LXC(Linux Container)

Lightweight virtual system mechanism

Gao feng

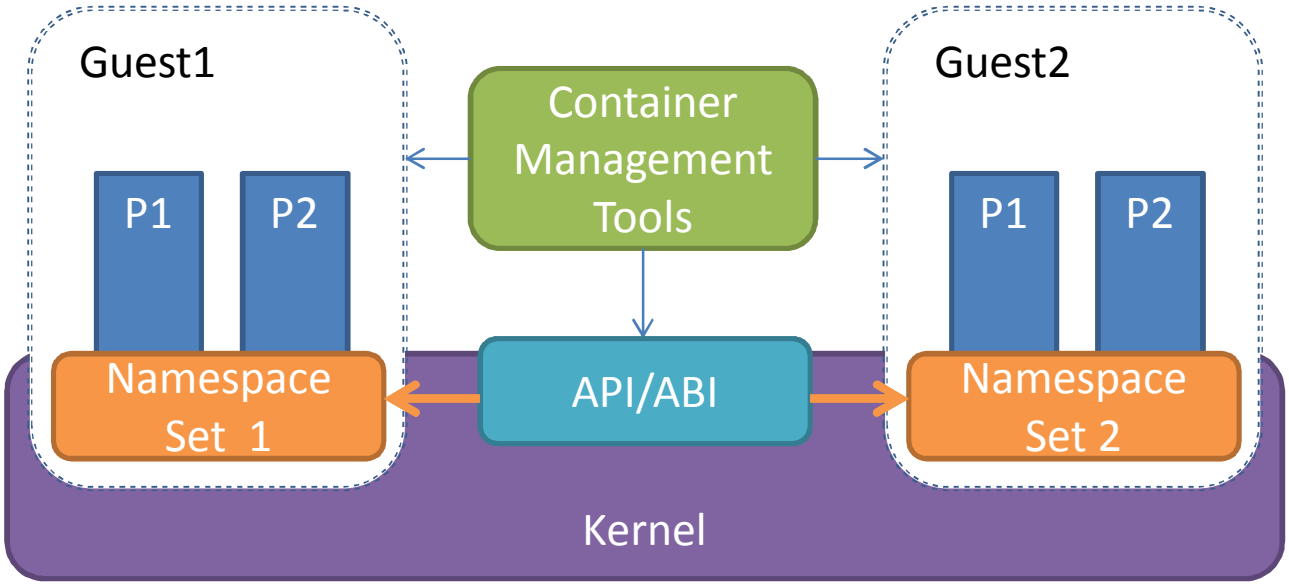
gaofeng@cn.fujitsu.com

Outline

- Introduction
  - Namespace
  - System API
  - Libvirt LXC
- Comparison
- Contribution
- Problems & Future work

