

# LXC(Linux Container)

Lightweight virtual system mechanism

Gao feng

gaofeng@cn.fujitsu.com

# Outline

## ■ Introduction

- Namespace
- System API
- Libvirt LXC

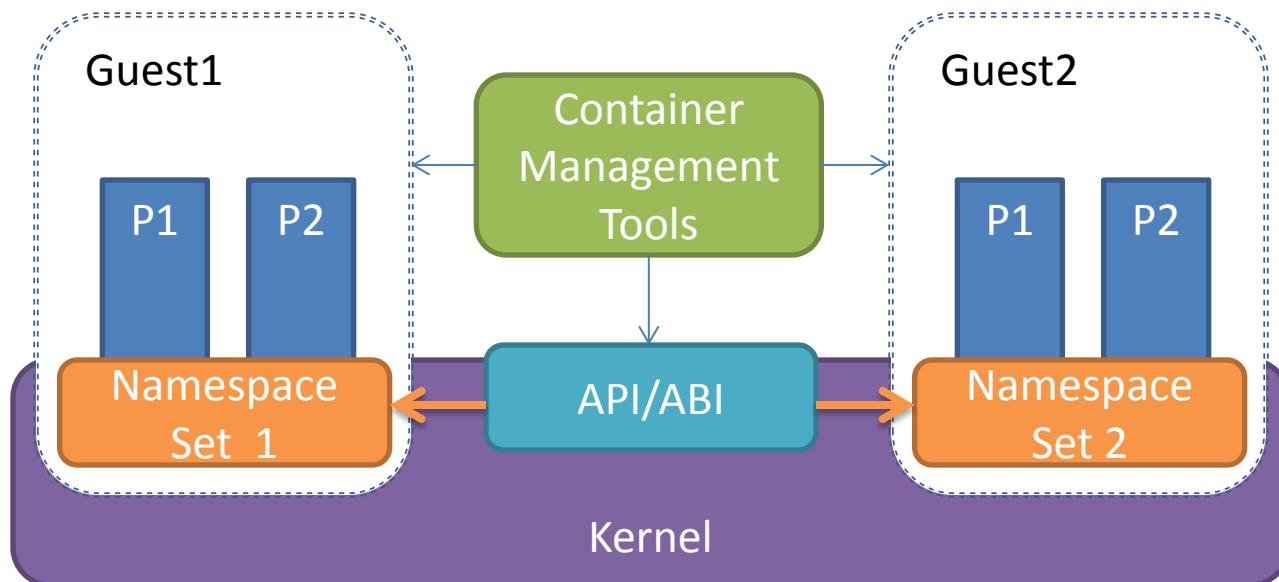
## ■ Comparison

## ■ Problems

## ■ Future work

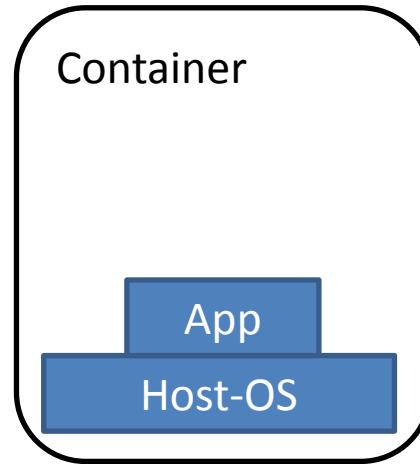
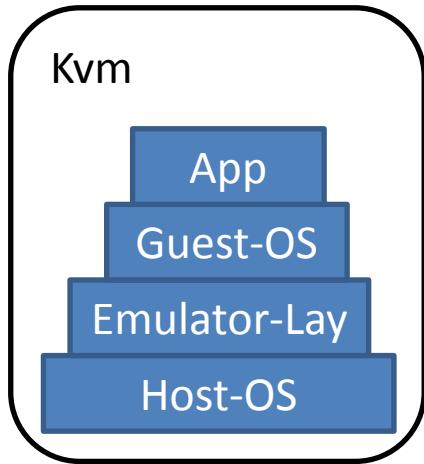
# Introduction

