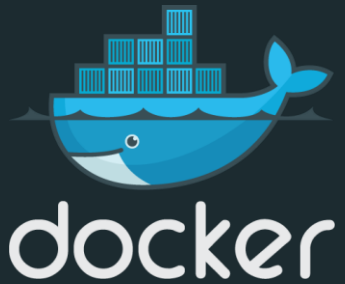
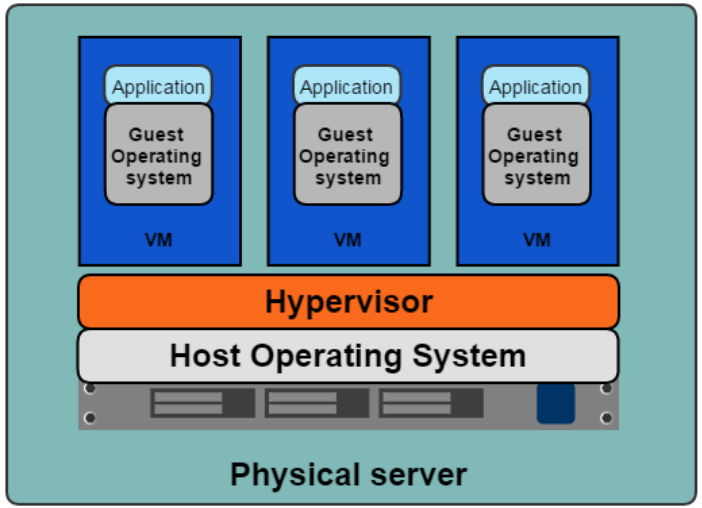
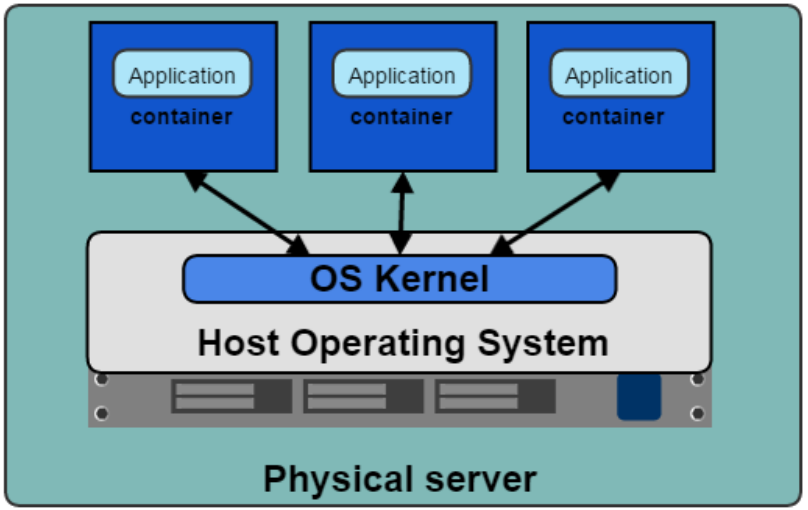
Seccomp, network and namespaces

