15th ANNUAL WORKSHOP 2019

REALWORLD HIGH PERFORMANCE NETWORKING ADVANTAGES, COMPARISON AND THE CHALLENGES IN THE FINANCIAL MARKETS

Sampath Tilakumara, Head of Technology / Indika Prasad Kumara, Network Architect

Millennium IT Software (LSEG Technology)

March 21, 2019

AGENDA

- How we became the fastest trading platform in 2009
- Fastest Trading System in production in 2010
- System upgrade after 9 years – Case Study
- MillenniumIT Benchmarking Tools
- RDMA vs Non-RDMA evaluation
- Questions for the OFA Community

HOW WE BECAME THE FASTEST TRADING PLATFORM IN YEAR 2009?

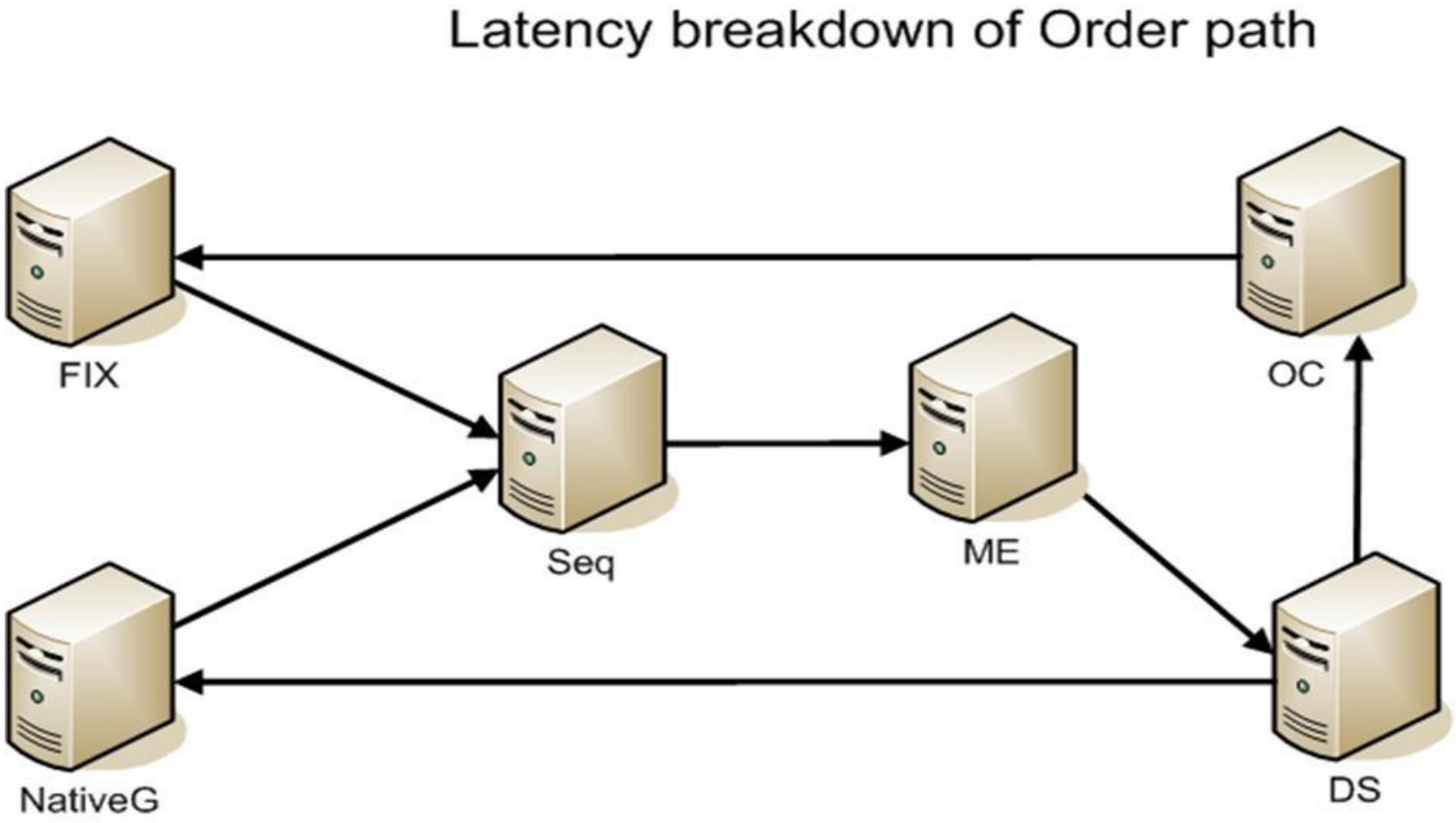
- MillenniumIT (now LSEG Technology) is a global FinTech solutions provider, specializing in ultra-low latency software systems.
- Millennium Exchange™ is a distributed system with multi threaded processes, optimized for performance, throughput and availability.
- The specific benchmark was run in Intel Labs, hosted on 9 Nehalem boxes (including order injector/measurement tool), and 1 Harpertown for Oracle.
- The inter-machine hop was supported via BladeNetwork's RackSwitch G8100 and iWarp capable NetEffect 10G RNICs (owned by Intel, but we were not utilizing the RNIC capabilities)
- We used AF_UNIX sockets with tight socket polling to minimize latency (and intra-machine hops used shared memory at times).