- Container: Operation System Level virtualization method for Linux



# Introduction

- Why Container
  - Better Performance



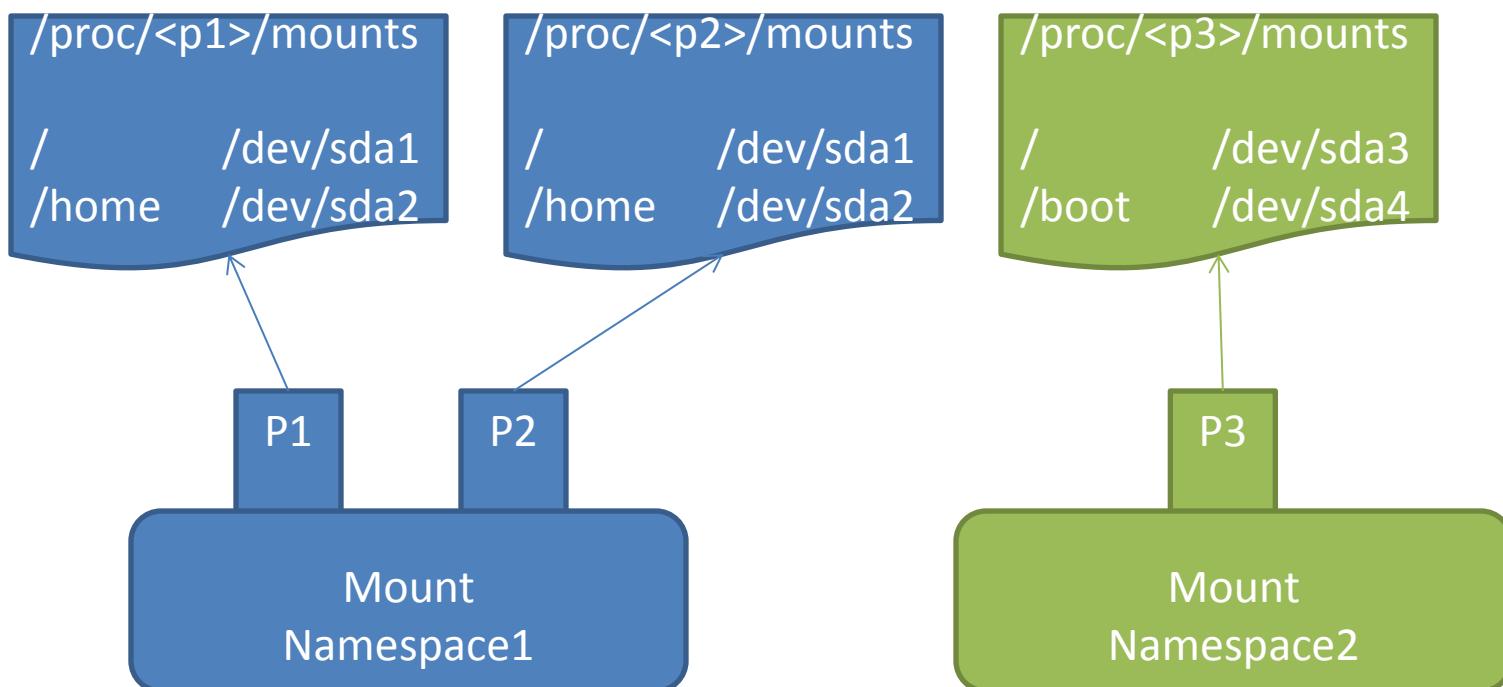
- Easy to set up Multi-Tenancy environment

# Namespace

- Namespace isolates the resources of system, currently there are 6 kinds of namespaces in linux kernel.
  - Mount namespace
  - UTS namespace
  - IPC namespace
  - Net namespace
  - Pid namespace
  - User namespace

