OPENFABRICS ALLIANCE

12th ANNUAL WORKSHOP 2016

INFINIBAND SELINUX SUPPORT

Dan Jurgens

Mellanox Technologies

[April 7th, 2016]

Mellanox® TECHNOLOGIES

LINUX SECURITY SUBSYSTEM

- **Linux has a modular security interface.**
  - Consists of hooks that are called from the rest of the kernel to enforce security policy
  - Provides default implementations that generally allow all access
- **SELinux and other security modules provide different hook implementations to enforce their own policy.**
- **Our Goal is to provide a security interface to control access to InfiniBand networks and enhance SELinux to enforce user defined policy.**

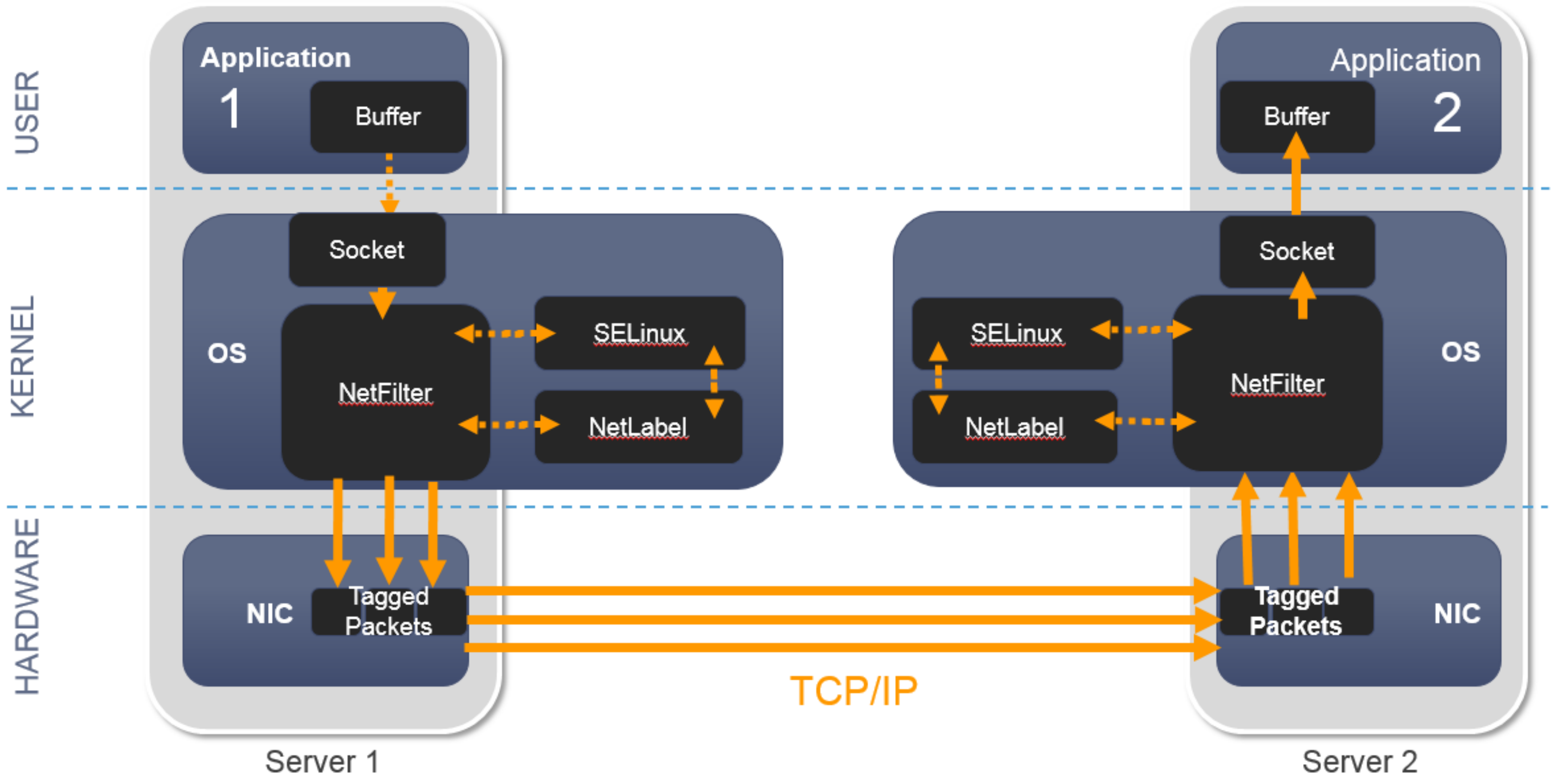
WHAT IS SELINUX?

- **SELinux is a Mandatory Access Control (MAC) scheme for Linux**
  - Central policy is loaded upfront into the kernel
    - Standard policies are typically provided by the Linux distribution
  - Applications cannot override or modify this policy

