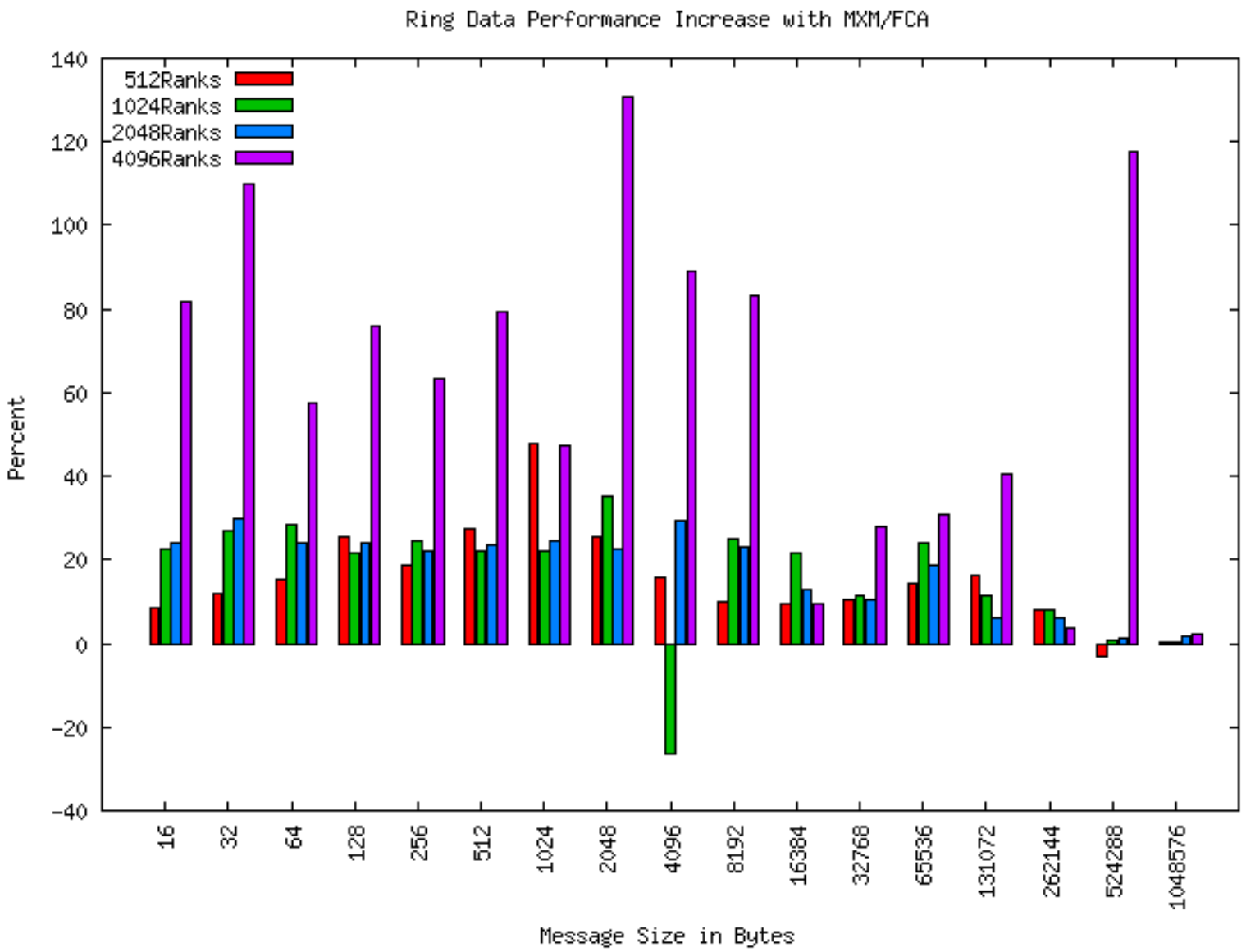