Introduction

- Container: Operation System Level virtualization method for Linux

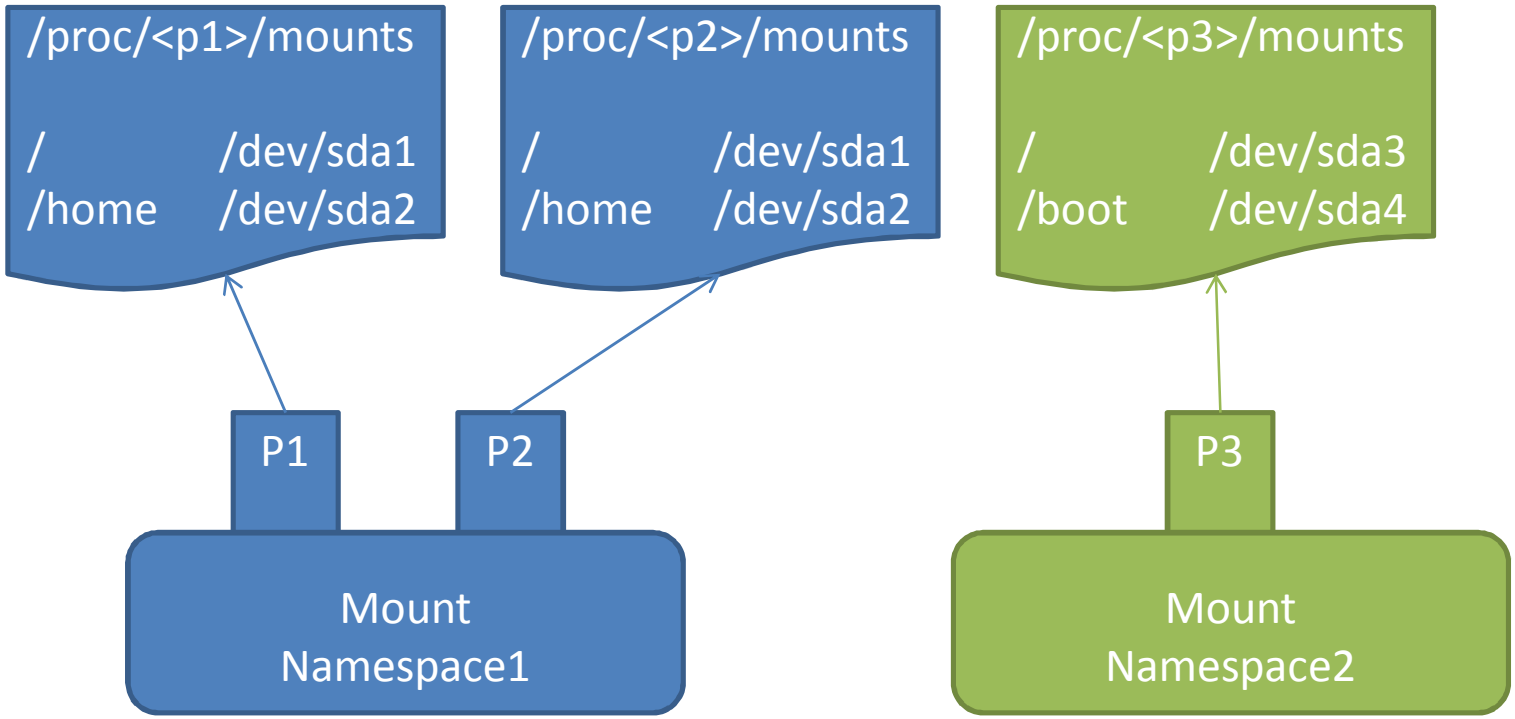


Namespace

- 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