Francesco Tornieri <francesco.tornieri AT kiratech.it>

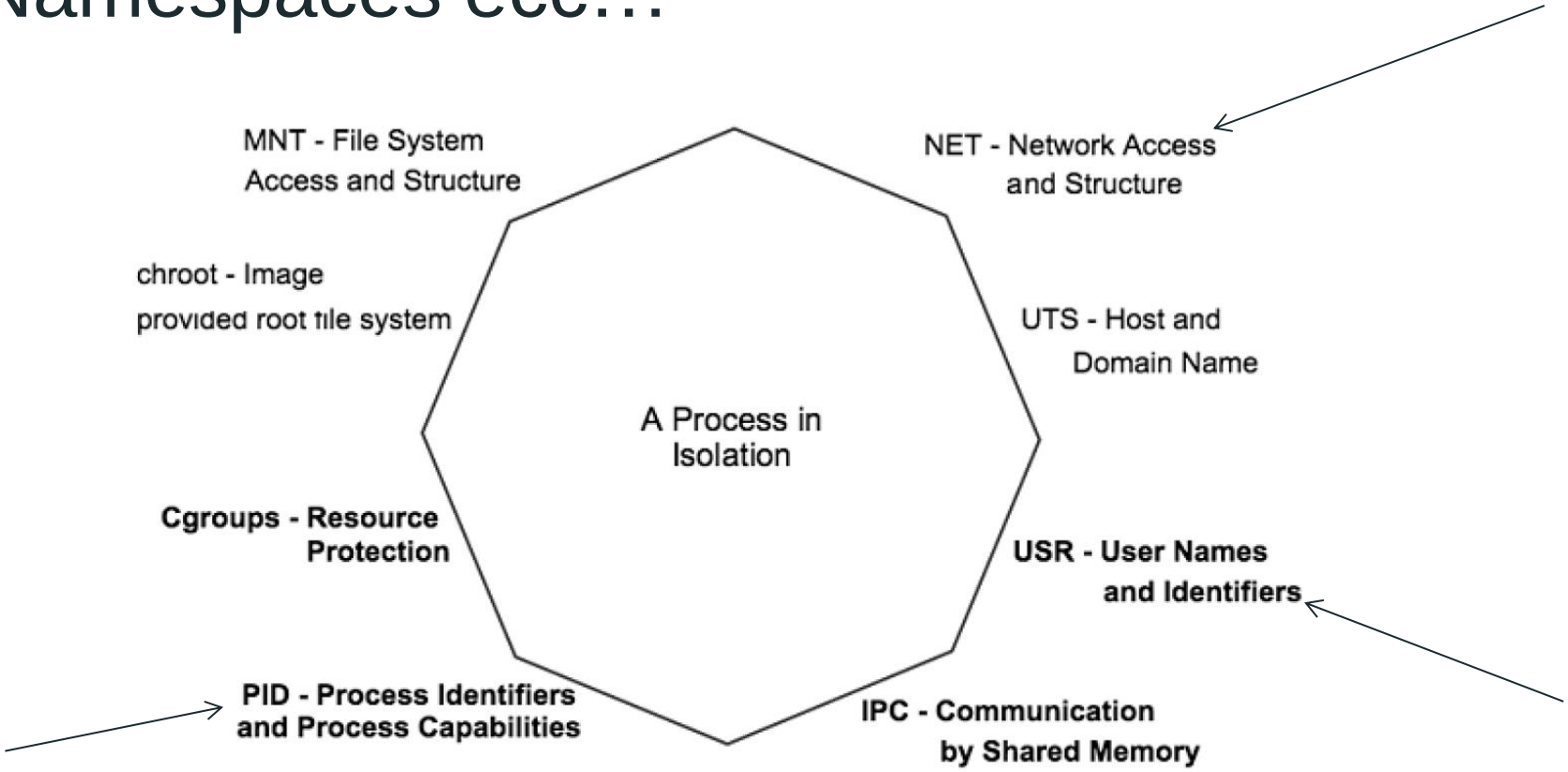


VM vs Container





Namespaces ecc…



Namespaces ecc…

- **man namespaces**:
  - *“A namespaces wraps a global system resource in a abstraction that makes it appear to the processes within the namespace that they have their own isolated instance of the global resource”*

PID & Capabilities (privileges associated with superuser)

- SETPCAP—Modify process capabilities
- SYS_MODULE—Insert/remove kernel modules
- SYS_RAWIO—Modify kernel memory
- SYS_PACCT—Configure process accounting
- SYS_NICE—Modify priority of processes
- SYS_RESOURCE—Override resource limits
- SYS_TIME—Modify the system clock
- SYS_TTY_CONFIG—Configure TTY devices
- AUDIT_WRITE—Write the audit log
- AUDIT_CONTROL—Configure audit subsystem
- MAC_OVERRIDE—Ignore kernel MAC policy
- MAC_ADMIN—Configure MAC configuration
- SYSLOG—Modify kernel print behavior
- NET_ADMIN—Configure the network
- SYS_ADMIN—Catchall for administrative functions

