

# Novell<sup>®</sup> Infiniband and XEN

XEN-IB project status

Patrick Mullaney

November 22, 2006



**Novell<sup>®</sup>**

# Infiniband and XEN

- Background
  - Client requirements:
    - > Guest OS access to Infiniband fabric
    - > Initial approach:
      - » L3 based solution using netfront/netback for IP
      - » Blkfront/Blkback based solution for storage (SRP initiator in domain0)
    - > Due to performance and scalability problems with the initial approach, a method for directly carrying out time critical data-path operations in the guest was required

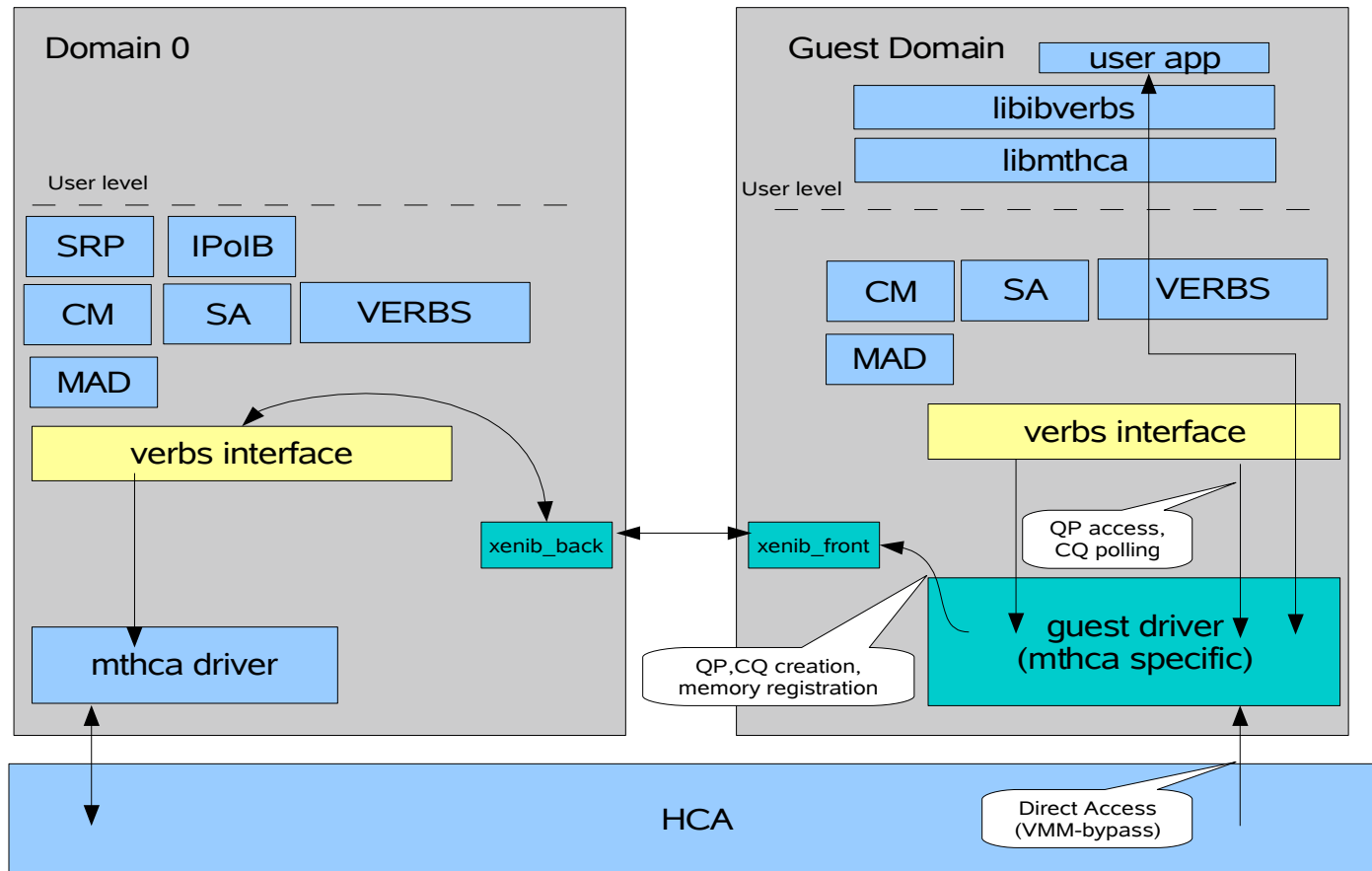
# Xensource: XEN-IB history

- XEN-IB tree at xensource is a proof-of-concept implementation of VMM-bypass
- VMM-bypass allows data-path operations to be done directly in the guest OS
- Impressive benchmark results for bandwidth and latency (on par with native OS)
- Work done at IBM Research and Ohio State
- Wiki: <http://wiki.xensource.com/xenwiki/XenSmartIO>
- VMM-bypass paper:  
<http://nowlab.cse.ohio-state.edu/publications/conf-papers/2006/huangwei-ics06.pdf>

# XEN-IB background

- Implementation
  - Split driver approach for control operations
    - > CQ, QP creation, memory registration, etc.
    - > event handling
  - Direct access operations (VMM-bypass)
    - > QP access
    - > CQ polling

# XEN-IB: block diagram



# Xensource: XEN-IB Status

- Problems and limitations
  - No support for guest access to IB management (SA, CM)
  - IPoIB and SRP unsupported
  - Limit of one virtual HCA per guest domain
  - IB diagnostic utilities non-functional in guest domain
  - Guest HCA driver is hardware dependent and currently only supports Mellanox MT23108 compatible adapters
  - Many guest HCA startup, shutdown and error path handling problems
  - Xensource wiki maintains a list of areas that need work.