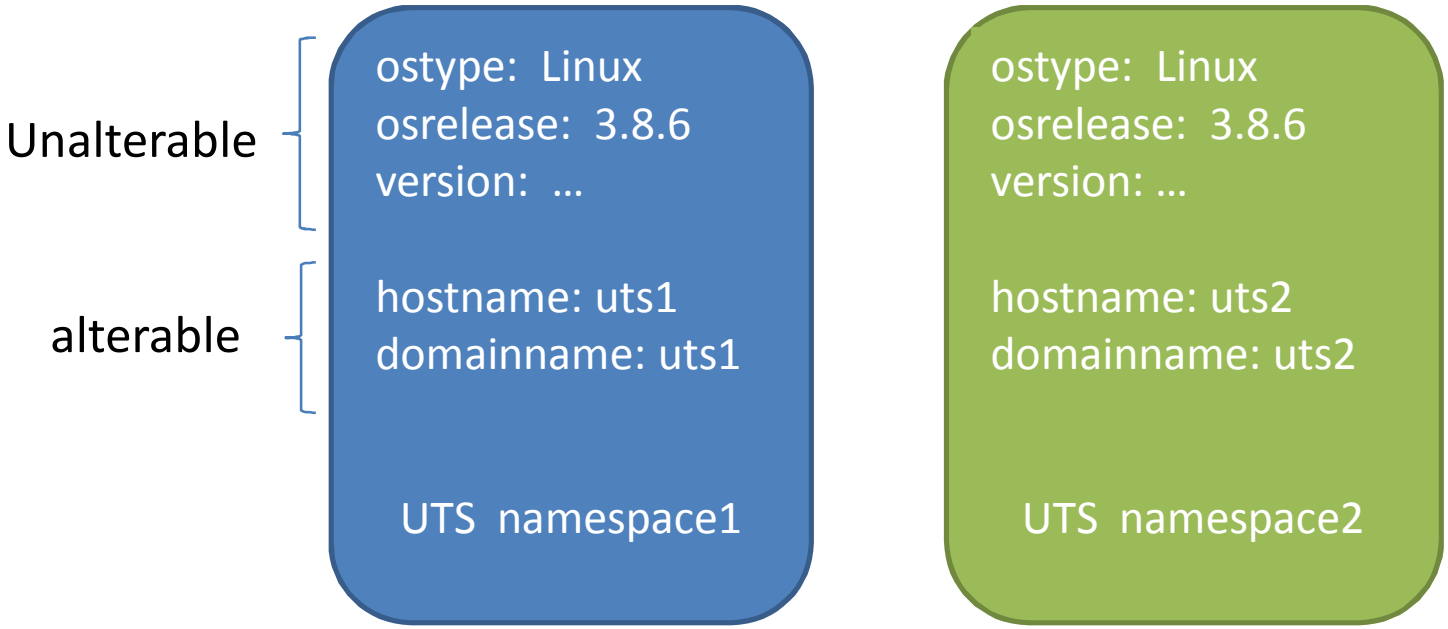
Mount Namespace

- Each mount namespace has its own filesystem layout.



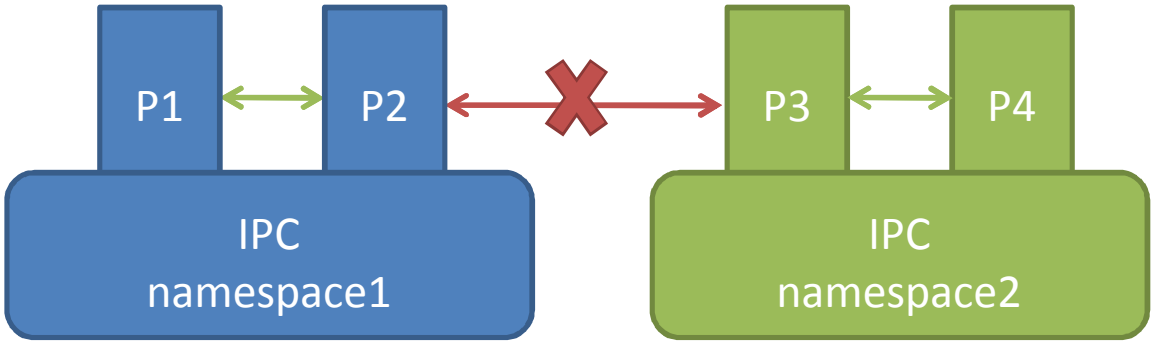
UTS Namespace

- Every uts namespace has its own uts related information.