HOW WE BECAME THE FASTEST TRADING PLATFORM IN YEAR 2009? (CONTD.)

- The inter-machine RTT latency we observed dropped from 120us (in Colombo) to 34us in the Intel labs (Nehalem+10G).
- Started working with Mellanox and Voltaire to get switch kits down to Colombo to work in native RNIC and IB mode comms.
- IB Native stack and the new 40G QDR switch with switching latency of 100-300ns, combined with QDR HCAs capable network switching system of 5GT/s provided us the boost to get overall system latency to 100+ us.

SYSTEM LATENCY BREAKDOWN

- Reduction of hop-to-hop network latency multiplies the effect



Component	Time (us)	Time(us)
OC		20
FIX		70
NativeG	20	
Seq	11	11
ME	60	60
DS	10	10
NW Hops	4x4	5x4
	117	191

These latency numbers are intended to be used as reference data for latency troubleshooting.
These number may vary depending on the hardware and application configurations
These number may be changed by MillenniumIT without any prior notice.

FASTEST TRADING SYSTEM IN PRODUCTION (2010)

IB Network and Servers

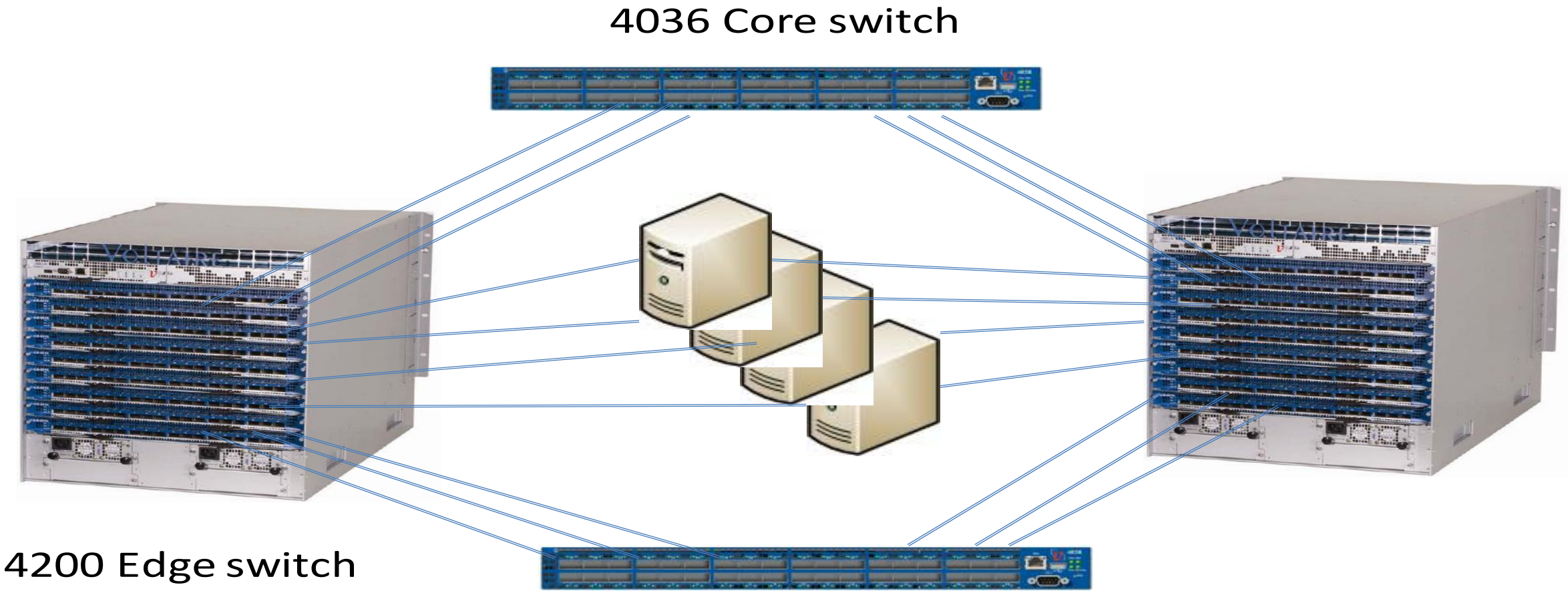
- IB QDR 40Gb/s network
- IB Switches: Chassis Based Spine-Leaf CLOS Architecture