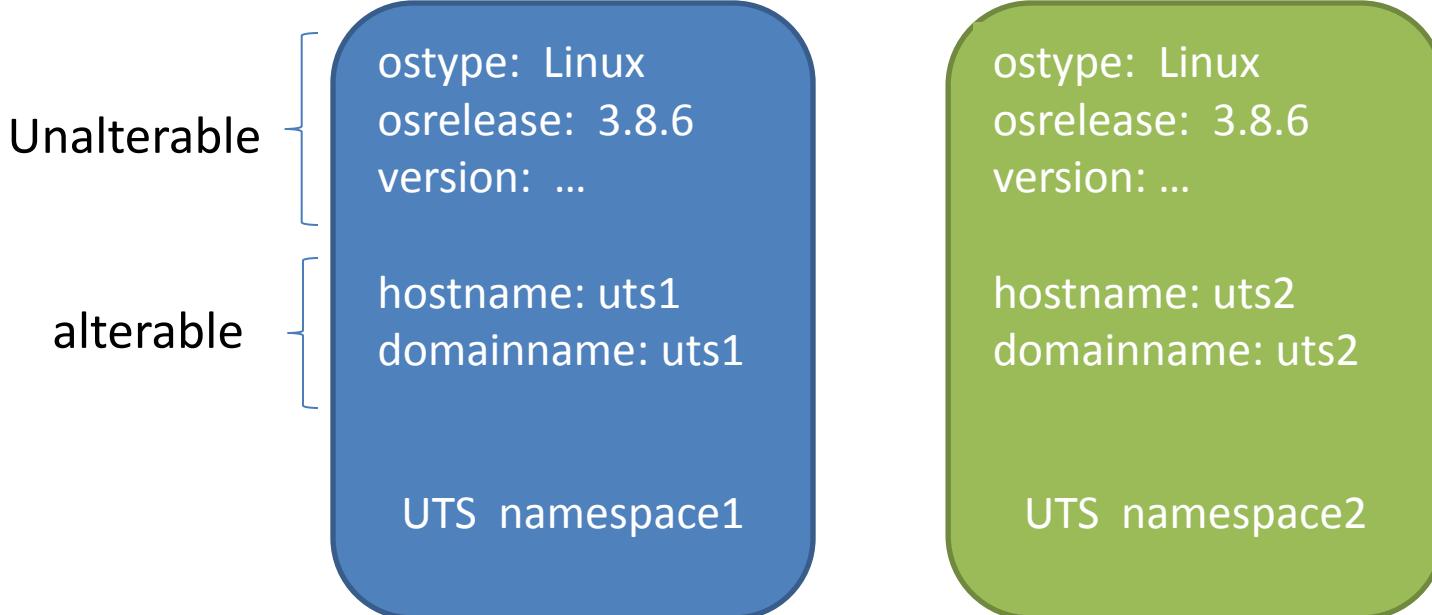
# Mount Namespace

- Each mount namespace has its own filesystem layout.



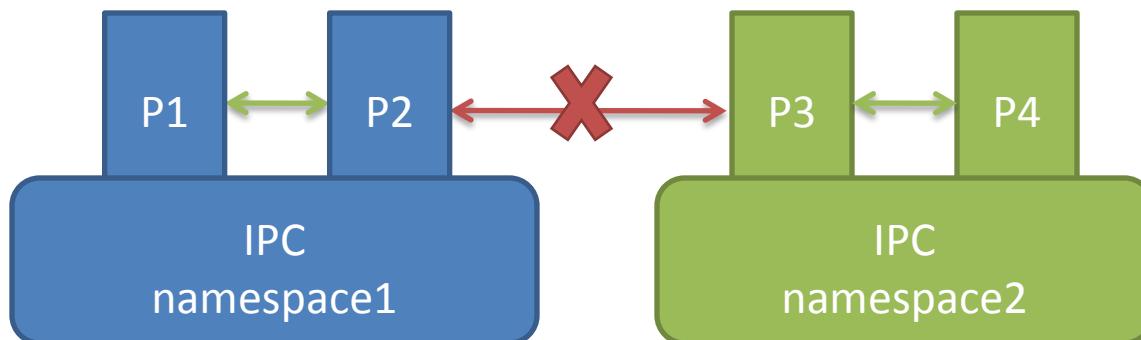
# UTS Namespace

- Every uts namespace has its own uts related information.



# IPC Namespace

- IPC namespace isolates the interprocess communication resource(shared memory, semaphore, message queue)



# Net Namespace

- Net namespace isolates the networking related resources

Net devices: eth0  
IP address: 1.1.1.1/24  
Route  
Firewall rule  
Sockets  
Proc  
sysfs  
...