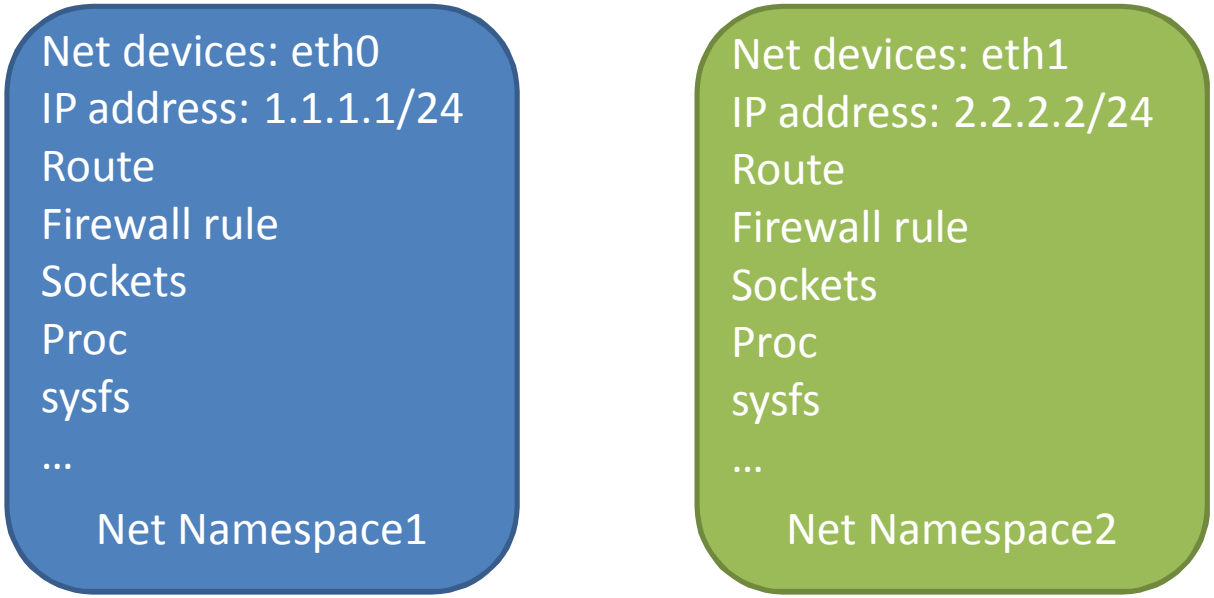
IPC Namespace

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



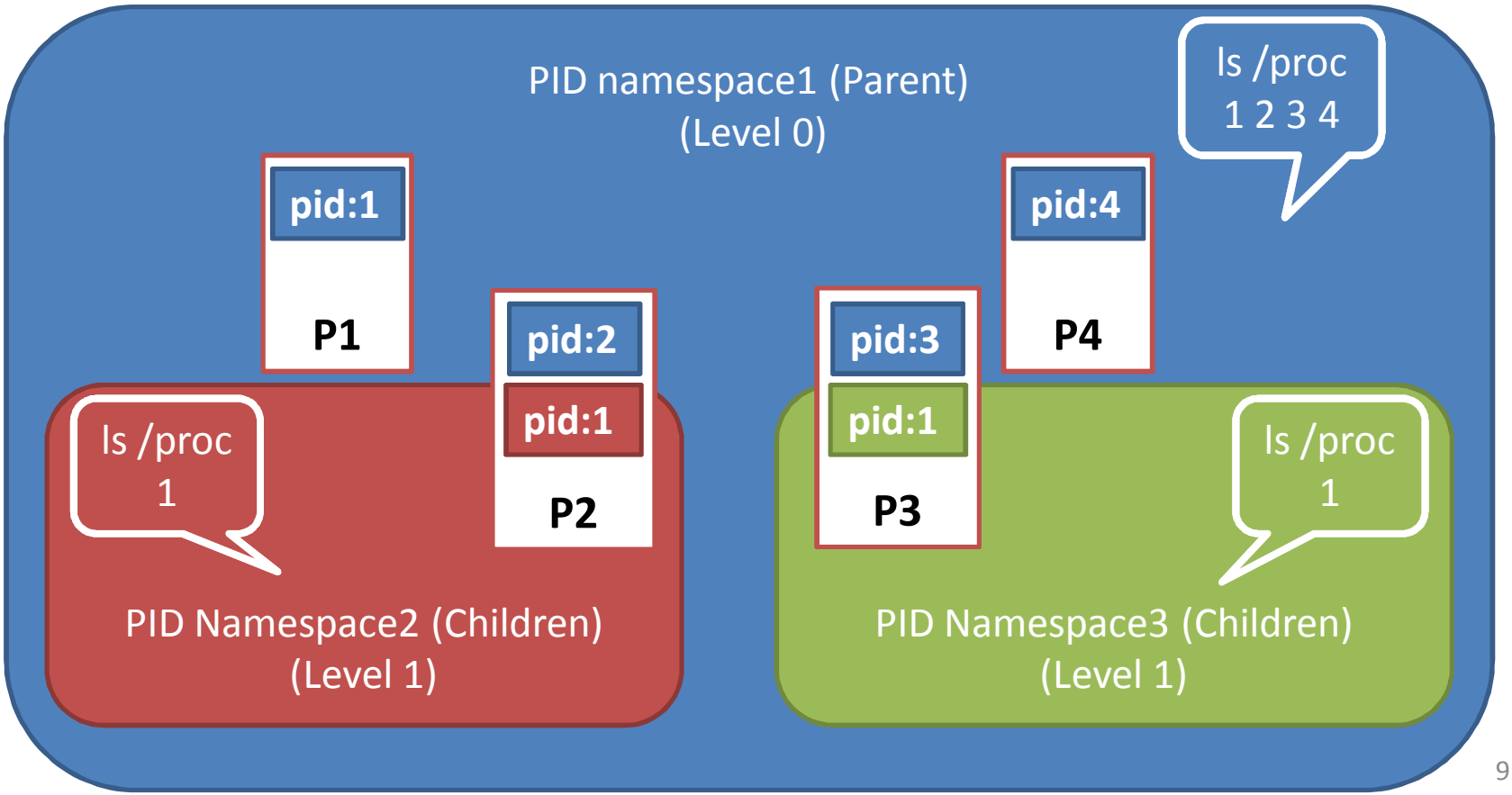
Net Namespace