  Voltaire 4036 36-Port Spine Switches

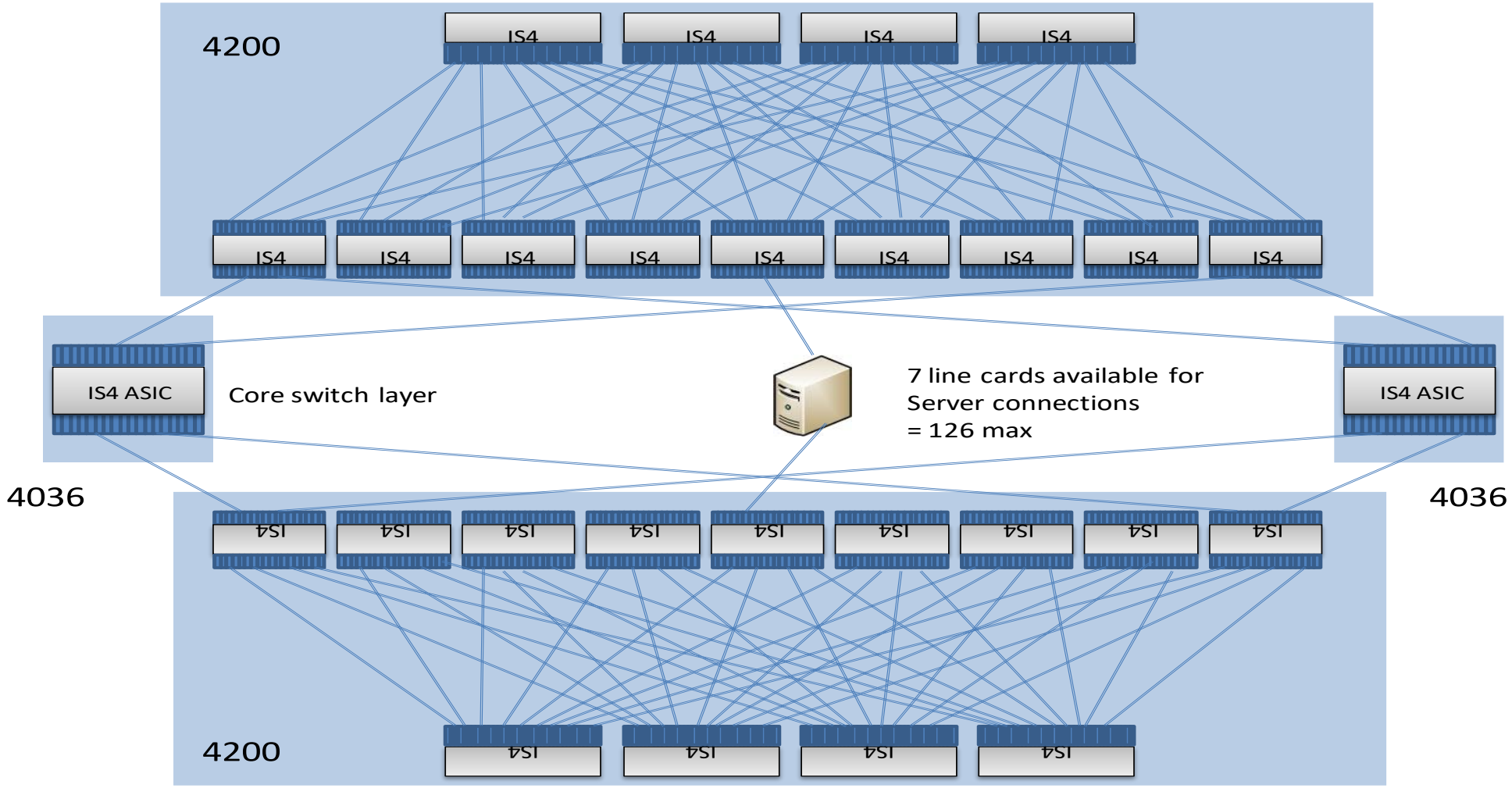
  Voltaire 4200 Director Class Leaf Switches