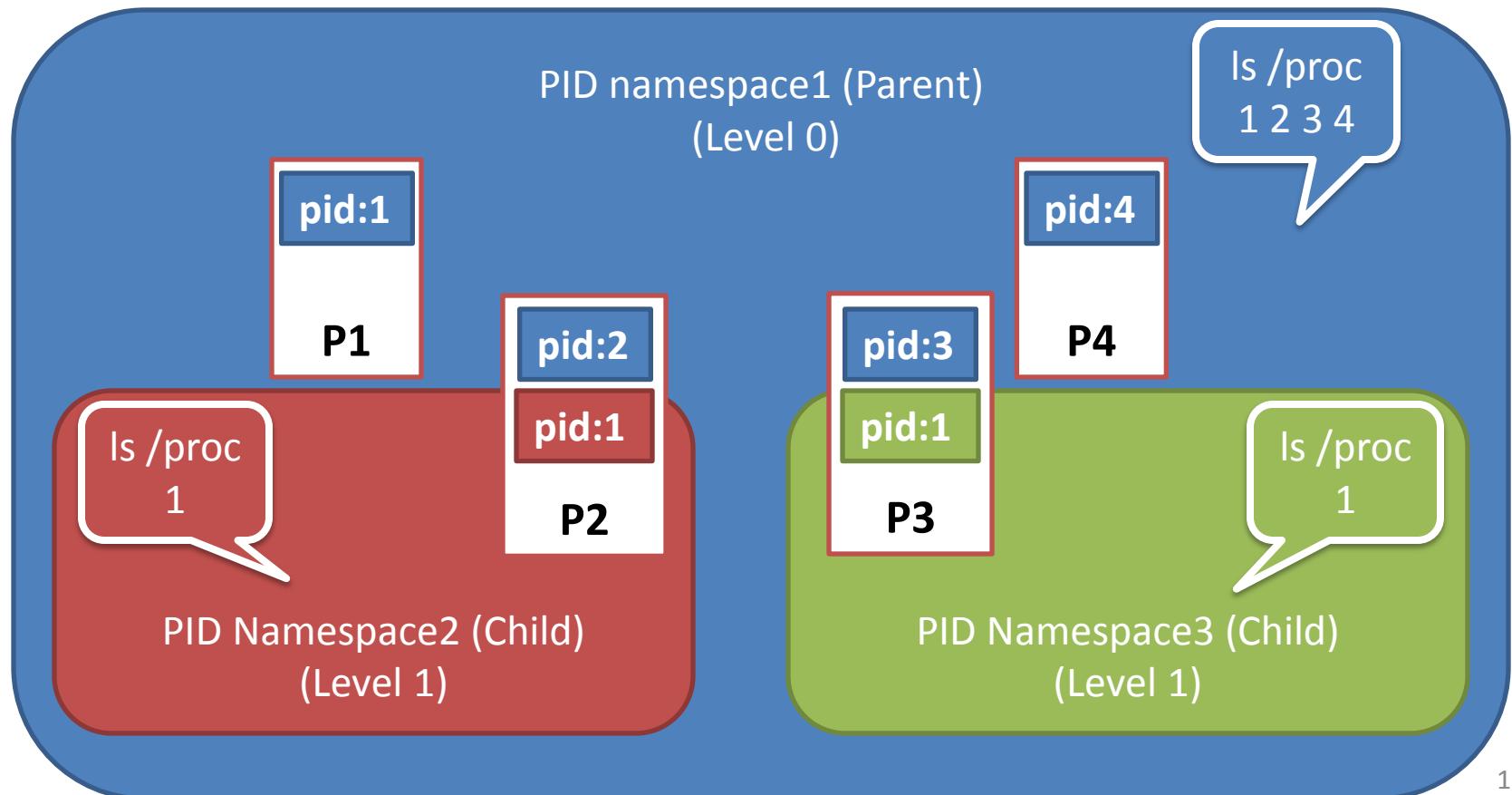
Net Namespace1

Net devices: eth1  
IP address: 2.2.2.2/24  
Route  
Firewall rule  
Sockets  
Proc  
sysfs  
...