- Net namespace isolates the networking related resources



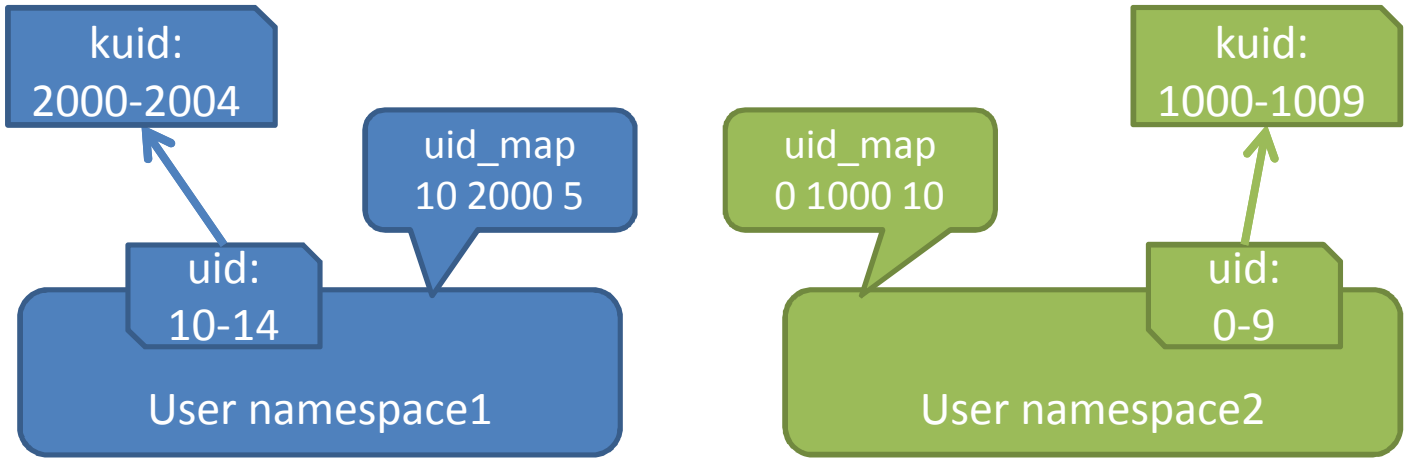
PID Namespace

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