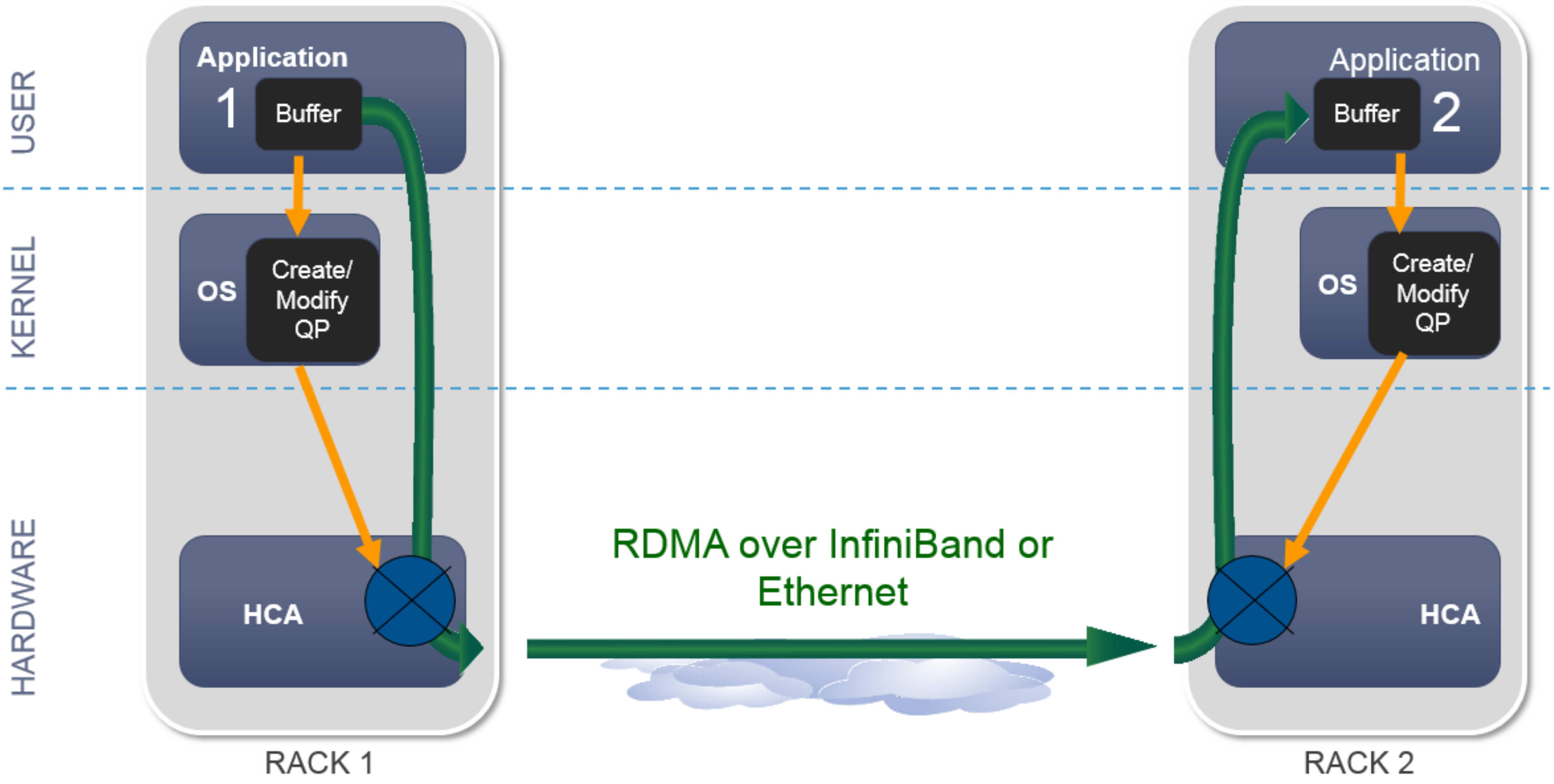
- **Benefits**
  - Differentiate a user from the applications that the user runs
  - Restrict application access only to what is required to perform its task
  - Allow granular policy segregation
  - Example
    - Run 2 instances of a Web Server: “top-secret” and “standard”
    - Each server can only
      - Receive traffic from specific network interfaces
      - Open sockets on specific ports
      - Serve files from specific directories
      - Communicate only with specific peer addresses

- **Type enforcement is the main security mechanism used by SELinux**

SELINUX IP/ETH NETWORKING

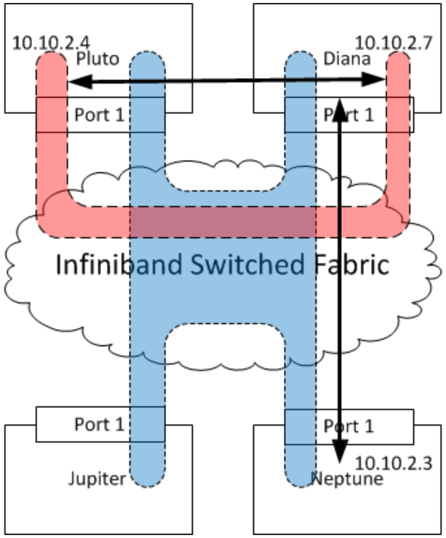


INFINIBAND RDMA FLOW



PARTITIONS AND QPS

- **Partition – Connects a subset of end nodes to a virtual fabric.**
  - Partition configuration of the nodes is managed by the Subnet Manager (SM) via the Subnet Management Interface (SMI).
  - Ports may be members of multiple partitions at once
  - PKey – partition label (a field in the BTH header)
  - Partitioning is enforced in the hardware.



- **Queue Pairs – The basic means of communication in InfiniBand**
  - Bound to a single partition prior to communication.

SELINUX AND KERNEL CHANGES

- **Several new LSM hooks are needed to enforce security for InfiniBand.**
  - Allocating security contexts. Security context are an opaque structure, they are stored by the QP and MAD agents and provided to the other security hooks for access verification.
  - Freeing security contexts.
  - Checking for SMI access. Take a device name, port number, and security context and verify the caller has permission to use the SMI.
  - Checking for PKey access. Takes a subnet prefix, PKey, and security context to verify the caller can access that PKey.
  - Registering and freeing a callback to be notified about security policy and enforcement changes.

