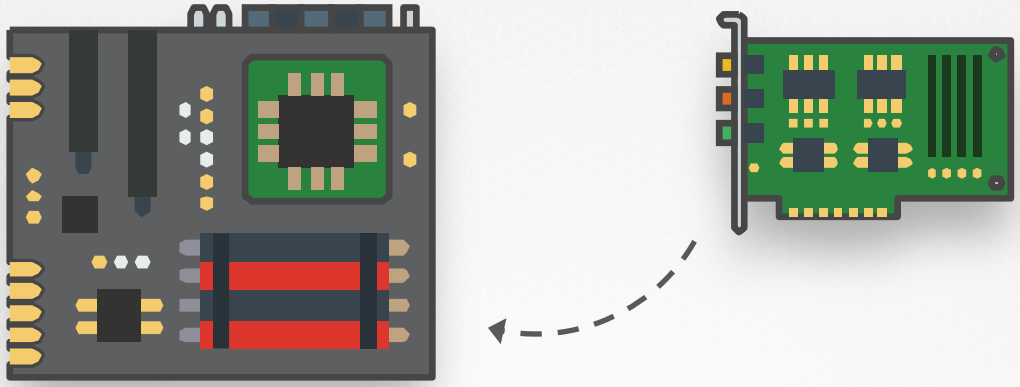
Hands-on Labs

cks.kodekloud.com


www.kodekloud.com

# Restrict Kernel Modules

# Restrict Kernel Modules



# Restrict Kernel Modules

lsmod

```
# lsmod
Module                  Size  Used by
floppy                 69417  0
xt_conntrack           16384  1
ipt_MASQUERADE         16384  1
nf_nat_masquerade_ipv4    16384  1 ipt_MASQUERADE
nf_conntrack_netlink    40960  0
nfnetlink              16384  2 nf_conntrack_netlink
xfrm_user              32768  1
xfrm_algo              16384  1 xfrm_user
xt_addrtype            16384  2
iptable_filter         16384  1
iptable_nat            16384  1
nf_conntrack_ipv4      16384  3
nf_defrag_ipv4         16384  1 nf_conntrack_ipv4
nf_nat_ipv4            16384  1 iptable_nat
nf_nat                 32768  2 nf_nat_masquerade_ipv4,nf_nat_ipv4
nf_conntrack          131072  7
bluetooth             544768  43 btrtl,btintel,btbcm,bnep,btusb,rfcomm
```

https://access.redhat.com/security/cve/cve-2019-3874

# Restrict Kernel Modules

```
cat /etc/modprobe.d/blacklist.conf
blacklist sctp
```

```
shutdown -r now
```

```
lsmod | grep sctp
```

https://access.redhat.com/security/cve/cve-2019-3874

# Restrict Kernel Modules

```
cat /etc/modprobe.d/blacklist.conf
blacklist sctp
blacklist dccp
```

```
shutdown -r now
```

```
lsmod | grep dccp
```

https://access.redhat.com/security/cve/cve-2019-3874

# CIS Benchmarks Reference

### *3.4 Uncommon Network Protocols*

The Linux kernel modules support several network protocols that are not commonly used. If these protocols are not needed, it is recommended that they be disabled in the kernel.

**Note:** This should not be considered a comprehensive list of uncommon network protocols, you may wish to consider additions to those listed here for your environment.

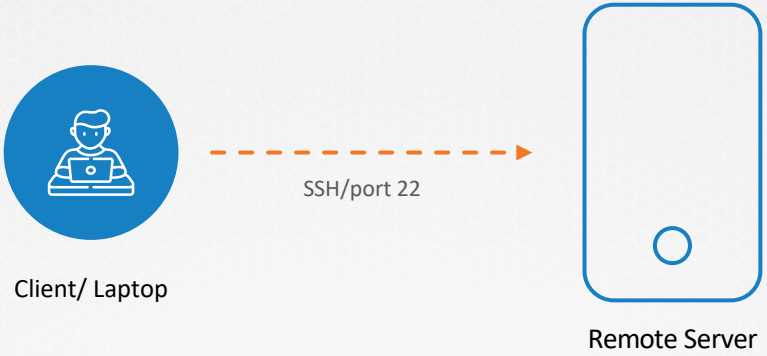
Hands-on Labs

cks.kodekloud.com


www.kodekloud.com

# Disable Open Ports

# Disable Open Ports




```
systemctl status ssh
```

```
● ssh.service - OpenBSD Secure Shell server
   Loaded: loaded (/lib/systemd/system/ssh.service; enabled; vendor preset: enabled)
   Active: active (running) since Wed 2021-03-17 08:33:29 UTC; 54min ago
 Main PID: 759 (sshd)
    Tasks: 1 (limit: 7372)
   CGroup: /system.slice/ssh.service
           └─759 /usr/sbin/sshd -D
```

# Disable Open Ports

```
netstat -an | grep -w LISTEN
tcp        0      0 127.0.0.1:10248         0.0.0.0:*               LISTEN
tcp        0      0 127.0.0.1:10249         0.0.0.0:*               LISTEN
tcp        0      0 127.0.0.1:2379          0.0.0.0:*               LISTEN
tcp        0      0 10.53.64.6:2379         0.0.0.0:*               LISTEN
tcp        0      0 10.53.64.6:2380         0.0.0.0:*               LISTEN
tcp        0      0 127.0.0.1:42893         0.0.0.0:*               LISTEN
tcp        0      0 127.0.0.1:2381          0.0.0.0:*               LISTEN
tcp        0      0 127.0.0.11:46607        0.0.0.0:*               LISTEN
tcp        0      0 0.0.0.0:8080            0.0.0.0:*               LISTEN
tcp        0      0 127.0.0.1:10257         0.0.0.0:*               LISTEN
tcp        0      0 127.0.0.1:10259         0.0.0.0:*               LISTEN
tcp        0      0 127.0.0.53:53           0.0.0.0:*               LISTEN
tcp        0      0 0.0.0.0:22              0.0.0.0:*               LISTEN
tcp6       0      0 :::10250                :::*                    LISTEN
tcp6       0      0 :::6443                 :::*                    LISTEN
tcp6       0      0 :::10256                :::*                    LISTEN
tcp6       0      0 :::22                   :::*                    LISTEN
tcp6       0      0 :::8888                 :::*                    LISTEN
```

# Disable Open Ports

```
cat /etc/services | grep -w 53
domain          53/tcp                          # Domain Name Server
domain          53/udp
```

# Disable Open Ports

## Control-plane node(s)

| Protocol | Direction | Port Range | Purpose | Used By |
|---|---|---|---|---|
| TCP | Inbound | 6443* | Kubernetes API server | All |
| TCP | Inbound | 2379-2380 | etcd server client API | kube-apiserver, etcd |
| TCP | Inbound | 10250 | kubelet API | Self, Control plane |
| TCP | Inbound | 10251 | kube-scheduler | Self |
| TCP | Inbound | 10252 | kube-controller-manager | Self |

## Worker node(s)

| Protocol | Direction | Port Range | Purpose | Used By |
|---|---|---|---|---|
| TCP | Inbound | 10250 | kubelet API | Self, Control plane |
| TCP | Inbound | 30000-32767 | NodePort Services† | All |

† Default port range for NodePort Services.

https://kubernetes.io/docs/setup/production-environment/tools/kubeadm/install-kubeadm/#check-required-ports

