13th ANNUAL WORKSHOP 2017

DEPLOYING OFS TECHNOLOGY IN THE WILD
A CASE STUDY

Susan Coulter / HPC-Design

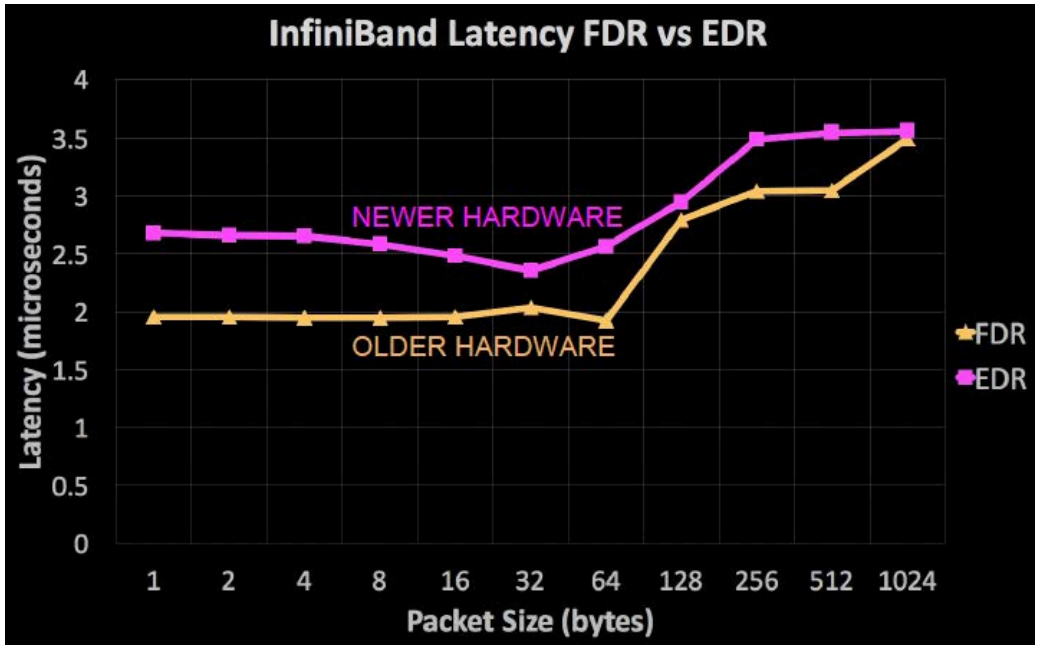
Los Alamos National Laboratory

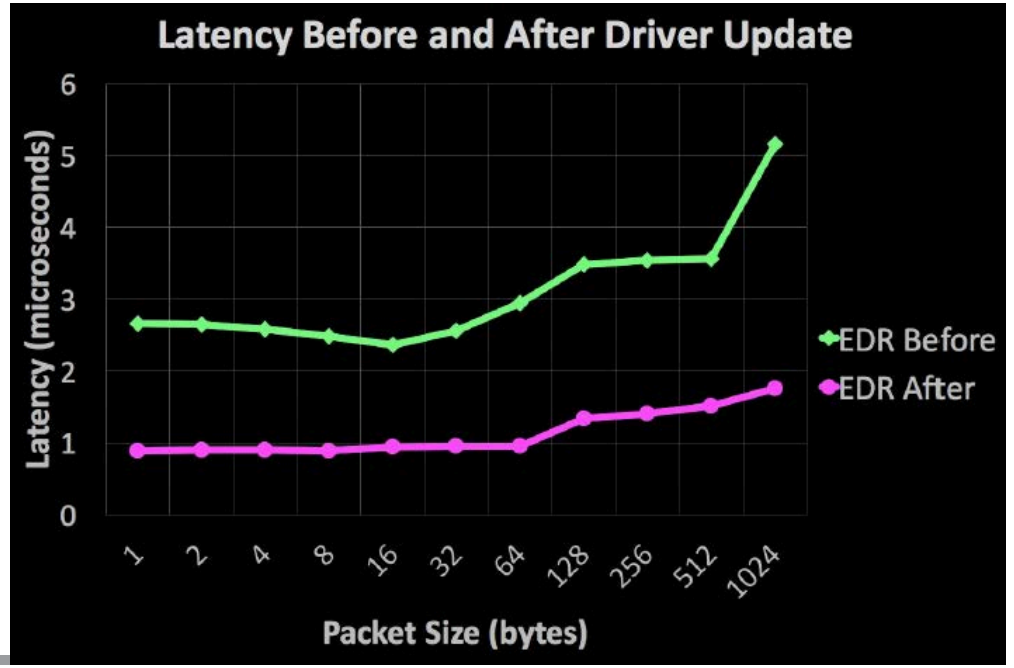
[March 31, 2017]

