LINUX CONTAINERS DIY

PŘEMEK PODLAHA DĚLÁ DOCKER

Vilibald Wanča - vilibald@wvi.cz

VILIBALD

- 15+ let ve vývoji
- Basic -> Pascal -> asm x86 -> C/C++ -> Python, Lisp, Go
- Unix/Linux uživatel od 1997 (SCO Unix a Slackware)

V současnosti digitální proletář v Apiary (apiary.io)

AGENDA

O ČEM BUDEME MLUVIT

- Co je to ten kontejner?
- Namespaces
- Cgroups
- Síť
- Image
- Lepíme to dohromady

Ptejte se hned a nečekejte na konec

CO JE TO TEN KONTEJNER?

Osekaná VM nebo něco jiného?

Obvyklý(é) Linux proces(y) s omezeným viděním světa.

OMEZOVÁNÍ PROCESŮ V LINUXU

Vlastnosti kernelu

- Namespaces
- Cgroups aka Control Groups
- Síť (bridge, iptables atd.)
- SELinux/AppArmor

NAMESPACES

Izolace na základě zdrojů

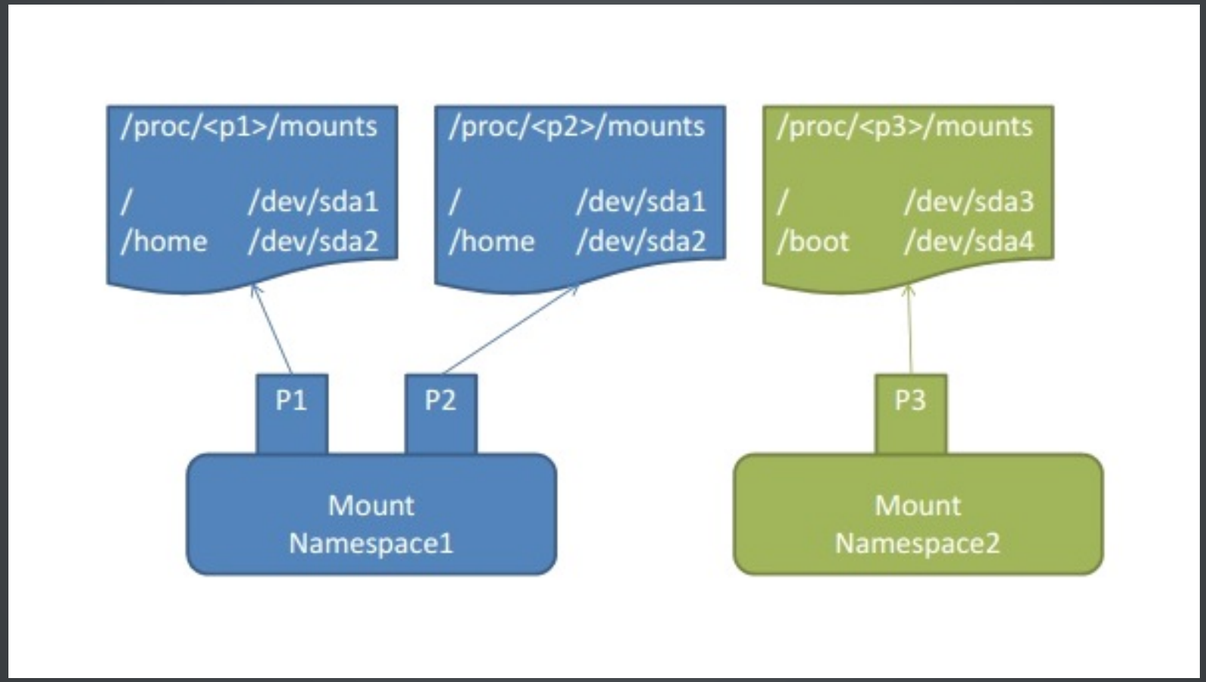
- Mount (2.4.19)
- UTS (2.6.19)
- IPC (2.6.19)
- PID (2.6.24)
- Network (2.6.29)
- User (3.8)
- Cgroups (4.6)