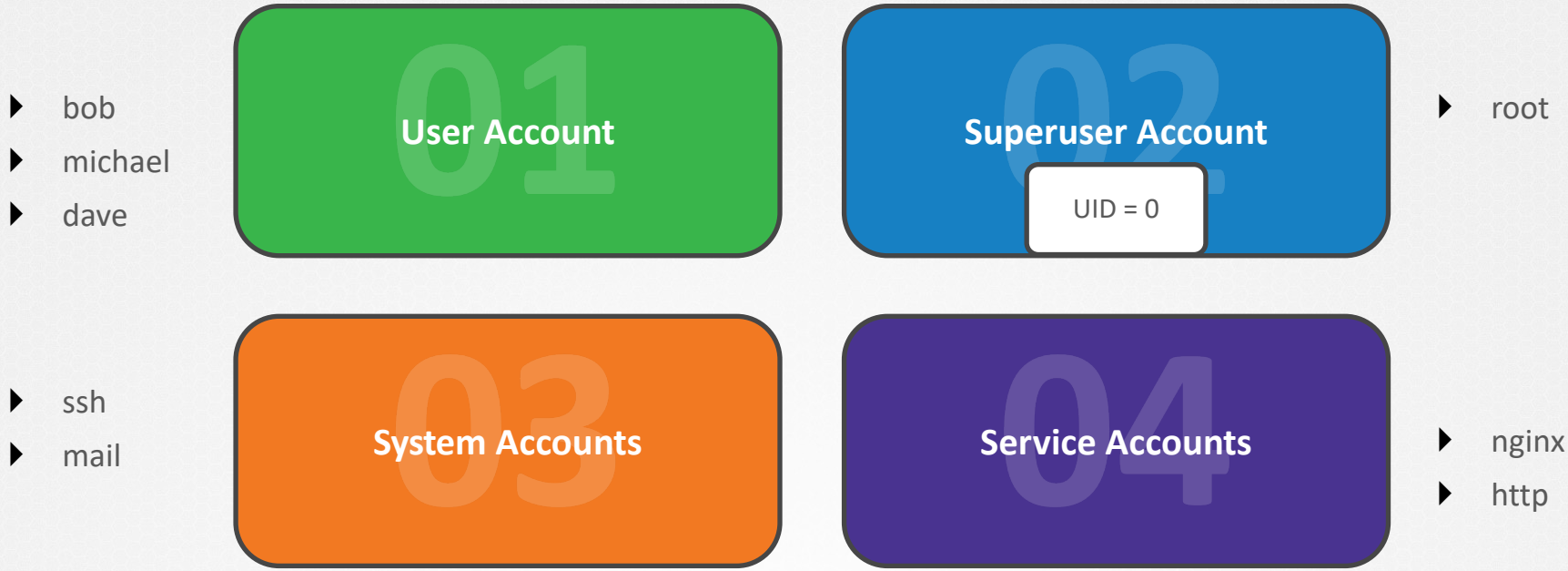
Hands-on Labs

cks.kodekloud.com

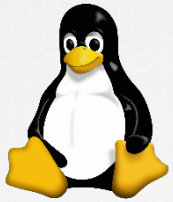

www.kodekloud.com

# Minimize IAM Policies and Roles

# Minimize IAM roles



# Minimize IAM roles



Linux Root User

Windows

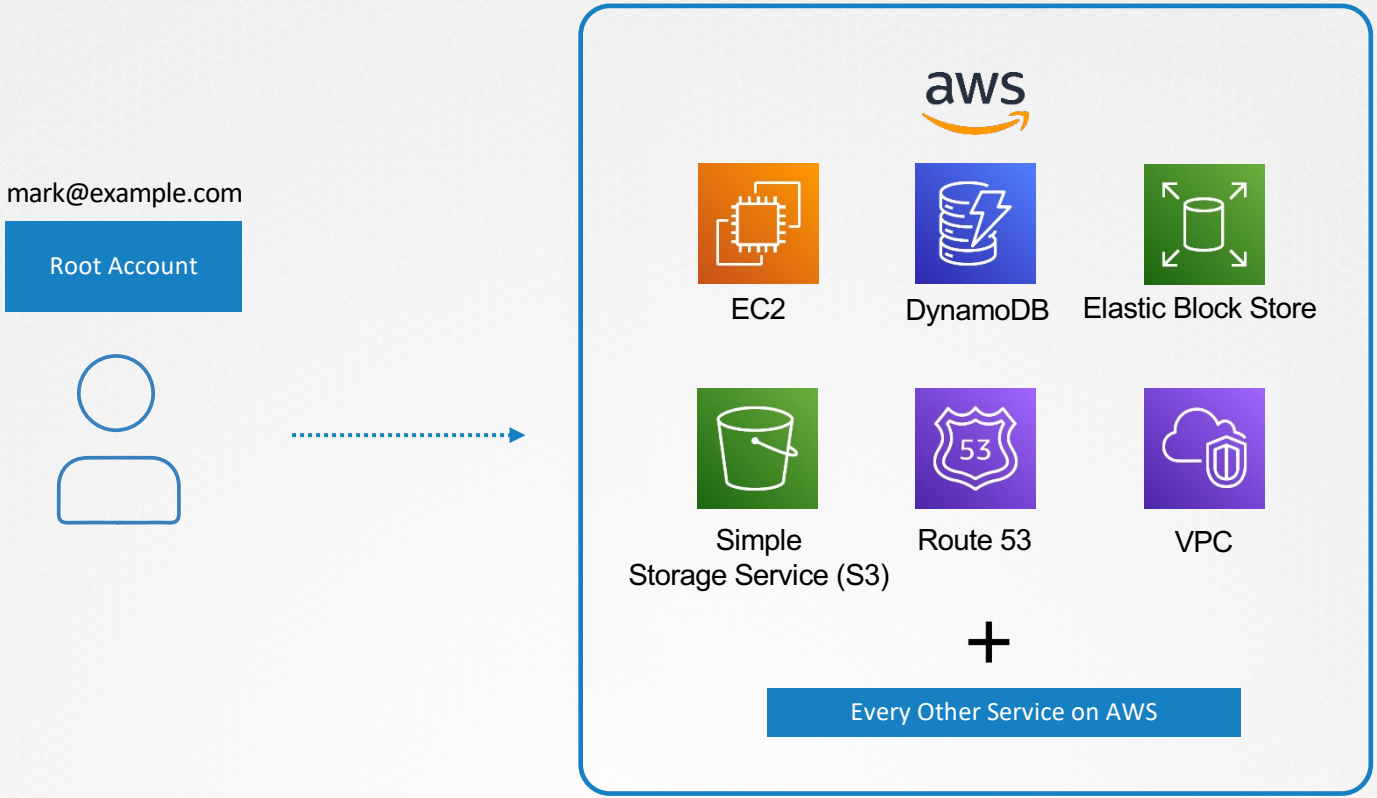
Windows Admin User

aws

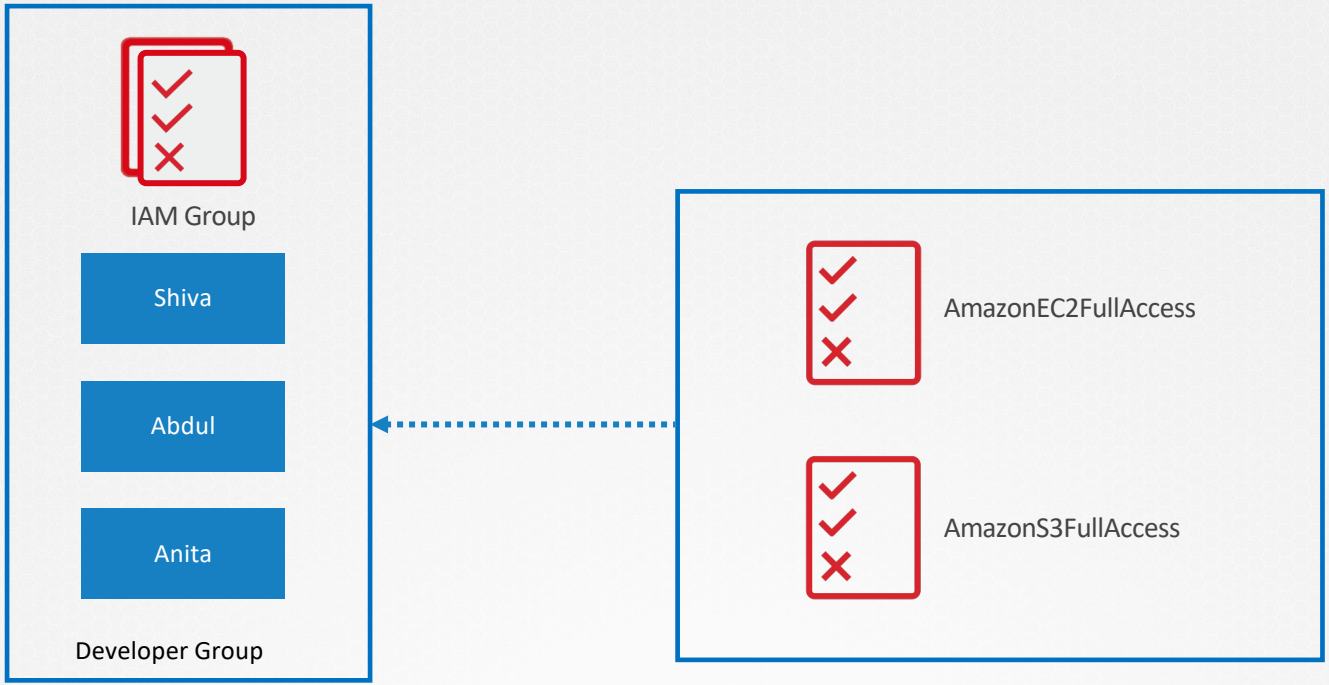
AWS Root Account

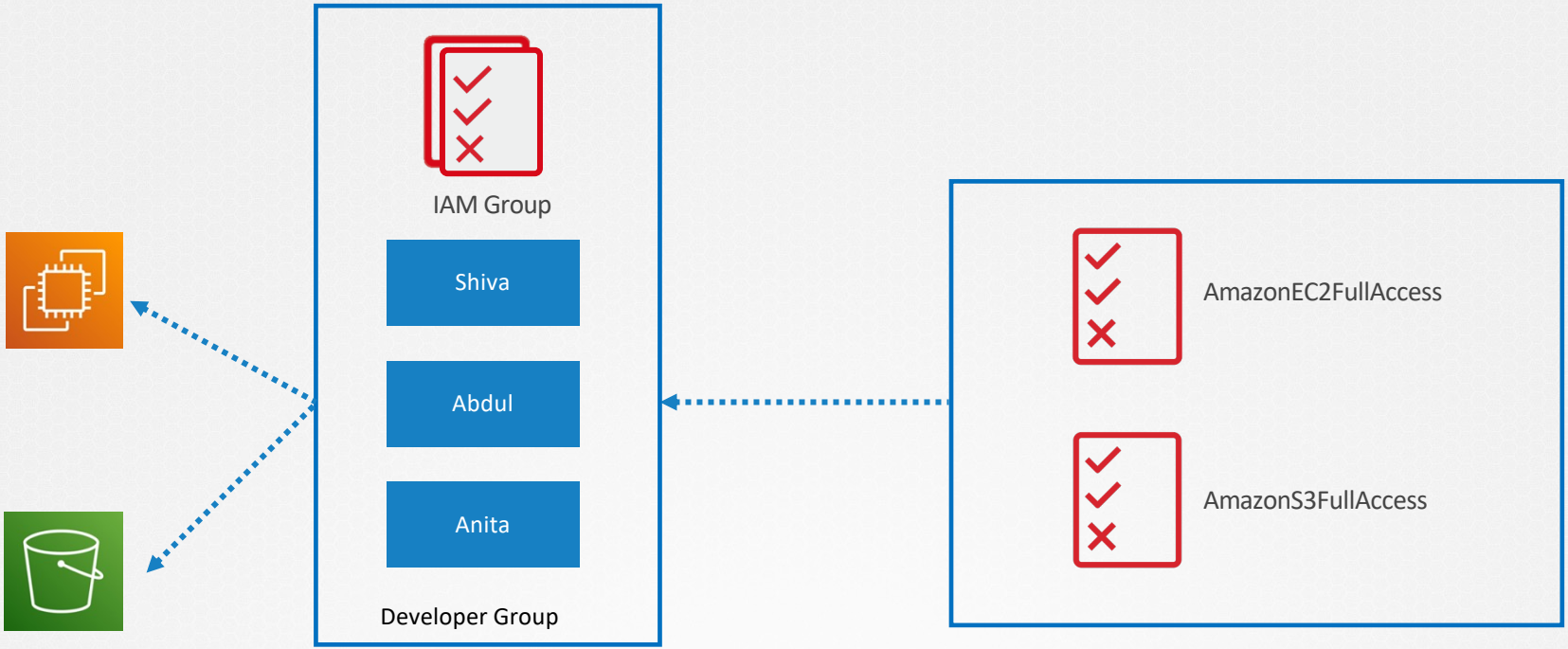
https://portal.aws.amazon.com/billing/signup#/start

# Identity and Access Management in AWS




https://portal.aws.amazon.com/billing/signup#/start





AWS Trusted Advisor

Security Command Center

Azure Advisor

Hands-on Labs

cks.kodekloud.com


www.kodekloud.com

# Restrict Access to External Networks

# Restrict Network Access

```
systemctl status ssh
```

```
● ssh.service - OpenBSD Secure Shell server
   Loaded: loaded (/lib/systemd/system/ssh.service; enabled; vendor preset: enabled)
   Active: active (running) since Wed 2021-03-17 08:33:29 UTC; 54min ago
 Main PID: 759 (sshd)
    Tasks: 1 (limit: 7372)
   CGroup: /system.slice/ssh.service
           └─759 /usr/sbin/sshd -D
```

```
cat /etc/services| grep ssh
```

```
ssh             22/tcp                          # SSH Remote Login Protocol
```

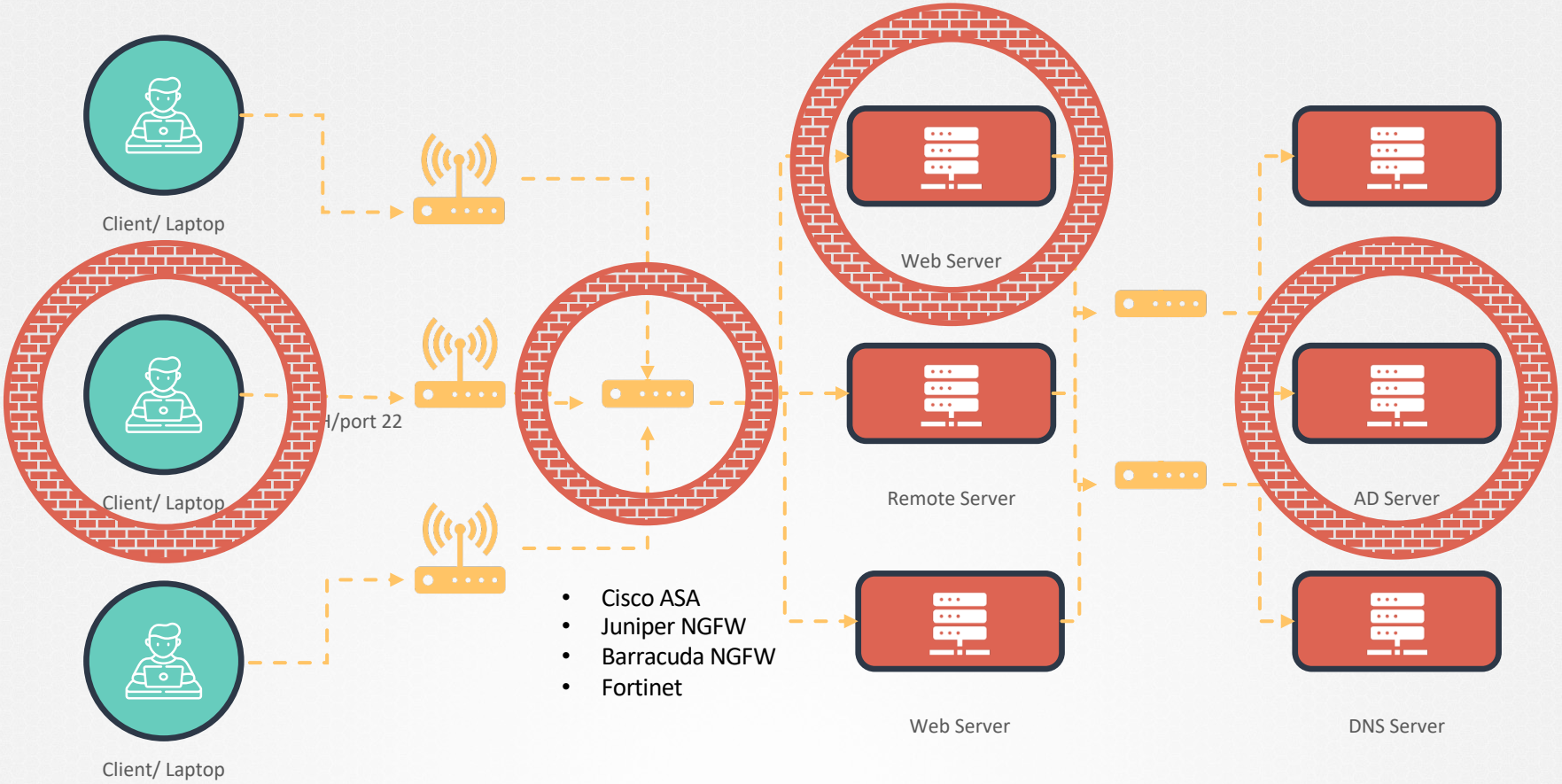
```
netstat -an | grep 22 | grep –w LISTEN
```

```
tcp        0      0 0.0.0.0:22              0.0.0.0:*               LISTEN
tcp6       0      0 :::22                   :::*                    LISTEN
```

