KVM on POWER®
Status update & IO Architecture

Benjamin Herrenschmidt–
benh@au1.ibm.com
IBM Linux Technology Center

November 2012

Reminders

- 2 different virtualization methods
  - “HV” KVM
    - Exploits HW virtualization facilities
    - 970 in “non-Apple” mode, POWER, FSL
    - Best performances
    - Requires bare metal access
      - No underlying hypervisor
  - “PR” KVM
    - Runs the guest with user priviledges
    - Emulation of most priviledge instructions
    - Runs on almost everything
      - Issues when running under PowerVM
        - HV calls caught by pHyp
    - Slower
      - Perf improved with paravirt tricks

Reminders

- Different host processor families
  - "server" processors
    - 970, POWER7
    - HV mode is paravirt only
      - pSeries paravirt platform (sPAPR)
      - Same guest ABI as pHyp
    - "PR" KVM supported
      - Could do full virt
      - Full virt Mac emulation
        - Bit rotting
      - Paravirt pSeries support (sPAPR)
  - "embedded" processors
    - More oriented toward full virt
    - ePAPR defines paravirt interfaces
    - Very different from sPAPR
    - No sPAPR guets support

“Server” & sPAPR

- Kernel HV KVM upstream status
  - Base functionality, MMU virtualization
  - All CPU threads in guest, one thread per core in host
    - Must use cpu hotplug to “unplug” secondaries before using KVM
  - MMIO emulation
    - No coalescing yet
  - MMU notifiers
    - POWER7 only, not 970
  - Support for all HW supported page sizes
  - Dirty tracking of memory regions
    - Framebuffer
    - Pre-req for migration
      - Migration not upstream yet

“Server” & sPAPR

- Kernel KVM not upstream yet
  - In-kernel interrupt controller
    - XICS emulation
      - ICP presentation controllers in real mode
        - No exits for IPIs
        - No exits for MSIs in most cases
      - ICS source controllers in kernel virtual mode
  - Migration
    - Complete CPU State save / restore
      - Mostly upstream, not complete
      - MMU Hash table save/restore
        - Ioctl API being discussed
        - “Chunky” API
  - VFIO support
    - Core is upstream
    - Some powerpc bits not quite yet