LA-UR-17-22449

HOW THE STORY STARTS…

- **LANL / CSCNSI**
  - Summer school for Junior/Senior Computer Science majors
  - Project: Compare 100G Ethernet to IB EDR
    - Cluster built with IB FDR
    - Preliminary test compared FDR to EDR

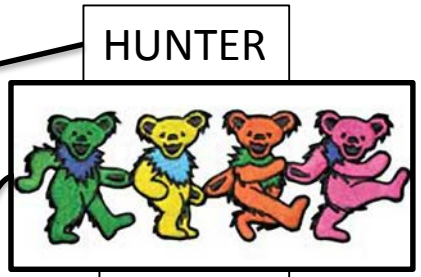


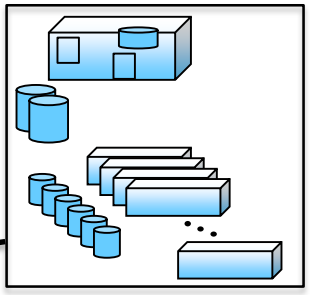


FIRST WRINKLE

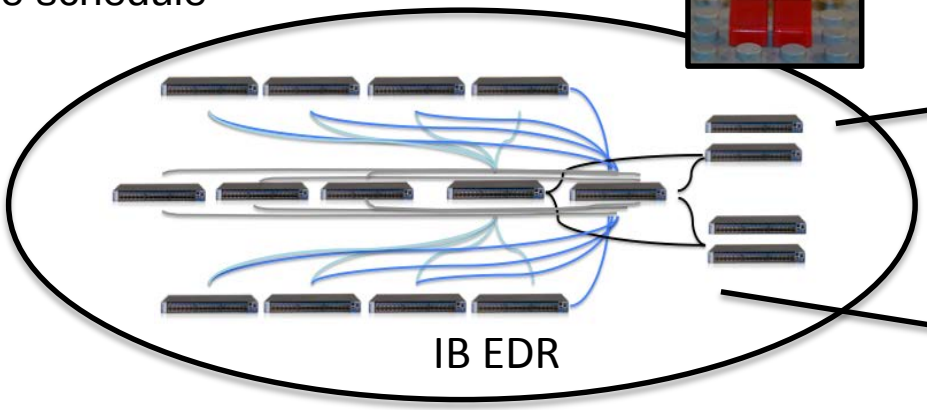
- **LANL deployed Damselfly IB backbone**
  - Only EDR systems in production
    - SM, slipknot cluster, redcap cluster
  - Most other systems FDR-connected
  - Built early with Mellanox-OFED
  - Replaced with TOSS(RedHat) bundled OFS
    - Tri-Lab Operating System Stack
      - TOSS2 -> RedHat6
      - TOSS3 -> RedHat7
    - LANL upgrade schedule slower than LLNL upgrade schedule
      - LANL running version(s) LLNL has frozen

Trinity Lustre

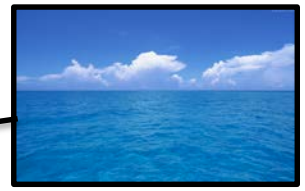




Common Lustre



IB EDR



Mid/Long Term Archive

HPSS
High Performance Storage System

WRINKLES WITHIN WRINKLES

- **Disk-ful / Disk-less / Configuration Management**
  - Install / test Mellanox OFED on TOSS standalone system – easy
    - Non-standard kernels use Mellanox script – easy
  - Cfengine controls cluster configuration
    - RPMs only – automation preferred except under extreme circumstances
      - Local updates repo (kernel RPMs and associated libraries)
      - Newer version number
      - depmod –a
        - » /etc/depmod.d/mlnx-ofa_kernel.conf
    - Hybrid images – RAM and NFS mount
      - Necessary kernel modules need to be in RAM
        - » rdma_cm requires configfs.ko

```
override ib_uverbs * weak-updates/mlnx-ofa_kernel/drivers/infiniband/core
override ib_addr * weak-updates/mlnx-ofa_kernel/drivers/infiniband/core
override ib_umad * weak-updates/mlnx-ofa_kernel/drivers/infiniband/core
override ib_core * weak-updates/mlnx-ofa_kernel/drivers/infiniband/core
```

SUCCESS !

