KVM Arch Introduction

version 0.01

Mark Wu

May 08, 2012

Contents

KVM Architecture Introduction

KVM API General Description

An excerpt from kernel doc: Documentation/virtual/kvm/api.txt:

- The kvm API is centered around file descriptors.
- An initial open("/dev/kvm") obtains a handle to the kvm subsystem; this handle can be used to issue system ioctls.
- A KVM_CREATE_VM ioctl on this handle will create a VM file descriptor which can be used to issue VM ioctls.
- A KVM_CREATE_VCPU ioctl on a VM fd will create a virtual cpu and return a file descriptor pointing to it.
- Finally, ioctls on a vcpu fd can be used to control the vcpu, including the important task of actually running guest code.

- KVM related file descriptors in qemu.

```
(gdb) p kvm_state->fd
$1 = 3
(gdb) p kvm_state->vmfd
$2 = 4
(gdb) info threads
  4 Thread 0x7f86a60f0700 (LWP 13455)  0x00007f86ad0803dc in pthread_cond_wait@@GLIBC_2.3.2 () from /lib64/libpthread.so.0
  3 Thread 0x7f86a56ef700 (LWP 13456)  0x00007f86ad0803dc in pthread_cond_wait@@GLIBC_2.3.2 () from /lib64/libpthread.so.0
  2 Thread 0x7f86a6af1700 (LWP 13960)  0x00007f86ad08075b in pthread_cond_timedwait@@GLIBC_2.3.2 () from /lib64/libpthread.so.0
* 1 Thread 0x7f86ae478940 (LWP 13453)  0x00007f86a97772f3 in select () from /lib64/libc.so.6
(gdb) t 3
[Switching to thread 3 (Thread 0x7f86a56ef700 (LWP 13456))]#0  0x00007f86ad0803dc in pthread_cond_wait@@GLIBC_2.3.2 () from /lib64/libpthread.so.0
(gdb) bt
#0  0x00007f86ad0803dc in pthread_cond_wait@@GLIBC_2.3.2 () from /lib64/libpthread.so.0
#1  0x00007f86ae60a2e9 in qemu_cond_wait (cond=<value optimized out>, mutex=<value optimized out>) at qemu-thread-posix.c:113
#2  0x00007f86ae67772f in qemu_kvm_wait_io_event (arg=0x7f86b10a0930) at /home/mark/Work/qemu/qemu/cpus.c:710
#3  qemu_kvm_cpu_thread_fn (arg=0x7f86b10a0930) at /home/mark/Work/qemu/qemu/cpus.c:745
#4  0x00007f86ad07c7f1 in start_thread () from /lib64/libpthread.so.0
#5  0x00007f86a977e70d in clone () from /lib64/libc.so.6
(gdb) p ((CPUX86State *)0x7f86b10a0930)->kvm_fd
$3 = 12
(gdb) t 4
[Switching to thread 4 (Thread 0x7f86a60f0700 (LWP 13455))]#0  0x00007f86ad0803dc in pthread_cond_wait@@GLIBC_2.3.2 () from /lib64/libpthread.so.0
(gdb) bt
#0  0x00007f86ad0803dc in pthread_cond_wait@@GLIBC_2.3.2 () from /lib64/libpthread.so.0
#1  0x00007f86ae60a2e9 in qemu_cond_wait (cond=<value optimized out>, mutex=<value optimized out>) at qemu-thread-posix.c:113
#2  0x00007f86ae67772f in qemu_kvm_wait_io_event (arg=0x7f86b1088a00) at /home/mark/Work/qemu/qemu/cpus.c:710
#3  qemu_kvm_cpu_thread_fn (arg=0x7f86b1088a00) at /home/mark/Work/qemu/qemu/cpus.c:745
#4  0x00007f86ad07c7f1 in start_thread () from /lib64/libpthread.so.0
#5  0x00007f86a977e70d in clone () from /lib64/libc.so.6
(gdb) p ((CPUX86State *)0x7f86b1088a00)->kvm_fd
$4 = 11
```