PID & Capabilities

- NET_RAW (man capabilities):
  - use RAW and PACKET sockets;
  - bind to any address for transparent proxying

- docker run --rm ubuntu capsh --print | grep net_raw
  - Current: = cap_chown,cap_dac_override,cap_fowner,cap_fsetid,cap_kill,cap_setgid,cap_setuid,cap_setpcap,cap_net_bind_service,cap_net_raw,cap_sys_chroot,cap_mknod,cap_audit_write,cap_setfcap+I
  - Bounding set =cap_chown,cap_dac_override,cap_fowner,cap_fsetid,cap_kill,cap_setgid,cap_setuid,cap_setpcap,cap_net_bind_service,cap_net_raw,cap_sys_chroot,cap_mknod,cap_audit_write,cap_setfcap
- docker run --rm --cap-drop net_raw ubuntu capsh --print | grep net_raw
  - ?

PID & Capabilities

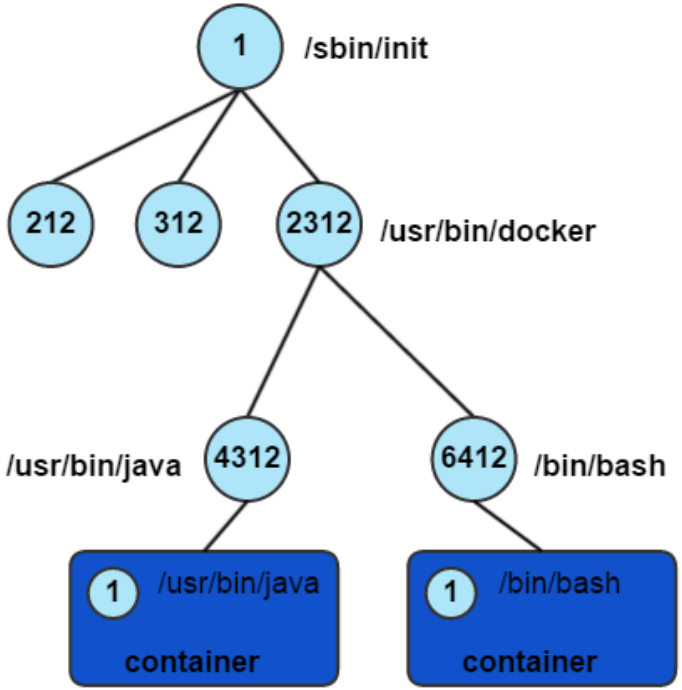
- NET_BIND_SERVICE (man capabilities):
  - Bind a socket to Internet domain privileged ports (port numbers less than 1024)

- docker run --rm -it --cap-drop ALL ubuntu bash
  - root@dba82716ff38:/# nc -l 127.0.0.1 25
  - nc: Permission denied
- docker run --rm -it --cap-drop ALL --cap-add net_bind_service ubuntu bash
  - root@e71400ba1c4e:/# nc -l 127.0.0.1 25
  - ^C