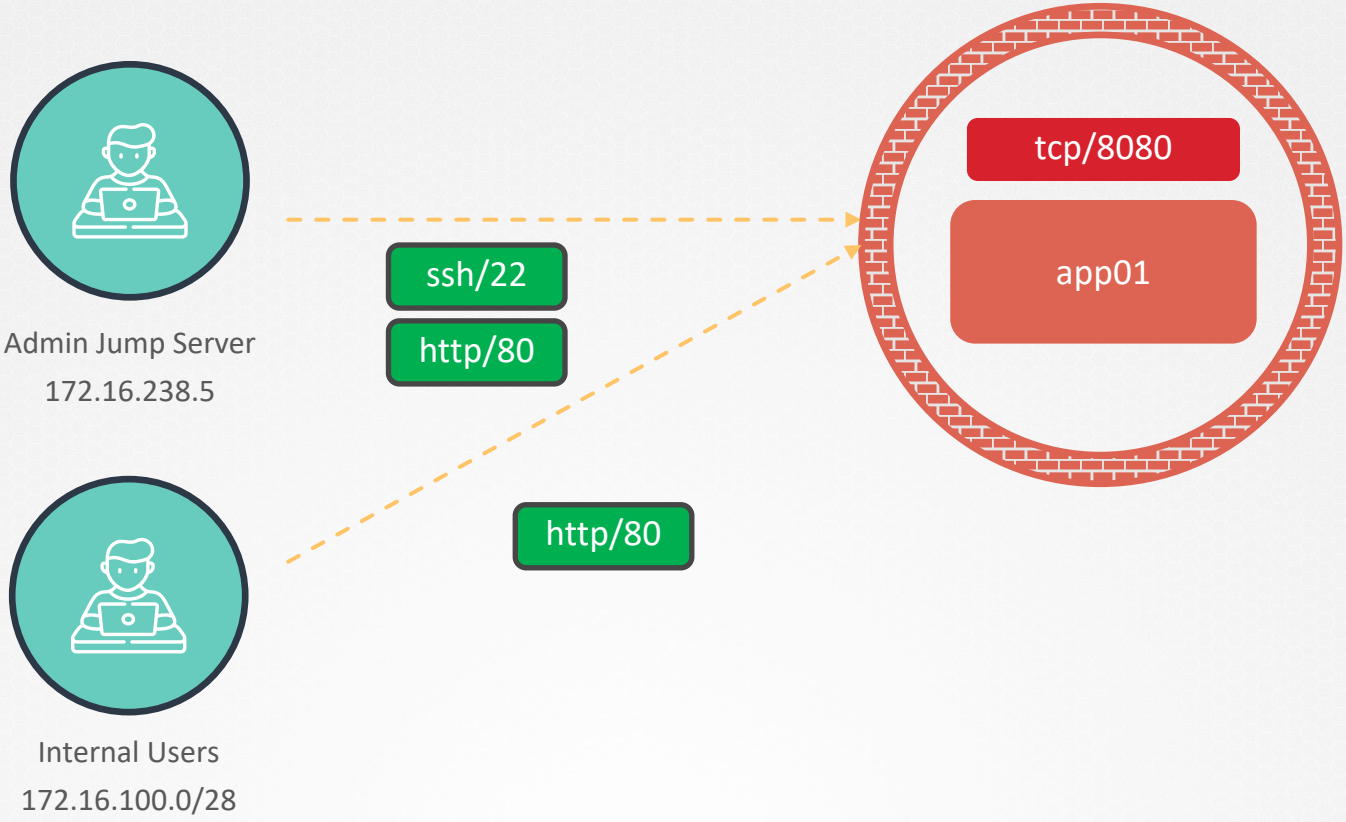
# Restrict Network Access

```
netstat -an | grep -w LISTEN
tcp        0      0 127.0.0.1:10248         0.0.0.0:*               LISTEN
tcp        0      0 127.0.0.1:10249         0.0.0.0:*               LISTEN
tcp        0      0 127.0.0.1:2379          0.0.0.0:*               LISTEN
tcp        0      0 10.53.64.6:2379         0.0.0.0:*               LISTEN
tcp        0      0 10.53.64.6:2380         0.0.0.0:*               LISTEN
tcp        0      0 127.0.0.1:42893         0.0.0.0:*               LISTEN
tcp        0      0 127.0.0.1:2381          0.0.0.0:*               LISTEN
tcp        0      0 127.0.0.11:46607        0.0.0.0:*               LISTEN
tcp        0      0 0.0.0.0:80              0.0.0.0:*               LISTEN
tcp        0      0 0.0.0.0:8080            0.0.0.0:*               LISTEN
tcp        0      0 127.0.0.1:10257         0.0.0.0:*               LISTEN
tcp        0      0 127.0.0.1:10259         0.0.0.0:*               LISTEN
tcp        0      0 127.0.0.53:53           0.0.0.0:*               LISTEN
tcp        0      0 0.0.0.0:22              0.0.0.0:*               LISTEN
tcp6       0      0 :::10250                :::*                    LISTEN
tcp6       0      0 :::6443                 :::*                    LISTEN
tcp6       0      0 :::10256                :::*                    LISTEN
tcp6       0      0 :::22                   :::*                    LISTEN
tcp6       0      0 :::8888                 :::*                    LISTEN
```

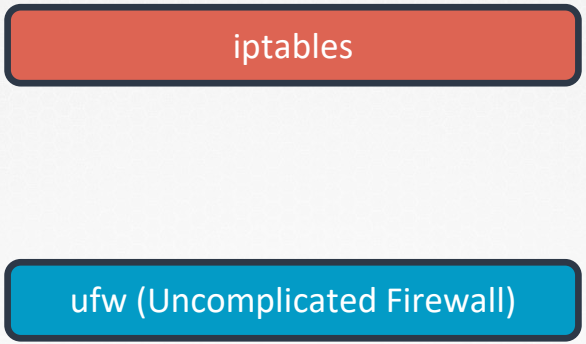
# Restrict Network Access



# UFW

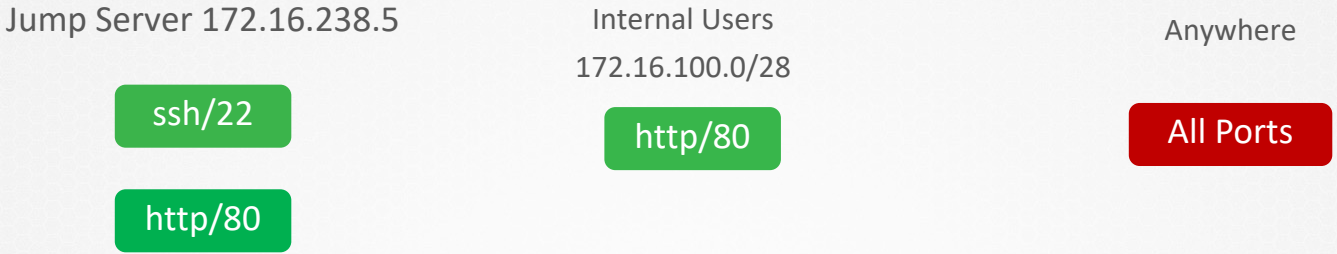


# Install UFW



# Install UFW

```
netstat -an | grep -w LISTEN
tcp        0      0 0.0.0.0:22              0.0.0.0:*               LISTEN
tcp        0      0 0.0.0.0:80              0.0.0.0:*               LISTEN
tcp        0      0 0.0.0.0:8080            0.0.0.0:*               LISTEN
```



# Install UFW

```
apt-get update
```

```
.
.
Fetched 22.3 MB in 6s (3449 kB/s)
Reading package lists... Done
Building dependency tree
Reading state information... Done
21 packages can be upgraded. Run 'apt list --upgradable' to see them.
```

```
apt-get install ufw
```

```
systemctl enable ufw
```

```
Synchronizing state of ufw.service with SysV service script with /lib/systemd/systemd-sysv-install.
Executing: /lib/systemd/systemd-sysv-install enable ufw
```

```
systemctl start ufw
```

# UFW Rules

```
ufw status
```

```
Status: inactive
```

```
ufw default allow outgoing
```

```
Default outgoing policy changed to 'allow'
(be sure to update your rules accordingly)
```

```
ufw default deny incoming
```

```
Default incoming policy changed to 'deny'
(be sure to update your rules accordingly)
```

# UFW Rules

```
ufw allow from 172.16.238.5 to any port 22 proto tcp
Rules updated
```

```
ufw allow from 172.16.238.5 to any port 80 proto tcp
Rules updated
```

```
ufw allow from 172.16.100.0/28 to any port 80 proto tcp
Rules updated
```

```
ufw deny 8080
Rules updated
```

# Enable UFW

```
ufw enable
```

```
Command may disrupt existing ssh connections. Proceed with operation (y|n)? y
Firewall is active and enabled on system startup
```

```
ufw status
```

```
Status: active

To                         Action      From
--                         ------      ----
22/tcp                     ALLOW       172.16.238.5
80/tcp                     ALLOW       172.16.238.5
80/tcp                     ALLOW       172.16.100.0/28
8080                       DENY        Anywhere
8080 (v6)                  DENY        Anywhere (v6)
```

# Delete Rules

```
ufw delete deny 8080
```

```
Rule deleted
Rule deleted (v6)
```

```
ufw status
```

```
Status: active

To                         Action      From
--                         ------      ----
22/tcp                     ALLOW       172.16.238.5      --> 1
80/tcp                     ALLOW       172.16.238.5      --> 2
80/tcp                     ALLOW       172.16.100.0/28   --> 3
8080                       DENY        Anywhere          --> 4
8080 (v6)                  DENY        Anywhere (v6)     --> 5
```

```
ufw delete 5
```

```
Deleting:
 deny 8080
Proceed with operation (y|n)? y
Rule deleted (v6)
```

# Delete Rules

```
ufw status
```

```
Status: active

To                         Action      From
--                         ------      ----
22/tcp                     ALLOW       172.16.238.5      --> 1
80/tcp                     ALLOW       172.16.238.5      --> 2
80/tcp                     ALLOW       172.16.100.0/28   --> 3
8080                       DENY        Anywhere          --> 4
```

```
ufw delete 4
```

```
Deleting:
 deny 8080
Proceed with operation (y|n)? y
Rule deleted
```