Net Namespace2

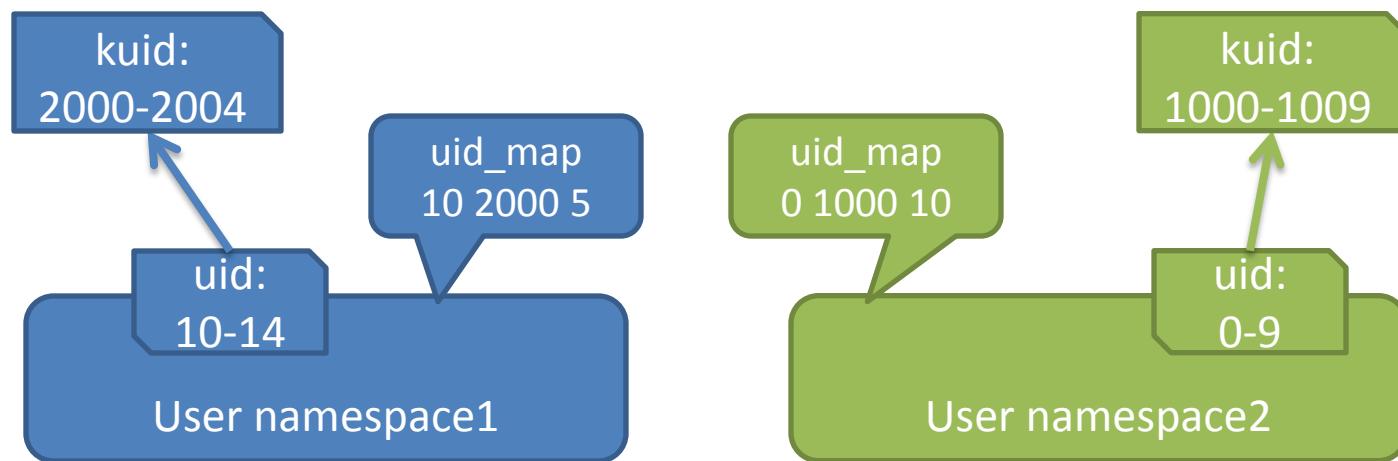
# PID Namespace

- PID namespace isolates the Process ID, implemented as a hierarchy.



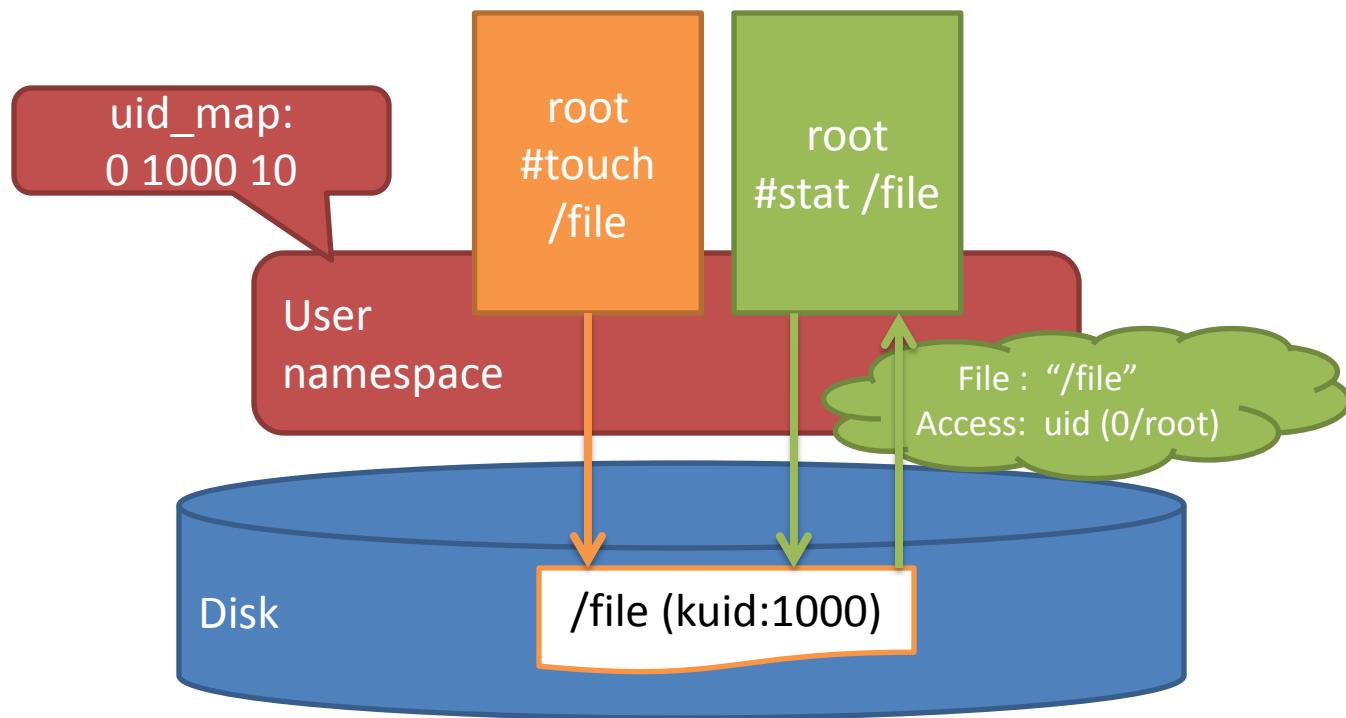
# User Namespace

- kuid/kgid: Original uid/gid, Global
- uid/gid: user id in user namespace, will be translated to kuid/kgid finally
- Only parent User NS has rights to set map



# User Namespace

## ■ Create and stat file in User namespace



# System API/ABI

- Proc

- /proc/<pid>/ns/

- System Call

- clone
  - unshare
  - setns

# Proc

- /proc/<pid>/ns ipc: ipc namespace
  - /proc/<pid>/ns/mnt: mount namespace
  - /proc/<pid>/ns/net: net namespace
  - /proc/<pid>/ns/pid: pid namespace
  - /proc/<pid>/ns/uts: uts namespace
  - /proc/<pid>/ns/user: user namespace
- 
- If the proc file of two processes is the same, these two processes must be in the same namespace.

