

Test Plan and Results for Voltaire 4036E

Main CHPC Web Site
Issue Tracking
CHPC Wiki Home

Contents
<ul style="list-style-type: none">• Overview• Test Environment• 10GbE Baseline Test Results• Basic Performance Test Results<ul style="list-style-type: none">• Topology• Summary of Results<ul style="list-style-type: none">• Unidirectional<ul style="list-style-type: none">• taildrop(IB) to fifo(Ethernet)• fifo(Ethernet) to taildrop(IB)• taildrop(IB) to wred(IB)• taildrop(IB) to wred(IB) - IB Layer2• Bidirectional<ul style="list-style-type: none">• taildrop(IB) < - > fifo(Ethernet)• Unidirectional with 2 sources and 2 destinations<ul style="list-style-type: none">• taildrop(IB) < - > fifo(Ethernet) AND wred(IB) < - > roundrobin(Ethernet)• Larger-scale Performance Test Results<ul style="list-style-type: none">• Topology• Summary of Results<ul style="list-style-type: none">• Unidirectional with 8 sources and 2 destinations<ul style="list-style-type: none">• taildrop,up012,up017,up139 - > fifo; wred,up142,up149,up241 - > roundrobin• Notes and Other Findings

Overview

This set of tests evaluates the IPoIB throughput performance of the Voltaire 4036E using basic iperf tests. All tests are performed using 1 stream, 4 streams, and 12 streams. See the Topology section below to understand name references. The basic test parameters are as follows:

- Unidirectional
 - taildrop(IB) to fifo(Ethernet)
 - fifo(Ethernet) to taildrop(IB)
 - taildrop(IB) to wred(IB)
 - taildrop(IB) to wred(IB) - IB Layer2
- Bidirectional
 - taildrop(IB) < - > fifo(Ethernet)
- Unidirectional with two sources and destinations
 - taildrop(IB) < - > fifo(Ethernet) AND wred(IB) < - > roundrobin(Ethernet)
- Bidirectional with two sources and destinations
 - taildrop(IB) < - > fifo(Ethernet) AND wred(IB) < - > roundrobin(Ethernet)

Test Environment

The following settings are used for all tests (unless otherwise noted):

- iperf version 2.0.4, using default settings
- 20 minute duration
- IPoIB MTU 65534 (default CM MTU)
- Infiniband MTU 4096
- Ethernet IP MTU 1500
- Ethernet MTU 9212

The Voltaire embedded SM provides subnet management for the IB fabric.