INFINIBAND DRIVER CHANGES

- **Control access to the SMI**
  - Prevents unauthorized modifications to virtual fabric topology.
  - Enforced during MAD agent registration, only authorized users can create an SMI MAD agent.

- **Control QP access to Partitions**
  - Only allow users authorized access to a partition to connect a QP on that partition.
  - When a QP is created it inherits the security ID of the process creating it.
  - Whenever the QP is modified with changes to PKey index, port, or alternate path a check is made to verify it has access for the new configuration.
    - If the QP is a shared QP all open handles must have permission for the new settings.
  - Maintain lists of which QPs are using each PKey index on a port. If the PKey table or GID changes walk the list and check that each QP has permission.
    - If not move the QP to error and raise a QP fatal event.

INFINIBAND DRIVER CHANGES

- **Implementation is not hardware dependent.**
  - Security is enforced in the ib_core, ib_mad, and ib_umad kernel modules.

- **Access control is in the control path.**
  - Users retain the normal performance characteristics of their InfiniBand fabric.

SAMPLE POLICY LABELING SYNTAX

```
attribute pkey_type;
type pkey_t, pkey_type;
sid pkey gen_context(system_u:object_r:pkey_t,s0)

type staff_allowed_pkey_t, pkey_type;
type admin_allowed_pkey_t, pkey_type;
type default_pkey_t, pkey_type;

pkeycon fe80:0:0:0:: 0xffff gen_context(system_u:object_r:default_pkey_t,s0)
pkeycon fe80:: 0x8001 gen_context(system_u:object_r:staff_allowed_pkey_t,s0)
pkeycon fe80:: 0x8002 gen_context(system_u:object_r:admin_allowed_pkey_t,s0)

attribute ibdev_type;
type admin_ibdev_t, ibdev_type;
type staff_ibdev_t, ibdev_type;

type ibdev_t, ibdev_type;
sid ibdev gen_context(system_u:object_r:ibdev_t,s0)

ibdevcon mlx4_0 1 gen_context(system_u:object_r:admin_ibdev_t,s0)
```

SAMPLE POLICY ALLOW SYNTAX

```
allow sysadm_t default_pkey_t:infiniband_pkey access;
allow sysadm_t admin_allowed_pkey_t:infiniband_pkey access;

allow staff_t default_pkey_t:infiniband_pkey access;
allow staff_t staff_allowed_pkey_t:infiniband_pkey access;

allow sysadm_t admin_ibdev_t:infiniband_device smi;
allow staff_t staff_ibdev_t:infiniband_device smi;
```

