Linux Virtualization Update

Chris Wright <chrisw@redhat.com>

Japan Linux Symposium, November 2007

redhat

Intro

- Virtualization mini-summit
- Paravirtualization
- Full virtualization
- Hardware changes
- Libvirt
- Xen

redhat

Virtualization Mini-summit

- June 25-27, 2007 – Just before OLS in Ottawa.
- 18 attendees
  - Xen, Vmware, KVM, lguest, UML, LinuxOnLinux
  - x86, ia64, PPC and S390
- Focused primarily on Linux as guest and areas of cooperation
  - paravirt_ops and virtio
- Common interfaces
  - Not the best group to design or discuss management interfaces
    - Defer to libvirt, CIM, etc...
  - CPUID 0x4000_00xx for hypervisor feature detection
  - Can we get to common ABI for paravirt hybrid guest?

redhat

Virtualization Mini-summit

- paravirt_ops
  - Make use of existing abstractions wherever possible (clocksource, clockevents or irqchip)
  - Could use a common lib/x86_emulate.c
  - Open question: performance benefit of shadow vs. direct paging?
- Distro Issues
  - Lack of feature parity between bare metal and Xen is difficult for distros
  - Single binary kernel image
  - Merge upstream
- Performance
  - NUMA awareness lacking in Xen – difficult for Altix
  - Static NUMA representation doesn't map well to dynamic virt environment
  - Cooperative memory management – guest memory hints

redhat

Virtualization Mini-summit

- Hardware
  - x86 and ia64 hardware virtualization roadmap
  - ppc virtualization is gaining in embedded market, realtime requirments
  - S390 “has an instruction for that”
- Virtio
  - Separate driver from transport
  - Makes driver small, looks like a Linux driver and reusable
  - Hypervisor specific transport
  - Open question: device discovery via PCI?
    - Doesn't work well for S390
    - The Linux Foundation is helping with possible PCI-SIG affiliation

redhat

Paravirt Guest

- paravirt_ops merged in Dec 2006 – 2.6.20-rc1
- VMI merged in Feb 2007 – 2.6.21-rc1
- lguest merged (UP only) in July 2007 – 2.6.23-rc1
- Xen merged (domU only) in July 2007 – 2.6.23-rc1
- virtio merged in Oct 2007 – 2.6.24-rc1
- paravirt_ops64 patches underway
- pv-on-hvm drivers (Linux and Windows)

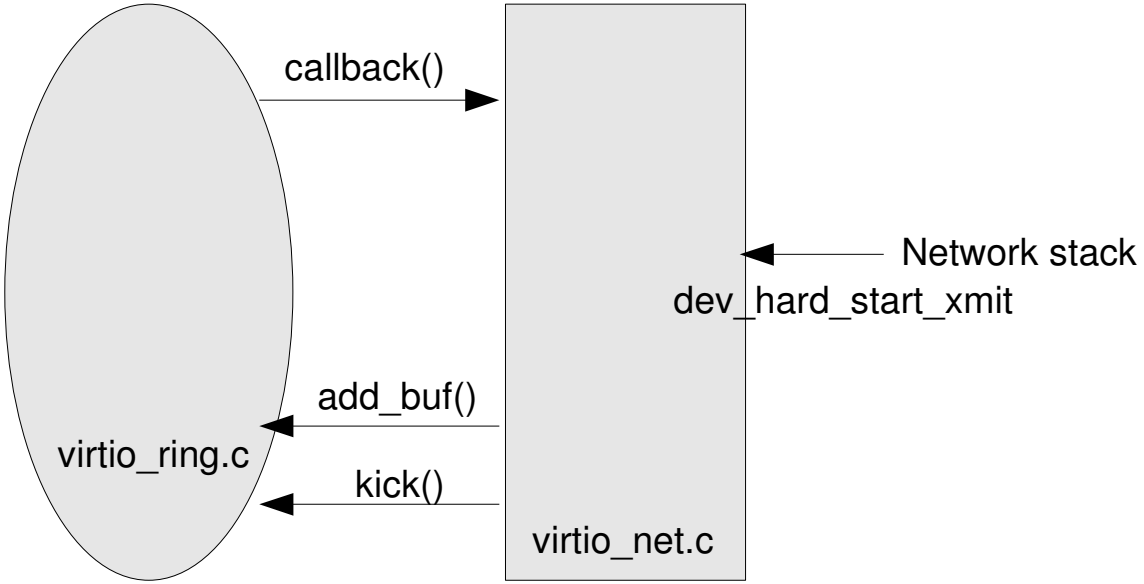
redhat

paravirt_ops

- 32-bit x86 only, 64-bit is a work in progress
- pv_info – random info, rather than function entrypoints
- pv_init_ops – functions used at boot time (some for module_init too)
- pv_time_ops – time-related functions
- pv_cpu_ops – various privileged instruction ops (includes lazy_mode)
- pv_irq_ops – operations for managing interrupt state
- pv_apic_ops – APIC operations
- pv_mmu_ops – operations for managing pagetables (includes lazy_mode)
- machine_ops – operations for controlling machine state
- smp_ops – operations for controlling multiprocessor machines
- Patches critical operations