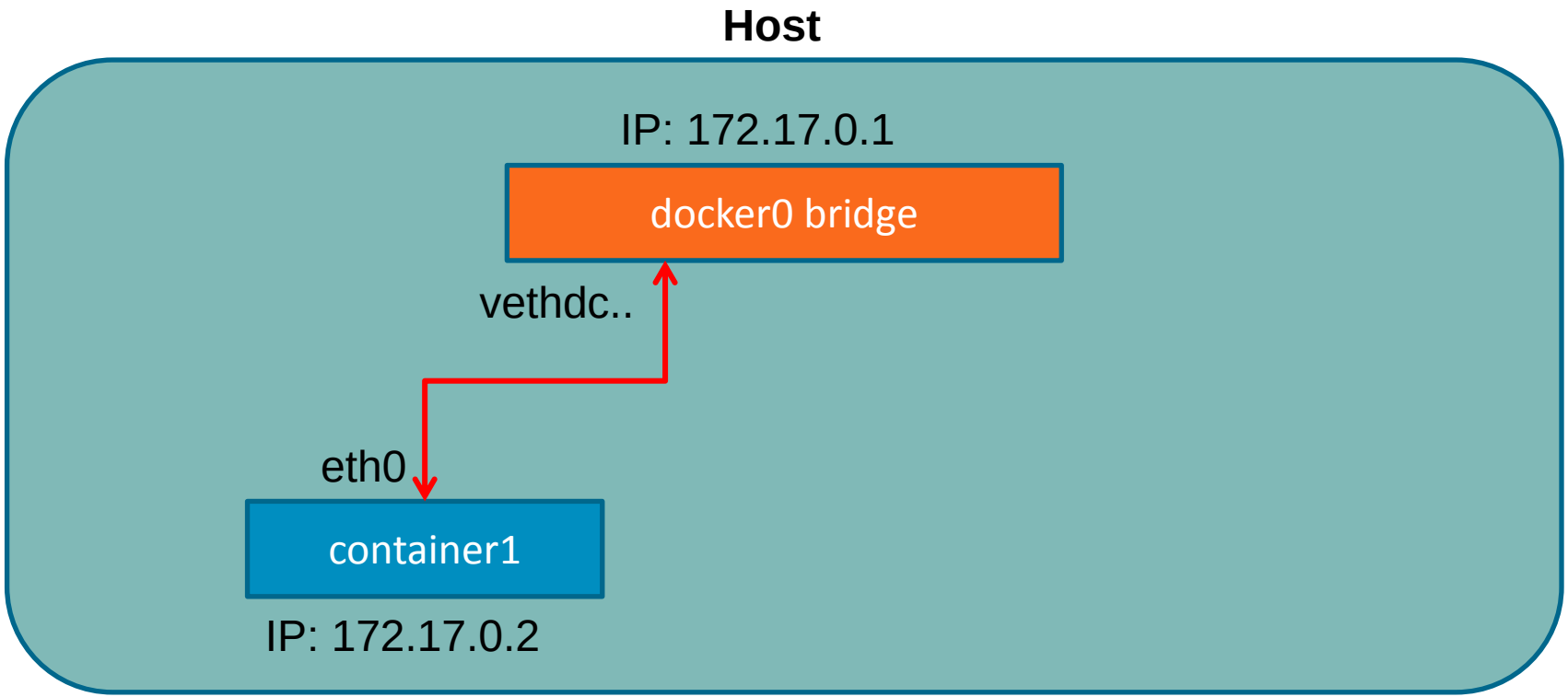
Container Processes

- Your command's process is always PID 1 inside the container
- docker run --rm ubuntu ps ax

```
PID TTY      STAT   TIME COMMAND
1 ?        Rs     0:00 ps ax
```