- Dump KVM related files via crash

```
crash> files 15011
PID: 15011  TASK: ffff880053ea0100  CPU: 0   COMMAND: "qemu-system-x86"
ROOT: /    CWD: /home/mark/Work/qemu/qemu
 FD       FILE            DENTRY           INODE       TYPE PATH
  0 ffff880050b8c8c0 ffff88000ad77a80 ffff880134d13318 CHR  /dev/pts/4
  1 ffff880050b8c8c0 ffff88000ad77a80 ffff880134d13318 CHR  /dev/pts/4
  2 ffff880050b8c8c0 ffff88000ad77a80 ffff880134d13318 CHR  /dev/pts/4
  3 ffff88008491fa80 ffff880134c9b0c0 ffff88013b372a78 CHR  /dev/kvm
  4 ffff88012eb52140 ffff8800ae376e40 ffff88013b71e2d8 REG  anon_inode:/kvm-vm
  5 ffff8801357e7180 ffff8800ae3760c0 ffff88013b71e2d8 REG  anon_inode:/[signalfd]
  6 ffff880014255a80 ffff8800ae376180 ffff88013b71e2d8 REG  anon_inode:/[eventfd]
  7 ffff880014255a80 ffff8800ae376180 ffff88013b71e2d8 REG  anon_inode:/[eventfd]
  8 ffff880136751bc0 ffff880089da2c80 ffff88003f6490c0 REG  /home/mark/Work/qemu/images/fedora.img
  9 ffff8800a3c4d480 ffff8800ae376300 ffff880134cb1358 FIFO
 10 ffff88008adc6980 ffff8800ae376300 ffff880134cb1358 FIFO
 11 ffff88008ae865c0 ffff88012256f440 ffff88013b71e2d8 REG  anon_inode:/kvm-vcpu
 12 ffff88007bb11ec0 ffff88012256f2c0 ffff88013b71e2d8 REG  anon_inode:/kvm-vcpu

crash> p ((struct file *)0xffff88008491fa80)->f_op
$5 = (const struct file_operations *) 0xffffffffa04f0e40
crash> sym 0xffffffffa04f0e40
ffffffffa04f0e40 (d) kvm_chardev_ops [kvm]

crash> px *((struct file*)0xffff88007bb11ec0)->f_op
$7 = {
  owner = 0xffffffffa05249a0,
  llseek = 0,
  read = 0,
  write = 0,
  :
  ioctl = 0,
  unlocked_ioctl = 0xffffffffa04bae00,
  compat_ioctl = 0xffffffffa04bae00,
  mmap = 0xffffffffa04b9220,
  open = 0,
  flush = 0,
  release = 0xffffffffa04bd830,
  fsync = 0,
  aio_fsync = 0,
  :
  setlease = 0
}
crash> sym 0xffffffffa04bae00
ffffffffa04bae00 (t) kvm_vcpu_ioctl [kvm]
crash> sym 0xffffffffa04b9220
ffffffffa04b9220 (t) kvm_vcpu_mmap [kvm]

crash> px ((struct file *)0xffff88012eb52140)->private_data
$15 = (void *) 0xffff880137c6c000
crash> px vm_list
vm_list = $16 = {
  next = 0xffff880137c6c280,
  prev = 0xffff880137c6c280
}
crash> sym vm_list
ffffffffa04f0aa0 (D) vm_list [kvm]
crash> px ((struct kvm*)0xffff880137c6c000)->vm_list
$17 = {
  next = 0xffffffffa04f0aa0,
  prev = 0xffffffffa04f0aa0
}
```

CPU Virtulization

vCPU initilization

- **qemu-kvm backtrace of vcpu initlizaton**

```
(gdb) bt
#0  qemu_init_vcpu (_env=0x7ffff8b18a00) at /home/mark/Work/qemu/qemu/cpus.c:936
#1  0x00007ffff7e9f869 in cpu_x86_init (cpu_model=0x7ffff7f8fca9 "qemu64") at /home/mark/Work/qemu/qemu/target-i386/helper.c:1263
#2  0x00007ffff7ee1de0 in pc_new_cpu (cpu_model=0x7ffff7f8fca9 "qemu64") at /home/mark/Work/qemu/qemu/hw/pc.c:936
#3  pc_cpus_init (cpu_model=0x7ffff7f8fca9 "qemu64") at /home/mark/Work/qemu/qemu/hw/pc.c:963
#4  0x00007ffff7ee297c in pc_init1 (system_memory=0x7ffff8b113f0, system_io=0x7ffff8b114f0, ram_size=536870912, boot_device=0x7fffffffdf10 "cad",
    kernel_filename=0x0, kernel_cmdline=0x7ffff7f668eb "", initrd_filename=0x0, cpu_model=0x0, pci_enabled=1, kvmclock_enabled=1)
    at /home/mark/Work/qemu/qemu/hw/pc_piix.c:103
#5  0x00007ffff7ee30d8 in pc_init_pci (ram_size=536870912, boot_device=0x7fffffffdf10 "cad", kernel_filename=0x0, kernel_cmdline=0x7ffff7f668eb "",
```

```
    initrd_filename=0x0, cpu_model=<value optimized out>) at /home/mark/Work/qemu/qemu/hw/pc_piix.c:245
#6  0x00007ffff7de57a9 in main (argc=<value optimized out>, argv=<value optimized out>, envp=<value optimized out>) at /home/mark/Work/qemu/qemu/vl.c:3351
```

```
qemu_init_vcpu
  qemu_kvm_start_vcpu
    qemu_thread_create(env->thread, qemu_kvm_cpu_thread_fn, env);  /* One qemu thread per vCPU */
      qemu_kvm_cpu_thread_fn
        kvm_init_vcpu
    +-->kvm_cpu_exec---+
    |                  |
    -------------------+

int kvm_init_vcpu(CPUState *env)
{
    KVMState *s = kvm_state;
    long mmap_size;
    int ret;

    DPRINTF("kvm_init_vcpu\n");

    ret = kvm_vm_ioctl(s, KVM_CREATE_VCPU, env->cpu_index);
    if (ret < 0) {
        DPRINTF("kvm_create_vcpu failed\n");
        goto err;
    }

    env->kvm_fd = ret;
    env->kvm_state = s;
    env->kvm_vcpu_dirty = 1;

    mmap_size = kvm_ioctl(s, KVM_GET_VCPU_MMAP_SIZE, 0);
    if (mmap_size < 0) {
        ret = mmap_size;
        DPRINTF("KVM_GET_VCPU_MMAP_SIZE failed\n");
        goto err;
    }

    env->kvm_run = mmap(NULL, mmap_size, PROT_READ | PROT_WRITE, MAP_SHARED,
                        env->kvm_fd, 0);
    :
}
```

