IBM Deutschland Entwicklung GmbH

eHCA Virtualization on System p

Christoph Raisch
Technical Lead eHCA Infiniband and HEA device drivers

2008-04-04

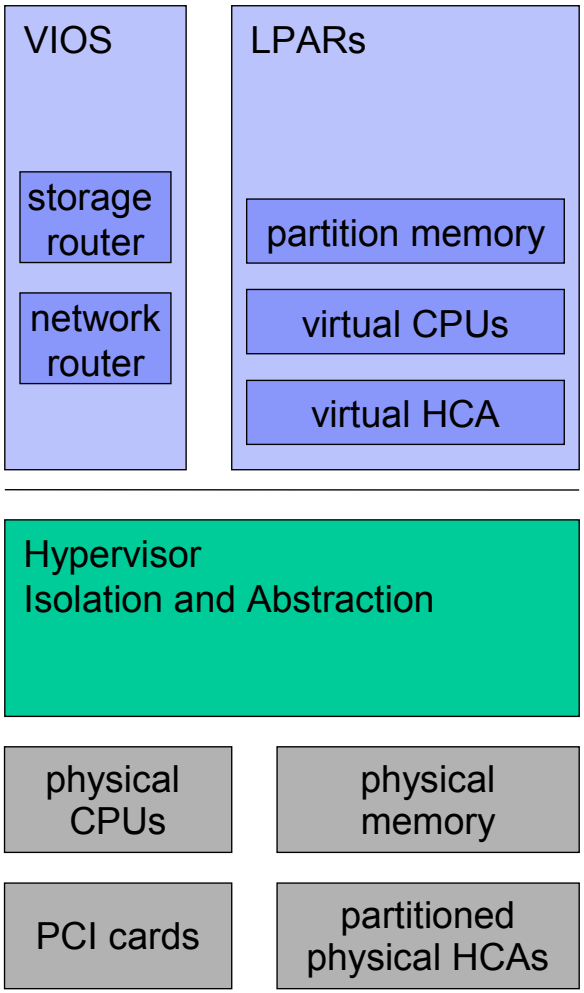
Trademark Statements

- **IBM, the IBM logo, ibm.com, System p, System p5 and POWER Hypervisor are registered trademarks of International Business Machines Corporation in the United States, other countries, or both.**
- **UNIX is a registered trademark of The Open Group in the United States and other countries.**
- **Linux is a registered trademark of Linus Torvalds in the United States, other countries, or both.**
- **Other company, product or service names may be trademarks or service marks of others.**

Overview

- **I/O Virtualization on System p**
- **Memory registration interface**
- **Send virtualization**
- **Receive virtualization**
- **QP 0/1 flow**
- **Additional eHCA capabilities**

I/O Virtualization on System p



architectured H_CALL Interface

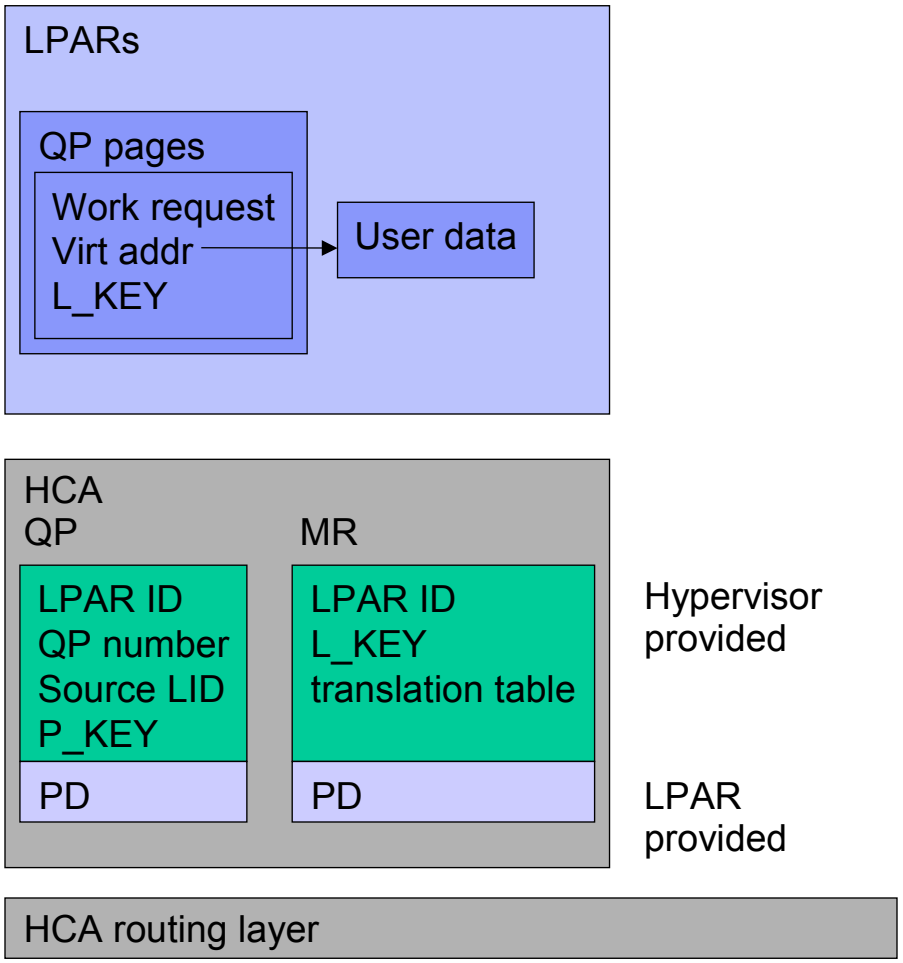
- **Paravirtualization**
  - Using H_CALLs (Hypervisor calls) instead of manipulating HW resources directly
- **Partitioning**
  - CPU
    Granularity: 1/10 of a CPU
  - Memory
    Granularity: 16MB blocks
- **I/O Hardware with partitioning support**
  - HCA, 16* InfiniBand
  - HEA (IVE), 32* Ethernet
- **I/O infrastructure for non-shareable PCI cards**
  - PCI slots are assigned to single partitions
  - PCI based ethernet and storage can be shared through VIOS