- **Campaign / Scality system upgraded**
  - ~25% increase in performance
    - Uses >lots< of small messages
- **Wiki page developed**
  - Kernel upgrades due to security vulnerabilities not uncommon

SusanCoulter Settings Logout

NetworkTeam/ EdrOnKernel2.x

FrontPage » Monitoring/analytics » network_oncall » WeeklyOncallSummary » EdrOnKernel2.x

RecentChanges | FindPage | HelpContents | SCR_Team_Page | opacables-CTS | Omnipath | NetworkTeam | network_oncall | RpmSigning | Monitoring/analytics

Edit (Text) Edit (GUI) Info Subscribe Remove Link Attachments More Actions:

Updating EDR Driver and dependencies

Note: This also applies to TOSS3, 3.x kernels until further notice.

Contents

1. Updating EDR Driver and dependencies
   1. Get/Build new driver
   2. To handle non-standard (TOSS) kernels
   3. Install the new driver and associated kernel symbol files and headers
      1. Replace some critical libraries that interact with the kernel bits
      2. Insure the right modules are loaded
   4. Build new Lustre Network Driver (ko2iblnd.ko)
      1. Grab Lustre source RPM
      2. Modifications to the source RPM
      3. Build
      4. New whitelist requirements
   5. History capture of commands to complete the build
   6. Location of pre-built RPMS for TOSS2 and TOSS3
   7. Extra RPMs that were needed on CTS-1 Master Node to complete the build

Get/Build new driver

- Grab newest MOFED from mellanox.com -> Products -> Software -> InfiniBand/VPI Drivers
  /tmp/MLNX_OFED_LINUX-3.4-1.0.0.0-rhel6.8-x86_64.tgz
- Untar it
- cd into the directory
  /tmp/MLNX_OFED_LINUX-3.4-1.0.0.0-rhel6.8-x86_64

To handle non-standard (TOSS) kernels

- run ./mlnx_add_kernel_support.sh -k ***uname -r*** -m /tmp/MLNX_OFED_LINUX-3.4-1.0.0.0-rhel6.8-x86_64 --make-tg
- The process above creates a new tar file, untar this file
- cd into the directory created when the file is untar'd

```
/tmp/MLNX_OFED_LINUX-3.4-1.0.0.0-rhel6.8-x86_64-ext
```

Install the new driver and associated kernel symbol files and headers

- cd into the RPMS subdirectory and install the kernel bits

```
/tmp/MLNX_OFED_LINUX-3.4-1.0.0.0-rhel6.8-x86_64-ext/RPMS

rpm -ivh kmod-mlnx-ofa_kernel-3.4-OFED.3.4.1.0.0.1.g2ed8a21.rhel6u8.x86_64.rpm
         mlnx-ofa_kernel-3.4-OFED.3.4.1.0.0.1.g2ed8a21.rhel6u8.x86_64.rpm
         mlnx-ofa_kernel-devel-3.4-OFED.3.4.1.0.0.1.g2ed8a21.rhel6u8.x86_64.rpm
```

Replace some critical libraries that interact with the kernel bits

- Update critical packages with RPMS from that same directory
  - these conflict with existing packages, so manual intervention is necessary

LUSTRE COMPLICATION

- **Lustre**
  - Lustre comes from TOSS – relatively old version
    - Rebuild from source RPM
    - Manual modification of Module.symvers – bad idea
    - Grab source and rebuild

```
< --disable-doc --enable-panic_dumplog --with-ldiskfsprogs --with-o2ib=yes
> --disable-doc --enable-panic_dumplog --with-ldiskfsprogs --with-o2ib=/usr/src/ofa_kernel/default
```

```
Packager: Susan K Coulter <skc@lanl.gov>
```

```
< %{!?downstream_release: %global downstream_release "10chaos"}
> %{!?downstream_release: %global downstream_release "9chaos"}
```

```
Disable koji checks – LANL does not use koji
```