JAK VZNIKAJÍ

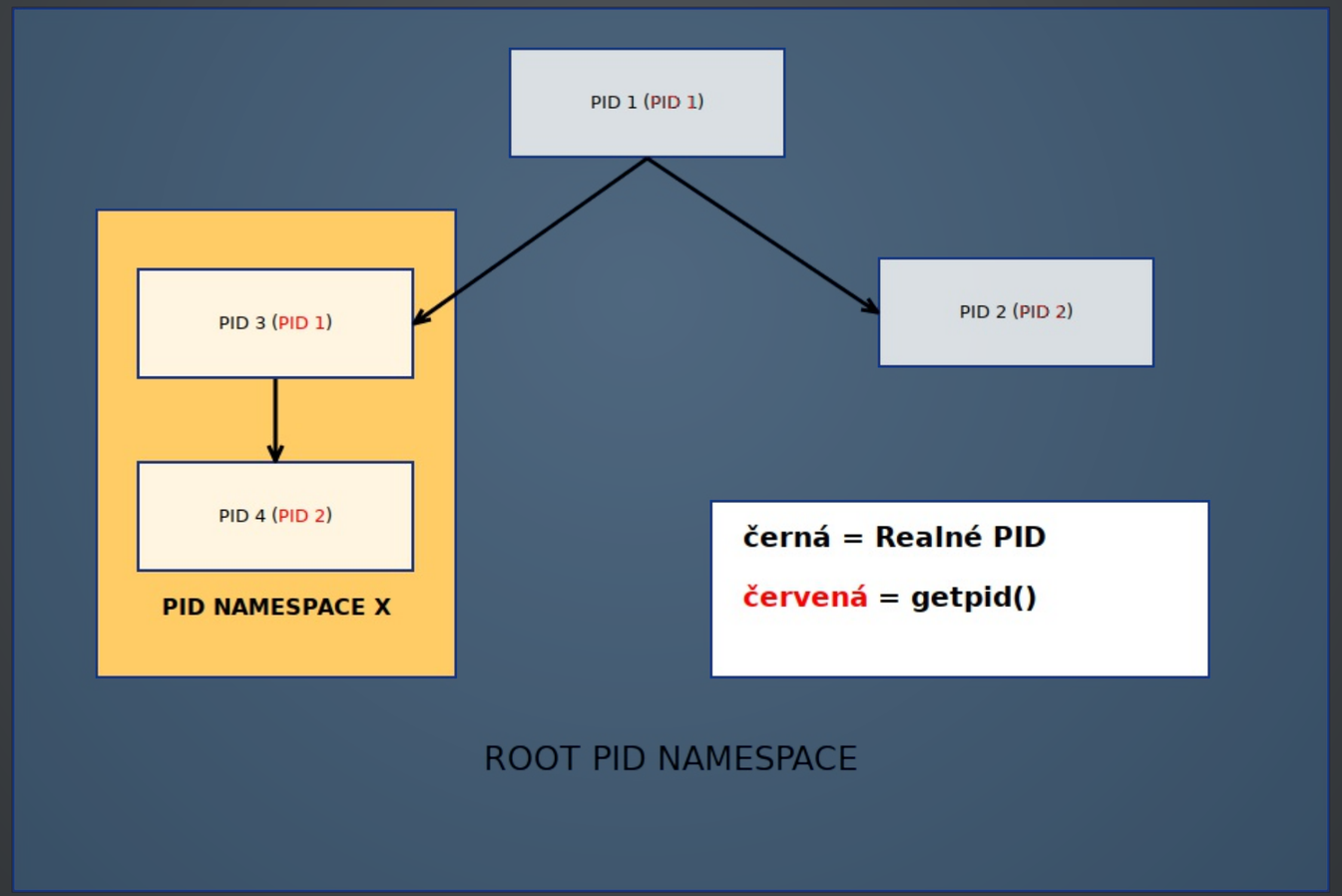
Systémová volání

- clone()
- unshare()
- setns()

MOUNT NAMESPACE



PID NAMESPACE



CONTROL GROUPS

Mechanismus pro kontrolu,
prioritizaci a účetnictví procesů

- *blkio* - limity na IO
- *cpu* - cpu scheduling
- *cpuset* - přiřazování CPU na multicore
- *devices* - přístup k zařízením
- *memory* - paměťové limity (rss, swap atd.)

/sys/fs/cgroup/

SÍŤ

Virtuální zařízení (veth), Linux bridge

a hlavně:

IMAGE

Co je ve skutečnosti image kontejneru?

