Subject: Re: OverlayFS (was Re: CUDA support inside containers)
Posted by abufrejoval on Wed, 23 Nov 2016 19:17:43 GMT
View Forum Message <> Reply to Message

khorenko wrote on Wed, 23 November 2016 16:57Hi,

sorry for the delay, just a quick note:
i've file a dev task for making CUDA management in VZ Containers easier:
https://bugs.openvz.org/browse/OVZ-6834

and the first issue fixed: we've allowed to show content of /proc/modules inside Containers.

Kernel 3.10.0-327.36.1.vz7.20.2 should appear tomorrow at
https://download.openvz.org/virtuozzo/factory/x86_64/os/Pack ages/v/

Ooops, didn't expect you to react so fast or I would have prepared better...

Unfortunately I haven't found a way to attach strace logs into my messages (any file is always too big), so you'll have to do with manual edits and this being terribly long...

Here are the /proc and /sys access I've been able to find from a *successful* execution of the basic deviceQuery utility on the host.
I have not check (yet) which of those would work inside the container because they are being reflected/translated.

It starts with a lot of dynamic loading stuff, which I'll leave out mostly

```
execve("./deviceQuery", ["./deviceQuery"], [/* 18 vars */]) = 0
brk(0)                                  = 0x1601000
mmap(NULL, 4096, PROT_READ|PROT_WRITE, MAP_PRIVATE|MAP_ANONYMOUS, -1, 0) =
0x7fcdaba38000
access("/etc/ld.so.preload", R_OK)      = -1 ENOENT (No such file or directory)
open("/etc/ld.so.cache", O_RDONLY|O_CLOEXEC) = 3
fstat(3, {st_mode=S_IFREG|0644, st_size=120462, ...}) = 0
mmap(NULL, 120462, PROT_READ, MAP_PRIVATE, 3, 0) = 0x7fcdaba1a000
close(3)                                = 0
open("/lib64/librt.so.1", O_RDONLY|O_CLOEXEC) = 3
read(3, "\177ELF\2\1\1\3\0\0\0\0\0\0\0\0\3\0>\0\1\0\0\0\300\"\0\0\0\0\0\0 "..., 832) = 832
fstat(3, {st_mode=S_IFREG|0755, st_size=44096, ...}) = 0

... //lots more libraries

stat("/etc/sysconfig/64bit_strstr_via_64bit_strstr_sse2_unaligned ", 0x7fffbdd53980) = -1 ENOENT
(No such file or directory)
stat("/etc/sysconfig/64bit_strstr_via_64bit_strstr_sse2_unaligned ", 0x7fffbdd53980) = -1 ENOENT
(No such file or directory)
sched_get_priority_max(SCHED_RR)        = 99
sched_get_priority_min(SCHED_RR)        = 1
```

```
munmap(0x7fcdab9f6000, 120462)          = 0
clock_gettime(CLOCK_MONOTONIC_RAW, {19326, 808928358}) = 0
open("/proc/sys/vm/mmap_min_addr", O_RDONLY) = 3
fstat(3, {st_mode=S_IFREG|0644, st_size=0, ...}) = 0
mmap(NULL, 4096, PROT_READ|PROT_WRITE, MAP_PRIVATE|MAP_ANONYMOUS, -1, 0) =
0x7fcdaba36000
read(3, "4096\n", 1024)                 = 5
close(3)                                = 0
munmap(0x7fcdaba36000, 4096)            = 0

// I haven't checked this one, I just noticed that there was also /sys access but currently this
seems to be the only one

open("/sys/devices/system/cpu/online", O_RDONLY|O_CLOEXEC) = 3

read(3, "0-23\n", 8192)                 = 5
close(3)                                = 0
statfs("/dev/shm/", {f_type=0x1021994, f_bsize=4096, f_blocks=16493769, f_bfree=16493590,
f_bavail=16493590, f_files=16493769, f_ffree=16493759, f_fsid={0, 0}, f_namelen=255,
f_frsize=4096}) = 0
futex(0x7fcdab817330, FUTEX_WAKE_PRIVATE, 2147483647) = 0
open("/dev/shm/cuda_injection_path_shm", O_RDWR|O_NOFOLLOW|O_CLOEXEC) = -1
ENOENT (No such file or directory)

... // some configuration file access stuff

// this stuff I believe is working

readlink("/proc/201468/exe", "/home/thomas/NVIDIA_CUDA-8.0_Sam"..., 4095) = 72
geteuid()                               = 0
socket(PF_LOCAL, SOCK_SEQPACKET|SOCK_CLOEXEC, 0) = 3
setsockopt(3, SOL_SOCKET, SO_PASSCRED, [1], 4) = 0
connect(3, {sa_family=AF_LOCAL, sun_path="/tmp/nvidia-mps/control"}, 26) = -1 ENOENT (No
such file or directory)
close(3)                                = 0
lstat("/proc", {st_mode=S_IFDIR|0555, st_size=0, ...}) = 0
lstat("/proc/self", {st_mode=S_IFLNK|S_ISVTX|0777, st_size=0, ...}) = 0
readlink("/proc/self", "201468", 4095)  = 6
lstat("/proc/201468", {st_mode=S_IFDIR|0555, st_size=0, ...}) = 0
lstat("/proc/201468/exe", {st_mode=S_IFLNK|0777, st_size=0, ...}) = 0
readlink("/proc/201468/exe", "/home/thomas/NVIDIA_CUDA-8.0_Sam"..., 4095) = 72
lstat("/home", {st_mode=S_IFDIR|0755, st_size=4096, ...}) = 0
lstat("/home/thomas", {st_mode=S_IFDIR|0700, st_size=4096, ...}) = 0
lstat("/home/thomas/NVIDIA_CUDA-8.0_Samples", {st_mode=S_IFDIR|0775, st_size=4096, ...}) =
0
lstat("/home/thomas/NVIDIA_CUDA-8.0_Samples/1_Utilities", {st_mode=S_IFDIR|0755,
st_size=4096, ...}) = 0
lstat(" /home/thomas/NVIDIA_CUDA-8.0_Samples/1_Utilities/deviceQuery ",
```