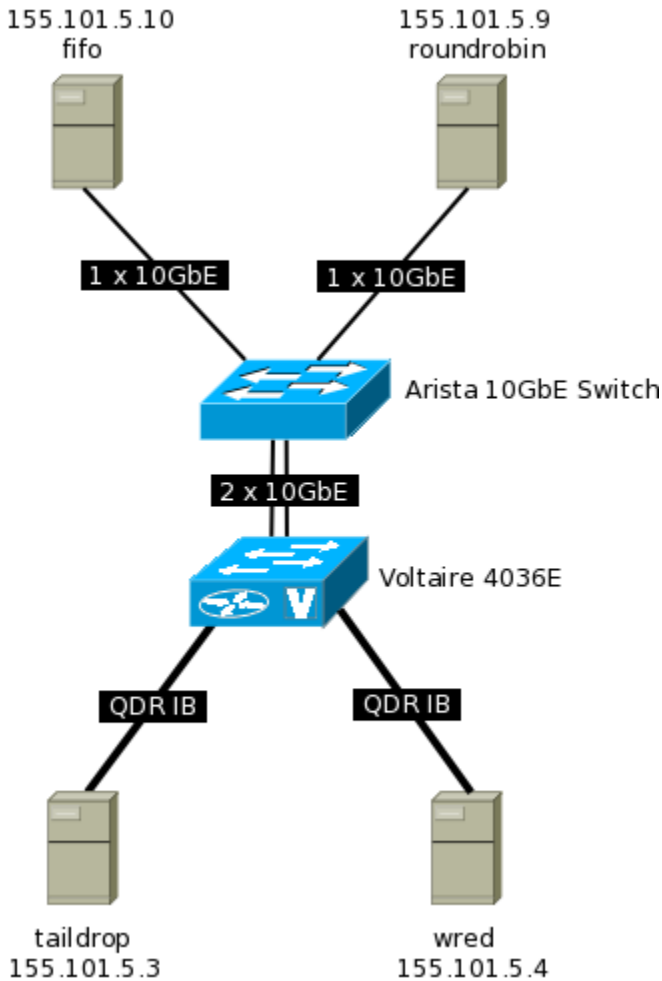
10GbE Baseline Test Results

We're using Myricom 10GbE NICs, which offer nearly line-rate performance across the Arista switch. This is a unidirectional test from fifo, across the Arista switch, to roundrobin:

```
[root@fifo ~]# iperf -c 155.101.5.9 -t 60 -P4
-----
Client connecting to 155.101.5.9, TCP port 5001
TCP window size: 64.0 KByte (default)
-----
[ 3] local 155.101.5.10 port 34042 connected with 155.101.5.9 port 5001
[ 4] local 155.101.5.10 port 34043 connected with 155.101.5.9 port 5001
[ 6] local 155.101.5.10 port 34045 connected with 155.101.5.9 port 5001
[ 5] local 155.101.5.10 port 34044 connected with 155.101.5.9 port 5001
[ ID] Interval      Transfer      Bandwidth
[ 3] 0.0-60.0 sec  15.5 GBytes  2.22 Gbits/sec
[ ID] Interval      Transfer      Bandwidth
[ 6] 0.0-60.0 sec  20.6 GBytes  2.96 Gbits/sec
[ ID] Interval      Transfer      Bandwidth
[ 4] 0.0-60.0 sec  13.3 GBytes  1.91 Gbits/sec
[ ID] Interval      Transfer      Bandwidth
[ 5] 0.0-60.0 sec  16.8 GBytes  2.40 Gbits/sec
[SUM] 0.0-60.0 sec  66.3 GBytes  9.48 Gbits/sec
```

Basic Performance Test Results

Topology



Summary of Results

Unidirectional

taildrop(IB) to fifo(Ethernet)

Streams	Run 1	Run 2	Run 3	Average
1	7.05 Gb/s	7.02 Gb/s	7.03 Gb/s	7.03 Gb/s
4	6.64 Gb/s	6.54 Gb/s	6.69 Gb/s	6.62 Gb/s
12	6.45 Gb/s	6.62 Gb/s	6.65 Gb/s	6.57 Gb/s

fifo(Ethernet) to taildrop(IB)

Streams	Run 1	Run 2	Run 3	Average
1	3.80 Gb/s	4.84 Gb/s	4.88 Gb/s	4.50 Gb/s
4	4.51 Gb/s	4.54 Gb/s	4.48 Gb/s	4.51 Gb/s
12	5.61 Gb/s	5.55 Gb/s	N/A	5.58 Gb/s

taildrop(IB) to wred(IB)

Streams	Run 1	Run 2	Run 3	Average
1	13.5 Gb/s	16.0 Gb/s	15.0 Gb/s	14.83 Gb/s
4	22.3 Gb/s	22.3 Gb/s	22.3 Gb/s	22.3 Gb/s
12	22.3 Gb/s	22.3 Gb/s	22.3 Gb/s	22.3 Gb/s

taildrop(IB) to wred(IB) - IB Layer2

```
[root@taildrop sysconfig]# ib_write_bw -m 4096 -a -n 10000 -q 4 155.101.5.4
```

```
-----
RDMA_Write BW Test
```

```
Number of qp's running 4
```

```
Connection type : RC
```

```
Each Qp will post up to 100 messages each time
```

```
Inline data is used up to 0 bytes message
```

```
local address: LID 0x04 QPN 0x3c004e PSN 0x90123e RKey 0x58042000 VAddr 0x002af6e0cfc000
```

```
local address: LID 0x04 QPN 0x3c004f PSN 0xaae250 RKey 0x58042000 VAddr 0x002af6e0cfc000
```

```
local address: LID 0x04 QPN 0x3c0050 PSN 0x47bd5a RKey 0x58042000 VAddr 0x002af6e0cfc000
```

```
local address: LID 0x04 QPN 0x3c0051 PSN 0xc91aa1 RKey 0x58042000 VAddr 0x002af6e0cfc000
```

```
remote address: LID 0x03 QPN 0x7c004e PSN 0x1d6bbd RKey 0x60042000 VAddr 0x002b8429984000
```

```
remote address: LID 0x03 QPN 0x7c004f PSN 0xcc8a9a RKey 0x60042000 VAddr 0x002b8429984000
```

```
remote address: LID 0x03 QPN 0x7c0050 PSN 0x906da6 RKey 0x60042000 VAddr 0x002b8429984000
```

```
remote address: LID 0x03 QPN 0x7c0051 PSN 0xb9ffc6 RKey 0x60042000 VAddr 0x002b8429984000
```

```
Mtu : 4096
```

```
-----
#bytes #iterations    BW peak[MB/sec]    BW average[MB/sec]
      2           10000             7.26              7.24
      4           10000            14.51             14.49
      8           10000            29.10             29.02
     16           10000            57.97             57.04
     32           10000           120.52            115.43
     64           10000           234.43            234.11
    128           10000           447.38            437.93
    256           10000           823.61            791.71
    512           10000           893.00            790.86
   1024           10000           796.96            786.27
   2048           10000          2803.05           2802.78
   4096           10000          2862.47           2861.99
   8192           10000          2895.43           2895.22
  16384           10000          2917.25           2916.65
  32768           10000          2923.78           2923.71
  65536           10000          2927.35           2927.31
 131072           10000          2928.65           2928.65
 262144           10000          2929.27           2929.27
 524288           10000          2929.71           2929.71
1048576           10000          2929.89           2929.89
2097152           10000          2929.99           2929.99
4194304           10000          2929.84           2929.84
8388608           10000          2929.76           2929.76
-----
```