# System Call

## ■ clone

```
int clone(int (*fn)(void *), void *child_stack,  
          int flags, void *arg, ...);
```

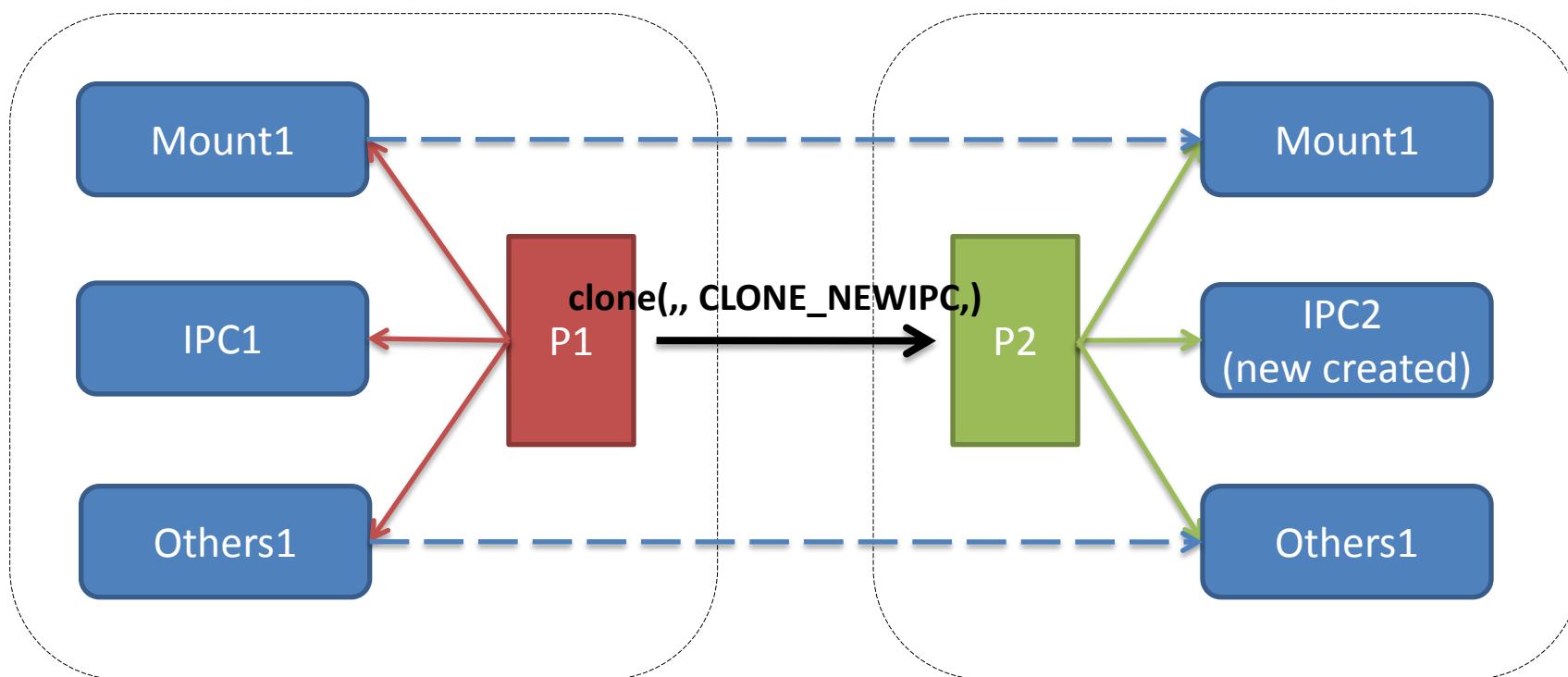
6 new flags:

- CLONE\_NEWIPC, CLONE\_NEWNET,
- CLONE\_NEWNS, CLONE\_NEWPID,
- CLONE\_NEWUTS, CLONE\_NEWUSER

# System Call

## clone

create process2 and IPC namespace2



# System Call

## ■ unshare

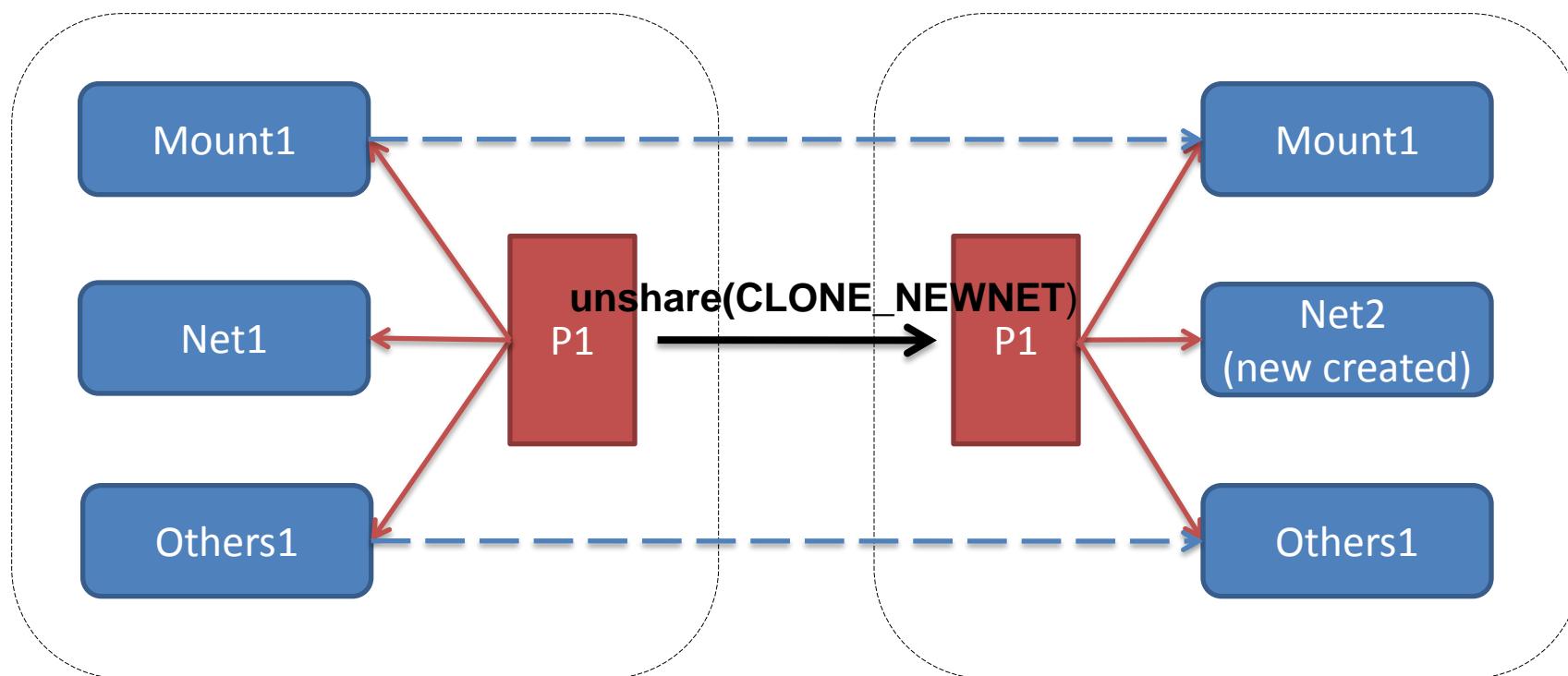
```
int unshare(int flags);
```

Namespace extends the system call unshare too. User space can use unshare to create new namespace and the caller will run in this new created namespace.

# System Call

## ■ unshare

create net namespace2



# System Call

## ■ setns

```
int setns(int fd, int nstype);
```

setns is a new added system call for namespace.

Process can use setns to set which namespace the process will belong to.