Guest execution

- **qemu function kvm_cpu_exec**

```
int kvm_cpu_exec(CPUState *env)
{
    struct kvm_run *run = env->kvm_run;
    int ret, run_ret;

    DPRINTF("kvm_cpu_exec()\n");

    if (kvm_arch_process_async_events(env)) {
        env->exit_request = 0;
        return EXCP_HLT;
    }

    cpu_single_env = env;

    do {
        if (env->kvm_vcpu_dirty) {
            kvm_arch_put_registers(env, KVM_PUT_RUNTIME_STATE);
            env->kvm_vcpu_dirty = 0;
        }

        kvm_arch_pre_run(env, run);
        if (env->exit_request) {
            DPRINTF("interrupt exit requested\n");
            /*
             * KVM requires us to reenter the kernel after IO exits to complete
             * instruction emulation. This self-signal will ensure that we
             * leave ASAP again.
             */
            qemu_cpu_kick_self();
        }
        cpu_single_env = NULL;
        qemu_mutex_unlock_iothread();

        run_ret = kvm_vcpu_ioctl(env, KVM_RUN, 0);

        qemu_mutex_lock_iothread();
        cpu_single_env = env;
        kvm_arch_post_run(env, run);

        kvm_flush_coalesced_mmio_buffer();

        if (run_ret < 0) {
            if (run_ret == -EINTR || run_ret == -EAGAIN) {
                DPRINTF("io window exit\n");
                ret = EXCP_INTERRUPT;
                break;
            }
            DPRINTF("kvm run failed %s\n", strerror(-run_ret));
            abort();
        }

        switch (run->exit_reason) {
        case KVM_EXIT_IO:
            DPRINTF("handle_io\n");
            kvm_handle_io(run->io.port,
                          (uint8_t *)run + run->io.data_offset,
                          run->io.direction,
                          run->io.size,
                          run->io.count);
            ret = 0;
            break;
        case KVM_EXIT_MMIO:
            DPRINTF("handle_mmio\n");
            cpu_physical_memory_rw(run->mmio.phys_addr,
                                   run->mmio.data,
                                   run->mmio.len,
                                   run->mmio.is_write);
            ret = 0;
            break;
            :
        }

    } while (ret == 0);

    :
    return ret;
}
```

- kernel code path

```
sys_ioctl
  do_vfs_ioctl
    vfs_ioctl
```

```
      kvm_vcpu_ioctl  /* kvm_vcpu_fops.unlocked_ioctl */
        kvm_arch_vcpu_ioctl_run
          __vcpu_run
           vcpu_enter_guest
             vmx_vcpu_run /* kvm_x86_ops->run */
                    |
                    v  vm entry
             +-----------------+
             |   guest code    |
             |  on this cpu    |
             +-----------------+
                    |  vm exit
                    v
             vmx_handle_exit /* kvm_x86_ops->handle_exit */
               return kvm_vmx_exit_handlers[exit_reason](vcpu)
```