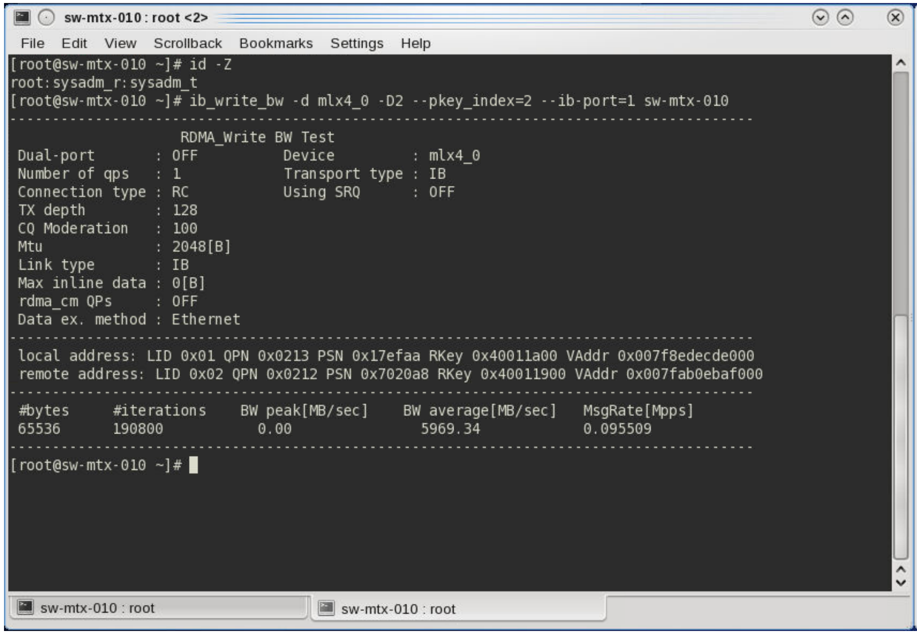
DEMO CONFIGURATION

- **Two roles**
  - Staff_r
  - Admin_r

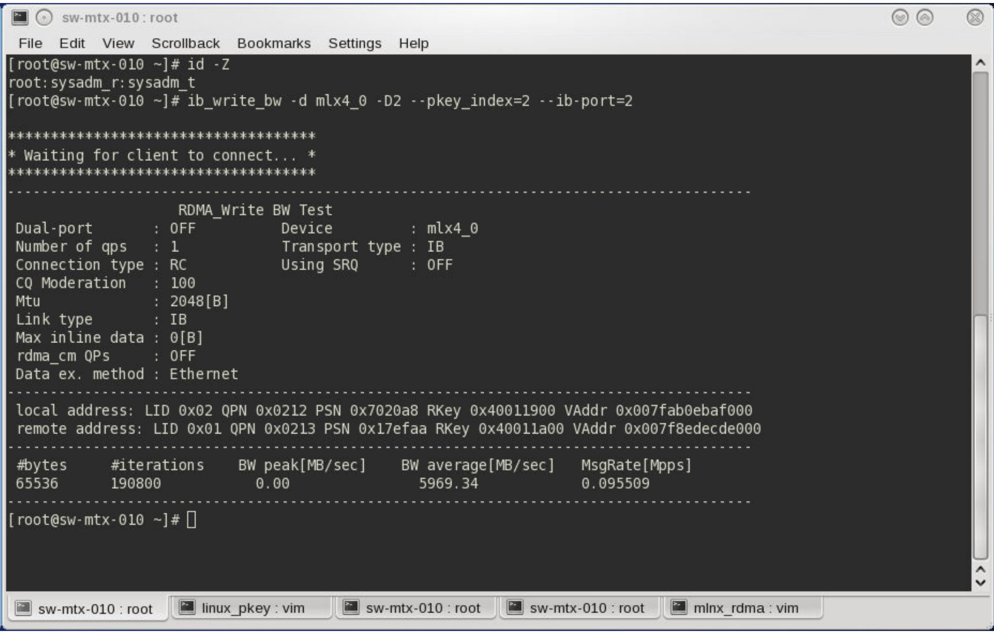
- **Four available partitions**
  - Default (0xFFFF) – both allowed
  - Staff allowed (0x8001)
  - Admin allowed (0x8002)
  - Neither allowed (0x8003)

- **SMI**
  - Admin allowed on mlx4_0 port 1.

ADMIN TO ADMIN ON ADMIN PARTITION

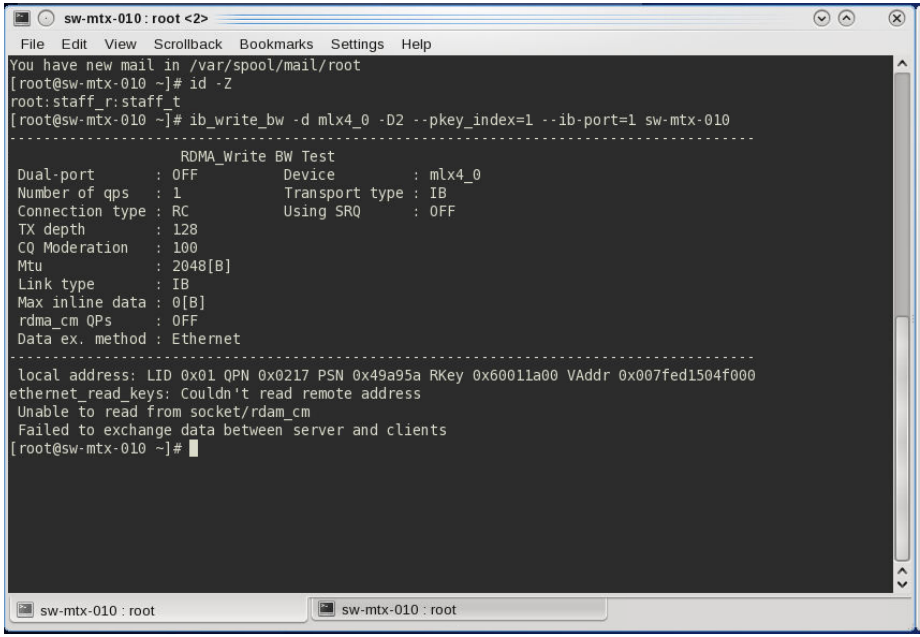


```
[root@sw-mtx-010 ~]# id -Z
root:sysadm_r:sysadm_t
[root@sw-mtx-010 ~]# ib_write_bw -d mlx4_0 -D2 --pkey_index=2 --ib-port=1 sw-mtx-010
---------------------------------------------------------------------------------------
                    RDMA_Write BW Test
 Dual-port       : OFF          Device         : mlx4_0
 Number of qps   : 1            Transport type : IB
 Connection type : RC           Using SRQ      : OFF
 TX depth        : 128
 CQ Moderation   : 100
 Mtu             : 2048[B]
 Link type       : IB
 Max inline data : 0[B]
 rdma_cm QPs     : OFF
 Data ex. method : Ethernet
---------------------------------------------------------------------------------------
 local address: LID 0x01 QPN 0x0213 PSN 0x17efaa RKey 0x40011a00 VAddr 0x007f8edecde000
 remote address: LID 0x02 QPN 0x0212 PSN 0x7020a8 RKey 0x40011900 VAddr 0x007fab0ebaf000
---------------------------------------------------------------------------------------
 #bytes     #iterations    BW peak[MB/sec]    BW average[MB/sec]   MsgRate[Mpps]
 65536      190800           0.00               5969.34            0.095509
---------------------------------------------------------------------------------------
[root@sw-mtx-010 ~]# █
```