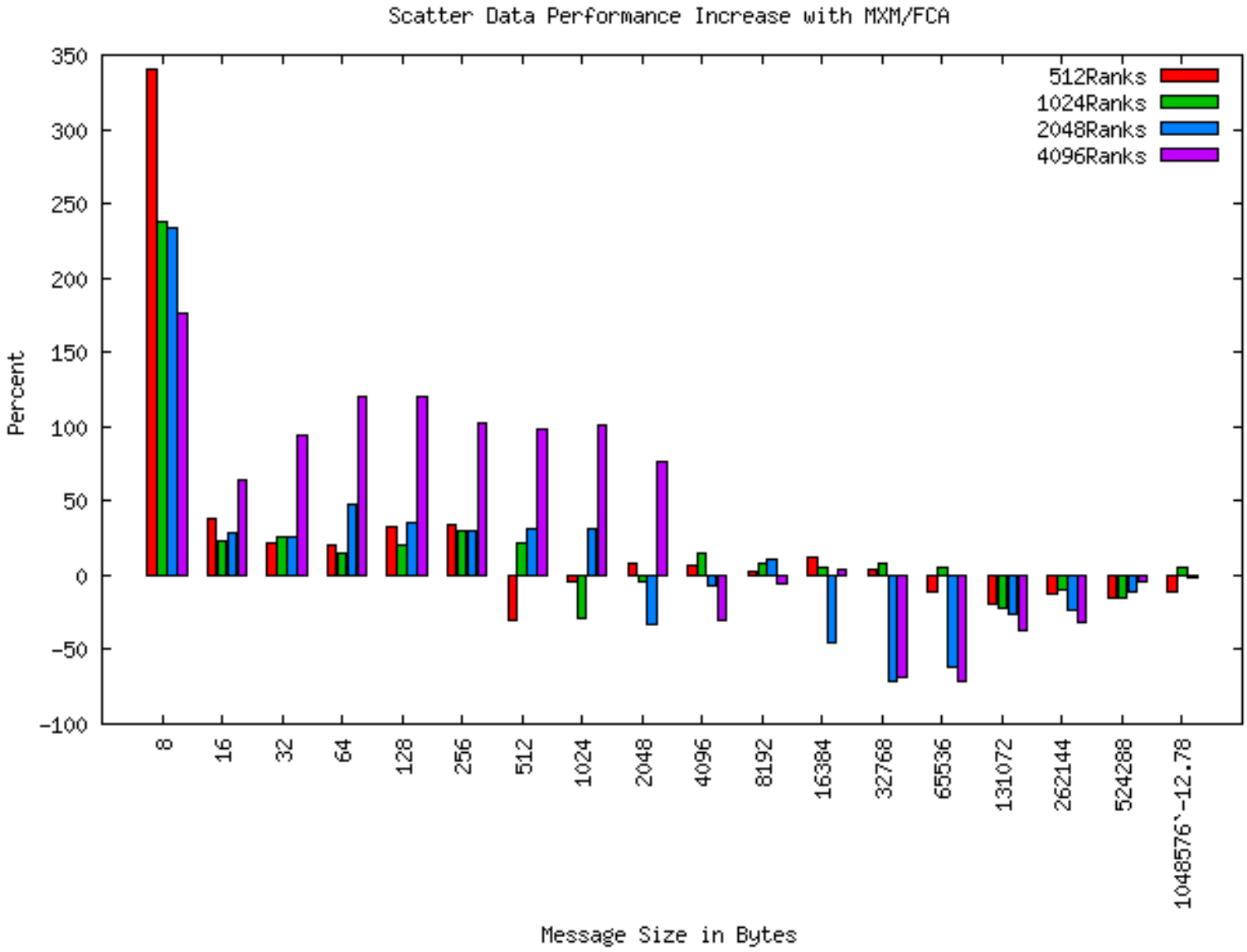
L M N O P Keys