# Our Efforts

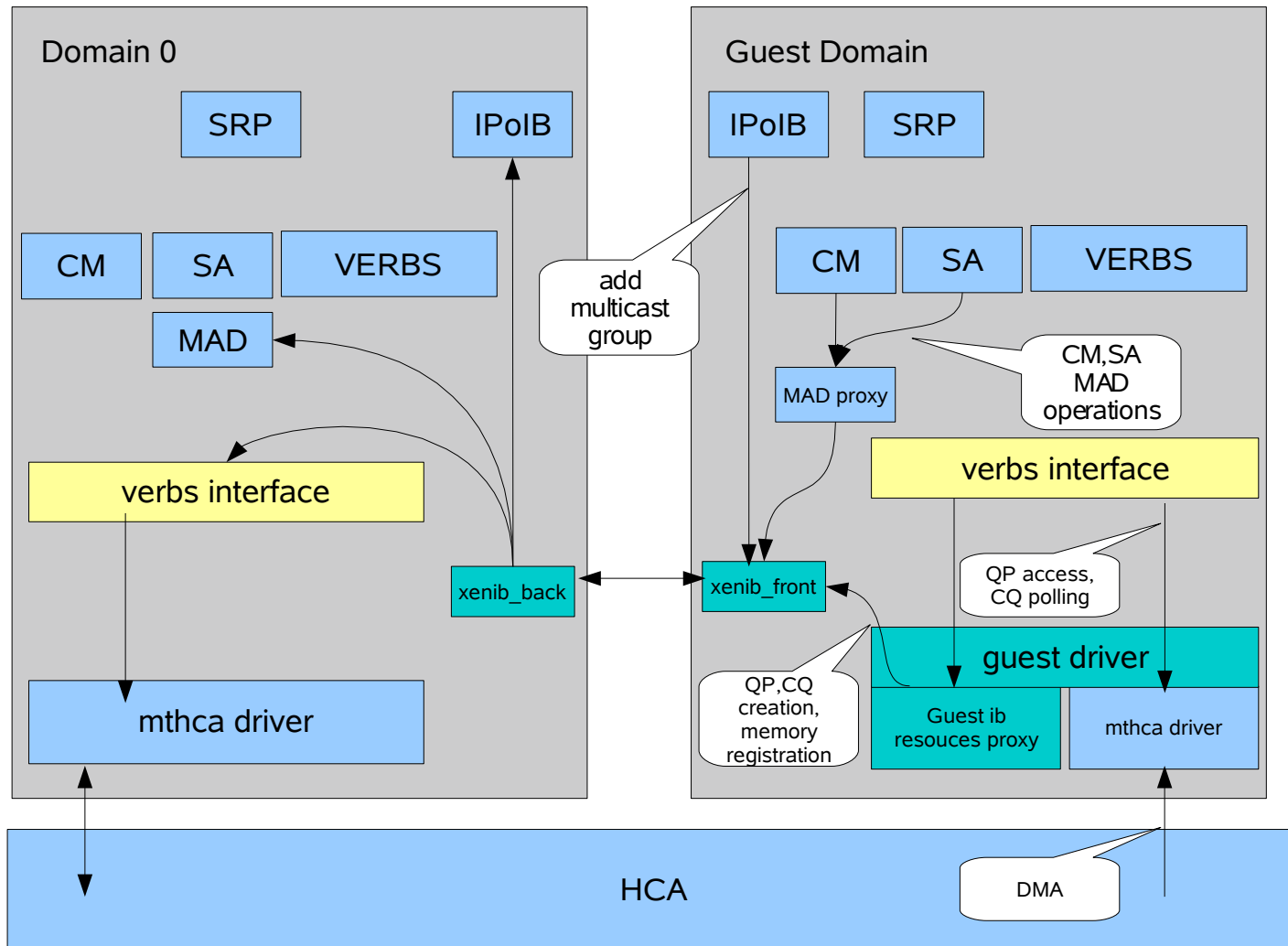
- Project status
  - Re-baselined code on OFED 1.1
  - Adding support for guest IB management access via an interdomain proxy mechanism for MADs
  - IPoIB working
    - > Added multicast group reference counting in domain 0 that is based off of Sean Hefty's multicast code
    - > SA path lookup via interdomain proxy
  - Reorganizing guest HCA driver to separate common code from hardware dependent parts
  - support all Mellanox adapters(currently support by mthca)
  - numerous problems with startup,cleanup,error handling fixed

# Our Efforts

- Project status cont.
  - Added kernel CONFIG options for selectively building dom0 or domU support
    - > Kernel without XEN-IB remains unchanged
  - IB diags will be functional in guest



# Our Efforts: Block Diagram



# Our Efforts

- Planned work
  - SRP initiator in guest
  - Enable XEN-IB in Windows guests
  - checkpointing/migration support

# New Problems

- Issues and Future work
  - Guest OS as the passive side of CM
    - > suffers from overlapping service id space
    - > solutions:
      - » partition service id space for guests
      - » infer guest by inspecting higher layers (example: private data field in SDP)
      - » other ideas?
  - Memory registration: domain0 ownership check of physical pages registered from guest

**Novell®**

## **Unpublished Work of Novell, Inc. All Rights Reserved.**

This work is an unpublished work and contains confidential, proprietary, and trade secret information of Novell, Inc. Access to this work is restricted to Novell employees who have a need to know to perform tasks within the scope of their assignments. No part of this work may be practiced, performed, copied, distributed, revised, modified, translated, abridged, condensed, expanded, collected, or adapted without the prior written consent of Novell, Inc. Any use or exploitation of this work without authorization could subject the perpetrator to criminal and civil liability.

## **General Disclaimer**

This document is not to be construed as a promise by any participating company to develop, deliver, or market a product. Novell, Inc., makes no representations or warranties with respect to the contents of this document, and specifically disclaims any express or implied warranties of merchantability or fitness for any particular purpose. Further, Novell, Inc., reserves the right to revise this document and to make changes to its content, at any time, without obligation to notify any person or entity of such revisions or changes. All Novell marks referenced in this presentation are trademarks or registered trademarks of Novell, Inc. in the United States and other countries. All third-party trademarks are the property of their respective owners.