- kernel exit handlers

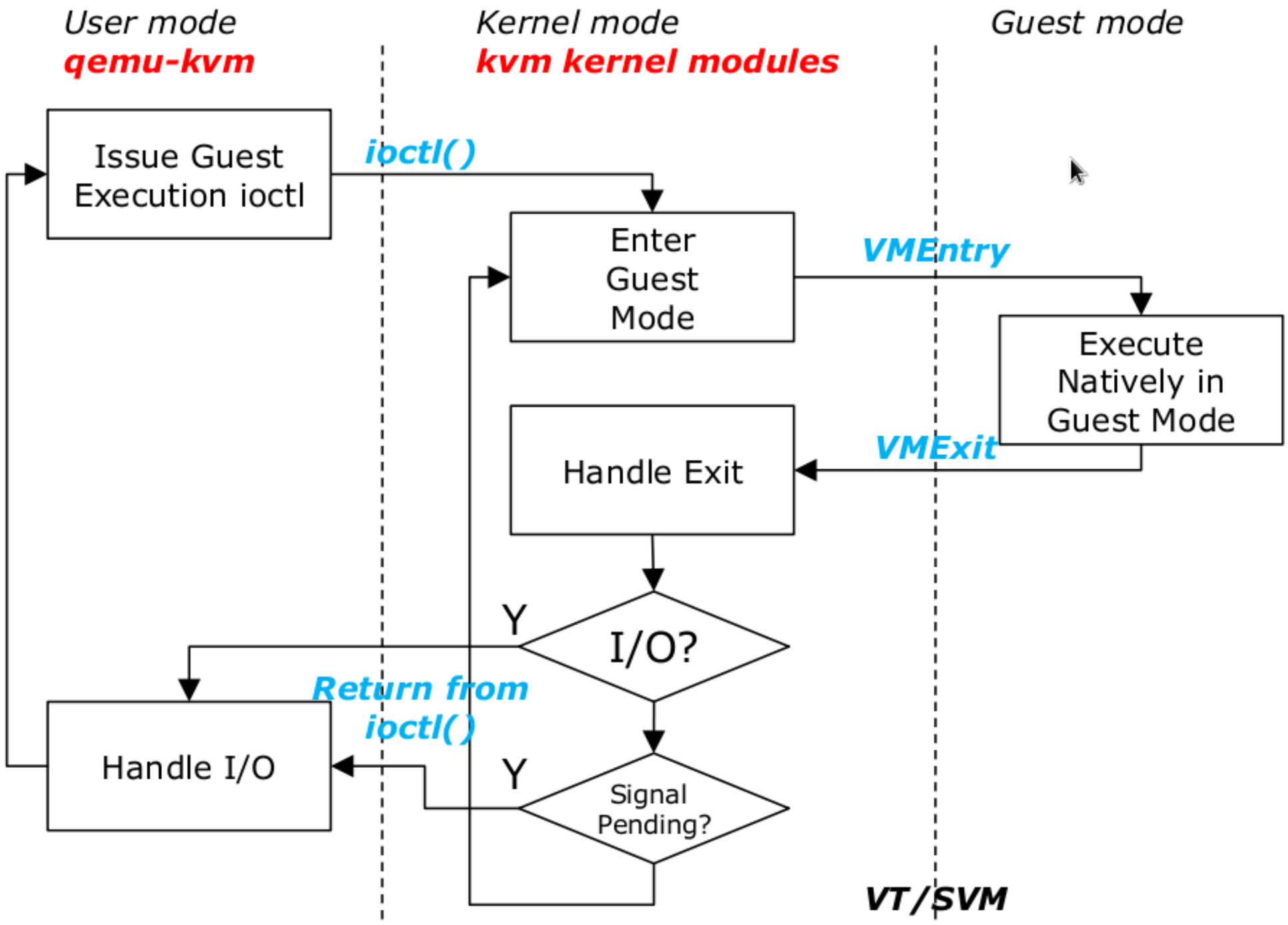
```
/*
 * The exit handlers return 1 if the exit was handled fully and guest execution
 * may resume.  Otherwise they set the kvm_run parameter to indicate what needs
 * to be done to userspace and return 0.
 */
static int (*kvm_vmx_exit_handlers[])(struct kvm_vcpu *vcpu) = {
        [EXIT_REASON_EXCEPTION_NMI]           = handle_exception,
        [EXIT_REASON_EXTERNAL_INTERRUPT]      = handle_external_interrupt,
        [EXIT_REASON_TRIPLE_FAULT]            = handle_triple_fault,
        [EXIT_REASON_NMI_WINDOW]              = handle_nmi_window,
        [EXIT_REASON_IO_INSTRUCTION]          = handle_io,
        :
```