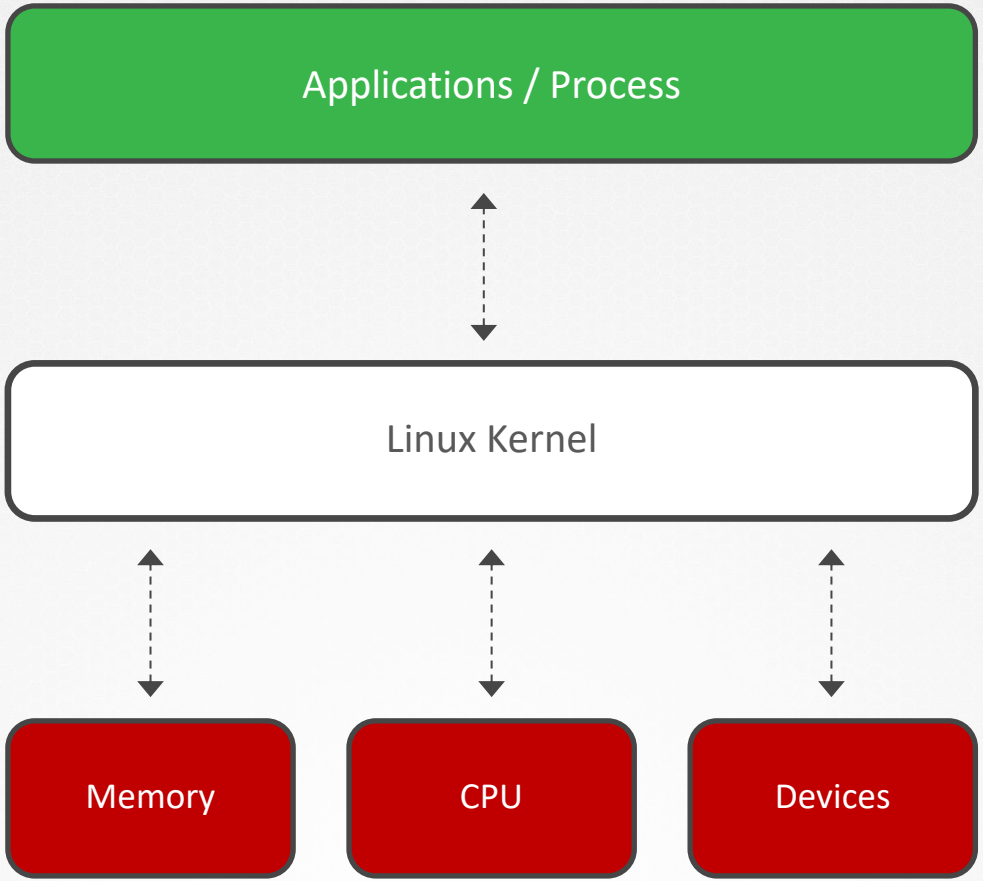
Hands-on Labs

cks.kodekloud.com

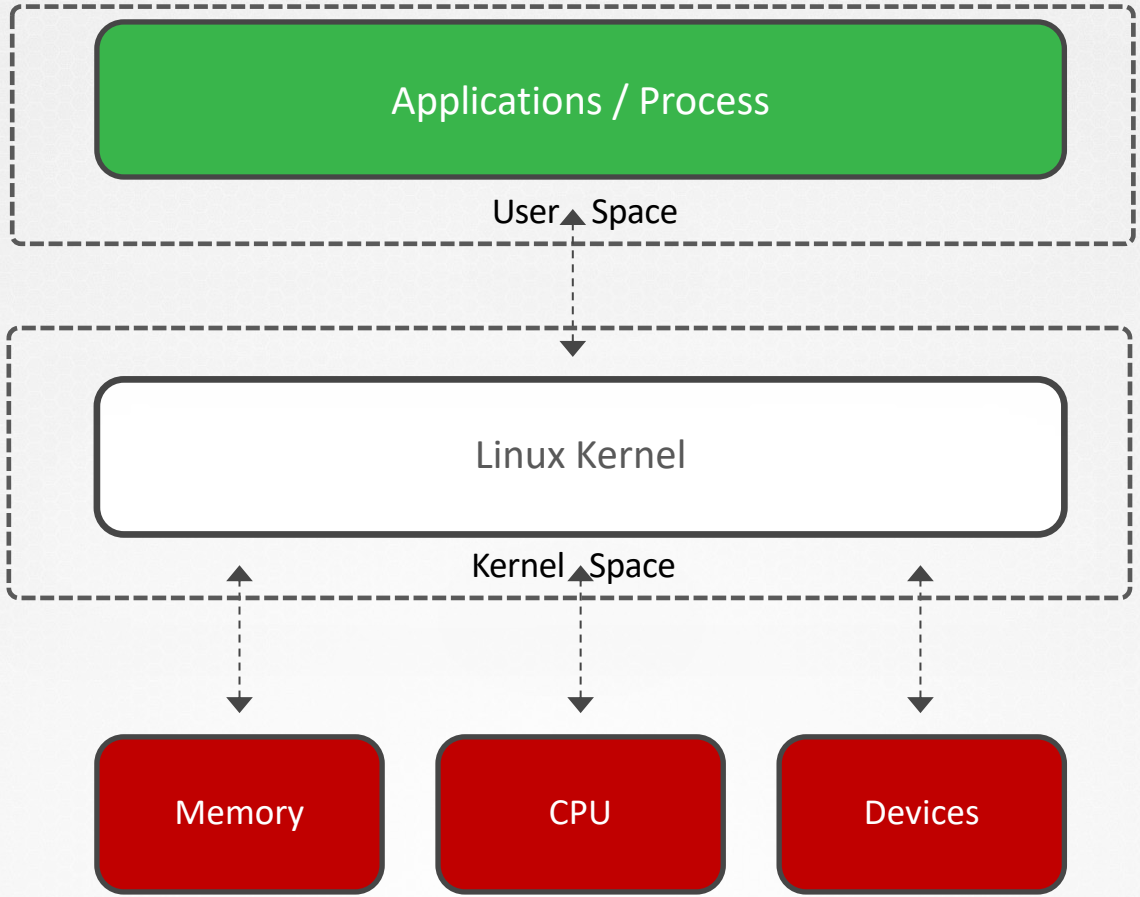

www.kodekloud.com

# LINUX SYSCALLS

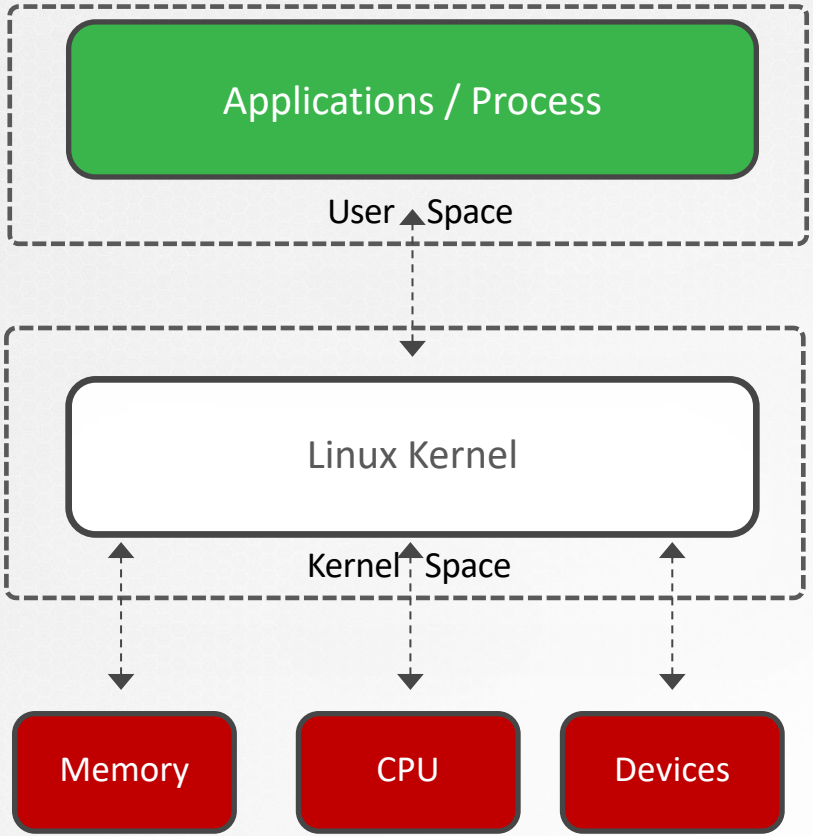
# Linux Kernel



# Linux Kernel



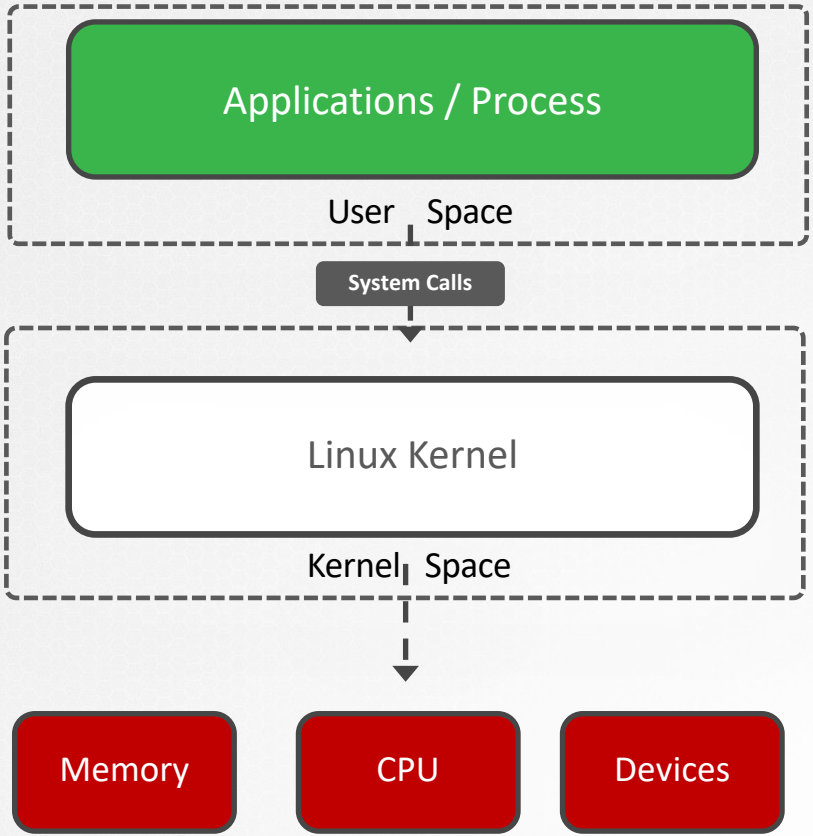
# Linux Kernel




- C
- Java
- Python
- Ruby
- Containers

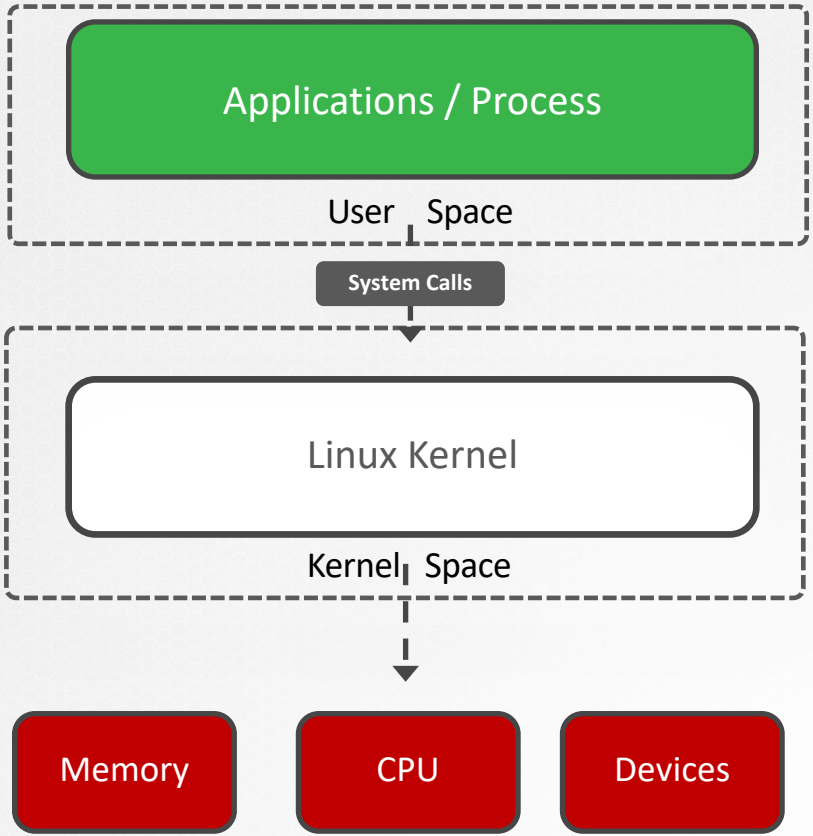
- Kernel Code
- Kernel Extensions
- Device Drivers

# Linux Kernel




- Open a File
- Write to a file
- List Processes
- Defining a variable

# Linux Kernel




```
touch /tmp/error.log
```

- open()
- close()
- execve()
- readdir()
- strlen()
- closedir()

# TRACING SYSCALLS

```
which strace
/usr/bin/strace
```

```
strace touch /tmp/error.log
execve("/usr/bin/touch", ["touch", "/tmp/error.log"], 0x7ffce8f874f8 /* 23 vars */) =
0
.
.
[Output Truncated]
```

# TRACING SYSCALLS

```
strace touch /tmp/error.log
```

```
execve("/usr/bin/touch", ["touch", "/tmp/error.log"], 0x7ffce8f874f8 /* 23 vars */) = 0
.
.
.
[Output Truncated]
```

```
env | wc -l
```

```
23
```

# TRACING SYSCALLS

```
pidof etcd
3596
```

```
strace -p 3596
strace: Process 3596 attached
futex(0x1ac6be8, FUTEX_WAIT_PRIVATE, 0, NULL) = 0
futex(0xc000540bc8, FUTEX_WAKE_PRIVATE, 1) = 1
```