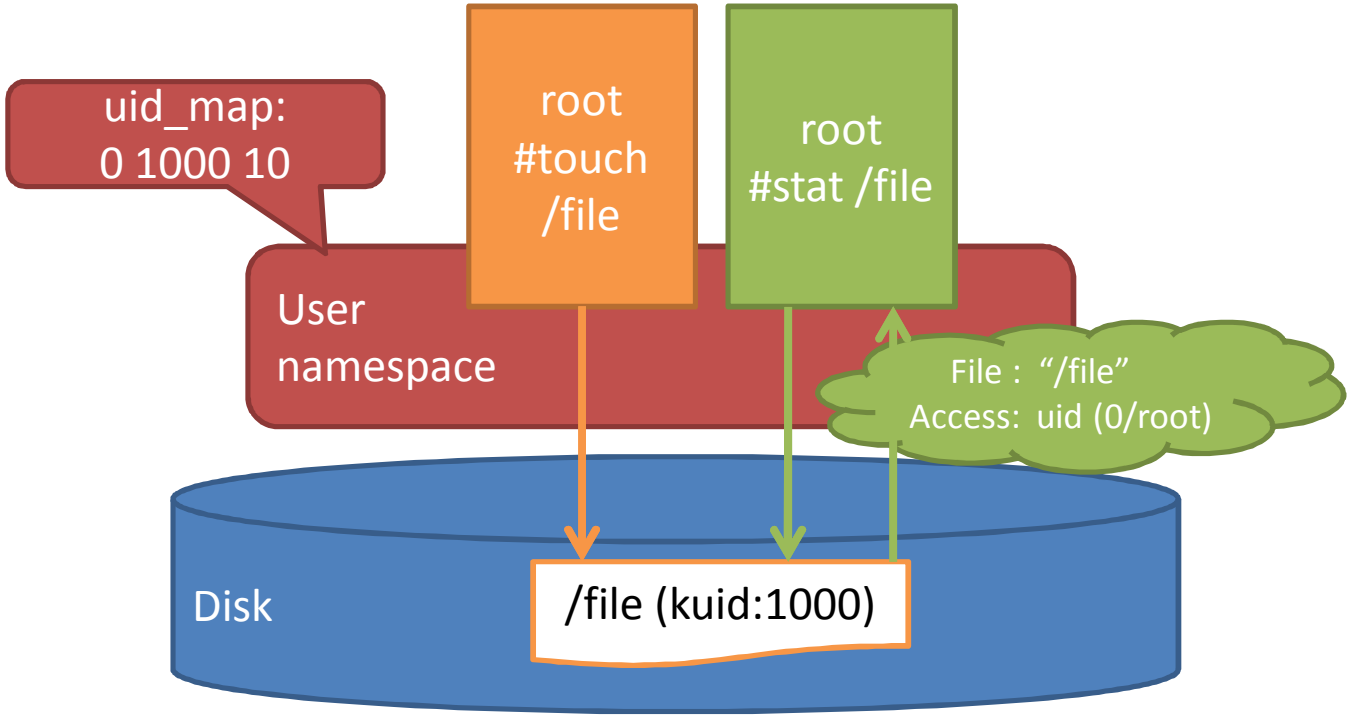
User Namespace

- kuid/kgid: Original uid/gid, Global
- uid/gid: user id in user namespace, will be translated to kuid/kgid finally