LUSTRE COMPLICATION - CONTINUED

- **Lustre**
  - Grab source and rebuild – no joy
  - Debug the failure, write a patch
  - Rebuild source RPM with patch

```
--- libcfs/include/libcfs/curproc.h.orig
+++ libcfs/include/libcfs/curproc.h
@@ -43,7 +43,7 @@
 #ifndef __LIBCFS_CURPROC_H__
 #define __LIBCFS_CURPROC_H__

-#if !defined(HAVE_UIDGID_HEADER) || !defined(__KERNEL__)
+#if (!defined(HAVE_UIDGID_HEADER) || !defined(__KERNEL__))
                                     && !defined(_LINUX_UIDGID_H)

 typedef uid_t kuid_t;
 typedef gid_t kgid_t;
```

```
--- lnet/klnds/o2iblnd/o2iblnd.c.orig
+++ lnet/klnds/o2iblnd/o2iblnd.c
@@ -728,7 +728,7 @@
        kib_dev_t           *dev;
        struct ib_qp_init_attr *init_qp_attr;
        struct kib_sched_info   *sched;
-#ifdef HAVE_IB_CQ_INIT_ATTR
+#if defined(FORWARD_PORT_FOR_MOFED3) && defined(HAVE_IB_CQ_INIT_ATTR)
        struct ib_cq_init_attr  cq_attr = {};
 #endif
        kib_conn_t          *conn;
@@ -826,7 +826,7 @@

        kiblnd_map_rx_descs(conn);

-#ifdef HAVE_IB_CQ_INIT_ATTR
+#if defined(FORWARD_PORT_FOR_MOFED3) && defined(HAVE_IB_CQ_INIT_ATTR)
        cq_attr.cqe = IBLND_CQ_ENTRIES(version);
        cq_attr.comp_vector = kiblnd_get_completion_vector(conn, cpt);
        cq = ib_create_cq(cmid->device,
```

LUSTRE / REDHAT 7.X COMPLICATION

- **Lustre v2.5 / RHEL 7.3**
  - Several Lustre file systems on single IB Storage Fabric
    - Multiple common Lustre file systems
    - Cray Sonexion Lustre file system – requires very high peer-credits
  - Older Lustre requires all IB parameters be identical
    - Tested Sonexion with peer-credits = 16
      - Estimated performance drop of 10-40%
  - Build new mlx5 drivers for TOSS3 – no dice

```
Dec 20 11:45:19 pippin kernel: mlx5_core 0000:04:00.0: firmware version: 12.16.1020
Dec 20 11:45:19 pippin kernel: BUG: sleeping function called from invalid context at mm/slub.c:941
Dec 20 11:45:19 pippin kernel: in_atomic(): 1, irqs_disabled(): 0, pid: 25152, name: modprobe
Dec 20 11:45:19 pippin kernel: CPU: 10 PID: 25152 Comm: modprobe Tainted: G   OE  -------   3.10.0-514.0.0.2chaos.ch6.x86_64 #1
Dec 20 11:45:19 pippin kernel: Hardware name: Dell Inc. PowerEdge R530/03XKDV, BIOS 1.2.6 06/08/2015
Dec 20 11:45:19 pippin kernel: ffff880fc7a80000 0000000093578ace ffff88103b3cb848 ffffffff8169c385
Dec 20 11:45:19 pippin kernel: ffff88103b3cb858 ffffffff810c0059 ffff88103b3cb8a0 ffffffff811e5d2a
Dec 20 11:45:19 pippin kernel: ffff88018fc07b00 ffffffffa07d9125 ffff880fc7a80000 ffff88103b3cb9f0
Dec 20 11:45:19 pippin kernel: Call Trace:
Dec 20 11:45:19 pippin kernel: [<ffffffff8169c385>] dump_stack+0x19/0x1b
Dec 20 11:45:19 pippin kernel: [<ffffffff810c0059>] __might_sleep+0xd9/0x100
Dec 20 11:45:19 pippin kernel: [<ffffffff811e5d2a>] kmem_cache_alloc_trace+0x4a/0x250
```