# TRACING SYSCALLS

```
strace -c touch /tmp/error.log
```

```
% time     seconds  usecs/call     calls    errors syscall
------ ----------- ----------- --------- --------- ----------------
  0.00    0.000000           0         1           read
  0.00    0.000000           0         6           close
  0.00    0.000000           0         2           fstat
  0.00    0.000000           0         5           mmap
  0.00    0.000000           0         4           mprotect
  0.00    0.000000           0         1           munmap
  0.00    0.000000           0         3           brk
  0.00    0.000000           0         3         3 access
  0.00    0.000000           0         1           dup2
  0.00    0.000000           0         1           execve
  0.00    0.000000           0         1           arch_prctl
  0.00    0.000000           0         3           openat
  0.00    0.000000           0         1           utimensat
------ ----------- ----------- --------- --------- ----------------
100.00    0.000000                    32         3 total
```


www.kodekloud.com

# Aquasec Tracee

# TRACING SYSCALLS


| Bind Mounts | Purpose |
|---|---|
| /tmp/tracee | Default workspace |
| /lib/modules | Kernel Headers |
| /usr/src | Kernel Headers |

| Additional Capabilities | |
|---|---|
| Privileged | |

# TRACING SYSCALLS

```
docker run --name tracee --rm --privileged --pid=host \
-v /lib/modules/:/lib/modules/:ro -v /usr/src:/usr/src:ro \
-v /tmp/tracee:/tmp/tracee aquasec/tracee:0.4.0 --trace comm=ls
```

```
TIME(s)        UID    COMM             PID     TID     RET    EVENT               ARGS
1263.457188    0      ls               27461   27461   -2     access              pathname: /etc/ld.so.nohwc
1263.457218    0      ls               27461   27461   -2     access              pathname: /etc/ld.so.prelo
1263.457238    0      ls               27461   27461   0      security_file_open  pathname: /etc/ld.so.cache
dev: 265289728, inode: 788102
263.457361     0      ls               27461   27461   0      fstat               fd: 3, statbuf: 0x7FFD469A
1263.457429    0      ls               27461   27461   0      close               fd: 3
.
.
.
[output truncated]
```

# TRACING SYSCALLS

```
sudo docker run --name tracee --rm --privileged --pid=host \
  -v /lib/modules/:/lib/modules/:ro -v /usr/src:/usr/src:ro \
  -v /tmp/tracee:/tmp/tracee aquasec/tracee:0.4.0 --trace pid=new
```

```
1613.769845    0      wc              1619    1619    -2              openat              dirfd: -100, pathname: /usr/sha
langpack/en/LC_MESSAGES/coreutils.mo, flags: O_RDONLY, mode: 0
1613.846148    0      kubectl         1617    1621    -2              openat              dirfd: -100, pathname: /root/.k
flags: O_RDONLY|O_CLOEXEC, mode: 0
1613.848222    0      kubectl         1617    1621    -2              openat              dirfd: -100, pathname: /root/.k
1613.849934    0      kubectl         1617    1621    0               security_file_open  pathname: /etc/nsswitch.conf, f
O_RDONLY|O_LARGEFILE, dev: 265289728, inode: 786661
1613.850907    0      kubectl         1617    1621    6               openat              dirfd: -100, pathname: /etc/nss
flags: O_RDONLY|O_CLOEXEC, mode: 0
1613.852260    0      kubectl         1617    1621    0               close               fd: 6
1613.853264    0      kubectl         1617    1621    0               security_file_open  pathname: /run/systemd/resolve/
resolv.conf, flags: O_RDONLY|O_LARGEFILE, dev: 23, inode: 543
1613.854583    0      kubectl         1617    1621    6               openat              dirfd: -100, pathname: /etc/res
flags: O_RDONLY|O_CLOEXEC, mode: 0
.
.
.
```

# TRACING SYSCALLS

```
sudo docker run --name tracee --rm --privileged --pid=host \
  -v /lib/modules/:/lib/modules/:ro -v /usr/src:/usr/src:ro \
  -v /tmp/tracee:/tmp/tracee aquasec/tracee:0.4.0 --trace container=new
```

```
.
.
.
821.928334     3b392a8f3c57     0      echo             1      /12719   1      /12719   0              fstat                fd: 3, statbuf: 0
821.928354     3b392a8f3c57     0      echo             1      /12719   1      /12719   0              close                fd: 3
821.928551     3b392a8f3c57     0      echo             1      /12719   1      /12719   0              security_file_open   pathname: /usr/li
gnu/libc-2.31.so, flags: O_RDONLY|O_LARGEFILE, dev: 265289728, inode: 3151213
821.928576     3b392a8f3c57     0      echo             1      /12719   1      /12719   3              openat               dirfd: -100, path
/lib/x86_64-linux-gnu/libc.so.6, flags: O_RDONLY|O_CLOEXEC, mode: 0
821.929035     3b392a8f3c57     0      echo             1      /12719   1      /12719   0              fstat                fd: 3, statbuf: 0
821.929690     3b392a8f3c57     0      echo             1      /12719   1      /12719   0              close                fd: 3
821.945346     3b392a8f3c57     0      echo             1      /12719   1      /12719   0              sched_process_exit
.
.
[output truncated]
```

```
docker run ubuntu echo hi
```

```
hi
```


www.kodekloud.com

# RESTRICTING SYSCALLS WITH SECCOMP

# Restricting SYSCALLS

```
strace -c touch /tmp/error.log
```

| % time | seconds | usecs/call | calls | errors | syscall |
|---|---|---|---|---|---|
| 0.00 | 0.000000 | 0 | 1 | | read |
| 0.00 | 0.000000 | 0 | 6 | | close |
| 0.00 | 0.000000 | 0 | 2 | | fstat |
| 0.00 | 0.000000 | 0 | 5 | | mmap |
| 0.00 | 0.000000 | 0 | 4 | | mprotect |
| 0.00 | 0.000000 | 0 | 1 | | munmap |
| 0.00 | 0.000000 | 0 | 3 | | brk |
| 0.00 | 0.000000 | 0 | 3 | 3 | access |
| 0.00 | 0.000000 | 0 | 1 | | dup2 |
| 0.00 | 0.000000 | 0 | 1 | | execve |
| 0.00 | 0.000000 | 0 | 1 | | arch_prctl |
| 0.00 | 0.000000 | 0 | 3 | | openat |
| 0.00 | 0.000000 | 0 | 1 | | utimensat |
| 100.00 | 0.000000 | | 32 | 3 | total |

# Restricting SYSCALLS

ptrace()

CVE List ▾ CNAs ▾ WGs ▾ Board ▾ About ▾ News & Blog ▾

NVD
Go to for:
CVSS Scores
CPE Info

Search CVE List | Downloads | Data Feeds | Update a CVE Record | Request CVE IDs

TOTAL CVE Records: 150428

HOME > CVE > CVE-2016-5195

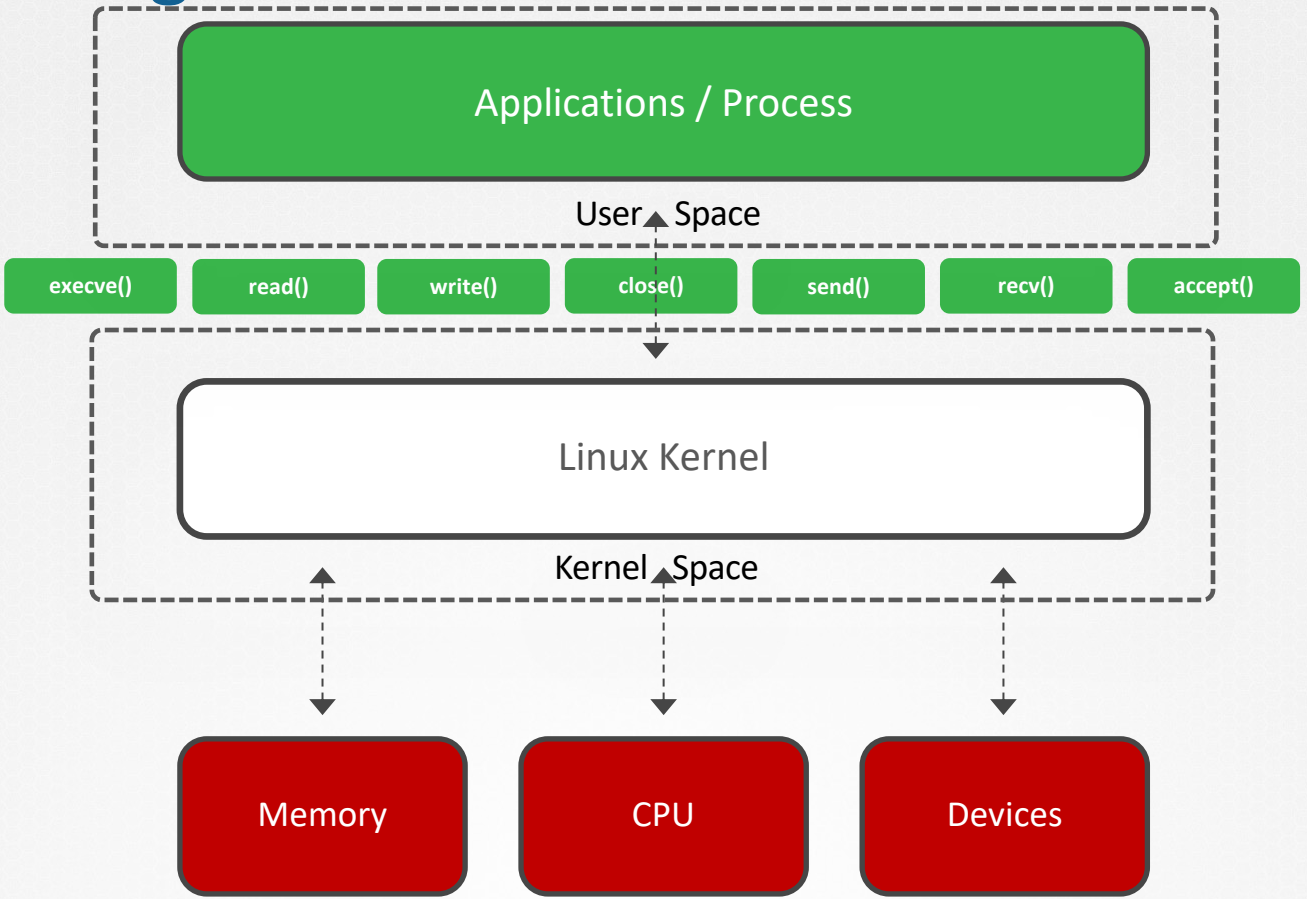
Printer-Friendly View

| CVE-ID | |
|---|---|
| CVE-2016-5195 | Learn more at National Vulnerability Database (NVD)<br>• CVSS Severity Rating • Fix Information • Vulnerable Software Versions • SCAP Mappings • CPE Information |

**Description**

Race condition in mm/gup.c in the Linux kernel 2.x through 4.x before 4.8.3 allows local users to gain privileges by leveraging incorrect handling of a copy-on-write (COW) feature to write to a read-only memory mapping, as exploited in the wild in October 2016, aka "Dirty COW."

# Restricting SYSCALLS



# Restricting SYSCALLS

```
grep -i seccomp /boot/config-$(uname -r)
CONFIG_HAVE_ARCH_SECCOMP_FILTER=y
CONFIG_SECCOMP_FILTER=y
CONFIG_SECCOMP=y
```

```
docker run docker/whalesay cowsay  hello!
 _______
< hello! >
 -------
    \
     \
      \
                    ##        .
              ## ## ##       ==
           ## ## ## ##      ===
       /""""""""""""""""___/ ===
  ~~~ {~~ ~~~~ ~~~ ~~~~ ~~ ~ /  ===- ~~~
       \______ o          __/
        \    \        __/
          \____\______/
```

# Restricting SYSCALLS

```
docker run -it --rm docker/whalesay /bin/sh
```

```
#
# date -s '19 APR 2012 22:00:00'
date: cannot set date: Operation not permitted
```

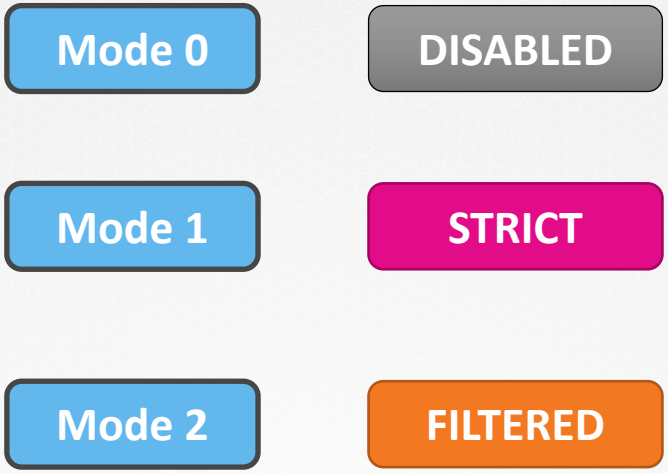
```
ps -ef
```

```
UID          PID    PPID  C STIME TTY          TIME CMD
root           1       0  0 15:44 pts/0    00:00:00 /bin/sh
root          12       1  0 15:47 pts/0    00:00:00 ps -ef
```

```
grep Seccomp /proc/1/status
```

```
Seccomp:        2
```

# Restricting SYSCALLS



https://github.com/moby/moby/blob/master/profiles/seccomp/default.json



# Restricting SYSCALLS

default.json