{st_mode=S_IFDIR|0755, st_size=4096, ...}) = 0
lstat(" /home/thomas/NVIDIA_CUDA-8.0_Samples/1_Utilities/deviceQuery /deviceQuery ",
{st_mode=S_IFREG|0775, st_size=582882, ...}) = 0

//this would be where things start failing inside a container

open("/proc/modules", O_RDONLY)   = 3
fstat(3, {st_mode=S_IFREG|0444, st_size=0, ...}) = 0
mmap(NULL, 4096, PROT_READ|PROT_WRITE, MAP_PRIVATE|MAP_ANONYMOUS, -1, 0) =
0x7fcdaba36000
read(3, "arc4 12608 0 - Live 0xffffffffa1"..., 1024) = 1024
close(3)   = 0
munmap(0x7fcdaba36000, 4096)   = 0

// inside a container this is a zero size file not a directory

open("/proc/devices", O_RDONLY)   = 3
fstat(3, {st_mode=S_IFREG|0444, st_size=0, ...}) = 0
mmap(NULL, 4096, PROT_READ|PROT_WRITE, MAP_PRIVATE|MAP_ANONYMOUS, -1, 0) =
0x7fcdaba36000
read(3, "Character devices:\n 1 mem\n 2 p"..., 1024) = 652
close(3)   = 0
munmap(0x7fcdaba36000, 4096)   = 0

// that should work, because I declared those devices as visible (can I can even open them)

stat("/dev/nvidia-uvm", {st_mode=S_IFCHR|0666, st_rdev=makedev(243, 0), ...}) = 0
stat("/dev/nvidia-uvm-tools", {st_mode=S_IFCHR|0666, st_rdev=makedev(243, 1), ...}) = 0
open("/dev/nvidia-uvm", O_RDWR|O_CLOEXEC) = 3

fcntl(3, F_GETFD)   = 0x1 (flags FD_CLOEXEC)
ioctl(3, FIBMAP, 0x7fffbdd531c0)   = 0
clock_gettime(CLOCK_MONOTONIC_RAW, {19326, 812779374}) = 0
ioctl(3, 0x27, 0x7fffbdd531e0)   = 0
ioctl(3, 0x7ff, 0x7fffbdd531e0)   = 0
mmap(NULL, 135168, PROT_READ|PROT_WRITE, MAP_PRIVATE|MAP_ANONYMOUS, -1, 0)
= 0x7fcdab9f3000
open("/proc/modules", O_RDONLY)   = 4
fstat(4, {st_mode=S_IFREG|0444, st_size=0, ...}) = 0
mmap(NULL, 4096, PROT_READ|PROT_WRITE, MAP_PRIVATE|MAP_ANONYMOUS, -1, 0) =
0x7fcdaba36000
read(4, "arc4 12608 0 - Live 0xffffffffa1"..., 1024) = 1024
read(4, "13a0000\ntun 27141 1 - Live 0xfff"..., 1024) = 1024
read(4, "persistent_data, Live 0xffffffff"..., 1024) = 1024
close(4)   = 0
munmap(0x7fcdaba36000, 4096)   = 0