- guest runtime information shared between kvm mod and qemu-kvm

```
env->kvm_run = mmap(NULL, mmap_size, PROT_READ | PROT_WRITE, MAP_SHARED,
                env->kvm_fd, 0);

(gdb) p ((struct CPUX86State*)0x7fcdbe63f930)->kvm_run
$2 = (struct kvm_run *) 0x7fcdbcfa2000
(gdb) p *((struct CPUX86State*)0x7fcdbe63f930)->kvm_run
$3 = {request_interrupt_window = 0 '\000', padding1 = "\000\000\000\000\000\000", exit_reason = 10, ready_for_interrupt_injection = 0 '\000', if_flag =
    0 '\000', padding2 = "\000", cr8 = 0, apic_base = 4276094976, {hw = {hardware_exit_reason = 4276093104}, fail_entry = {hardware_entry_failure_reason =
    4276093104}, ex = {exception = 4276093104, error_code = 0}, io = {direction = 176 '\260', size = 0 '\000', port = 65248, count = 0, data_offset =
    513418191540584448}, debug = {arch = {exception = 4276093104, pad = 0, pc = 513418191540584448, dr6 = 4294967300, dr7 = 0}}, mmio = {phys_addr =
    4276093104, data = "\000\000\000\000 \a \a", len = 4, is_write = 1 '\001'}, hypercall = {nr = 4276093104, args = {513418191540584448, 4294967300, 0, 0,
    0, 0}, ret = 0, longmode = 0, pad = 0}, tpr_access = {rip = 4276093104, is_write = 0, pad = 119539488}, s390_sieic = {icptcode = 176 '\260', ipa =
    65248, ipb = 0}, s390_reset_flags = 4276093104, dcr = {dcrn = 4276093104, data = 0, is_write = 0 '\000'}, internal = {suberror = 4276093104, ndata = 0,
      data = {513418191540584448, 4294967300, 0 <repeats 14 times>}}, osi = {gprs = {4276093104, 513418191540584448, 4294967300, 0 <repeats 29 times>}},
    papr_hcall = {nr = 4276093104, ret = 513418191540584448, args = {4294967300, 0, 0, 0, 0, 0, 0, 0}}, padding =
    "\260\000\340\376\000\000\000\000\000\000\000\000 \a \a\004\000\000\000\001", '\000' <repeats 234 times>}}
```

```
crash> vtop 7fcdbcfa2000
VIRTUAL     PHYSICAL
7fcdbcfa2000  12eb3c000

   PML: 137dfd7f8 => 136ff7067
   PUD: 136ff79b0 => 134069067
   PMD: 134069f38 => 13671c067
   PTE: 13671cd10 => 800000012eb3c067
  PAGE: 12eb3c000

      PTE         PHYSICAL   FLAGS
800000012eb3c067  12eb3c000  (PRESENT|RW|USER|ACCESSED|DIRTY|NX)

      VMA           START       END     FLAGS FILE
ffff8800aac39b70 7fcdbcfa2000 7fcdbcfa5000     fb anon_inode:/kvm-vcpu

      PAGE       PHYSICAL      MAPPING       INDEX CNT FLAGS
ffffea0004237520 12eb3c000                0 ffff8800b72c9980  2 40000000000014

crash> px ((struct kvm*)0xffff880137c6c000)->vcpus[1]->run
$23 = (struct kvm_run *) 0xffff88012eb3c000
crash> vtop 0xffff88012eb3c000
VIRTUAL           PHYSICAL
ffff88012eb3c000  12eb3c000

PML4 DIRECTORY: ffffffff81a85000
PAGE DIRECTORY: 1a86063
   PUD: 1a86020 => a067
   PMD: aba8 => 800000012ea001e3
  PAGE: 12ea00000  (2MB)

      PTE         PHYSICAL   FLAGS
800000012ea001e3  12ea00000  (PRESENT|RW|ACCESSED|DIRTY|PSE|GLOBAL|NX)

      PAGE       PHYSICAL      MAPPING       INDEX CNT FLAGS
ffffea0004237520 12eb3c000                0 ffff8800b72c9980  2 40000000000014

crash> px ((struct file*)0xffff88007bb11ec0)->private_data
$30 = (void *) 0xffff88013860c2b8
crash> px ((struct kvm_vcpu *)0xffff88013860c2b8)->run
$31 = (struct kvm_run *) 0xffff88012eb3c000
```

- Summary: Guest Execution Loop



Guest Execution Loop (Copied from other slides)