  (Previously planned for Mellanox part numbers but later changed to rebranded Voltaire part numbers, considering the support levels)
- IB Network cards: Mellanox ConnectX-2 QDR (40Gb/s) dual port adapters
- IB Network Card Driver: Mellanox OFED 1.5.1
- Servers: IBM X3650 Intel Xeon based servers
- Operating System: SUSE Linux Enterprise 11

IB NETWORK - YEAR 2010



IB NETWORK - YEAR 2010



IB NETWORK - YEAR 2010

Part Number	Description	Quantity
Voltaire 4200	Voltaire Director class InfiniBand Switches	8
Voltaire 4036	Voltaire 36 port InfiniBand Switches	10
Ordered pre-manufactured in fixed lengths	Voltaire/Mellanox Optical QFSP cables	Approx. 300
Mellanox MHQH29B-XTR	Mellanox ConnectX-2 Dual Port 4x QFSP 40Gb/s (QDR) InfiniBand Host Channel Adapters	1 per server (Approx 300)

Grid Director 4036E - Front Panel

Grid Director 4036E - Rear Panel

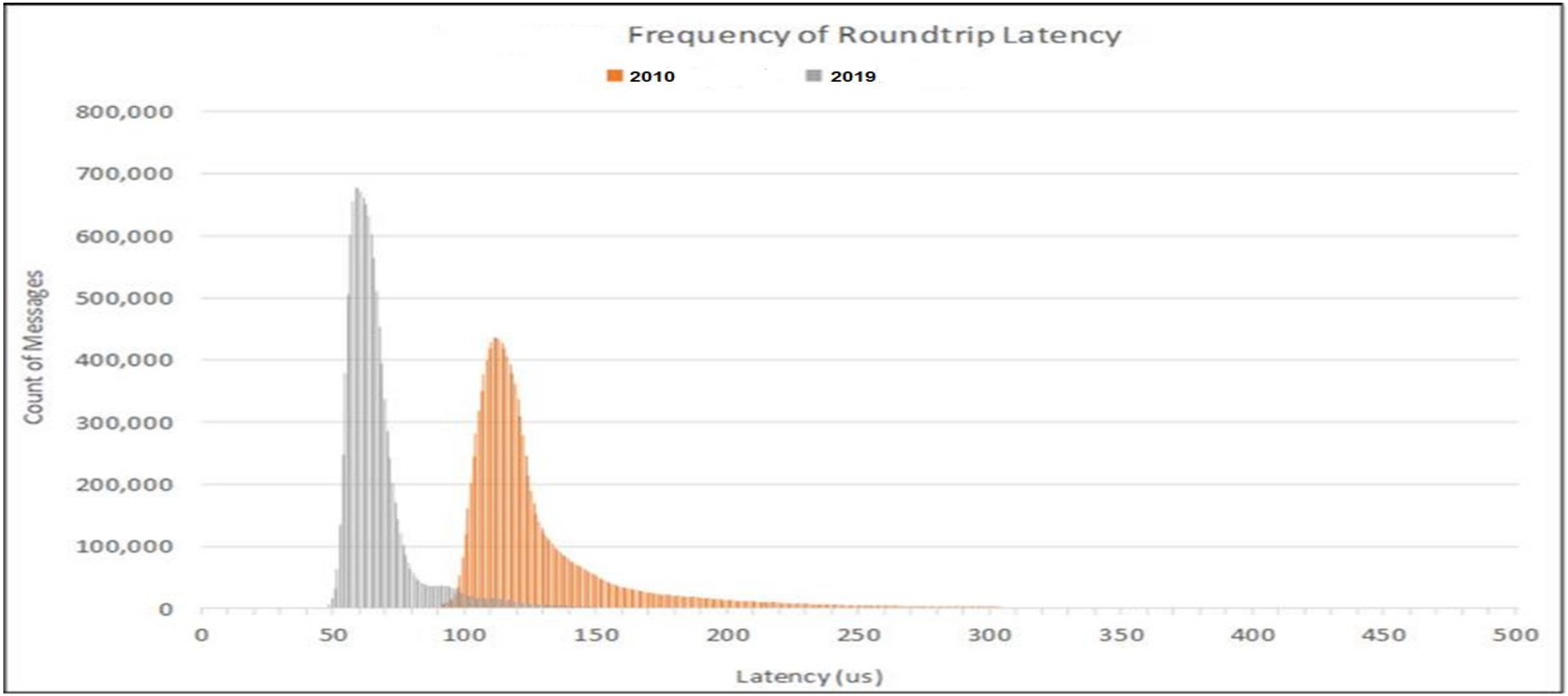
IB QDR VS EDR


InfiniBand Roadmap
INFINIBAND TRADE ASSOCIATION
Link Bandwidth per direction, Gb/s
10,000
1,000
100
10
QDR
FDR
EDR
HDR
NDR
XDR
12X
4X
1X
2X
168G
56G
14G
300G
100G
25G
600G
200G
100G
50G
1.2T
400G
200G
100G
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
©2018 InfiniBand® Trade Association

IB NETWORK AND SERVERS - YEAR 2019