JE TO JENOM TARBALL

nebo tarball plný dalších tarballů v sofistikované verzi

SLEPÍME TO DOHROMADY I

```
main():
    flags =  SIGCHLD | CLONE_NEWNS | CLONE_NEWPID ....
    pid = clone(container_exec, stack.ptr, flags, args);
    setup_network_and_cgroups();
    waitpid(pid);
    exit();
```

SLEPÍME TO DOHROMADY II

```
container_exec(args):

    umount("/proc");
    pivot_root("/tmp/container", "/tmp/container/.pivot_root");
    chdir("/");
    copy_files("/.pivot_root/etc/resolv.conf", ...);
    umount("/.pivot_root");
    mount("/proc", "proc");
    mount("/dev", "devtmpfs")
    sethostname("container");
    setup_network();
    rc = execvp(args[0], args);
    return rc;
```

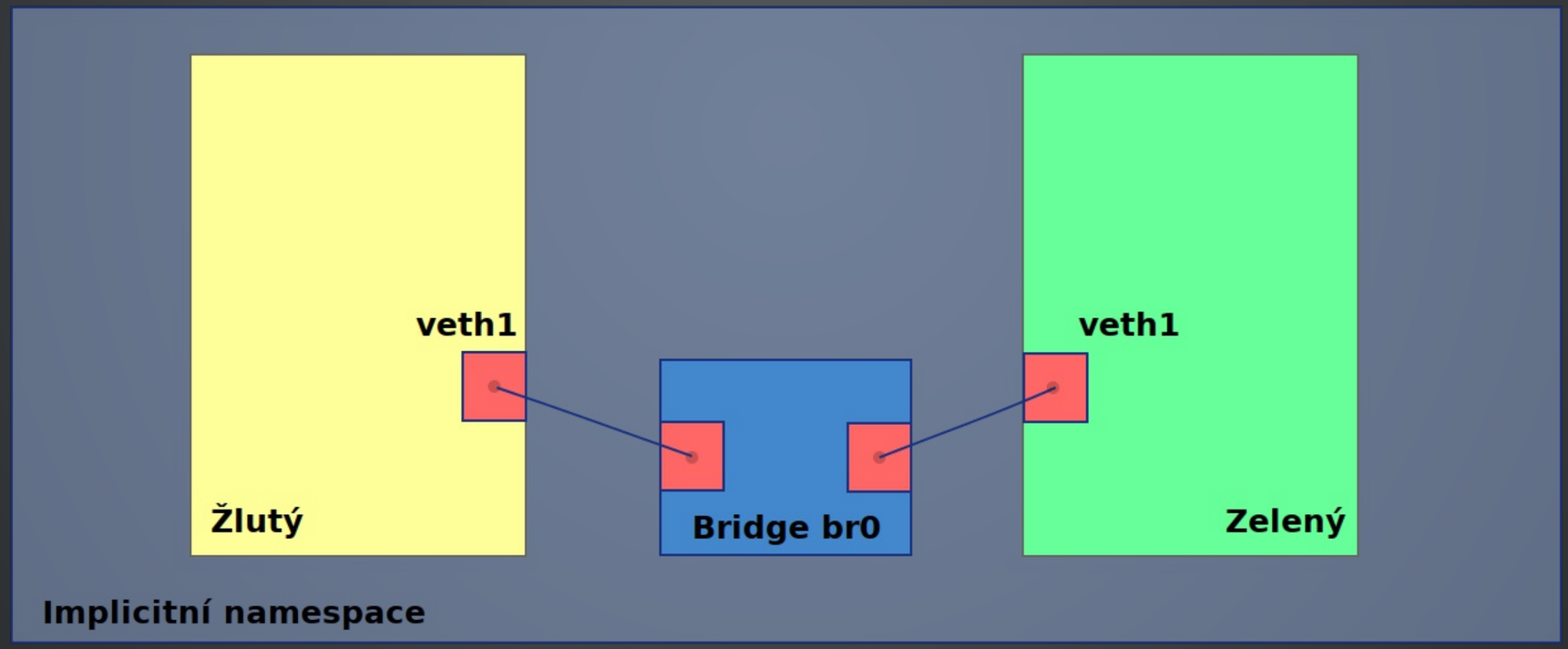
SLEPÍME TO DOHROMADY (FILESYSTEM)

Image + tmp = union fs

- aufs
- overlayFS
- vfs
- btrfs
- devicemapper

SLEPÍME TO DOHROMADY (SÍŤ)

Propojení více namespaces přes bridge



ČAS NA DEMO

https://github.com/w-vi/diyc

DÍKY ZA POZORNOST

Vilibald Wanča

vilibald@wvi.cz