Physical Memory Virtualization

Physical memory intialization

- Qemu backtrace

```
(gdb) bt
#0  kvm_set_user_memory_region (s=0x7ffff8b100a0, slot=0x7ffff8b100a0) at /home/mark/Work/qemu/qemu/kvm-all.c:168
#1  0x00007ffff7ea3fae in kvm_set_phys_mem (client=<value optimized out>, start_addr=<value optimized out>, size=<value optimized out>,
    phys_offset=<value optimized out>, log_dirty=false) at /home/mark/Work/qemu/qemu/kvm-all.c:650
#2  kvm_client_set_memory (client=<value optimized out>, start_addr=<value optimized out>, size=<value optimized out>, phys_offset=<value optimized out>,
    log_dirty=false) at /home/mark/Work/qemu/qemu/kvm-all.c:663
#3  0x00007ffff7e8405a in cpu_notify_set_memory (start_addr=0, size=134217728, phys_offset=0, region_offset=0, log_dirty=false)
    at /home/mark/Work/qemu/qemu/exec.c:1742
#4  cpu_register_physical_memory_log (start_addr=0, size=134217728, phys_offset=0, region_offset=0, log_dirty=false)
    at /home/mark/Work/qemu/qemu/exec.c:2675
#5  0x00007ffff7eaac70 in address_space_update_topology_pass (as=0x7ffff82f31e0, old_view=..., new_view=..., adding=true)
    at /home/mark/Work/qemu/qemu/memory.c:731
#6  0x00007ffff7eacf31 in address_space_update_topology (as=0x7ffff82f31e0) at /home/mark/Work/qemu/qemu/memory.c:746
#7  0x00007ffff7ead514 in memory_region_update_topology () at /home/mark/Work/qemu/qemu/memory.c:760
#8  0x00007ffff7ee1787 in pc_memory_init (system_memory=0x7ffff8b11430, kernel_filename=<value optimized out>, kernel_cmdline=0x7ffff7f668eb "",
    initrd_filename=0x0, below_4g_mem_size=134217728, above_4g_mem_size=0, rom_memory=0x7ffff8b32240, ram_memory=0x7fffffffe188)
    at /home/mark/Work/qemu/qemu/hw/pc.c:996
#9  0x00007ffff7ee2d96 in pc_init1 (system_memory=0x7ffff8b11430, system_io=0x7ffff8b11530, ram_size=134217728, boot_device=0x7fffffffe500 "cad",
    kernel_filename=0x0, kernel_cmdline=0x7ffff7f668eb "", initrd_filename=0x0, cpu_model=0x0, pci_enabled=1, kvmclock_enabled=1)
    at /home/mark/Work/qemu/qemu/hw/pc_piix.c:128
#10 0x00007ffff7ee30d8 in pc_init_pci (ram_size=134217728, boot_device=0x7fffffffe500 "cad", kernel_filename=0x0, kernel_cmdline=0x7ffff7f668eb "",
    initrd_filename=0x0, cpu_model=<value optimized out>) at /home/mark/Work/qemu/qemu/hw/pc_piix.c:245
#11 0x00007ffff7de57a9 in main (argc=<value optimized out>, argv=<value optimized out>, envp=<value optimized out>) at /home/mark/Work/qemu/qemu/vl.c:3351
-----------------------------------------------
kvm_set_user_memory_region
  kvm_vm_ioctl
    ioctl(kvm_context->vm_fd, KVM_SET_USER_MEMORY_REGION, ...)
```

