Связь реальности и виртуальности

Курс «Программное моделирование вычислительных систем»

Григорий Речистов
grigory.rechistov@phystech.edu

18 апреля 2016 г.

На прошлых лекциях

Изоляция + эквивалентность $\neq$ эффективность

- Мы стремились к наиболее точной симуляции. Программы в гостевом окружении не должны догадываться, что они исполняются не на реальной аппаратуре
- Они не должны знать о мониторе ВМ
- Цена этому — скорость работы

Вопросы по предыдущей лекции

- Какие три признака эффективного монитора ВМ сформулированы Г. и П.?
- Что такое безвредные (innocuous) инструкции?
- Что такое и как используется SLAT (second level address translation)?

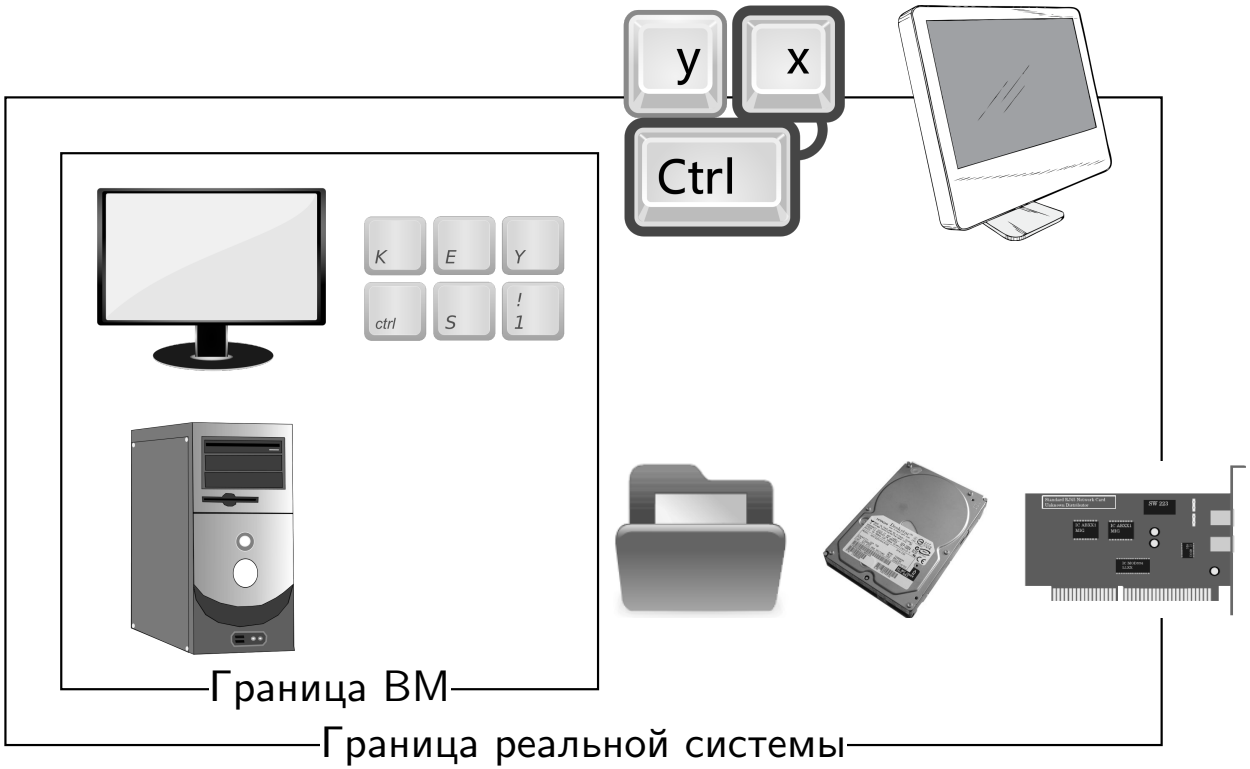
На этой лекции

- Взаимодействие виртуальности с реальностью
- Паравиртуализация

Зачем

- Иметь возможность передавать данные в/из симулируемой системы
- Более эффективно *совместно* работать над управлением ресурсами машины: память, время, устройства
- Иметь возможность передавать аппаратные ресурсы частично или полностью в VM

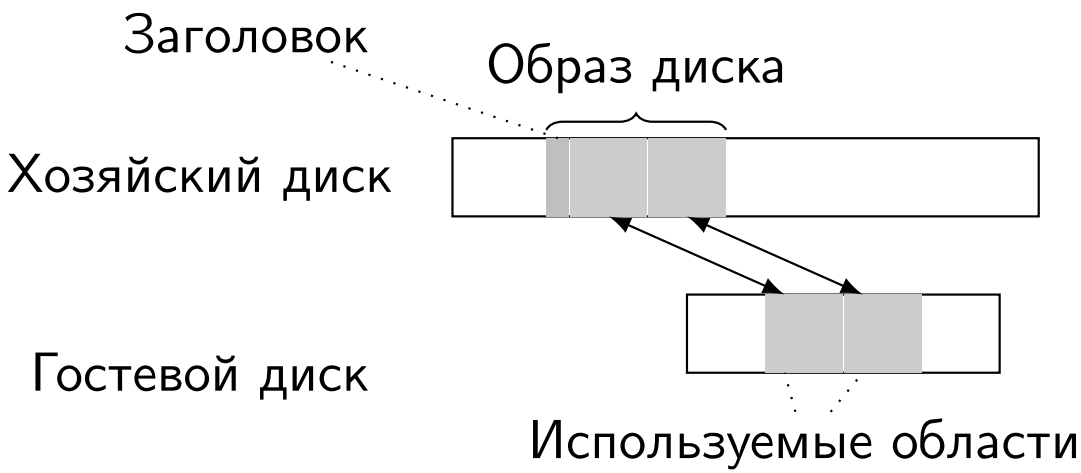
Связи между гостем и хозяином



Образы дисков

Жёсткие диски

- RAW
- VMDK, VDI, Qcow2, CRAFF, HDD, VHD...