```
{
    "defaultAction": "SCMP_ACT_ERRNO",
    "architectures": [
        "SCMP_ARCH_X86_64",
        "SCMP_ARCH_X86",
        "SCMP_ARCH_X32"
    ],
    "syscalls": [
        {
            "names": [
                "arch_prctl",
                "brk",
                "capget",
                "capset",
                "mkdir",
                "close",
                "execve",

                .
                .
                "clone"
            ],
            "action": "SCMP_ACT_ALLOW"
        }
```

# Restricting SYSCALLS

whitelist.json

```
{
    "defaultAction": "SCMP_ACT_ERRNO",
    "architectures": [
        "SCMP_ARCH_X86_64",
        "SCMP_ARCH_X86",
        "SCMP_ARCH_X32"
    ],
    "syscalls": [
        {
            "names": [
                "<syscall-1>",
                "<syscall-2>",
                "<syscall-3>"
            ],
            "action": "SCMP_ACT_ALLOW"
        }
    ]
}
```

blacklist.json

```
{
    "defaultAction": "SCMP_ACT_ALLOW",
    "architectures": [
        "SCMP_ARCH_X86_64",
        "SCMP_ARCH_X86",
        "SCMP_ARCH_X32"
    ],
    "syscalls": [
        {
            "names": [
                "<syscall-1>",
                "<syscall-2>",
                "<syscall-3>"
            ],
            "action": "SCMP_ACT_ERRNO"
        }
    ]
}
```

https://docs.docker.com/engine/security/seccomp/#significant-syscalls-blocked-by-the-default-profile



# Restricting SYSCALLS

- clock_adjtime
- settimeofday
- clock_settime
- create_module
- reboot
- swapoff
- mount
- stime
- umount
- delete_module

```
docker run -it --rm docker/whalesay /bin/sh
#
#date -s '19 APR 2012 22:00:00'
date: cannot set date: Operation not permitted
```

# Restricting SYSCALLS

custom.json

```
{
    "defaultAction": "SCMP_ACT_ERRNO",
    "architectures": [
        "SCMP_ARCH_X86_64",
        "SCMP_ARCH_X86",
        "SCMP_ARCH_X32"
    ],
    "syscalls": [
        {
            "names": [
                "arch_prctl",
                "brk",
                "capget",
                "capset",

                "close",
                "execve",

                .
                .
                "clone"
            ],
            "action": "SCMP_ACT_ALLOW"
        }
```

```
docker run -it --rm --security-opt seccomp=/root/custom.json \
docker/whalesay /bin/sh
```

```
/ #

/ # mkdir test

mkdir: can't create directory 'test': Operation not permitted
```

# Restricting SYSCALLS

```
docker run -it --rm --security-opt seccomp=unconfined docker/whalesay  /bin/sh
# date -s '19 APR 2012 22:00:00'
date: cannot set date: Operation not permitted
```

Hands-on Labs

cks.kodekloud.com


www.kodekloud.com

# Implementing Seccomp In Kubernetes

# Seccomp in Kubernetes

```
docker run r.j3ss.co/amicontained amicontained
Container Runtime: docker
Has Namespaces:
        pid: true
        user: false
AppArmor Profile: docker-default (enforce)
Capabilities:
        BOUNDING -> chown dac_override fowner fsetid kill setgid setuid setpcap net_bind_service net_raw
sys_chroot mknod audit_write setfcap
Seccomp: filtering

Blocked Syscalls (64):
        MSGRCV SYSLOG SETPGID SETSID USELIB USTAT SYSFS VHANGUP PIVOT_ROOT _SYSCTL ACCT SETTIMEOFDAY MOUNT
UMOUNT2 SWAPON SWAPOFF REBOOT SETHOSTNAME SETDOMAINNAME IOPL IOPERM CREATE_MODULE INIT_MODULE
DELETE_MODULE GET_KERNEL_SYMS QUERY_MODULE QUOTACTL NFSSERVCTL GETPMSG PUTPMSG AFS_SYSCALL TUXCALL
SECURITY LOOKUP_DCOOKIE CLOCK_SETTIME VSERVER MBIND SET_MEMPOLICY GET_MEMPOLICY KEXEC_LOAD ADD_KEY
REQUEST_KEY KEYCTL MIGRATE_PAGES UNSHARE MOVE_PAGES PERF_EVENT_OPEN FANOTIFY_INIT NAME_TO_HANDLE_AT
OPEN_BY_HANDLE_AT CLOCK_ADJTIME SETNS PROCESS_VM_READV PROCESS_VM_WRITEV KCMP FINIT_MODULE KEXEC_FILE_LOAD
BPF USERFAULTFD MEMBARRIER PKEY_MPROTECT PKEY_ALLOC PKEY_FREE RSEQ
Looking for Docker.sock
```

https://github.com/genuinetools/amicontained

# Seccomp in Kubernetes

```
kubectl run amicontained --image r.j3ss.co/amicontained amicontained -- amicontained
pod/test created
```

```
kubectl logs amicontained
Container Runtime: docker
Has Namespaces:
        pid: true
        user: false
AppArmor Profile: docker-default (enforce)
Capabilities:
        BOUNDING -> chown dac_override fowner fsetid kill setgid setuid setpcap
net_bind_service net_raw sys_chroot mknod audit_write setfcap
Seccomp: disabled

Blocked Syscalls (21):
        SYSLOG SETPGID SETSID VHANGUP PIVOT_ROOT ACCT SETTIMEOFDAY UMOUNT2 SWAPON
SWAPOFF REBOOT SETHOSTNAME SETDOMAINNAME INIT_MODULE DELETE_MODULE LOOKUP_DCOOKIE
KEXEC_LOAD FANOTIFY_INIT OPEN_BY_HANDLE_AT FINIT_MODULE KEXEC_FILE_LOAD
Looking for Docker.sock
```

https://github.com/genuinetools/amicontained

# Seccomp in Kubernetes

pod-definition.yaml

```yaml
apiVersion: v1
kind: Pod
metadata:
  labels:
    run: amicontained
  name: amicontained
spec:
  securityContext:
    seccompProfile:
      type: RuntimeDefault
  containers:
  - args:
    - amicontained
    image: r.j3ss.co/amicontained
    name: amicontained
    securityContext:
      allowPrivilegeEscalation: false
```

# Seccomp in Kubernetes

```
kubectl apply -f pod-definition.yaml
```

```
pod/amicontained created
```

```
kubectl logs amicontained
```

```
Container Runtime: docker
Has Namespaces:
        pid: true
        user: false
AppArmor Profile: docker-default (enforce)
Capabilities:
        BOUNDING -> chown dac_override fowner fsetid kill setgid setuid setpcap net_bind_service net_raw
sys_chroot mknod audit_write setfcap
Seccomp: filtering

Blocked Syscalls (64):
        SYSLOG SETPGID SETSID USELIB USTAT SYSFS VHANGUP PIVOT_ROOT _SYSCTL ACCT SETTIMEOFDAY MOUNT
UMOUNT2 SWAPON SWAPOFF REBOOT SETHOSTNAME SETDOMAINNAME IOPL CREATE_MODULE INIT_MODULE DELETE_MODULE
GET_KERNEL_SYMS QUERY_MODULE QUOTACTL NFSSERVCTL GETPMSG PUTPMSG AFS_SYSCALL TUXCALL SECURITY
LOOKUP_DCOOKIE CLOCK_SETTIME VSERVER MBIND SET_MEMPOLICY GET_MEMPOLICY KEXEC_LOAD ADD_KEY REQUEST_KEY
KEYCTL MIGRATE_PAGES UNSHARE MOVE_PAGES PERF_EVENT_OPEN FANOTIFY_INIT NAME_TO_HANDLE_AT OPEN_BY_HANDLE_AT
CLOCK_ADJTIME SETNS PROCESS_VM_READV PROCESS_VM_WRITEV KCMP FINIT_MODULE KEXEC_FILE_LOAD BPF USERFAULTFD
MEMBARRIER PKEY_MPROTECT PKEY_ALLOC PKEY_FREE RSEQ
Looking for Docker.soc
```

# Seccomp in Kubernetes

pod-definition.yaml

```yaml
apiVersion: v1
kind: Pod
metadata:
  labels:
    run: amicontained
  name: amicontained
spec:
  securityContext:
    seccompProfile:
      type: Unconfined
  containers:
  - args:
    - amicontained
    image: r.j3ss.co/amicontained
    name: amicontained
    securityContext:
      allowPrivilegeEscalation: false
```

# Seccomp in Kubernetes

pod-definition.yaml

```yaml
apiVersion: v1
kind: Pod
metadata:
  name:  test-audit
spec:
  securityContext:
    seccompProfile:
      type: Localhost
      localhostProfile: <path to the custom JSON file>
  containers:
  - command: ["bash", "-c", "echo 'I just made some syscalls' && sleep 100"]
    image:  ubuntu
    name: ubuntu
    securityContext:
      allowPrivilegeEscalation: false
```

/var/lib/kubelet/seccomp

# Seccomp in Kubernetes

```
mkdir -p /var/lib/kubelet/seccomp/profiles
```

/var/lib/kubelet/seccomp/profiles/audit.json

```
{
    "defaultAction": "SCMP_ACT_LOG"
}
```

test-audit.yaml

```
apiVersion: v1
kind: Pod
metadata:
  name: test-audit
spec:
  securityContext:
    seccompProfile:
      type: Localhost
      localhostProfile: profiles/audit.json
  containers:
  - command: ["bash", "-c", "echo 'I just made some syscalls' && sleep 100"]
    image: ubuntu
    name: ubuntu
    securityContext:
      allowPrivilegeEscalation: false
```

# Seccomp in Kubernetes

```
grep syscall /var/log/syslog
Mar 19 23:53:45 node01 kernel: [  264.340952] audit: type=1326 audit(1616198025.076:14): auid=4294967295
uid=0 gid=0 ses=4294967295 pid=8816 comm="runc:[2:INIT]" exe="/" sig=0 arch=c000003e syscall=257 compat=0
ip=0x56428101a0aa code=0x7ffc0000
Mar 19 23:53:45 node01 kernel: [  264.340954] audit: type=1326 audit(1616198025.076:15): auid=4294967295
uid=0 gid=0 ses=4294967295 pid=8816 comm="runc:[2:INIT]" exe="/" sig=0 arch=c000003e syscall=35 compat=0
ip=0x564280fc662d code=0x7ffc0000
Mar 19 23:53:45 node01 kernel: [  264.340970] audit: type=1326 audit(1616198025.076:16): auid=4294967295
uid=0 gid=0 ses=4294967295 pid=8816 comm="runc:[2:INIT]" exe="/" sig=0 arch=c000003e syscall=233 compat=0
ip=0x564280fc6d48 code=0x7ffc0000
Mar 19 23:53:45 node01 kernel: [  264.340991] audit: type=1326 audit(1616198025.076:18): auid=4294967295
uid=0 gid=0 ses=4294967295 pid=8816 comm="runc:[2:INIT]" exe="/" sig=0 arch=c000003e syscall=138 compat=0
ip=0x56428101a030 code=0x7ffc0000
Mar 19 23:53:45 node01 kernel: [  264.341026] audit: type=1326 audit(1616198025.076:19): auid=4294967295
uid=0 gid=0 ses=4294967295 pid=8816 comm="runc:[2:INIT]" exe="/" sig=0 arch=c000003e syscall=217 compat=0
ip=0x56428101a030 code=0x7ffc0000
Mar 19 23:53:45 node01 kernel: [  264.341027] audit: type=1326 audit(1616198025.076:20): auid=4294967295
uid=0 gid=0 ses=4294967295 pid=8816 comm="runc:[2:INIT]" exe="/" sig=0 arch=c000003e syscall=35 compat=0
ip=0x564280fc662d code=0x7ffc0000
Mar 19 23:55:25 node01 kernel: [  364.361771] audit: type=1326 audit(1616198125.096:410): auid=4294967295
uid=0 gid=0 ses=4294967295 pid=8816 comm="sleep" exe="/usr/bin/sleep" sig=0 arch=c000003e syscall=231
compat=0 ip=0x7f74114032c6 code=0x7ffc0000
Mar 19 23:55:25 node01 kernel: [  364.361704] audit: type=1326 audit(1616198125.096:408): auid=4294967295
uid=0 gid=0 ses=4294967295 pid=8816 comm="sleep" exe="/usr/bin/sleep" sig=0 arch=c000003e syscall=3
compat=0 ip=0x7f74114334ab code=0x7ffc0000
```

# Seccomp in Kubernetes