Guest physical memory mapping

- dump gpa <-> hva <-> hpa mapping via crash

```
crash> px vm_list
vm_list = $7 = {
  next = 0xffff880080cb4280,
  prev = 0xffff880080cb4280
}
crash> struct kvm.vm_list
struct kvm {
   [640] struct list_head vm_list;
}
crash> px 0xffff880080cb4280-640
$8 = 0xffff880080cb4000
crash> pd ((struct kvm *)0xffff880080cb4000)->memslots
$9 = (struct kvm_memslots *) 0xffff880139326000
crash> px *((struct kvm *)0xffff880080cb4000)->memslots
$6 = {
  nmemslots = 0x23,
  memslots = {{
      base_gfn = 0x0,
      npages = 0xa0,
      flags = 0x0,
      rmap = 0xffffc90016aac000,
      dirty_bitmap = 0x0,
      lpage_info = {0xffffc900175d6000, 0xffffc900175d9000},
      userspace_addr = 0x7f30dbe00000,
      user_alloc = 0x1,
      id = 0x0
    }, {
      base_gfn = 0xfffe0,
      npages = 0x20,
      flags = 0x0,
      rmap = 0xffffc90016a82000,
      dirty_bitmap = 0x0,
      lpage_info = {0xffffc90016a85000, 0xffffc90016a88000},
```

```
    userspace_addr = 0x7f310b1f0000,
    user_alloc = 0x1,
    id = 0x1
  }, {
    base_gfn = 0xc0,
    npages = 0xc,
    flags = 0x0,
    rmap = 0xffffc9001787f000,
    dirty_bitmap = 0x0,
    lpage_info = {0xffffc90017882000, 0xffffc90017885000},
    userspace_addr = 0x7f30dbec0000,
    user_alloc = 0x1,
    id = 0x2
  }, {
    base_gfn = 0xfc000,
    npages = 0x800,
    flags = 0x1,
    rmap = 0xffffc90017b39000,
    dirty_bitmap = 0xffffc90017b45000,
    lpage_info = {0xffffc90017b3f000, 0xffffc90017b42000},
    userspace_addr = 0x7f3101c00000,
    user_alloc = 0x1,
    id = 0x3
  }, {
    base_gfn = 0xcc,
    npages = 0x24,
    flags = 0x0,
    rmap = 0xffffc90017990000,
    dirty_bitmap = 0x0,
    lpage_info = {0xffffc90017993000, 0xffffc90017996000},
    userspace_addr = 0x7f30dbecc000,
    user_alloc = 0x1,
    id = 0x4
  }, {
    base_gfn = 0xf0,
    npages = 0x10,
    flags = 0x0,
    rmap = 0xffffc90017999000,
    dirty_bitmap = 0x0,
    lpage_info = {0xffffc9001799c000, 0xffffc9001799f000},
    userspace_addr = 0x7f30dbef0000,
    user_alloc = 0x1,
    id = 0x5
  }, {
    base_gfn = 0x100,
    npages = 0x1ff00,
    flags = 0x0,
    rmap = 0xffffc900179a2000,
    dirty_bitmap = 0x0,
    lpage_info = {0xffffc90017aa4000, 0xffffc90017aa7000},
    userspace_addr = 0x7f30dbf00000,
    user_alloc = 0x1,
    id = 0x6
  }, {
    base_gfn = 0x0,
    npages = 0x0,
    flags = 0x0,
    rmap = 0x0,
    dirty_bitmap = 0x0,
    lpage_info = {0x0, 0x0},
    userspace_addr = 0x0,
    user_alloc = 0x0,
    id = 0x0
  },
```

```
On Guest:
[root@localhost ~]# ./hello
[0x400638]: Hello, world

crash> ps \|grep hello
   2203   2112   0  ffff88001d68ae60  IN   0.1    4124    356  hello
crash> set 2203
    PID: 2203
COMMAND: "hello"
   TASK: ffff88001d68ae60  [THREAD_INFO: ffff88001da6c000]
    CPU: 0
  STATE: TASK_INTERRUPTIBLE
crash> rd 0x400638 2
          400638:  77202c6f6c6c6548 255b000a646c726f   Hello, world..[%
crash> vtop 0x400638
VIRTUAL     PHYSICAL
400638      30b2638

   PML: 1d669000 => 1dbfa067
   PUD: 1dbfa000 => 1c82d067
   PMD: 1c82d010 => 1ab49067
   PTE: 1ab49000 => 30b2025
  PAGE: 30b2000

On Host:
crash> px 0x7f30dbf00000+0x30b2638-0x100000
$7 = 0x7f30deeb2638
crash> rd 0x7f30deeb2638 2
    7f30deeb2638:  77202c6f6c6c6548 255b000a646c726f   Hello, world..[%
```

MMU Virtualization

Extended Page Table

- Overview

```
Guest CR3                       EPT Base Pointer
   |                                 |
   +-->+-------------------+        +---->+---------------------+
GVA--->|  Guest Page Table | ---> GPA ---> | Extended Page Table | ---> HPA
       +-------------------+               +---------------------+
```