Registration interface for CQs/EQs/MRs/QPs

1. H_ALLOCATE_RESOURCE_MR (size, attributes,…)
   - Allocate a MR + LR_KEY for partition, allocate memory for MR page pointers, set partition number in MR
2. H_REGISTER_RPAGE(mr_pages)
   - Translate partition address to physical address
     add physical address to MR page tables
3. H_REGISTER_RPAGE(last_mr_page)
   - Enable MR

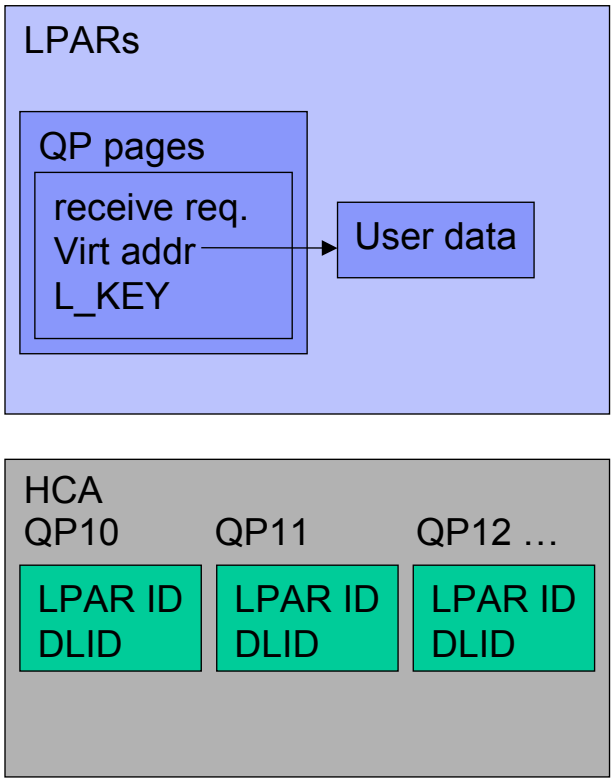
- **This allows to work with a single level of address translation**
- **Hypervisor has more information how pages are used**