Bidirectional

taildrop(IB) < - > fifo(Ethernet)

Streams	Run 1	Run 2	Run 3	Average
1	3.64 Gb/s / 487 Mb/s	757 Mb/s / 3.57 Gb/s	3.58 Gb/s / 553 Mb/s	3.59 Gb/s / 599 Mb/s

Unidirectional with 2 sources and 2 destinations

taildrop(IB) < - > fifo(Ethernet) AND wred(IB) < - > roundrobin(Ethernet)

taildrop

Streams	Run 1	Run 2	Run 3	Average
1	4.17 Gb/s	3.80 Gb/s	4.13 Gb/s	4.03 Gb/s
4	4.10 Gb/s	4.10 Gb/s	4.14 Gb/s	4.11 Gb/s

12	4.13 Gb/s	3.82 Gb/s	4.21 Gb/s	5.04 Gb/s
----	-----------	-----------	-----------	-----------

wred

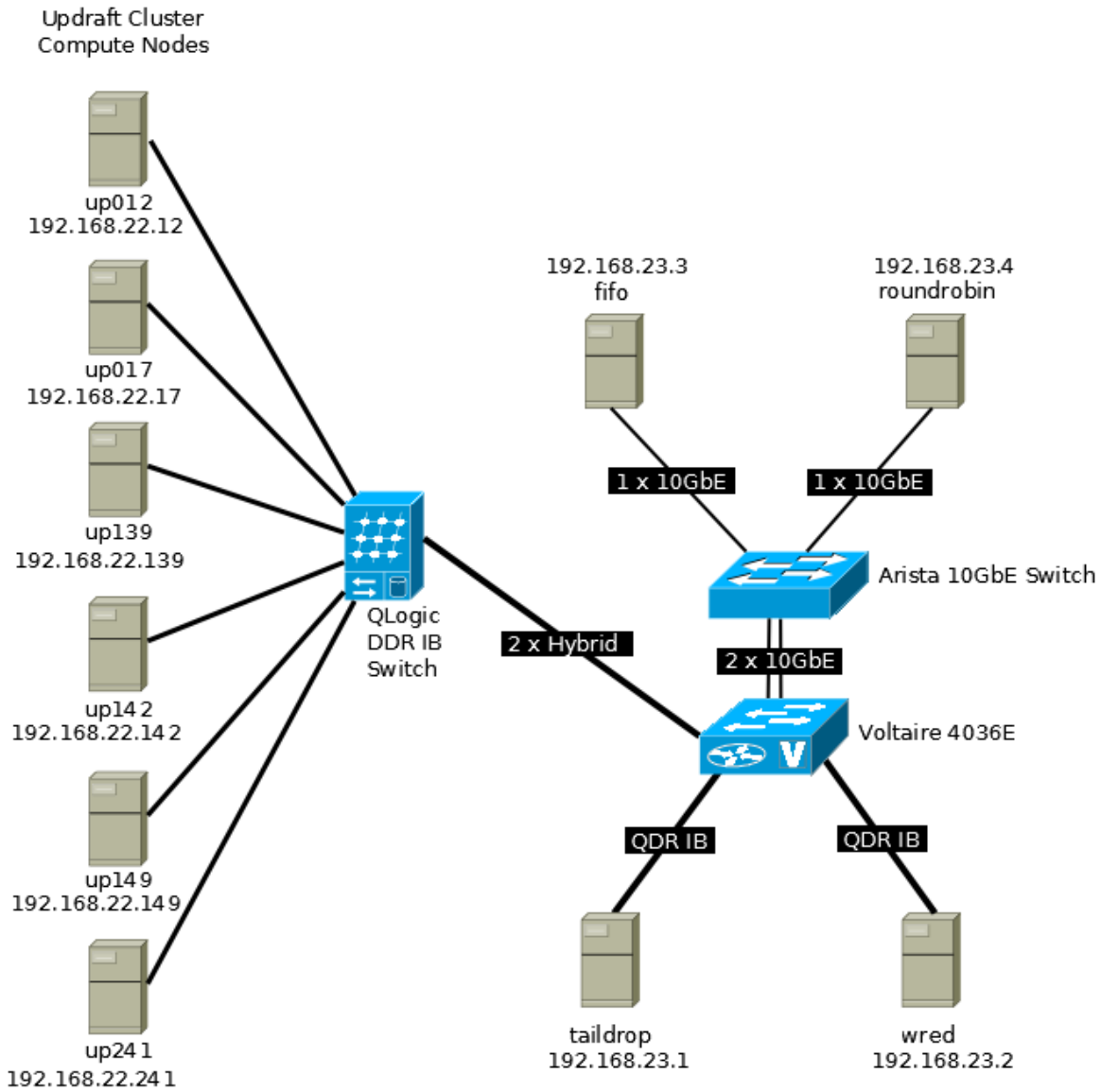
Streams	Run 1	Run 2	Run 3	Average
1	3.97 Gb/s	3.99 Gb/s	3.94 Gb/s	3.96 Gb/s
4	3.90 Gb/s	3.91 Gb/s	3.85 Gb/s	3.88 Gb/s
12	4.18 Gb/s	4.20 Gb/s	4.20 Gb/s	4.19 Gb/s