User Namespace

- Create and stat file in User namesapce



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,

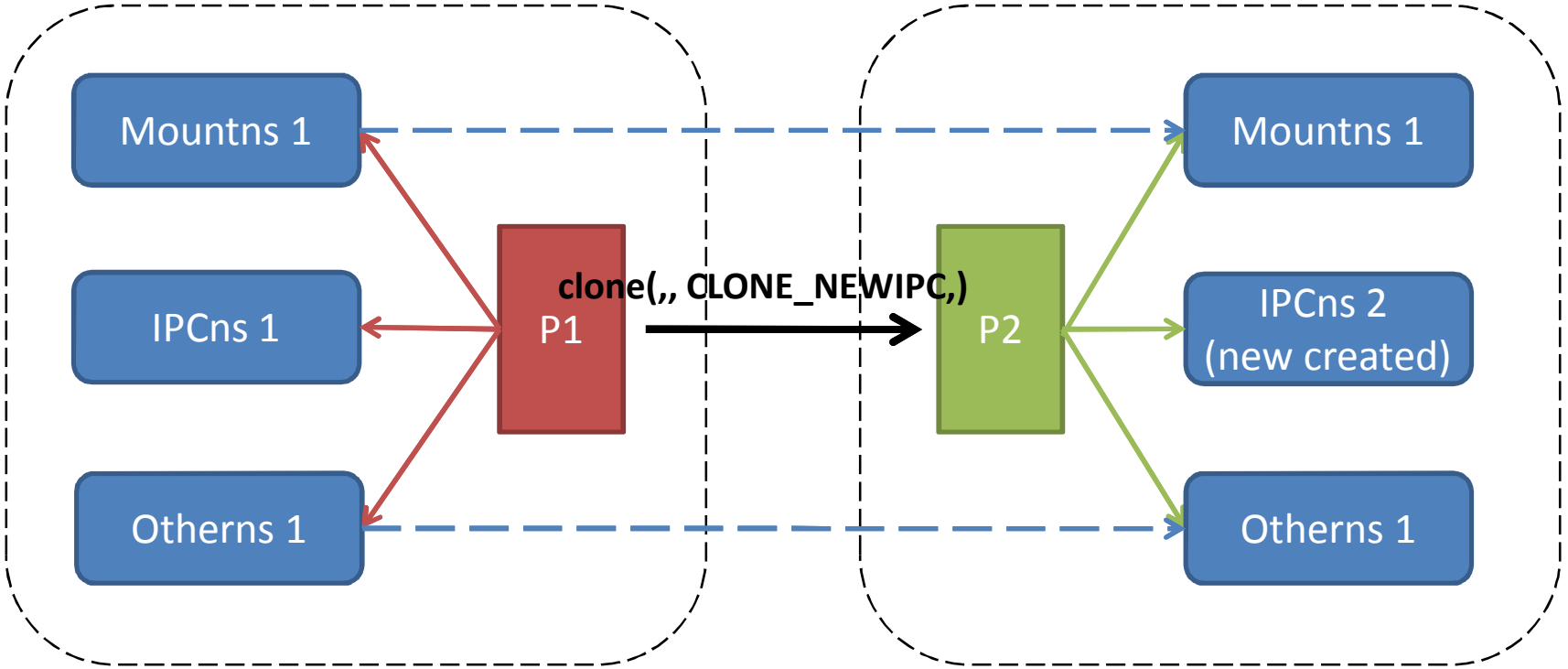
CLONE_NEWNS,CLONE_NEWPID,

CLONE_NEWUTS,CLONE_NEWUSER

System Call

- clone

  create process2 and IPC namespace2



System Call

- unshare

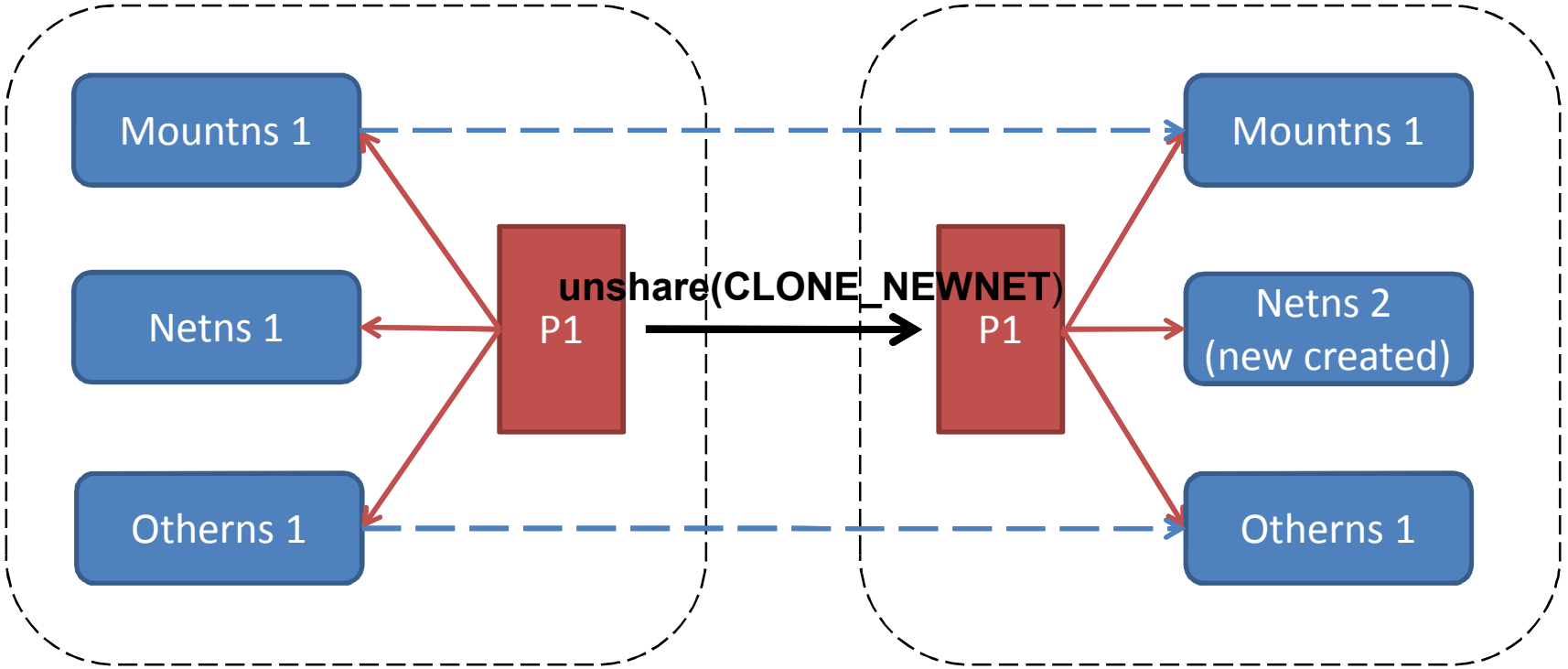
  int unshare(int flags);

  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);