// and here the bind mount fails because I have a zero size file in /proc that blocks me from

duplicating the directory tree below

```
open("/proc/driver/nvidia/params", O_RDONLY) = 4
fstat(4, {st_mode=S_IFREG|0444, st_size=0, ...}) = 0
mmap(NULL, 4096, PROT_READ|PROT_WRITE, MAP_PRIVATE|MAP_ANONYMOUS, -1, 0) =
0x7fcdaba36000
read(4, "Mobile: 4294967295\nResmanDebugLe"..., 1024) = 441
close(4)                                = 0
munmap(0x7fcdaba36000, 4096)            = 0
stat("/dev/nvidiactl", {st_mode=S_IFCHR|0666, st_rdev=makedev(195, 255), ...}) = 0
open("/dev/nvidiactl", O_RDWR)          = 4
fcntl(4, F_SETFD, FD_CLOEXEC)           = 0
ioctl(4, 0xc04846d2, 0x7fffbdd53d80)    = 0
ioctl(4, 0xc00446ca, 0x7fcdaa60c060)    = 0
ioctl(4, 0xca0046c8, 0x7fcdaa60b660)    = 0
ioctl(4, 0xc020462b, 0x7fffbdd53dd0)    = 0
ioctl(4, 0xc020462a, 0x7fffbdd53ba0)    = 0
ioctl(4, 0xc020462a, 0x7fffbdd53ba0)    = 0

... // lots of stuff going on

// here are some /proc access which I believe you support already including PID translation

getrlimit(RLIMIT_AS, {rlim_cur=RLIM64_INFINITY, rlim_max=RLIM64_INFINITY}) = 0
open("/proc/self/maps", O_RDONLY)       = 8
fstat(8, {st_mode=S_IFREG|0444, st_size=0, ...}) = 0
mmap(NULL, 4096, PROT_READ|PROT_WRITE, MAP_PRIVATE|MAP_ANONYMOUS, -1, 0) =
0x7fcdaba36000
read(8, "00400000-00469000 r-xp 00000000 "..., 1024) = 1024
close(8)                                = 0
munmap(0x7fcdaba36000, 4096)            = 0
mmap(0x200000000, 4297064448, PROT_NONE, MAP_PRIVATE|MAP_ANONYMOUS, -1, 0) =
0x200000000
open("/proc/self/maps", O_RDONLY)       = 8
fstat(8, {st_mode=S_IFREG|0444, st_size=0, ...}) = 0
mmap(NULL, 4096, PROT_READ|PROT_WRITE, MAP_PRIVATE|MAP_ANONYMOUS, -1, 0) =
0x7fcdaba36000
read(8, "00400000-00469000 r-xp 00000000 "..., 1024) = 1024
close(8)                                = 0
munmap(0x7fcdaba36000, 4096)            = 0
mmap(0x10000000000, 4294967296, PROT_NONE, MAP_PRIVATE|MAP_ANONYMOUS, -1, 0)
= 0x10000000000
stat("/proc/201468/ns/pid", {st_mode=S_IFREG|0444, st_size=0, ...}) = 0
stat("/proc/201468/ns/pid", {st_mode=S_IFREG|0444, st_size=0, ...}) = 0
socket(PF_LOCAL, SOCK_SEQPACKET|SOCK_CLOEXEC, 0) = 8
unlink("")                              = -1 ENOENT (No such file or directory)
bind(8, {sa_family=AF_LOCAL, sun_path=@"cuda-uvmfd-4026531836-201468"}, 32) = 0
listen(8, 128)                          = 0
```

mmap(NULL, 8392704, PROT_READ|PROT_WRITE,
MAP_PRIVATE|MAP_ANONYMOUS|MAP_STACK, -1, 0) = 0x7fcda31d9000
mprotect(0x7fcda31d9000, 4096, PROT_NONE) = 0
clone(child_stack=0x7fcda39d8fb0,
flags=CLONE_VM|CLONE_FS|CLONE_FILES|CLONE_SIGHAND|CLONE_THRE
AD|CLONE_SYSVSEM|CLONE_SETTLS|CLONE_PARENT_SETTID|CLONE_CHIL
D_CLEARTID, parent_tidptr=0x7fcda39d99d0, tls=0x7fcda39d9700,
child_tidptr=0x7fcda39d99d0) = 201470
futex(0x160b9c8, FUTEX_WAKE_PRIVATE, 1) = 1
futex(0x66c260, FUTEX_WAKE_PRIVATE, 2147483647) = 0