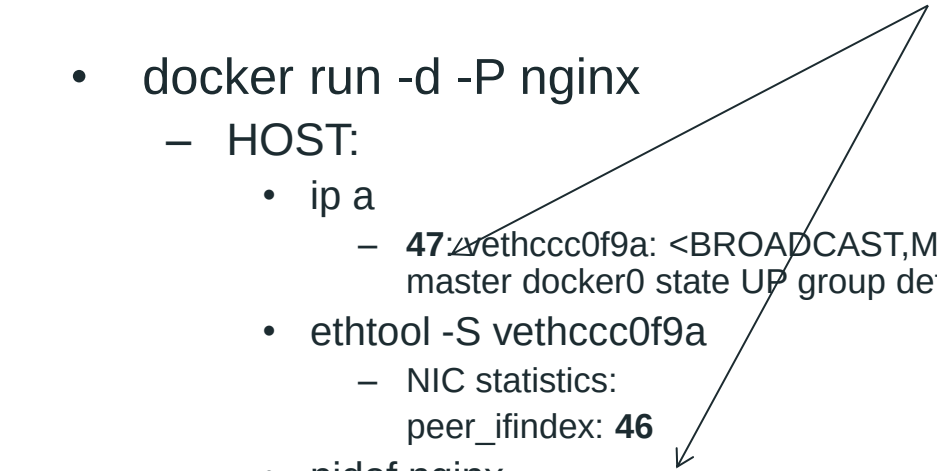


Bridge Network Model



Network namespaces

- docker run -d -P nginx
  - HOST:
    - ip a
      - **47**: vethccc0f9a: <BROADCAST,MULTICAST,UP,LOWER_UP> mtu 1500 qdisc noqueue master docker0 state UP group default
    - ethtool -S vethccc0f9a
      - NIC statistics:
        peer_ifindex: **46**
    - pidof nginx
      - 15363 15349
    - sudo ls /proc/15363/ns/net
    - sudo mkdir -p /var/run/netns
    - sudo ln -sf /proc/15363/ns/net /var/run/netns/myproc
    - ip netns
      - myproc



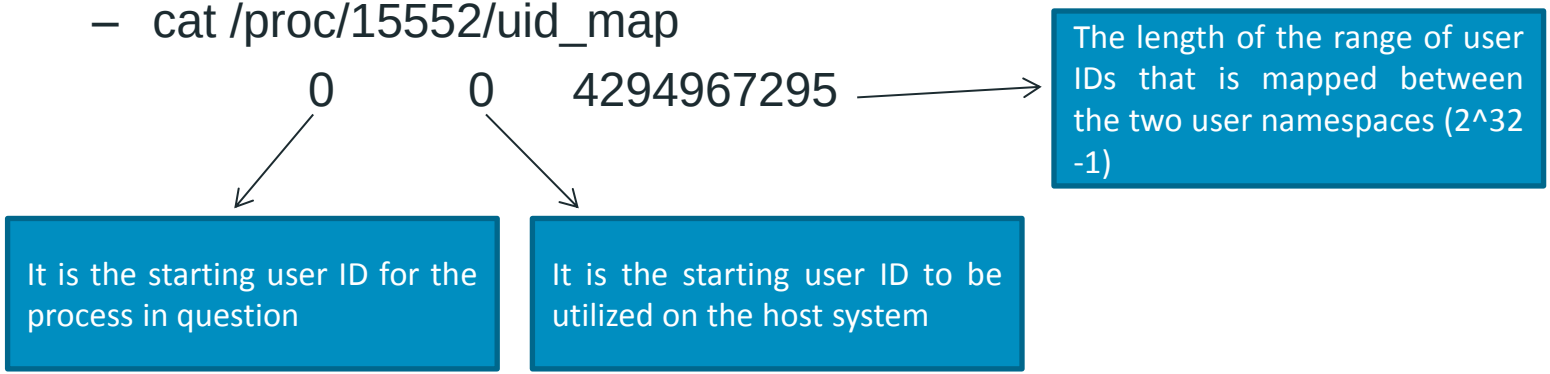
Network namespace

- HOST:
  - sudo ip netns exec myproc ip a

```
1: lo: <LOOPBACK,UP,LOWER_UP> mtu 65536 qdisc noqueue state UNKNOWN group default
    link/loopback 00:00:00:00:00:00 brd 00:00:00:00:00:00
    inet 127.0.0.1/8 scope host lo
       valid_lft forever preferred_lft forever
    inet6 ::1/128 scope host
       valid_lft forever preferred_lft forever
46: eth0: <BROADCAST,MULTICAST,UP,LOWER_UP> mtu 1500 qdisc noqueue state UP
group default
    link/ether 02:42:ac:11:00:02 brd ff:ff:ff:ff:ff:ff
    inet 172.17.0.2/16 scope global eth0
       valid_lft forever preferred_lft forever
    inet6 fe80::42:acff:fe11:2/64 scope link
       valid_lft forever preferred_lft forever
```

