OPENFABRICS ALLIANCE

12th ANNUAL WORKSHOP 2016

MAINSTREAMING RDMA API CHANGES FOR REDHAT 7.2

Christoph Lameter

GenTwo

[April 7th, 2016]

INTRODUCTION

Overview

- Issue: Stable, Q&Aed and reliable system software for our growing IT infrastructure.
- Initially using Ubuntu 10.04 with OFED 1.5.4.
  - Instability due to OFED issues. Weird unusual build procedure.
  - Problems of patching custom kernel with OFED patches.
  - Bizarre and fragile build process with lots of custom patches.
  - Stuck for years on the same release.
- Decision to move to stable business class distribution -> Redhat
- Work with upstream, Mellanox and Redhat to get required functionality standardized and supported through Redhat. RH was working on a new major release.
- Still custom patches but aim to continual reduce the amount of custom patches in each release.
- Continually upgrading production environment following minor releases of Redhat. Dynamically changing and upgrading software.
- Issues get addressed these days and do not stay around for years.

PHASES OF GETTING THIS DONE

- Upstreaming (with Roland) 2009-2012
  - Ensure functionality is merged upstream that is needed
  - Stay in touch with RDMA subsystem maintainer
- Initial meetings with Redhat regarding 7.0 release plans
  - Agreement on fully functional RDMA subsystem for our needs.
  - Scoping out which patches were needed to get upstream up and functional.
- Redhat 7.0 beta
  - Feedback. Additional patches requested via Roland/Redhat.
- Redhat 7.0 release
  - Establish patches that are still needed.
  - Work with upstream on additional issues discovered. In particular Ambient Capabilities.
- Redhat 7.1
  - First time doing a minor update in production environment.
- Redhat 7.2
  - Enablement for 100G (ConnectX4). EDR and Ethernet. First 100G Fabric 1Q 2016
  - Ethernet deficiencies. Custom patches (~70 patches) to enable 100G Ethernet support in ConnectX4. Patches used are those scheduled for Rh 7.3

SECURITY: AMBIENT CAPABILITIES

- Raw ethernet access and access to scheduling parameters are privileged operations in Linux and require capabilities
- All our mission critical apps require these security capabilities.
- We run with patches kernels that enable these capabilities in general. But this is not acceptable in a regular distribution kernel.
- Capabilities subsystem exists but capabilities are not easily inheritable.
- Work with Linux security developers and maintainers to come up with a solution.
- Get other Algo Trading company developers involved to write patches to implement inheritable capabilities.
- Merged upstream in Linux 4.3. Waiting for pickup in Redhat 7.3

RAW ETHERNET QP

- Already in OFED 1.5.4
- Reworked using “Flows” in MOFED then upstream. In Redhat 7.0 but only last minute
- RH 7.1 changed the RAW QP flow API.
- RH 7.2 has no support for ConnectX4. Got RH kernel patches from Mellanox to implement this.
- RH 7.3 is planned support for ConnectX4.
- Sniffing mode upstream in Linux 4.6 and planned for RH 7.3
- Timestamp support
  - Various approaches exist
  - CQ format needs API agreement
  - Unstable clocks. Need more precise and synchronized clocks in order to be useful.
  - Can use other network devices to create timestamps.

INFINIBAND FEATURES

- IPoIB sendonly join does not fully join Multicast group.
  - Worked in OFED 1.5.4
  - Not upstream. Therefore not in RH 7.X
  - RH 7.3 planned to fix the issues.
  - Effect: Ethernet IB gateways do not forward multicast traffic for send only joined Multicast Groups.
  - Thus custom patches now on top of RH 7.2
- Multicast loopback prevention
  - Issue fixed with hacks to simply switch off loopback
  - This is a socket level option that has been available since the 70s.
  - Patches upstream but not userspace portions (libraries). Expect it to be available for RH 7.3
  - Keeping custom patches around for now.
- IPoIB offloads. We need kernel bypass for IPoIB traffic as well. Benefit from control of hardware timestamp support.
- Unidirectional multicast streams. IBTA spec change done. Expected later than RH 7.3.
- Timestamps

INFINIBAND SRP STORAGE TARGET SUPPORT

- Large storage devices that are submitted high volumes of data have issues when using the maximum scatter gather list size.
- Fixed in MOFED but not upstream. Some storage systems now have to run MOFED on top of Redhat. But here we can use an unpatched kernel.
- Expected to be addressed in Redhat 7.3.

API CHANGES IN THE LINUX KERNEL

API CHANGES IN THE LINUX KERNEL

- Communicate the use case. Why is this change needed and why can it not be done without API change?
- Develop a prototype patch. Often no answer to posting a simple use case.
- Prototype requires for a discussion basis. A variety of approaches may be explored over a couple of months.
- Consensus reached and maintainers step in to confirm that the approach is going to be accepted.
- Productionize patchset. Provide documentation etc etc.
- Patches gets merged upstream in Linux
- Request Redhat to merge upstream feature.
- Test results in lab.
- Deploy in production

OPENFABRICS ALLIANCE

12th ANNUAL WORKSHOP 2016

THANK YOU

Christoph Lameter

GenTwo

OFED

- Maintaining a code base usable for old distributions. But this is really a task that the individual enterprise distributions should undertake. Redhat and Canonical do this in the last years because RDMA has become important. OFED is brittle.
- Energy of OFA is directed to legacy support instead of working on improvement to canonical code base.
- EWG should get involved in upstream discussions and deal with conflicts etc.
- Project either linux kernel and similarly organized user space projects using git tree’s. Do not focus on OFED releases but push development in the respective upstream communities.

HISTORY OF THE RDMA SUBSYSTEM IN LINUX

- **2004-2015 Roland Dreier submits RDMA subsystem into the kernel**
- **2006-2014 Roland packages RDMA user space tools for Debian. Debian and Ubuntu allow use of RDMA without OFED. Slow emergence of RAW ETH QP support.**
- **2011 Roland is hired by Pure Storage to focus on storage.**
- **2014 Redhat prepares Redhat 7.0 with functional RDMA subsystem from upstream sources in cooperation with Roland. First time integrated working RDMA stack in Redhat. Rewrite of RAW ETH QP support by Mellanox.**
- **2015 Doug Ledford takes over maintenance of the RDMA subsystem.**
- **2016 Redhat releases RHEL 7.2 with initial 100G EDR support**
- **2016? Redhat working on RHEL 7.3 with mature 100G EDR/ Ethernet support.**

7 YEARS OF RDMA WORK

- 2009-2013 OFED/Ubuntu
  - Dark ages: Patches kept secret. Work with Ubuntu to get things into distro. Work with Mellanox restricted to firmware and driver updates. Work with Roland to ensure OFED code becomes upstreamed. Special requests to update packages in Debian and upstream but custom patches kept private.
- 2015 Redhat 7.0 preparation phase. Test MOFED and push of MOFED features into upstream/Redhat beta. Dark ages begin to end. Patches can be contributed. Identify ways to make hacks acceptable.
- 2016 Redhat 7.2 EDR support but we wanted 100G Ethernet support.

REDHAT HISTORY

- RH 5.X/6.X not used. Redhat put some work into RH 6.X
- 2014 7.0 first usable RDMA stack for us
- 2015 7.1 Reduction of patches
- 2016 7.2 Support for EDR IB
- 2016/17? 7.3 Support for 100G Ethernet