virtio

- Separate driver core from setup and transport
- Includes lguest backend
- Drivers for block, net and console



redhat

Full Virt

- KVM
  - First posted in Oct 2007
  - Merged in Dec 2007 – 2.6.20-rc1

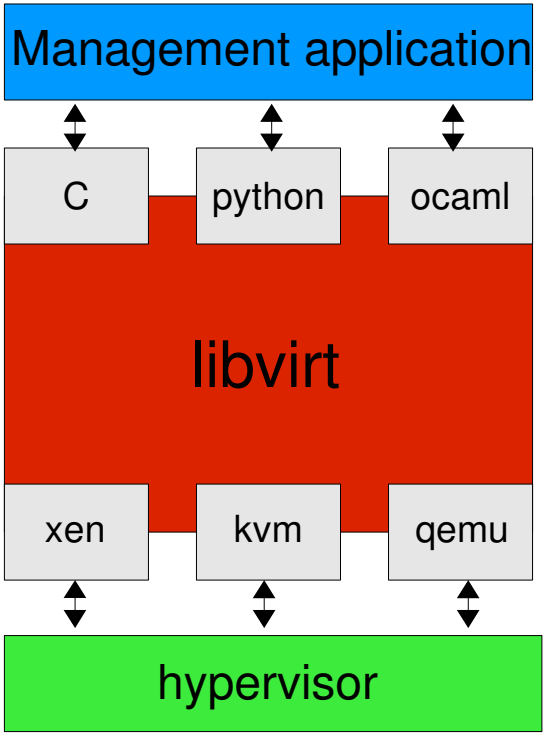
redhat

Hardware Virtualization

- ongoing reduction of VMEXIT costs – general performance improvement
- VMCALL cost approaches syscall cost
- vTPR – both 32 and 64-bit
- ASID, VPID – Tagged TLB
- NPT, EPT – Hardware solution to shadow page tables
- VT-d, IOMMU – safe and efficient direct access to i/o devices
- PCI-SIG IOV – device level virtual functions

libvirt

- Management API
  - create, destroy, start, stop, ...
- Stable Interface
- Hypervisor neutral
- Many language bindings



redhat

libvirt

- Current
  - Inactive domain management
  - KVM/qemu support
  - OpenVZ experimental driver
  - Secure remote management, with TLS + x509 certificates
  - Device hotplug for disks & NICs
  - VCPU pinning, NUMA, and scheduler param tuning
  - Virtual network management
  - Save / restore / migration
- Future
  - Storage management APIs for files, disks, LVM, & ISCSI
  - Host device enumeration
  - SASL authentication (in particular for Kerberos)
  - CIM providers based upon libvirt

redhat

Xen 3.0.3 – Oct 2006

- Credit scheduler
- xenoprofile
- Improved HVM support (usable)
- blktap
- Network segmentation offload
- Improved IA64 support
- Initial Power support
- 3.0.2 to 3.0.3
  - ~13.5% changes from Japanese developers

redhat

Xen 3.0.4 – Dec 2006

- kexec/kdump for Xen and dom0
- Paravirt framebuffer
- Introduction of new XenAPI
- More improvements for IA64
- Improvements for Power
- 3.0.3 to 3.0.4
  - ~15.8% changes from Japanese developers

redhat

Xen 3.1.0 – May 2007

- XenAPI 1.0 is official management API
  - XML config files for VMs
  - VM life-cycle management
  - Secure XML-RPC
  - Many language bindings
- HVM save/restore/migrate support
- HVM ballooning
- 32-on-64 PV
- 3.0.4 to 3.1.0
  - ~7% changes from Japanese developers

redhat

Xen Future

- Qemu update
- AMD ASID support
- vTPR, NMIp
- AMD NPT support
- ACPI S3 support
- XSM (supports ACM and Flask)
- Intel VT-d support
- AMD IOMMU support
- MTRR/PAT virtualization
- cpufreq
- qemu-dm pv support
- MSI support
- Intel TXT – Verified Launch
- Native client support
- Rebase to current Linux
- XenAPI Japanese translation
- pvSCSI
- 3.1.0 to current
  - >18% changes from Japanese developers

redhat

Future?

- 64-bit hypervisor
- Hardware that supports virtualization
  - CPU, platform, and I/O devices
- Paravirtualization continues to be useful
  - I/O, time, dynamic memory management
- Linux guest adapts to environment dynamically
- Hypervisor is everywhere
  - Embedded Server
    - VMware ESX3i, XenServer OEM Edition
    - Small footprint, pre-installed
  - Embedded Client
    - Intel vPro
    - Small footprint, pre-installed