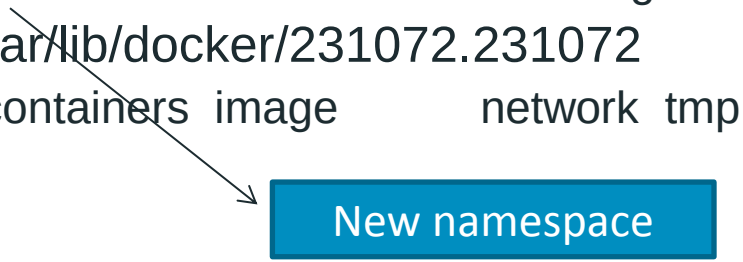

Username Namepaces

- DEFAULT DOCKER == NO USER NAMESPACE
  - docker run -d -it ubuntu
  - pidof bash
  - cat /proc/15552/uid_map



Username Namepaces

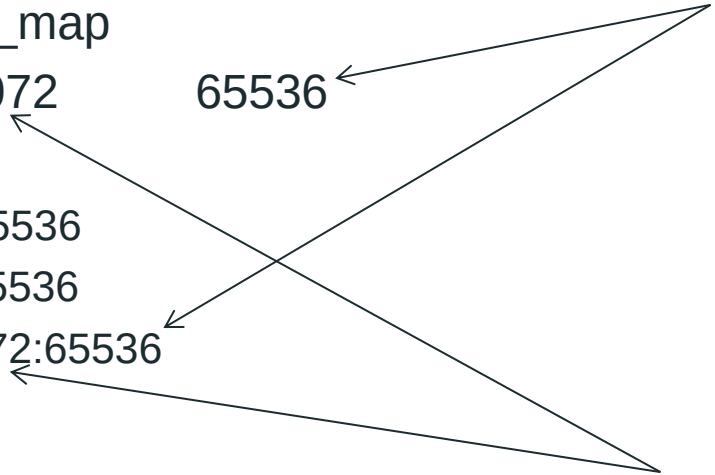
- docker daemon –help
  - …
  - --userns-remap
- Modify /etc/default/docker (DOCKER_OPTS):
  - DOCKER_OPTS="-H unix:///var/run/docker.sock -H tcp://0.0.0.0:2375 **--userns-remap default**"
- docker run -d -it ubuntu
  - Unable to find image 'ubuntu:latest' locally
- sudo ls /var/lib/docker/
  - 231072.231072 aufs containers image network tmp trust volumes
- sudo ls /var/lib/docker/231072.231072
  - aufs containers image network tmp trust volumes



Username Namepaces

- pidof bash
  - cat /proc/16003/uid_map

    0 231072 65536
  - cat /etc/subuid
    - ubuntu:100000:65536
    - docker:165536:65536
    - dockremap:231072:65536



Seccomp

- From **kernel.org** documentation (https://www.kernel.org/doc/Documentation/prctl/seccomp_filter.txt):

  “A large number of system calls are exposed to every userland process with many of them going unused for the entire lifetime of the process…. The resulting set reduces the total kernel surface exposed to the application… Seccomp filtering provides a means for a process to specify a filter for incoming system calls.”
- docker engine >= 1.10
- kernel with CONFIG_SECCOMP enabled
  - cat /boot/config-`uname -r` | grep CONFIG_SECCOMP=
  - CONFIG_SECCOMP=y

Seccomp

- Create a policy.json:

```
{
        "defaultAction": "SCMP_ACT_ALLOW",
        "syscalls": [
                {
                    "name": "mkdir",
                    "action": "SCMP_ACT_ERRNO"
                }
                ]
}
```

The seccomp filter will have no effect on the thread calling the syscall if it does not match any of the configured seccomp filter rules

Seccomp

- docker run -it --security-opt seccomp:policy.json ubuntu
  - mkdir test
    - mkdir: cannot create directory ‘test’: Permission denied

THANK YOU