eHCA send virtualization



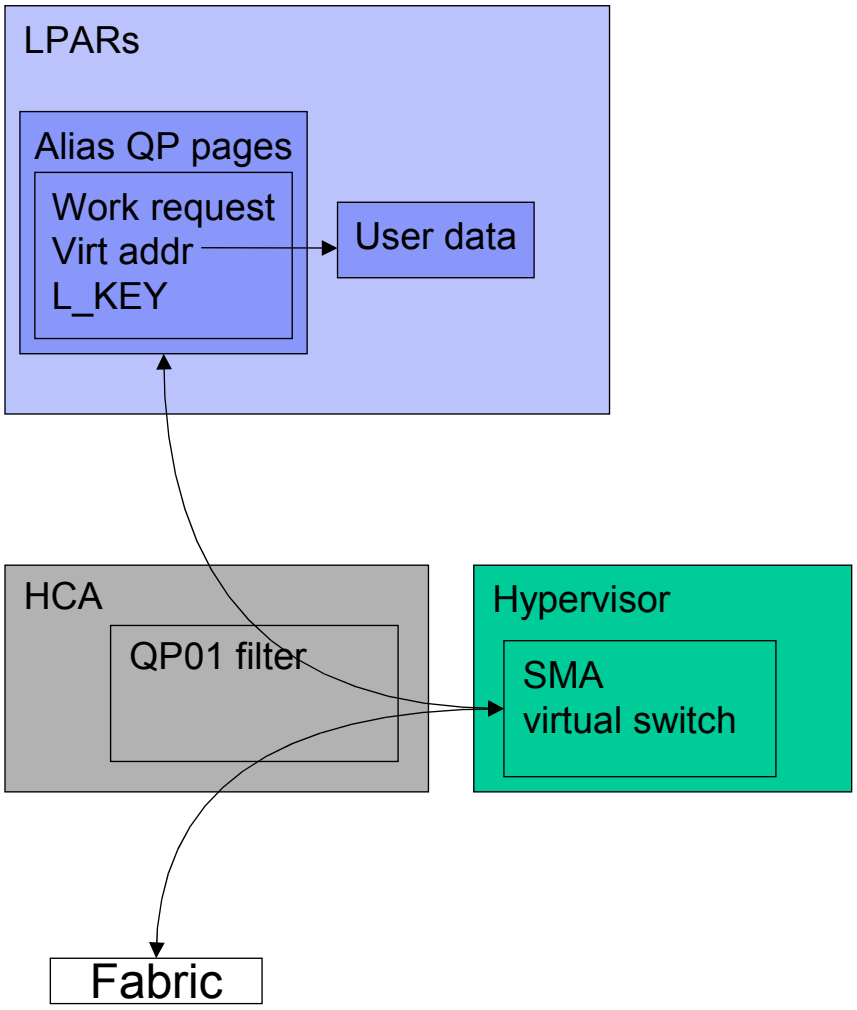
- LPAR triggers HCA
- HCA enforces hypervisor provided parameters
  - QP number
  - P_KEY
  - Source LID
  - QP type
- HCA uses L_KEY for MR, cross checks LPAR ID, PD
- HCA creates IB frames
- HCA routing layer identifies LIDs assigned to LPARs,
  IB frames either routed to external port or to receive processing

eHCA receive virtualization



- HCA routing layer identifies LIDs assigned to LPARs,
  IB frames either routed to receive processing or dropped
- HCA uses QP as primary selector
- HCA verifies destination LID
- HCA uses L_KEY for MR, cross checks LPAR ID, PD

QP 0/1 traffic



- **Inbound processing**
  - HCA identifies packets on QP 0/1
  - HCA forwards packets to Hypervisor
  - Hypervisor responds to QP0/1 requests or forwards them through “alias QP” to LPAR
- **Outbound processing**
  - “alias QP” sends packets
  - HCA redirects packets to Hypervisor if necessary
  - Hypervisor forwards packets through HCA to external fabric

eHCA capabilities

- **eHCA 1, GA in 2006**
  - 2k MTU
  - Low latency RC queues (user data in work request)
  - used in System p570 for PCI-E expansion network, 12x SDR mode
- **eHCA 2**
  - 4k MTU
  - Low latency UD/RC queues
  - Shared receive queue
  - System p Linux drivers available in OFED-1.3
  - currently available in System z for I/O expansion network and System z coupling, 12x DDR mode

eHCA summary

- **LPAR isolation designed in right from the start**
- **Protection against address spoofing**
  - no unfiltered packet stream access
  - IB Parameter enforcement on all send operations for LPARs
- **Each virtual HCA is visible on the subnet, has own GUID**
- **Paravirtualized driver interface**
  - Allows to change hardware while keeping the same device driver stack
- **Zero overhead for main path operations**
- **Full integration in OFED-1.1, 1.2, 1.2.5, 1.3**

For more Information

- **Implementing InfiniBand on IBM System p (including OFED)**
  http://www.redbooks.ibm.com/redbooks/pdfs/sg247351.pdf
- **IBM High Performance Computing for Linux on POWER Using InfiniBand**
  http://www-03.ibm.com/systems/p/software/whitepapers/hpc_linux.html
- **Getting Started with InfiniBand on System z10 and System z9**
  http://www.redbooks.ibm.com/Redbooks.nsf/RedpieceAbstracts/sg247539.html?Open

Questions?

Backup