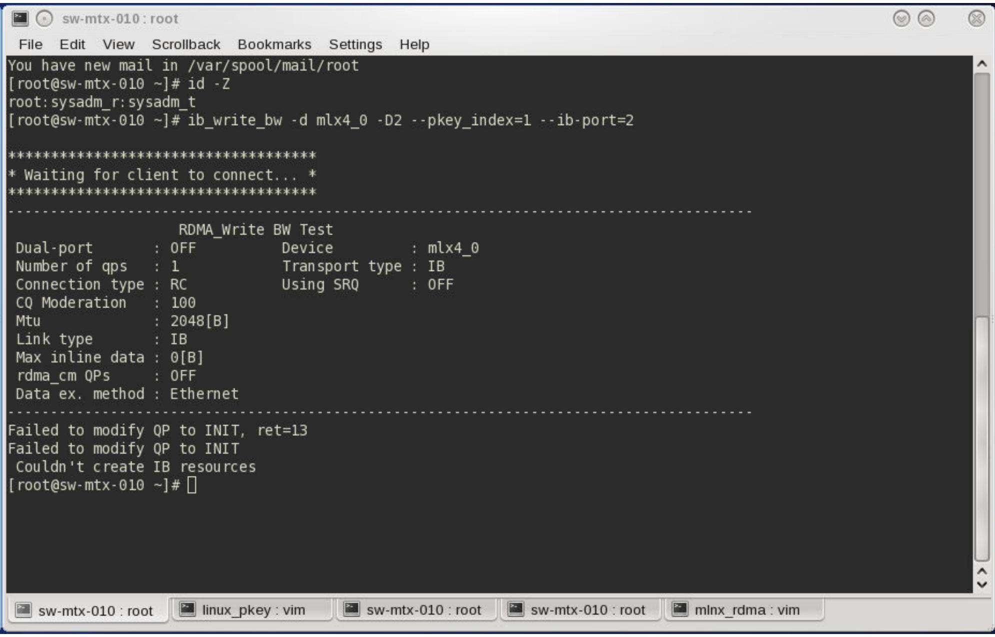


```
[root@sw-mtx-010 ~]# id -Z
root:sysadm_r:sysadm_t
[root@sw-mtx-010 ~]# ib_write_bw -d mlx4_0 -D2 --pkey_index=2 --ib-port=2

************************************
* Waiting for client to connect... *
************************************
---------------------------------------------------------------------------------------
                    RDMA_Write BW Test
 Dual-port       : OFF          Device         : mlx4_0
 Number of qps   : 1            Transport type : IB
 Connection type : RC           Using SRQ      : OFF
 CQ Moderation   : 100
 Mtu             : 2048[B]
 Link type       : IB
 Max inline data : 0[B]
 rdma_cm QPs     : OFF
 Data ex. method : Ethernet
---------------------------------------------------------------------------------------
 local address: LID 0x02 QPN 0x0212 PSN 0x7020a8 RKey 0x40011900 VAddr 0x007fab0ebaf000
 remote address: LID 0x01 QPN 0x0213 PSN 0x17efaa RKey 0x40011a00 VAddr 0x007f8edecde000
---------------------------------------------------------------------------------------
 #bytes     #iterations    BW peak[MB/sec]    BW average[MB/sec]   MsgRate[Mpps]
 65536      190800           0.00               5969.34            0.095509
---------------------------------------------------------------------------------------
[root@sw-mtx-010 ~]# □
```

ADMIN TO STAFF ON STAFF PARTITION



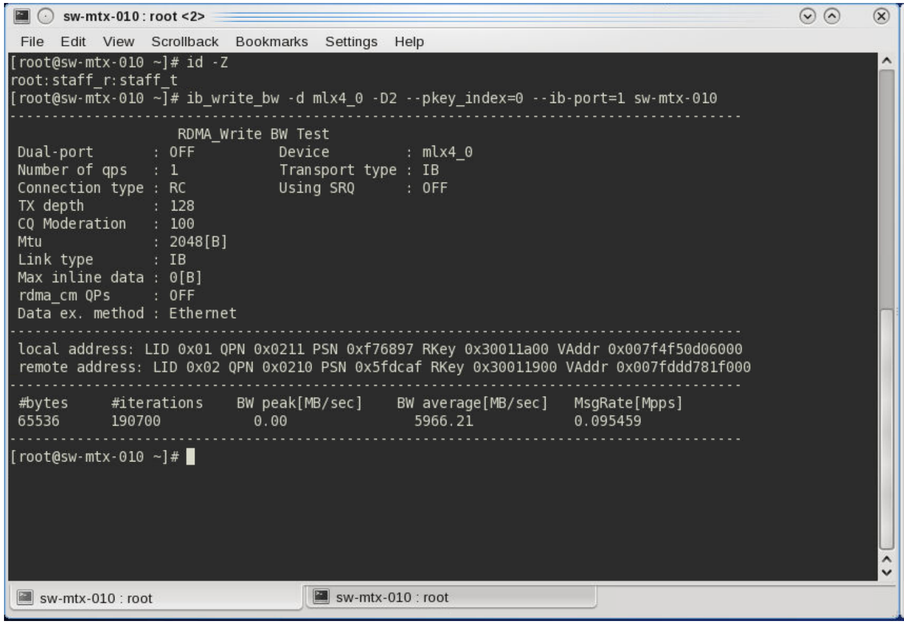
```
You have new mail in /var/spool/mail/root
[root@sw-mtx-010 ~]# id -Z
root:staff_r:staff_t
[root@sw-mtx-010 ~]# ib_write_bw -d mlx4_0 -D2 --pkey_index=1 --ib-port=1 sw-mtx-010
---------------------------------------------------------------------------------------
                    RDMA_Write BW Test
 Dual-port       : OFF          Device         : mlx4_0
 Number of qps   : 1            Transport type : IB
 Connection type : RC           Using SRQ      : OFF
 TX depth        : 128
 CQ Moderation   : 100
 Mtu             : 2048[B]
 Link type       : IB
 Max inline data : 0[B]
 rdma_cm QPs     : OFF
 Data ex. method : Ethernet
---------------------------------------------------------------------------------------
 local address: LID 0x01 QPN 0x0217 PSN 0x49a95a RKey 0x60011a00 VAddr 0x007fed1504f000
ethernet_read_keys: Couldn't read remote address
 Unable to read from socket/rdam_cm
 Failed to exchange data between server and clients
[root@sw-mtx-010 ~]# 
```