```
grep -w 35 /usr/include/asm/unistd_64.h
#define __NR_nanosleep 35
```

| SYSCALL NUMBER | SYSCALL NAME |
|---|---|
| 3 | close |
| 35 | nanosleep |
| 72 | fcntl |
| 138 | fstatfs |
| 217 | getdents64 |
| 231 | exit_group |
| 233 | epoll_ctl |
| 257 | openat |

# Seccomp in Kubernetes


```
sudo docker run --name tracee --rm --privileged --pid=host \
-v /lib/modules/:/lib/modules/:ro -v /usr/src:/usr/src:ro \
-v /tmp/tracee:/tmp/tracee aquasec/tracee:0.4.0 --trace container=new
```

```
1788.624896   test-audit      0      bash            1      /23676   1      /23676   0              fstat
1788.624916   test-audit      0      bash            1      /23676   1      /23676   0              close
1788.624969   test-audit      0      bash            1      /23676   1      /23676   0              fstat
1788.625085   test-audit      0      bash            1      /23676   1      /23676   0              close
1788.625160   test-audit      0      bash            1      /23676   1      /23676   0              security_file_open
1788.625172   test-audit      0      bash            1      /23676   1      /23676   3              openat
1788.625193   test-audit      0      bash            1      /23676   1      /23676   0              fstat
1788.625218   test-audit      0      bash            1      /23676   1      /23676   0              close
1788.625382   test-audit      0      bash            1      /23676   1      /23676   0              fstat
1788.625422   test-audit      0      bash            1      /23676   1      /23676   0              stat
1788.625437   test-audit      0      bash            1      /23676   1      /23676   -2             stat
.
.
.
[output truncated]
```

# Seccomp in Kubernetes

/var/lib/kubelet/seccomp/profiles /violation.json

```
{
    "defaultAction": "SCMP_ACT_ERRNO"
}
```

test-violation.yaml

```
apiVersion: v1
kind: Pod
metadata:
  name: test-violation
spec:
  securityContext:
    seccompProfile:
      type: Localhost
      localhostProfile: profiles/violation.json
  containers:
  - command: ["bash", "-c", "echo 'I just made some syscalls' && sleep 100"]
    image: ubuntu
    name: ubuntu
  restartPolicy: Never
```

# Seccomp in Kubernetes

```
kubectl get pods
pod/test-violation created
```

```
kubectl apply -f test-violation.yaml
NAME              READY   STATUS                 RESTARTS   AGE
test-violation    0/1     ContainerCannotRun     0          2m2s
```

# Seccomp in Kubernetes

test-custom.yaml

```yaml
apiVersion: v1
kind: Pod
metadata:
  name: test-custom
spec:
  securityContext:
    seccompProfile:
      type: Localhost
      localhostProfile: profiles/custom.json
  containers:
  - command: ["bash", "-c", "echo 'I just made some syscalls' && sleep 100"]
    image: ubuntu
    name: ubuntu
  restartPolicy: Never
```

kubectl get pods

```
NAME          READY   STATUS    RESTARTS   AGE
test-custom   1/1     Running   0          2m2s
```

# Exam Tip

https://kubernetes.io/docs/tutorials/clusters/seccomp/

Hands-on Labs

cks.kodekloud.com


www.kodekloud.com

# AppArmor

# AppArmor

profile.json

```
{
    "defaultAction": "SCMP_ACT_ERRNO",
    "architectures": [
        "SCMP_ARCH_X86_64",
        "SCMP_ARCH_X86",
        "SCMP_ARCH_X32"
    ],
    "syscalls": [
        {
            "names": [
                "execve",
                "close",

                "brk",
                .
                .
              ],
            "action": "SCMP_ACT_ALLOW"
        }
    ]
}
```

```
docker run -it --security-opt seccomp=/root/custom.json docker/whalesay /bin/sh
/ #
/ # mkdir test
mkdir: can't create directory 'test': Operation not permitted
```

# AppArmor

```
systemctl status apparmor
```

```
● apparmor.service - AppArmor initialization
   Loaded: loaded (/lib/systemd/system/apparmor.service; enabled; vendor preset:
enabled)
   Active: active (exited) since Mon 2021-03-22 03:12:41 UTC; 2min 29s ago
     Docs: man:apparmor(7)
           http://wiki.apparmor.net/
 Main PID: 313 (code=exited, status=0/SUCCESS)
    Tasks: 0 (limit: 4678)
   CGroup: /system.slice/apparmor.service
```

# AppArmor

```
cat /sys/module/apparmor/parameters/enabled
Y
```

```
cat /sys/kernel/security/apparmor/profiles
docker-default (enforce)
/usr/sbin/tcpdump (enforce)
/usr/sbin/ntpd (enforce)
/usr/lib/snapd/snap-confine (enforce)
/usr/lib/snapd/snap-confine//mount-namespace-capture-helper
(enforce)
/usr/lib/connman/scripts/dhclient-script (enforce)
/usr/lib/NetworkManager/nm-dhcp-helper (enforce)
/usr/lib/NetworkManager/nm-dhcp-client.action (enforce)
/sbin/dhclient (enforce)
man_groff (enforce)
man_filter (enforce)
/usr/bin/man (enforce)
```

# AppArmor Profile

apparmor-deny-write

```
profile apparmor-deny-write flags=(attach_disconnected) {
  file,
  # Deny all file writes.
  deny /** w,
}
```

apparmor-deny-proc-write

```
profile apparmor-deny-proc-write flags=(attach_disconnected) {
  file,
  # Deny all file writes to /proc.
  deny /proc/* w,
}
```

# AppArmor Profile

apparmor-deny-remount-root

```
profile apparmor-deny-remount-root flags=(attach_disconnected) {

  # Deny remount readonly the root filesystem.
  deny mount options=(ro, remount) ->/,
}
```

# AppArmor

```
aa-status
apparmor module is loaded.
12 profiles are loaded.
12 profiles are in enforce mode.
   /sbin/dhclient
   /usr/bin/man
   /usr/lib/NetworkManager/nm-dhcp-client.action
   /usr/lib/NetworkManager/nm-dhcp-helper
.
.
/usr/sbin/tcpdump
   docker-default
   man_filter
   man_groff
0 profiles are in complain mode.
11 processes have profiles defined.
11 processes are in enforce mode.
   /sbin/dhclient (621)
   docker-default (3970)
   docker-default (4025)
docker-default (9853)
   docker-default (9964)
0 processes are in complain mode.
0 processes are unconfined but have a profile defined.
```

enforce

complain

unconfined

# Creating AppArmor Profiles

# AppArmor

add_data.sh

```
 #!/bin/bash
data_directory=/opt/app/data
mkdir -p ${data_directory}
echo "=>  File created at `date`" | tee ${data_directory}/create.log
```

```
./add_data.sh
=>  File created at Mon Mar 12 03:29:22 UTC 2021
```

```
cat /opt/app/data/create.log
=>  File created at Mon Mar 12 03:29:22 UTC 2021
```

# AppArmor

```
apt-get install -y apparmor-utils
Reading package lists... Done
Building dependency tree
Reading state information... Done
The following packages were automatically installed and are no longer
required:
  libc-ares2 libhttp-parser2.7.1 libnetplan0 libuv1 nodejs-doc python3-
netifaces
Use 'apt autoremove' to remove them.
The following additional packages will be installed:
  python3-apparmor python3-libapparmor
Suggested packages:
  vim-addon-manager
The following NEW packages will be installed:
  apparmor-utils python3-apparmor python3-libapparmor
.
.
Unpacking apparmor-utils (2.12-4ubuntu5.1) ...
Setting up python3-libapparmor (2.12-4ubuntu5.1) ...
Setting up python3-apparmor (2.12-4ubuntu5.1) ...
Setting up apparmor-utils (2.12-4ubuntu5.1) ...
Processing triggers for man-db (2.8.3-2ubuntu0.1) ...
```

# AppArmor

```
./add_data.sh
=>  File created at Mon Mar 12 03:42:22 UTC 2021
```

```
aa-genprof /root/add_data.sh
```

```
root@kodekloud:~# aa-genprof /root/add_data.sh
Writing updated profile for /root/add_data.sh.
Setting /root/add_data.sh to complain mode.

Before you begin, you may wish to check if a
profile already exists for the application you
wish to confine. See the following wiki page for
more information:
https://gitlab.com/apparmor/apparmor/wikis/Profiles

Profiling: /root/add_data.sh

Please start the application to be profiled in
another window and exercise its functionality now.

Once completed, select the "Scan" option below in
order to scan the system logs for AppArmor events.

For each AppArmor event, you will be given the
opportunity to choose whether the access should be
allowed or denied.

[(S)can system log for AppArmor events] / (F)inish
```

```
Once completed, select the "Scan" option below in
order to scan the system logs for AppArmor events.

For each AppArmor event, you will be given the
opportunity to choose whether the access should be
allowed or denied.

[(S)can system log for AppArmor events] / (F)inish



.
.
.
Reading log entries from /var/log/syslog.
Updating AppArmor profiles in /etc/apparmor.d.

Profile:  /root/add_data.sh
Execute:  /usr/bin/mkdir
Severity: unknown

(I)nherit / (C)hild / (N)amed / (X) ix On / (D)eny / Abo(r)t / (F)inish
```

```
.
Reading log entries from /var/log/syslog.
Updating AppArmor profiles in /etc/apparmor.d.

Profile:  /root/add_data.sh
Execute:  /usr/bin/mkdir
Severity: unknown

(I)nherit / (C)hild / (N)amed / (X) ix On / (D)eny / Abo(r)t / (F)inish



.
.
.
Profile:  /root/add_data.sh
Execute:  /usr/bin/tee
Severity: 3

(I)nherit / (C)hild / (P)rofile / (N)amed / (U)nconfined / (X) ix On /
(D)eny / Abo(r)t / (F)inish
```

```
.
.
.
Profile:  /root/add_data.sh
Path:     /dev/tty
New Mode: owner rw
Severity: 9
(A)llow / [(D)eny] / (I)gnore / (G)lob / Glob with (E)xtension / (N)ew /
Audi(t) / (O)wner permissions off / Abo(r)t / (F)inish
Adding #include <abstractions/consoles> to profile.
```

```
.
.
.
Profile:  /root/add_data.sh
Path:     /proc/filesystems
New Mode: owner r
Severity: 6
[1 - owner /proc/filesystems r,]
(A)llow / [(D)eny] / (I)gnore / (G)lob / Glob with (E)xtension / (N)ew /
Audi(t) / (O)wner permissions off / Abo(r)t / (F)inish
Adding deny owner /proc/filesystems r, to profile.
```

```
= Changed Local Profiles =

The following local profiles were changed. Would you like to save them?

 [1 - /root/add_data.sh]
(S)ave Changes / Save Selec(t)ed Profile / [(V)iew Changes] / View Changes
b/w (C)lean profiles / Abo(r)t
Writing updated profile for /root/add_data.sh.

Profiling: /root/add_data.sh

.
.
For each AppArmor event, you will be given the
opportunity to choose whether the access should be
allowed or denied.

[(S)can system log for AppArmor events] / (F)inish
Setting /root/add_data.sh to enforce mode.