combined

Streams	Average
1	7.99 Gb/s
4	7.99 Gb/s
12	9.23 Gb/s

Larger-scale Performance Test Results

Topology



Summary of Results

Unidirectional with 8 sources and 2 destinations

taildrop,up012,up017,up139 -> fifo; wred,up142,up149,up241 -> roundrobin

1-stream Tests, Run 1

node	result
taildrop	724 Mb/s
wred	833 Mb/s
up012	1.21 Gb/s
up017	1.22 Gb/s
up139	1.67 Gb/s
up142	1.23 Gb/s
up149	1.21 Gb/s

up241	1.22 Gb/s
TOTAL	9.26 Gb/s

1-stream Tests, Run 2

node	result
taildrop	799 Mb/s
wred	811 Mb/s
up012	1.13 Gb/s
up017	1.12 Gb/s
up139	1.31 Gb/s
up142	1.66 Gb/s
up149	1.65 Gb/s
up241	299 Mb/s
TOTAL	8.67 Gb/s

12-stream Tests, Run 1

node	result
taildrop	2.83 Gb/s
wred	2.87 Gb/s
up012	610 Mb/s
up017	607 Mb/s
up139	1.13 Gb/s
up142	602 Mb/s
up149	619 Mb/s
up241	601 Mb/s
TOTAL	9.85 Gb/s

12-stream Tests, Run 2

node	result
taildrop	2.77 Gb/s
wred	2.87 Gb/s
up012	653 Mb/s
up017	655 Mb/s
up139	1.18 Gb/s
up142	657 Mb/s
up149	(619 Mb/s)
up241	653 Mb/s
TOTAL	10.05 Gb/s

Note: the number for up149 in this test is taken from the previous test, because up149 crashed near the beginning of the run*

Notes and Other Findings

- ARP queries take a long time when crossing the IB/Ethernet boundary:

```
[root@roundrobin ~]# arp -a
gw-chpc-lab.chpc.utah.edu (155.101.4.1) at 00:16:9C:54:80:00 [ether] on eth0
[root@roundrobin ~]# ping tailldrop
PING tailldrop (155.101.5.3) 56(84) bytes of data.
64 bytes from tailldrop (155.101.5.3): icmp_seq=1 ttl=64 time=1001 ms
64 bytes from tailldrop (155.101.5.3): icmp_seq=2 ttl=64 time=1.85 ms
64 bytes from tailldrop (155.101.5.3): icmp_seq=3 ttl=64 time=0.150 ms

[root@tailldrop ~]# arp -a
gw-chpc-lab.chpc.utah.edu (155.101.4.1) at 00:16:9C:54:80:00 [ether] on eth0
[root@tailldrop ~]# ping fifo
PING fifo (155.101.5.10) 56(84) bytes of data.
64 bytes from fifo (155.101.5.10): icmp_seq=1 ttl=64 time=1001 ms
64 bytes from fifo (155.101.5.10): icmp_seq=2 ttl=64 time=2.26 ms
64 bytes from fifo (155.101.5.10): icmp_seq=3 ttl=64 time=0.116 ms
```

- To set up pMTU discovery on the 4036E:
Set an IP address on the if0 interface. This address must be in the same subnet as the servers on both sides of the gateway:

```
4036E-0188# io

Welcome to Voltaire 4036E IO 4036E-6253

4036E-0188# config
4036E-0188(config)# interface
4036E-0188(config-if)# ip-address set if0 155.101.5.20 255.255.255.0
4036E-0188(config-if)# interface show all
entry alias      ip                mask              broadcast         vlan pkey(hex) trunk
|----|-----|-----|-----|-----|----|-----|-----|
1    if0          155.101.5.20     255.255.255.0    155.101.5.255   no  0x7fff    P1-2
```

For questions about this document, contact Tom Ammon at tom.ammon@utah.edu