“Server” & sPAPR

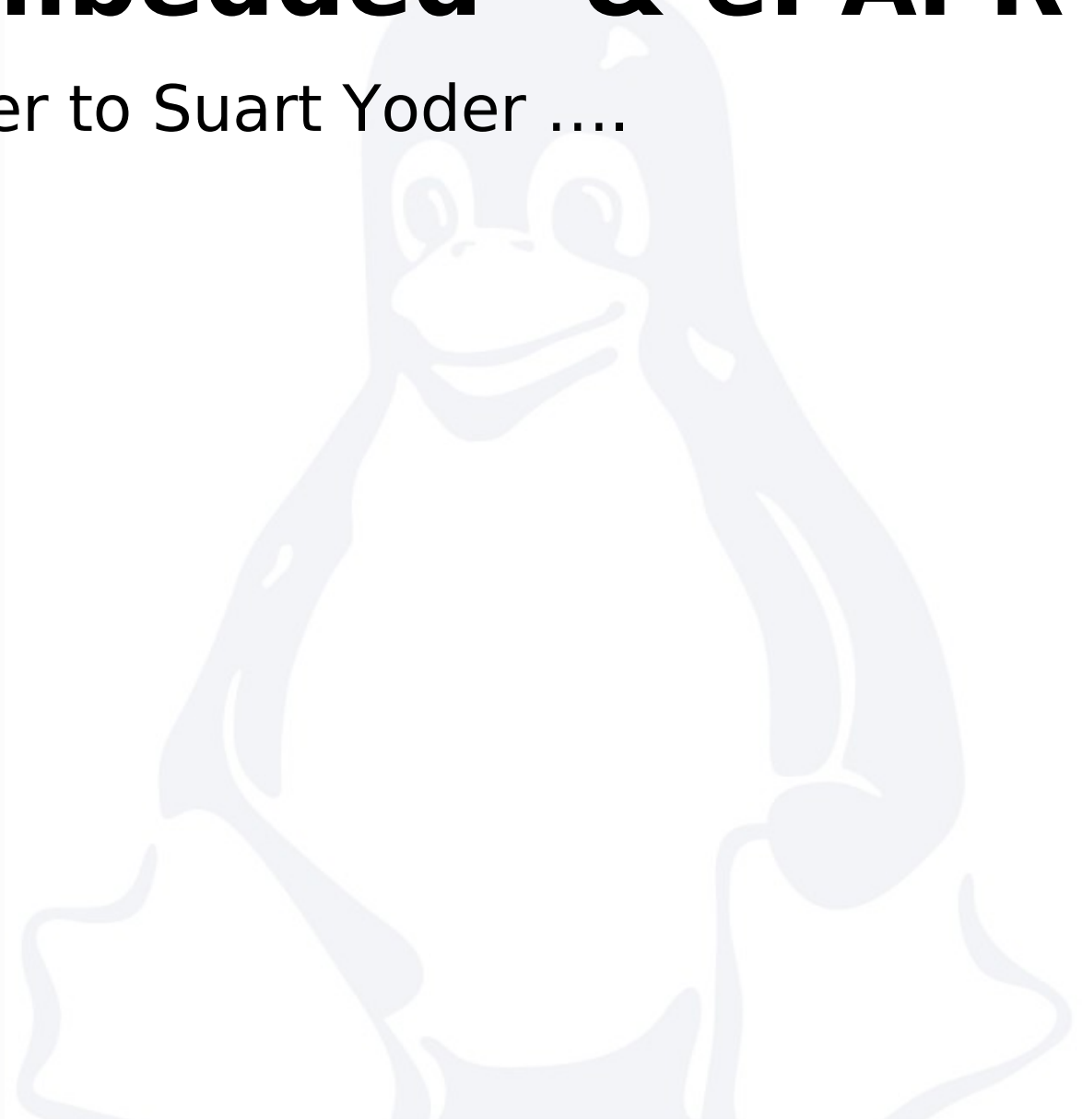
- Qemu upstream status
  - IOMMU infrastructure
  - sPAPR “VIO” (vscsi, veth, vterm)
    - Default for pSeries platform
  - PCI Emulation
    - Emulated devices
      - OHCI, e1000, ...
    - Virtio-pci
  - VGA supported
    - std-vga, cirrus is busted by design
  - SLOF FW boot device support
    - vscsi, veth
    - virtio-blk, virtio-net

“Server” & sPAPR

- Qemu not upstream yet
  - Virtio-scsi support in SLOF
    - Done, need review & merging
  - Nvram support
    - Basic support done
      - Including SLOF side
    - Still sorting out boot device selection
  - Migration
    - Working internally
    - Some state save/restore upstream
    - Waiting on finalizing kernel interfaces
      - MMU hash table save/restore
  - VFIO
    - Working, cleaning up
  - Some xHCI fixes

“Embedded” & ePAPR

- Hand over to Suart Yoder ....



PCI/PCIe IO Arch “IODA”

- Server IO architecture
- Error handling, recovery and isolation
  - Prevent propagation of bad data
- A “Partitionable Endpoint” ties device with state
  - Configuration space
  - MMIO/IO accesses
  - DMA
  - MSIs
- PE states
  - MMIO frozen
  - DMA frozen
- Any error triggers freeze
  - Defreeze under driver control (recovery)

IODA today

- Config accesses & PCIe errors
  - Matched on RID (bus/dev/fn)
- DMA
  - Matched on address
    - Segmented DMA space
  - Validation on RID
  - Device sees “windows” of DMA space
    - Remapped by TCEs
      - Translation Control Entries
      - Aka iommu
- MSIs
  - Matched on message address & value
  - Validation on RID

IODA today

- MMIO / IO (outbound)
  - Outbound “windows” from CPU to PCI bus
    - M32 (32-bit PCI MMIO)
    - IO (32-bit PCI IO)
    - M64 (64-bit PCI MMIO)
  - Windows are segmented
    - Equal size segments (up to 128 for M32)
    - Each segment associated with a PE
      - Dynamic association
        - N segments → same PE is possible
  - Resources for a device need to be grouped
    - Or aligned within exclusive segments
  - We use bridges
    - A PE is a set of busses
    - Issues with SR-IOV

Legal Statement

This work represents the view of the author and does not necessarily represent the view of IBM.

IBM, IBM (logo), AIX, POWER, POWER6, POWER7 and PowerVM are trademarks or registered trademarks of International Business Machines Corporation in the United States and/or other countries.

Linux is a registered trademark of Linus Torvalds.

Other company, product and service names may be trademarks or service marks of others.