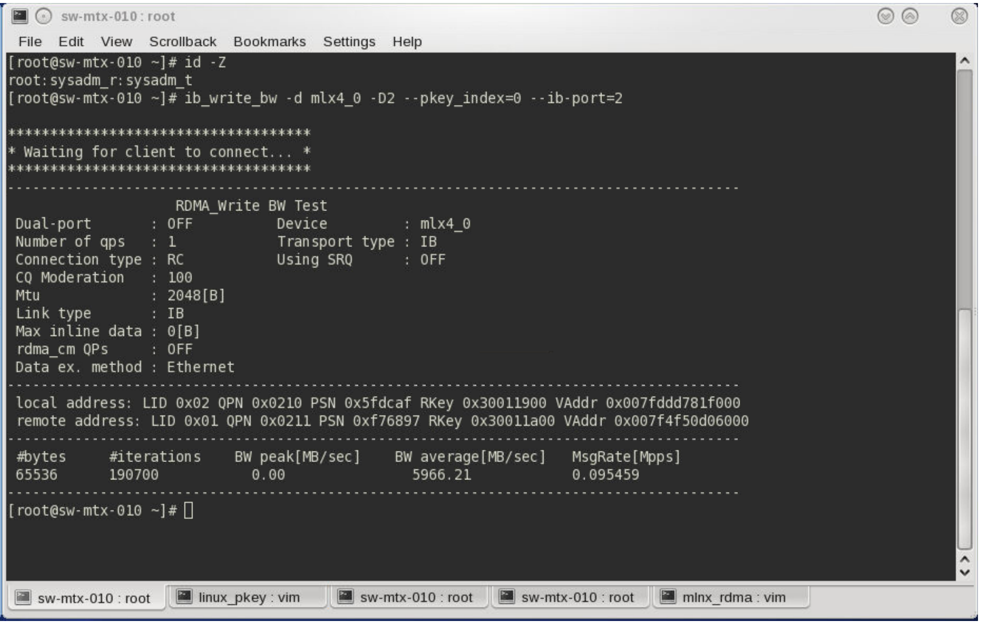
```
You have new mail in /var/spool/mail/root
[root@sw-mtx-010 ~]# id -Z
root:sysadm_r:sysadm_t
[root@sw-mtx-010 ~]# ib_write_bw -d mlx4_0 -D2 --pkey_index=1 --ib-port=2

************************************
* Waiting for client to connect... *
************************************
---------------------------------------------------------------------------------------
                    RDMA_Write BW Test
 Dual-port       : OFF          Device         : mlx4_0
 Number of qps   : 1            Transport type : IB
 Connection type : RC           Using SRQ      : OFF
 CQ Moderation   : 100
 Mtu             : 2048[B]
 Link type       : IB
 Max inline data : 0[B]
 rdma_cm QPs     : OFF
 Data ex. method : Ethernet
---------------------------------------------------------------------------------------
Failed to modify QP to INIT, ret=13
Failed to modify QP to INIT
 Couldn't create IB resources
[root@sw-mtx-010 ~]# 
```

Note in this case only the Admin side encounters an EACCESS error. The Staff side just has an error connecting.