WRINKLES – THE SEQUEL

- **SGI cluster for OPA and ConnectX-5**
  - Build and test install of new driver – success
    - simple (RedHat 7.2)
  - On boot … no luck
  - /etc/depmod.d/zz01-mlnx-ofa_kernel.conf
  - extras vs weak-updates

```
override ib_uverbs * extras/mlnx-ofa_kernel/drivers/infiniband/core
override ib_addr * extras/mlnx-ofa_kernel/drivers/infiniband/core
override ib_umad * extras/mlnx-ofa_kernel/drivers/infiniband/core
override ib_core * extras/mlnx-ofa_kernel/drivers/infiniband/core
```

```
[Wed Mar  8 15:33:55 2017] hfi1: disagrees about version of symbol ib_umem_release
[Wed Mar  8 15:33:55 2017] hfi1: Unknown symbol ib_umem_release (err -22)
[Wed Mar  8 15:33:55 2017] hfi1: disagrees about version of symbol ib_modify_qp_is_ok
[Wed Mar  8 15:33:55 2017] hfi1: Unknown symbol ib_modify_qp_is_ok (err -22)
[Wed Mar  8 15:33:55 2017] hfi1: disagrees about version of symbol ib_unregister_device
[Wed Mar  8 15:33:55 2017] hfi1: Unknown symbol ib_unregister_device (err -22)
```

```
[Wed Mar  8 15:33:55 2017] Request for unknown module key 'Mellanox Technologies signing key:
                                                    61feb074fc7292f958419386ffdd9d5ca999e403' err -11
```

```
[Wed Mar  8 15:33:58 2017] ib_uverbs: Unknown symbol rdma_port_get_link_layer (err -22)
[Wed Mar  8 15:33:58 2017] ib_uverbs: disagrees about version of symbol ib_dealloc_pd
[Wed Mar  8 15:33:58 2017] ib_uverbs: Unknown symbol ib_dealloc_pd (err -22)
[Wed Mar  8 15:33:58 2017] ib_uverbs: disagrees about version of symbol ib_attach_mcast
[Wed Mar  8 15:33:58 2017] ib_uverbs: Unknown symbol ib_attach_mcast (err -22)
```

WRINLKLE – SEQUEL RESOLUTION !

```
[root@r02n02 3.10.0-327.el7.x86_64]# modinfo ib_uverbs
filename:       /lib/modules/3.10.0-327.el7.x86_64/updates/ib_uverbs.ko
license:        Dual BSD/GPL
description:    InfiniBand userspace verbs access
author:         Roland Dreier
rhelversion:    7.2
```

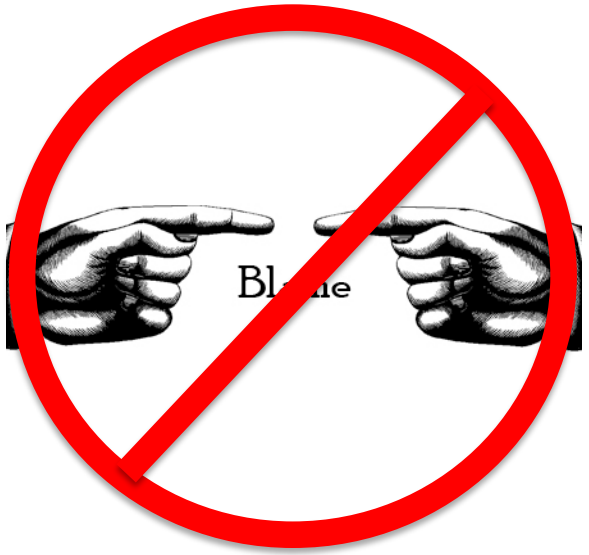
```
[root@r02n02 3.10.0-327.el7.x86_64]# rpm -qf /usr/lib/modules/3.10.0-327.el7.x86_64/updates/ib_uverbs.ko
ifs-kernel-updates-3.10.0_327.el7.x86_64-5.x86_64
```

```
[root@r02n02 3.10.0-327.el7.x86_64]# rpm -qi ifs-kernel-updates
Name        : ifs-kernel-updates
Version     : 3.10.0_327.el7.x86_64
Release     : 5
Architecture: x86_64

Build Host  : phbldprivrhel7-2.ph.intel.com
Relocations : (not relocatable)
Summary     : Extra kernel modules for IFS
Description :
Updated kernel modules for OPA IFS
```

I KNOW YOU ARE - BUT WHAT AM I

- **Not assigning blame**
  - Not criticizing LANL
    - Upgrade schedule
  - Not criticizing LLNL
    - RHEL / Lustre modifications
  - Not criticizing Mellanox
    - What is upstream and what is not
  - Not criticizing Intel
    - Special uverbs package



TAKEAWAYS

- **Complex, integrated systems are here to stay**
- **Managers**
  - Understand your environment
  - Understand the skill set of your people
  - Listen to and trust your technical people
- **Vendors**
  - Don’t assume single technology
  - Try to get a sense of the skill level of the customer
  - Don’t assume we can upgrade
- **Administrators**
  - Try to understand the underlying code
  - Learn the interdepencies of the kernel modules and libraries
  - Subscribe to linux-rdma mailing list
  - Build relationships in the community
- **Developers**
  - Try to put yourselves in administrators shoes
  - Reach out to administrators / deployments

To: Aleksey Senin @ Mellanox

OPENFABRICS ALLIANCE

13th ANNUAL WORKSHOP 2017

THANK YOU

Susan Coulter/ HPC-Design

Los Alamos National Laboratory