- EPT walkthrough

```
crash> px ((struct kvm_vcpu *)0xffff88007768c078)->arch.mmu
$18 = {
  new_cr3 = 0xffffffffa04dca40 <nonpaging_new_cr3>,
  page_fault = 0xffffffffa04e4410 <tdp_page_fault>,
  free = 0xffffffffa04e0870 <nonpaging_free>,
  gva_to_gpa = 0xffffffffa04e4b70 <paging64_gva_to_gpa>,
  prefetch_page = 0xffffffffa04dc7a0 <nonpaging_prefetch_page>,
  sync_page = 0xffffffffa04dc7d0 <nonpaging_sync_page>,
  invlpg = 0xffffffffa04dc7e0 <nonpaging_invlpg>,
  root_hpa = 0x138457000,
  root_level = 0x4,
  shadow_root_level = 0x4,
  base_role = {
    word = 0x0,
    {
      glevels = 0x0,
      level = 0x0,
      quadrant = 0x0,
      pad_for_nice_hex_output = 0x0,
```

```
      direct = 0x0,
      access = 0x0,
      invalid = 0x0,
      cr4_pge = 0x0,
      nxe = 0x0,
      cr0_wp = 0x0,
      smep_andnot_wp = 0x0
    }
  },
  pae_root = 0xffff88000d2c2000,
  rsvd_bits_mask = {{0xfff0000000000, 0xfff0000000000, 0xfff0000000180, 0xfff0000000180}, {0x0, 0xfff00001fe000, 0xfff003fffe000, 0xfff0000000180}}
}
crash> px (0x30b2638>>39)&0x1ff
$19 = 0x0
crash> rd -p 0x138457000
       138457000:  0000000043138007                    ...C....
crash> px (0x30b2638>>30)&0x1ff
$20 = 0x0
crash> rd -p 0x43138000
        43138000:  0000000108c3c007                    ........
crash> px (0x30b2638>>21)&0x1ff
$21 = 0x18
crash> px (0x108c3c007 & ~0xfff)+ (8*0x18)
$22 = 0x108c3c0c0
crash> rd -p 0x108c3c0c0
       108c3c0c0:  0000000125713007                    .0q%....
crash> px (0x30b2638>>12)&0x1ff
$23 = 0xb2
crash> px (0x125713007 & ~0xfff) + (8*0xb2)
$24 = 0x125713590
crash> rd -p 0x125713590
       125713590:  000000011289a277                    w.......
crash> vtop 7f30deeb2638
VIRTUAL     PHYSICAL
7f30deeb2638  11289a638

   PML: 575e07f0 => 43236067
   PUD: 43236618 => 7ef3c067
   PMD: 7ef3c7b8 => 139495067
   PTE: 139495590 => 8000000011289a067
  PAGE: 11289a000

      PTE         PHYSICAL   FLAGS
8000000011289a067  11289a000  (PRESENT|RW|USER|ACCESSED|DIRTY|NX)

      VMA           START       END     FLAGS FILE
ffff88005d100788 7f30dbe00000 7f30fbe00000 80100073

      PAGE       PHYSICAL      MAPPING       INDEX CNT FLAGS
ffffea0003c0e1b0 11289a000 ffff88000d307f61 7f30deeb2  1 4000000010006c
```

Shadow Page Table

- Overview

```
Guest CR3
   |
   +-->+-------------------+
GVA--->|  Guest Page Table | ---> GPA
       +-------------------+

Host CR3
   |
   +-->+-------------------+
GVA--->|  Shadow Page Table | ---> HPA
       +-------------------+
```

- Shadow page table walkthrough (with option *ept=no* for kernel moduel kvm_intel)

```
crash> px ((struct kvm_vcpu *)0xffff88007768c078)->arch.mmu
    mmu = {
      new_cr3 = 0xffffffffa0914890 <paging_new_cr3>,
      page_fault = 0xffffffffa091a1b0 <paging64_page_fault>,
      free = 0xffffffffa0914880 <paging_free>,
      gva_to_gpa = 0xffffffffa0918b70 <paging64_gva_to_gpa>,
      prefetch_page = 0xffffffffa0915920 <paging64_prefetch_page>,
      sync_page = 0xffffffffa09177e0 <paging64_sync_page>,
      invlpg = 0xffffffffa0913b20 <paging64_invlpg>,
      root_hpa = 0x8886d000,
      root_level = 0x4,
      shadow_root_level = 0x4,
      base_role = {
        word = 0xe00004,
        {
          glevels = 0x4,
          level = 0x0,
          quadrant = 0x0,
          pad_for_nice_hex_output = 0x0,
          direct = 0x0,
          access = 0x0,
          invalid = 0x0,
          cr4_pge = 0x1,
          nxe = 0x1,
          cr0_wp = 0x1,
          smep_andnot_wp = 0x0
        }
      },
      pae_root = 0xffff88008893e000,
      rsvd_bits_mask = {{0xfff0000000000, 0xfff0000000000, 0xfff0000000180, 0xfff0000000180}, {0x0, 0xfff00001fe000, 0xfff003fffe000, 0xfff0000000180}}
    },

crash> px (0x400608 >> 39) & 0x1ff
$17 = 0x0
crash> rd -p 0x8886d000
        8886d000:  0000000081517027                    'pQ.....
crash> px (0x400608 >> 30) & 0x1ff
$18 = 0x0
crash> px (0x81517027 & ~0xfff)
$19 = 0x81517000
crash> rd -p 0x81517000
        81517000:  000000008159f027                    '.Y.....
crash> px (0x400608 >> 21) & 0x1ff
$20 = 0x2
crash> px (0x8159f027 & ~0xfff)+(8*0x2)
$21 = 0x8159f010
crash> rd -p 0x8159f010
        8159f010:  0000000069fd7027                    'p.i....
crash> px (0x400608 >> 12) & 0x1ff
$22 = 0x0
crash> rd -p 0x69fd7000
        69fd7000:  0000000055b99265                    e..U....
crash> rd -p 55b99608 2
        55b99608:  77202c6f6c6c6548 255b000a646c726f   Hello, world..[%
```