ADMIN TO STAFF ON THE DEFAULT PARTITION

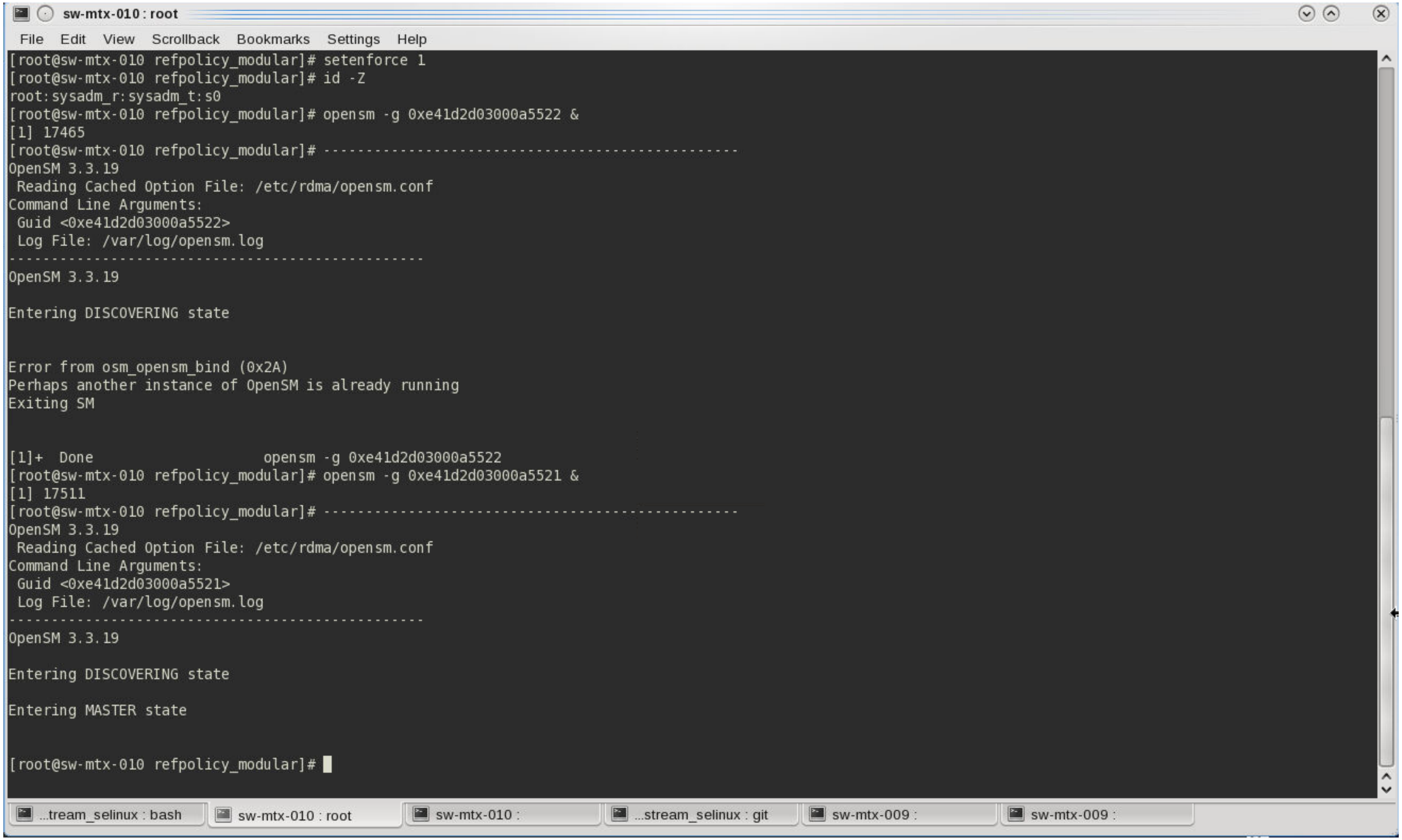


```
[root@sw-mtx-010 ~]# id -Z
root:staff_r:staff_t
[root@sw-mtx-010 ~]# ib_write_bw -d mlx4_0 -D2 --pkey_index=0 --ib-port=1 sw-mtx-010
---------------------------------------------------------------------------------------
                    RDMA_Write BW Test
 Dual-port       : OFF          Device         : mlx4_0
 Number of qps   : 1            Transport type : IB
 Connection type : RC           Using SRQ      : OFF
 TX depth        : 128
 CQ Moderation   : 100
 Mtu             : 2048[B]
 Link type       : IB
 Max inline data : 0[B]
 rdma_cm QPs     : OFF
 Data ex. method : Ethernet
---------------------------------------------------------------------------------------
 local address: LID 0x01 QPN 0x0211 PSN 0xf76897 RKey 0x30011a00 VAddr 0x007f4f50d06000
 remote address: LID 0x02 QPN 0x0210 PSN 0x5fdcaf RKey 0x30011900 VAddr 0x007fddd781f000
---------------------------------------------------------------------------------------
 #bytes     #iterations    BW peak[MB/sec]    BW average[MB/sec]   MsgRate[Mpps]
 65536      190700           0.00               5966.21             0.095459
---------------------------------------------------------------------------------------
[root@sw-mtx-010 ~]# 
```