Образы дисков

Оптические диски CD/DVD/Blueray

- RAW для ISO 9660
- Существует множество форматов, но они не используются в симуляторах или ВМ: NRG, MDF, ISZ, DMG, IMG...

Гибкие диски

- 360 кБ — 2,88 МБ; Формат — RAW

Волшебные инструкции I

- Инструкция процессора с побочными эффектами
- Остановка симуляции
- Вызов обработчика, имеющего доступ к состоянию симулируемой системы
- Изменение состояния
- Возобновление симуляции
- Для VM действия происходят «мгновенно»

Очень желательно, чтобы инструкция не встречалась в «обычном» коде

- Непривилегированная

Волшебные инструкции II

- Идеально, чтобы она вообще не имела эффектов вне симуляции
- NOP — хороший кандидат
- ... но она используется в программах очень часто (ложные срабатывания)

Варианты

- NOP с префиксами/аргументами
- В IA-32 есть «длинный NOP», от 1 до 9 байт
- CPUID — в Simics для IA-32 (т.к. вызывает VM-exit)
- Недокументированные инструкции (если ещё остались)

Использование:

- Инструментация приложений

Волшебные инструкции III

- Передача данных (по одному слову)
- Канал управления для механизма RPC

Не очень волшебные инструкции

Специальные привилегированные инструкции для явного взаимодействия гостя-хозяина или минимизации накладных расходов для частых операций

- Hypercall, по аналогии с SYSCALL. В IA-32 - VMCALL
- Кооперация по управлению виртуальной памятью. В IA-32: #VE и VMFUNC

Паравиртуальные устройства

- Объём передаваемых данных за один раз больше
- Ненастоящее устройство в карте памяти модели
- Требует модификации гостевой ОС: драйвера устройств
- Типичные кандидаты — устройства с большой пропускной способностью: диски, сетевые карты
- Пример: VirtIO [7]

Паравиртуальные сервисы

Взаимодействие с сервисами гостевой ОС:

- Управление питанием и жизненным циклом ОС
- Файловая система
- Потребление физической памяти
- Консольный доступ
- Синхронизация времени [3]

Примеры:

- Wind River Simics Agent/Matic, Simics hostfs
- Oracle Virtualbox Guest Additions
- VMWare ESX Tools
- Microsoft Hyper-V Integration Services [6]

Проброс устройств (passthrough) I

Передача хозяйского устройства в эксклюзивное использование гостя

- Графическая карта
- Сетевые карты
- Прочие устройства (PCIe/USB)

Проблемы: устройство или хозяйская ОС могут быть против!

- Доступы в регистры устройства
- Доступы устройства в память (DMA)
- Доставка прерываний от устройства

Дополнительные проблемы:

- Отключение устройства от хозяина

Проброс устройств (passthrough) II

- Графика и прочие legacy-особенности
- Сохранение-восстановление и миграция состояния

- Аппаратная поддержка (I/O Memory management unit)
- Intel VT-d, AMD IOMMU, IBM Translation Control Entry, Sun DVMA

Сетевое взаимодействие

- Изначально создано для связи систем различной природы
- Агностично к аппаратуре
- Можно выбирать уровень абстракции, на котором будет проходить граница миров

Недостаток: требуется рабочий сетевой стек в госте

Передача данных на разных уровнях модели OSI/ISO

- Прикладной уровень — сервис-точки внутри симуляции. Отвечают на запросы гостя по конкретному прикладному протоколу: NFS, FTP, Samba, DHCP...
- Представительный уровень
- Сеансовый уровень
- Транспортный уровень — NAT. Пакеты ретранслируются от имени хозяина. Входящие пакеты по умолчанию не доходят
- Сетевой уровень — TUN-драйвер хозяина. Создаёт IP туннель
- Канальный уровень — TAP-драйвер хозяина. Создаёт псевдоустройство Ethernet хозяина
- Физический уровень — модель карты внутри симуляции

Демо на VirtualBox

Что осталось за рамками лекции

- Chroot
- Containers
- PAL/SAL уровень
- paravirt_ops [4, 5]

На следущей лекции

Языки моделирования и описания аппаратуры

Литература I

- http://www.kernel.org/pub/linux/kernel/people/marcelo/linux-2.4/Documentation/networking/tuntap.txt
- http://wiki.xen.org/wiki/Xen_PCI_Passthrough
  http://wiki.xen.org/wiki/XenVGAPassthrough
  http://wiki.xen.org/wiki/XenUSBPassthrough
- Виртуальное время, часть 2: вопросы симуляции и виртуализации
  http://habrahabr.ru/company/intel/blog/260119/
- http://lwn.net/Articles/194339/
- http://lwn.net/Articles/225881/

Литература II

- https://technet.microsoft.com/en-us/library/dn798297.aspx
- http://www.linux-kvm.org/images/d/dd/KvmForum2007$kvm_pv_drv.pdf

Спасибо за внимание!

Слайды и материалы курса доступны по адресу
http://is.gd/ivuboc

Замечание: все торговые марки и логотипы, использованные в данном материале, являются собственностью их владельцев. Представленная точка зрения отражает личное мнение автора.