Susan Coulter
Los Alamos National Laboratory
LA-UR-14-22074
April 3rd, 2014

Why MKeys? FCA

- Fabric Collective Accelerator
- Implemented using Multicast GIDs
- No user space API in the kernel to manage these mgids
- SA MAD packets used to communicate this information
- No “SA class” vs “Non-SA class”
- No distinction between “read” and “write”
- Must allow user level access to /dev/infiniband/umad*


Ring Data Performance Increase with MXM/FCA
512Ranks
1024Ranks
2048Ranks
4096Ranks
Percent
140
120
100
80
60
40
20
0
-20
-40
16
32
64
128
256
512
1024
2048
4096
8192
16384
32768
65536
131072
262144
524288
1048576
Message Size in Bytes


Scatter Data Performance Increase with MXM/FCA
512Ranks
1024Ranks
2048Ranks
4096Ranks
Percent
350
300
250
200
150
100
50
0
-50
-100
8
16
32
64
128
256
512
1024
2048
4096
8192
16384
32768
65536
131072
262144
524288
1048576`-12.78
Message Size in Bytes

crw-rw-rw- 1 root root /dev/infiniband/umad0

What's the problem?

- Allowing this access exposes an attack vector
- Normal users can send and receive MAD packets
- MAD floods/DoS
- Disable ports – even the SM port
- Rogue SM could take over
- … the earth plummets into the sun …

The fix: MKey, etc

opensm.conf

```
# M_Key value sent to all ports qualifying all Set(PortInfo)
m_key 0x0000000000000003

# The lease period used for the M_Key on this subnet in [sec]
m_key_lease_period 604800

# Protection bits for the M_Key
m_key_protection_level 2

# sm_Key value of the SM used for SM authentication
sm_key 0x0000000000000003

# sa_key value to qualify rcv SA queries as 'trusted'
sa_key 0x0000000000000003
```

MKey – other issues

```
[root@mu-master ~]# smpquery pi 335 | grep Mkey

Mkey:.............................<not displayed>
MkeyLeasePeriod:.................14976
MkeyViolations:...................0
```

```
[root@lo2-sn ~]# smpquery pi 213 | grep Mkey
Mkey:.............................0x0000000000000000
MkeyLeasePeriod:.................0
MkeyViolations:...................0
```

/etc/infiniband-diags/ibdiag.conf:

```
# define a default m_key
m_key=0x03
```

Unexpected Fallout

Mar 14 17:31:36 832126 [2D62D700] 0x01 -> log_trap_info: Received Generic Notice type:2 num:256
(Bad M_Key) Producer:2 (Switch) from LID:22 TID:0x0000000000dbd808
Mar 14 17:31:36 832224 [2D62D700] 0x02 -> log_notice: Reporting Generic Notice type:2 num:256
(Bad M_Key) from LID:22 GID:fe80::2:c902:44:3f10
Mar 14 17:31:37 224169 [2D02A700] 0x01 -> log_trap_info: Received Generic Notice type:2 num:256
(Bad M_Key) Producer:2 (Switch) from LID:449 TID:0x0000000000db569b
Mar 14 17:31:37 224240 [2D02A700] 0x02 -> log_notice: Reporting Generic Notice type:2 num:256
(Bad M_Key) from LID:449 GID:fe80::2:c902:44:86f0
Mar 14 17:31:37 608874 [2E836700] 0x01 -> log_trap_info: Received Generic Notice type:2 num:256
(Bad M_Key) Producer:2 (Switch) from LID:465 TID:0x0000000000dbfa96
Mar 14 17:31:37 608941 [2E836700] 0x02 -> log_notice: Reporting Generic Notice type:2 num:256
(Bad M_Key) from LID:465 GID:fe80::2:c902:44:8920

Switch 36 "S-0002c90200443f10" # "muibcore2 Spine 01" enhanced port 0 lid 22 lmc 0
Switch 36 "S-0002c902004486f0" # "muibcore1 Spine 01" enhanced port 0 lid 449 lmc 0
Switch 36 "S-0002c90200448920" # "muibcore3 Spine 01" enhanced port 0 lid 465 lmc 0

Required software versions

- opensm-3.3.15-3, devel, lib, static
- libibmad-1.3.9-1, devel, static
- infiniband-diags-1.6.1-3
- pragmatic-infiniband-utilities-1.2.5-1
- ibutils-1.5.8-1

Questions?

Why PKeys?

- Effort to consolidate 2 unclassified networks
- Save on hardware and all associated maintenance/support
- Two separate projects/funding – similar data classification
- Separation currently enforced with symlinks and unix permissions

  - zorrillo test cluster
  - 8 compute nodes 4 IO nodes
  - QDR Mellanox
  - IO nodes are gateways to external file systems