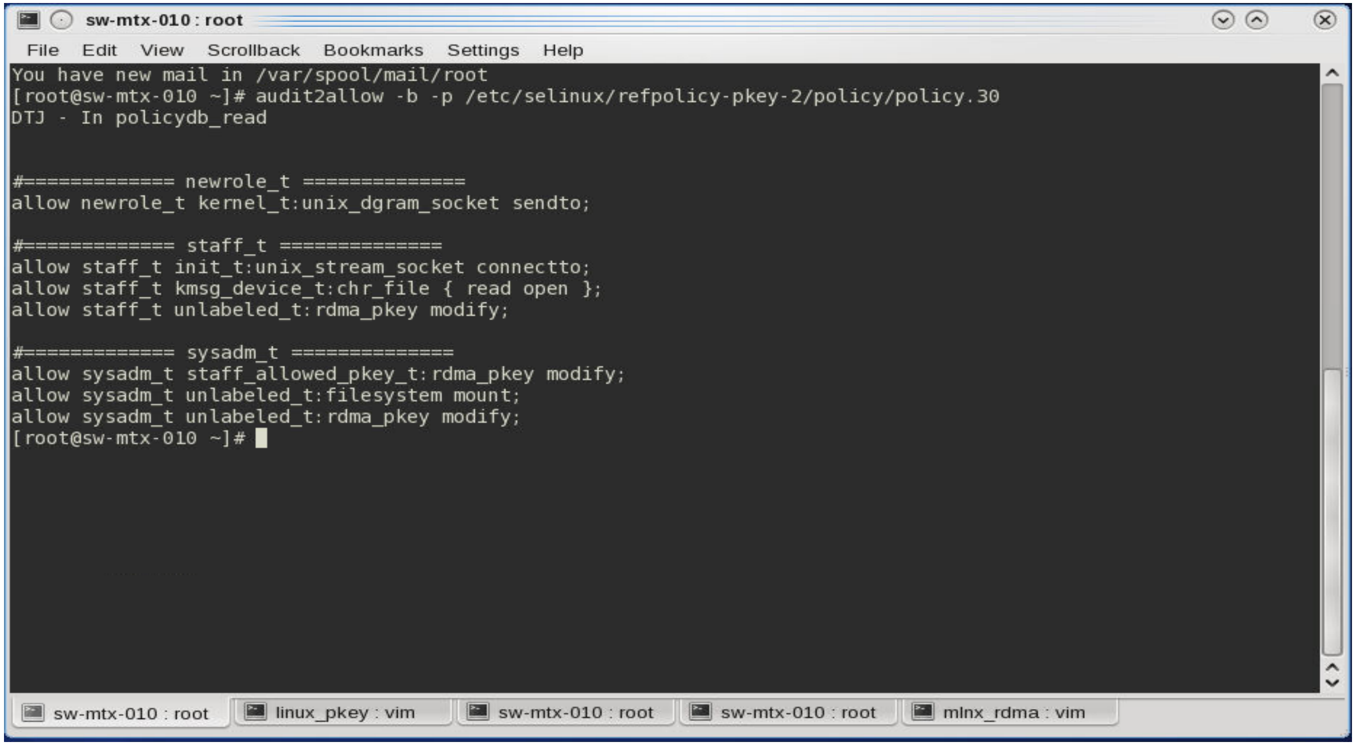
```
[root@sw-mtx-010 ~]# id -Z
root:sysadm_r:sysadm_t
[root@sw-mtx-010 ~]# ib_write_bw -d mlx4_0 -D2 --pkey_index=0 --ib-port=2

************************************
* Waiting for client to connect... *
************************************
---------------------------------------------------------------------------------------
                    RDMA_Write BW Test
 Dual-port       : OFF          Device         : mlx4_0
 Number of qps   : 1            Transport type : IB
 Connection type : RC           Using SRQ      : OFF
 CQ Moderation   : 100
 Mtu             : 2048[B]
 Link type       : IB
 Max inline data : 0[B]
 rdma_cm QPs     : OFF
 Data ex. method : Ethernet
---------------------------------------------------------------------------------------
 local address: LID 0x02 QPN 0x0210 PSN 0x5fdcaf RKey 0x30011900 VAddr 0x007fddd781f000
 remote address: LID 0x01 QPN 0x0211 PSN 0xf76897 RKey 0x30011a00 VAddr 0x007f4f50d06000
---------------------------------------------------------------------------------------
 #bytes     #iterations    BW peak[MB/sec]    BW average[MB/sec]   MsgRate[Mpps]
 65536      190700           0.00               5966.21             0.095459
---------------------------------------------------------------------------------------
[root@sw-mtx-010 ~]# 
```

SMI



```
[root@sw-mtx-010 refpolicy_modular]# setenforce 1
[root@sw-mtx-010 refpolicy_modular]# id -Z
root:sysadm_r:sysadm_t:s0
[root@sw-mtx-010 refpolicy_modular]# opensm -g 0xe41d2d03000a5522 &
[1] 17465
[root@sw-mtx-010 refpolicy_modular]# ------------------------------------------------
OpenSM 3.3.19
 Reading Cached Option File: /etc/rdma/opensm.conf
Command Line Arguments:
 Guid <0xe41d2d03000a5522>
 Log File: /var/log/opensm.log
-------------------------------------------------
OpenSM 3.3.19

Entering DISCOVERING state


Error from osm_opensm_bind (0x2A)
Perhaps another instance of OpenSM is already running
Exiting SM


[1]+  Done                    opensm -g 0xe41d2d03000a5522
[root@sw-mtx-010 refpolicy_modular]# opensm -g 0xe41d2d03000a5521 &
[1] 17511
[root@sw-mtx-010 refpolicy_modular]# ------------------------------------------------
OpenSM 3.3.19
 Reading Cached Option File: /etc/rdma/opensm.conf
Command Line Arguments:
 Guid <0xe41d2d03000a5521>
 Log File: /var/log/opensm.log
-------------------------------------------------
OpenSM 3.3.19

Entering DISCOVERING state

Entering MASTER state


[root@sw-mtx-010 refpolicy_modular]# █
```

POLICY UTILITIES



```
You have new mail in /var/spool/mail/root
[root@sw-mtx-010 ~]# audit2allow -b -p /etc/selinux/refpolicy-pkey-2/policy/policy.30
DTJ - In policydb_read


#============= newrole_t ==============
allow newrole_t kernel_t:unix_dgram_socket sendto;

#============= staff_t ==============
allow staff_t init_t:unix_stream_socket connectto;
allow staff_t kmsg_device_t:chr_file { read open };
allow staff_t unlabeled_t:rdma_pkey modify;

#============= sysadm_t ==============
allow sysadm_t staff_allowed_pkey_t:rdma_pkey modify;
allow sysadm_t unlabeled_t:filesystem mount;
allow sysadm_t unlabeled_t:rdma_pkey modify;
[root@sw-mtx-010 ~]#
```

- This tool generates policy code to allow violations in the audit log.
- If we added the three allow lines for “rdma_pkey” the access errors in the demo would be allowed.

CONCLUSION

- **Targeting submission for 4.7 Kernel**

OPENFABRICS ALLIANCE

12th ANNUAL WORKSHOP 2016

THANK YOU

Dan Jurgens

Mellanox Technologies

Mellanox® TECHNOLOGIES