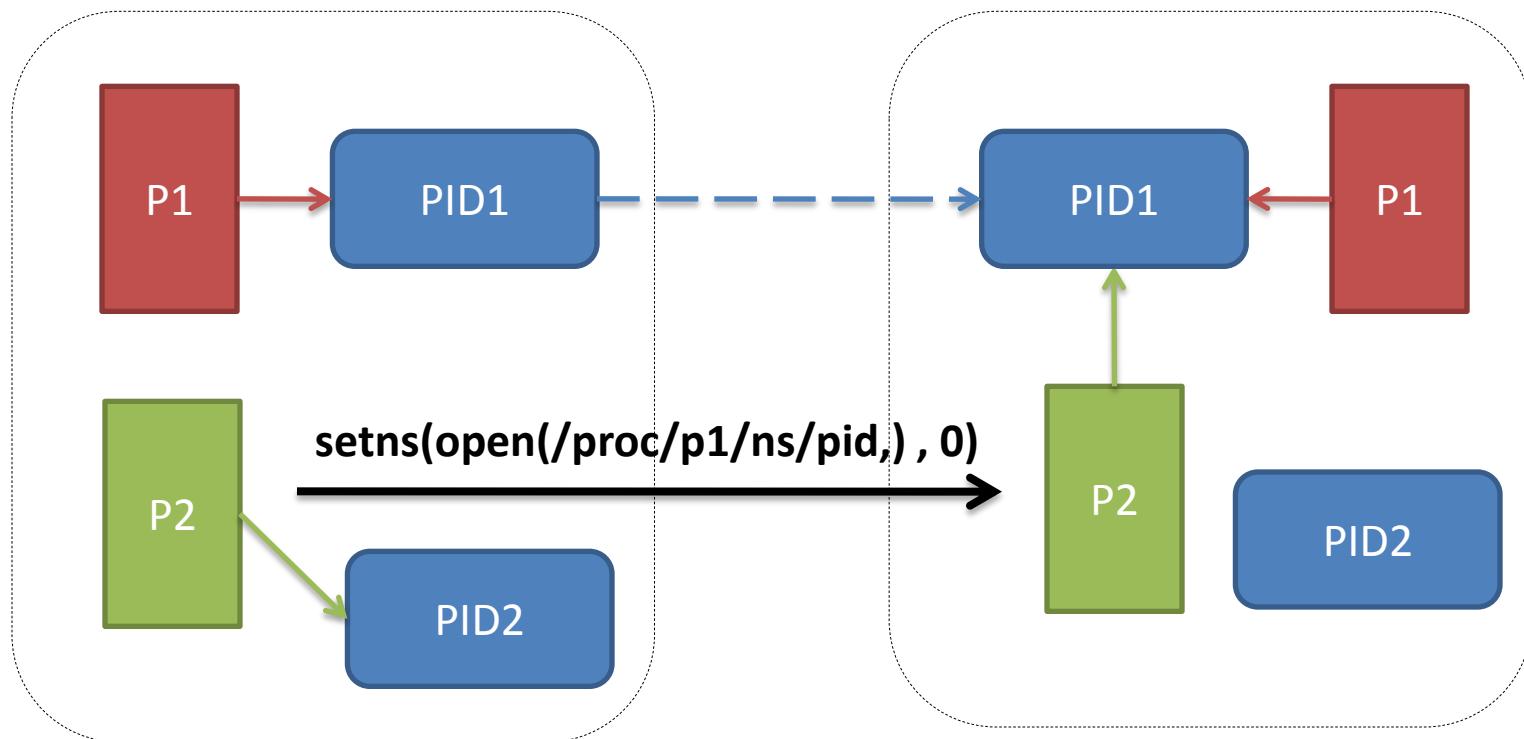
@fd: file descriptor of namespace(/proc/<pid>/ns/\*)

@nstype: type of namespace.

# System Call

## ■ setns

Change the PID namespace of P2



# Libvirt LXC

- Libvirt LXC: userspace container management tool,  
Implemented as one type of libvirt driver.
  - Manage containers
  - Create namespace
  - Create private filesystem layout for container
  - Create devices for container
  - Resources controller by cgroup

# Comparison

■ The feature that host share the same kernel with guest makes container different from other virtualization method

	Container	KVM
performance	Great	Normal
OS support	Linux Only	No Limit
Security	Normal	Great
Completeness	Low	Great

# Problems

## ■ /proc/meminfo, cpuinfo...

- Kernel space (relate to cgroup)
- User space (poor efficiency)

## ■ New namespace

- Audit (assign to user namespace?)
- Syslog (do we really need it?)

# Problems

## ■ Bandwidth control

### ■ TC Qdisc

- On host (How to handle setting nic to container?)
- On container (user can change it)

### ■ Netfilter

- How to control Ingress bandwidth

## ■ Disk quota

- Uid/Gid Quota (Many users )
- Project Quota (xfs only)

# Future Work

- Improve Libvirt LXC
- Unchanged systemd in Libvirt LXC
- Use interface of systemd to set cgroup
- Libvirt LXC based Docker
- Audit namespace

Thank you!  
Q&A