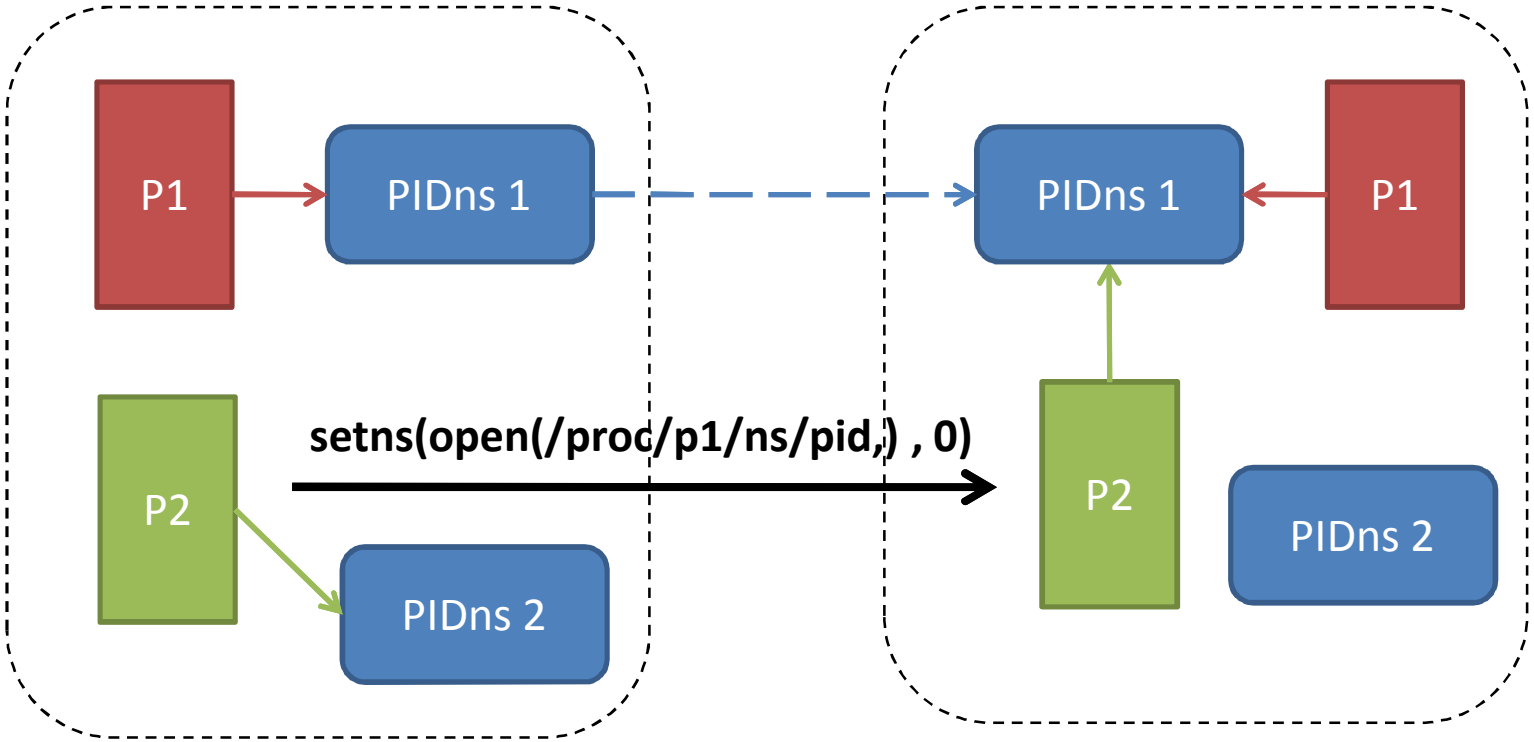
  setns is a new added system call for namespace. Process can use setns to set which namespace the process will belong to.

  @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
  - Create devices
  - Resources controller by cgroup

Comparison

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

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

Contribution

- Kernel
  - Add net namespace support for L3,L4 proto of netfilter.
  - Add net namespace support for nfqueue.
  - Add net namespace support for nflog.
  - Add net namespace support for inet_peer.
  - Net/User namespace related improvement.
  - Add user namespace support for Audit.
  - BUG FIX

Contribution

- Libvirt
  - Add fuse filesystem support for Libvirt LXC, virtualize meminfo,cpuinfo…through fuse filesystem.
  - Add cpuset cgroup support for Libvirt LXC.
  - Add User namespace support for Libvirt LXC.
  - BUG FIX

Problems & Future Work

- Problems
  - /proc/meminfo, cpuinfo…
  - New namespace
- Future Work
  - Syslog, crypto namespace
  - Improve Libvirt LXC