Finished generating profile for /root/add_data.sh.
```

# AppArmor

```
aa-status
apparmor module is loaded.
13 profiles are loaded.
13 profiles are in enforce mode.
   /root/add_data.sh
   /sbin/dhclient
   /usr/bin/man
   /usr/lib/NetworkManager/nm-dhcp-client.action
   /usr/lib/NetworkManager/nm-dhcp-helper
.
.
.
/usr/sbin/tcpdump
   docker-default
   man_filter
   man_groff
0 profiles are in complain mode.
11 processes have profiles defined.
11 processes are in enforce mode.
   /root/add_data.sh
   /sbin/dhclient (621)
   docker-default (3970)
   docker-default (4025)
docker-default (9853)
   docker-default (9964)
0 processes are in complain mode.
0 processes are unconfined but have a profile defined.
```

# AppArmor

```
cat /etc/apparmor.d/root.add_data.sh
# Last Modified: Mon Mar 22 11:21:42 2021
#include <tunables/global>

/root/add_data.sh {
  #include <abstractions/base>
  #include <abstractions/bash>
  #include <abstractions/consoles>

  deny owner /proc/filesystems r,

  /root/add_data.sh r,
  /usr/bin/bash ix,
  /usr/bin/date mrix,
  /usr/bin/mkdir mrix,
  /usr/bin/tee mrix,
  owner /opt/app/ rw,
  owner /opt/app/data/ w,
  owner /opt/app/data/create.log w,

}
```

# AppArmor

```
cat add_data.sh
#!/bin/bash
data_directory=/opt
mkdir -p ${data_directory}
echo "=>  File created at `date`" | tee ${data_directory}/create.log
```

```
./add_data.sh
./add_data.sh
tee: /opt/create.log: Permission denied
=>  File created at Mon 22 Mar 2021 04:04:47 PM EDT
```

# AppArmor

```
# apparmor_parser /etc/apparmor.d/root.add_data.sh
#
```

```
# apparmor_parser -R /etc/apparmor.d/root.add_data.sh
#
```

```
# ln -s /etc/apparmor.d/root.add_data.sh /etc/apparmor.d/disable/
#
```

Hands-on Labs

cks.kodekloud.com


www.kodekloud.com

# AppArmor in Kubernetes

# AppArmor in Kubernetes

K8 Version > 1.4

AppArmor Kernel Module Enabled

AppArmor Profile Loaded in the Kernel

Container Runtime should be Supported

# AppArmor in Kubernetes

ubuntu-sleeper.yaml

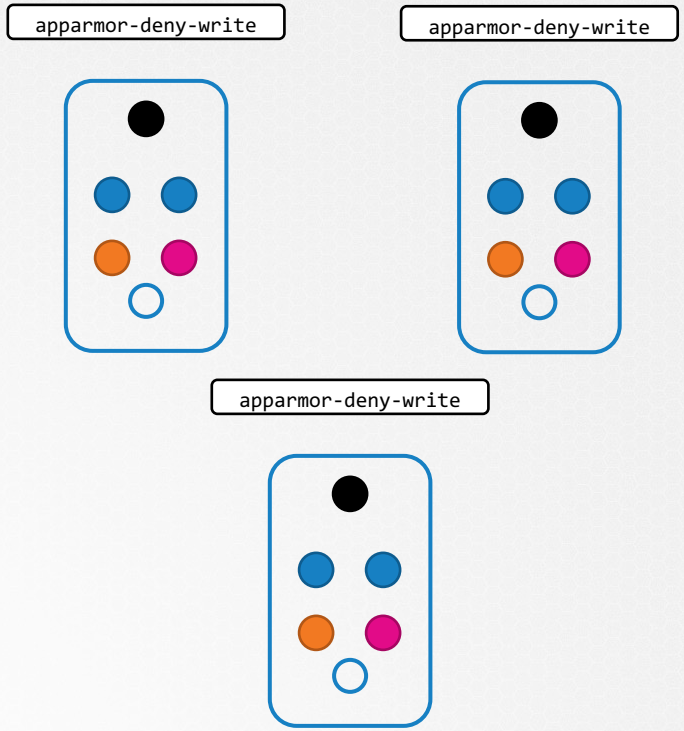
```yaml
apiVersion: v1
kind: Pod
metadata:
  name: ubuntu-sleeper
spec:
  containers:
  - name: hello
    image: ubuntu
    command: [ "sh", "-c", "echo 'Sleeping for an hour!' && sleep 1h" ]
```

apparmor-deny-write

```
profile apparmor-deny-write flags=(attach_disconnected) {
  file,
  # Deny all file writes.
  deny /** w,
}
```

# AppArmor in Kubernetes

```
aa-status
apparmor module is loaded.
13 profiles are loaded.
13 profiles are in enforce mode.
   apparor-deny-write
   /sbin/dhclient
   /usr/bin/man
   /usr/lib/NetworkManager/nm-dhcp-client.action
   /usr/lib/NetworkManager/nm-dhcp-helper
.
.
/usr/sbin/tcpdump
   docker-default
   man_filter
   man_groff
0 profiles are in complain mode.
11 processes have profiles defined.
11 processes are in enforce mode.
   /sbin/dhclient (621)
   docker-default (3970)
   docker-default (4025)
docker-default (9853)
   docker-default (9964)
0 processes are in complain mode.
0 processes are unconfined but have a profile defined.
```



# AppArmor in Kubernetes

ubuntu-sleeper.yaml

```yaml
apiVersion: v1
kind: Pod
metadata:
  name: ubuntu-sleeper
  annotations:
    container.apparmor.security.beta.kubernetes.io/<container_name>: localhost/<profile-name>
spec:
  containers:
  - name: ubuntu-sleeper
    image: ubuntu
    command: [ "sh", "-c", "echo 'Sleeping for an hour!' && sleep 1h" ]
```

# AppArmor in Kubernetes

ubuntu-sleeper.yaml

```yaml
apiVersion: v1
kind: Pod
metadata:
  name: ubuntu-sleeper
  annotations:
    container.apparmor.security.beta.kubernetes.io/ubuntu-sleeper: localhost/apparmor-deny-write
spec:
  containers:
  - name: ubuntu-sleeper
    image: ubuntu
    command: [ "sh", "-c", "echo 'Sleeping for an hour!' && sleep 1h" ]
```

# AppArmor in Kubernetes

```
kubectl create -f ubuntu-sleeper.yaml
pod/ubuntu-sleeper created
```

```
kubectl logs ubuntu-sleeper
Sleeping for an hour!
```

```
kubectl exec -ti ubuntu-sleeper -- touch /tmp/test
touch: cannot touch '/tmp/test': Permission denied
command terminated with exit code 1
```

Hands-on Labs

cks.kodekloud.com


www.kodekloud.com

# Linux Capabilities

# Linux Capabilities

```
docker run -it --rm --security-opt seccomp=unconfined docker/whalesay  /bin/sh
```

```
#  date -s '19 APR 2012 22:00:00'
date: cannot set date: Operation not permitted
Thu Apr 19 22:00:00 UTC 2012
```

```
kubectl run --rm -it ubuntu-sleeper --image=ubuntu -- bash
```

```
If you don't see a command prompt, try pressing enter.
root@ubuntu-sleeper:/#  date -s '19 APR 2012 22:00:00'
date: cannot set date: Operation not permitted
Thu Apr 19 22:00:00 UTC 2012
root@ubuntu-sleeper:/#

root@ubuntu-sleeper:/# whoami
root
root@ubuntu-sleeper:/#

root@ubuntu-sleeper:/# id
uid=0(root) gid=0(root) groups=0(root)
root@ubuntu-sleeper:/#
```

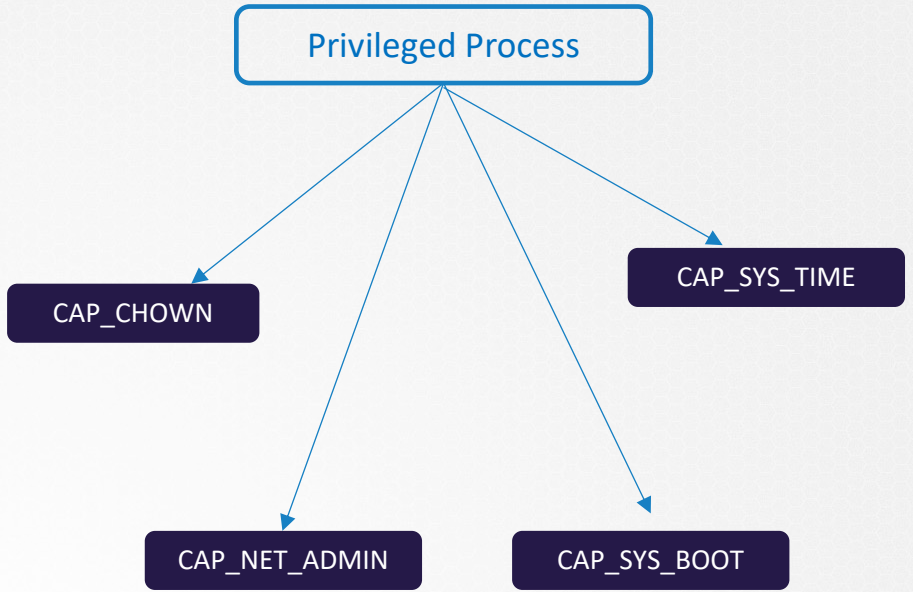
# Linux Capabilities

< Kernel 2.2

Privileged Process

Unprivileged Process

>= Kernel 2.2



https://man7.org/linux/man-pages/man7/capabilities.7.html



# Linux Capabilities

- CAP_CHOWN
- CAP_AUDIT_CONTROL
- CAP_BPF
- CAP_DAC_OVEREIDE
- CAP_SETGID
- CAP_FOWNER
- CAP_KILL
- CAP_NET_ADMIN
- CAP_NET_RAW
- CAP_SYS_TIME
- CAP_SETUID
- CAP_SYS_ADMIN
- CAP_SYS_BOOT
- CAP_SYS_PTRACE
- CAP_WAKE_ALARM

# Linux Capabilities

```
getcap /usr/bin/ping
/usr/bin/ping = cap_net_raw+ep
```

```
ps -ef | grep /usr/sbin/sshd | grep –v grep
root       779     1  0 03:55 ?        00:00:00 /usr/sbin/sshd -D
```

```
getpcaps 779
capabilities for `779': =
cap_chown,cap_dac_override,cap_dac_read_search,cap_fowner,cap_fsetid,cap_kill,cap_setgid
,cap_setuid,cap_setpcap,cap_linux_immutable,cap_net_bind_service,cap_net_broadcast,cap_n
et_admin,cap_net_raw,cap_ipc_lock,cap_ipc_owner,cap_sys_module,cap_sys_rawio,cap_sys_chr
oot,cap_sys_ptrace,cap_sys_pacct,cap_sys_admin,cap_sys_boot,cap_sys_nice,cap_sys_resourc
e,cap_sys_time,cap_sys_tty_config,cap_mknod,cap_lease,cap_audit_write,cap_audit_control,
cap_setfcap,cap_mac_override,cap_mac_admin,cap_syslog,cap_wake_alarm,cap_block_suspend,c
ap_audit_read+ep
```

https://github.com/moby/moby/blob/master/oci/caps/defaults.go#L6-L19



# Linux Capabilities

```
kubectl run --rm -it ubuntu-sleeper --image=ubuntu -- bash
```

```
If you don't see a command prompt, try pressing enter.
root@ubuntu-sleeper:/#  date -s '19 APR 2012 22:00:00'
date: cannot set date: Operation not permitted
Thu Apr 19 22:00:00 UTC 2012
root@ubuntu-sleeper:/#
```

```
3  // DefaultCapabilities returns a Linux kernel default capabilities
4  func DefaultCapabilities() []string {
5          return []string{
6                  "CAP_CHOWN",
7                  "CAP_DAC_OVERRIDE",
8                  "CAP_FSETID",
9                  "CAP_FOWNER",
10                 "CAP_MKNOD",
11                 "CAP_NET_RAW",
12                 "CAP_SETGID",
13                 "CAP_SETUID",
14                 "CAP_SETFCAP",
15                 "CAP_SETPCAP",
16                 "CAP_NET_BIND_SERVICE",
17                 "CAP_SYS_CHROOT",
18                 "CAP_KILL",
19                 "CAP_AUDIT_WRITE",
```

# Linux Capabilities

ubuntu-sleeper.yaml

```yaml
apiVersion: v1
kind: Pod
metadata:
  name: ubuntu-sleeper
spec:
  containers:
  - name: ubuntu-sleeper
    image: ubuntu
    command: [ "sleep", "1000" ]
    securityContext:
      capabilities:
        add: ["SYS_TIME"]
```

```
kubectl apply –f ubuntu-sleeper.yaml
pod/ubuntu-sleeper created
```

```
kubectl exec -ti ubuntu-sleeper -- bash
root@ubuntu-sleeper:/# date
Sat Apr  3 05:32:06 UTC 2021
root@ubuntu-sleeper:/# date -s '19 APR 2012 22:00:00'
Thu Apr 19 22:00:00 UTC 2012
root@ubuntu-sleeper:/# date
Thu Apr 19 22:00:02 UTC 2012
root@ubuntu-sleeper:/#
```

# Linux Capabilities

ubuntu-sleeper.yaml

```yaml
apiVersion: v1
kind: Pod
metadata:
  name: ubuntu-sleeper
spec:
  containers:
  - name: ubuntu-sleeper
    image: ubuntu
    command: [ "sleep", "1000" ]
    securityContext:
      capabilities:
        add: ["SYS_TIME"]
        drop: ["CHOWN"]
```

```
kubectl apply –f ubuntu-sleeper.yaml
pod/ubuntu-sleeper created
```

```
kubectl exec -ti ubuntu-sleeper -- bash
root@ubuntu-sleeper:~# touch /tmp/test
root@ubuntu-sleeper:~# ls -l /tmp/test
-rw-r--r-- 1 root root 0 Apr  3 05:46 /tmp/test
root@ubuntu-sleeper:~# chown backup /tmp/test
chown: changing ownership of '/tmp/test': Operation
not permitted
root@ubuntu-sleeper:~#
```

Hands-on Labs

cks.kodekloud.com


www.kodekloud.com