PKeys - Configuration

```
# Default partition, all full except compute nodes
Default = 0x7fff : ALL_SWITCHES=full , SELF=full ;
Default = 0x7fff , ipoib , defmember=full :
      0x0002c903000ab913 , 0x0002c903000b270b , 0x0002c903000b27b7 , 0x0002c903000b265b ;

# zoa001-zoa004
Sheba = 0x7ffa , ipoib , defmember=full :
      0x003048fffff4e579 , 0x002590ffff19d465 , 0x002590ffff1b4ef1 , 0x002590ffff19d3fd ;
Sheba , ipoib , defmember=full :
      0x0002c903000ab913 , 0x0002c903000b270b , 0x0002c903000b27b7 , 0x0002c903000b265b ;

# zoa005-zoa008
Rudi = 0x7ffb , ipoib , defmember=full :
      0x003048fffff4e885 , 0x003048fffff4e541 , 0x002590ffff19cc85 , 0x003048fffff625c1 ;
Rudi , ipoib , defmember=full :
      0x0002c903000ab913 , 0x0002c903000b270b , 0x0002c903000b27b7 , 0x0002c903000b265b ;
```

PKeys – The SM view

[zo-master ~]# for x in 17 9 11 7 10 8 5 4 13 14 12 6; do smpquery PKeys $x | grep " 0:"; done

0: 0xffff 0x0000 0x0000 0x0000 0x0000 0x0000 0x0000 0x0000

0: 0x7fff 0xfffa 0x0000 0x0000 0x0000 0x0000 0x0000 0x0000
0: 0x7fff 0xfffa 0x0000 0x0000 0x0000 0x0000 0x0000 0x0000
0: 0x7fff 0xfffa 0x0000 0x0000 0x0000 0x0000 0x0000 0x0000

0: 0x7fff 0xfffb 0x0000 0x0000 0x0000 0x0000 0x0000 0x0000
0: 0x7fff 0xfffb 0x0000 0x0000 0x0000 0x0000 0x0000 0x0000
0: 0x7fff 0xfffb 0x0000 0x0000 0x0000 0x0000 0x0000 0x0000
0: 0x7fff 0xfffb 0x0000 0x0000 0x0000 0x0000 0x0000 0x0000

0: 0xffff 0xfffa 0xfffb 0x0000 0x0000 0x0000 0x0000 0x0000
0: 0xffff 0xfffa 0xfffb 0x0000 0x0000 0x0000 0x0000 0x0000
0: 0xffff 0xfffa 0xfffb 0x0000 0x0000 0x0000 0x0000 0x0000
0: 0xffff 0xfffa 0xfffb 0x0000 0x0000 0x0000 0x0000 0x0000

PKeys – Security ? … no PathRecords

```
[root@zo-master partitions]# saquery -p --src-to-dst 4:9

[root@zo-master partitions]# saquery -p --src-to-dst 4:5
PathRecord dump:
                        service_id.....0x0000000000000000
                        dgid.....fe80::25:90ff:ff19:cc85
                        sgid.....fe80::30:48ff:fff6:25c1
                        dlid.....5
                        slid.....4
                        hop_flow_raw.....0x0
                        tclass.....0x0
                        num_path_revers.....0x80
                        pkey.....0xFFFB
                        qos_class.....0x0
                        sl.....0x0
                        mtu.....0x85
                        rate.....0x87
                        pkt_life.....0x92
                        preference.....0x0
                        resv2.....0x000000000000
```

PKeys – IPoIB

echo 0xfffa > /sys/class/net/ib0/create_child

```
ib0.fffa  Link encap:InfiniBand  HWaddr 80:00:00:68:FE:80:00:00:00:00:00:00:00
          inet addr:10.16.77.2  Bcast:10.16.77.255  Mask:255.255.255.0
          inet6 addr: fe80::225:90ff:ff19:d465/64 Scope:Link
          UP BROADCAST RUNNING MULTICAST  MTU:2044  Metric:
```

```
[root@zo-master ~]# saquery -g | grep MGID
                    MGID.....ff12:401b:ffff::1
                    MGID.....ff12:401b:ffff::1
                    MGID.....ff12:401b:ffff::1
                    …
                    MGID.....ff12:401b:ffff::1
                    MGID.....ff12:401b:ffff::ffff:ffff
                    MGID.....ff12:401b:ffff::ffff:ffff
                    MGID.....ff12:401b:ffff::ffff:ffff
                    …
                    MGID.....ff12:401b:ffff::ffff:ffff
```

duplicates

PKeys – Fixes / Workarounds

Perftest:

ib_read_bw --pkey_index=1 zoa008

Running w/out IPoIB between compute nodes:

mpirun -n 16 --mca btl_openib_pkey 0xfffa
--mca oob_tcp_if_include eth0 /users/markus/zorrillo/mpi-hello

Questions?