- Infiniband Enhanced Data Rate (EDR) 100Gb/s
- IB Network cards: HPE InfiniBand EDR 100Gb 841QSFP28 Adapter (Mellanox ConnectX-5 OEM)
- IB Network Card Driver: OS OFED
- Servers: HPE DL380 Gen10 Intel Xeon Gold (Xeon V5 Skylake) based servers
- Operating System: Red Hat Enterprise Linux 7.4

LATENCY DISTRIBUTIONS ON SYSTEM UPGRADE



BENCHMARKING TOOLS

- **NwTool**

  Network Latency (Min, Avg , Max)

- **MCastNwTool**

  Multicast Jitter and Drops

- **Dtool**

  Disk Writes/Read Latency (Min, Avg, Max)

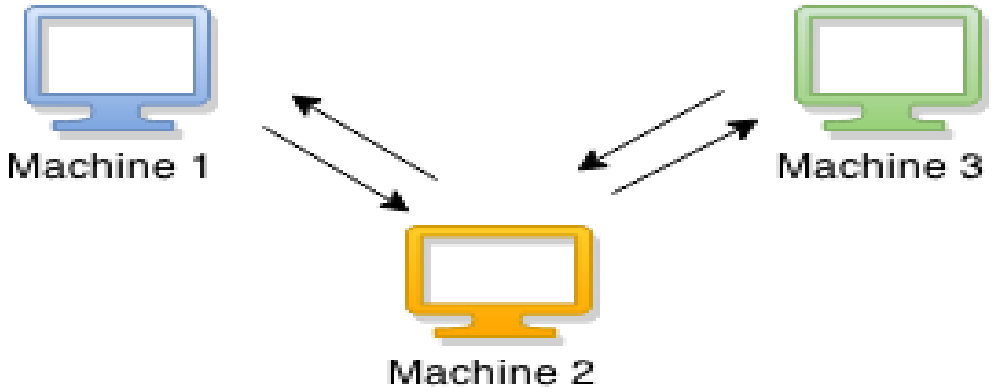
- **All-In-One Tool**

  All of above and application simulations

NWTOOL

NwTool Capabilities: ping-ping, message rate, different packet sizes, simulation of real inter-hop application behaviors.

Description: Responsible of tackling network related issues in TCP or RDMA communication. As an example; this enables running of three instances of the tool in three machines and connect them.




```
./NwTool -W - i 1 -p 23700 -s 1024 -r 10000 -L 1 -t 1
sizeof 32
No -i <stat interval> option specified, taking -i as 1
Switch file: 0
Test Duration: 1 minutes.
b_InstructFileSwitch 0
SendRes  =100us
SendCount=1
Thread was successfully entered to event loop: 13591:13591:<NwTool>
MHPC::Register() -> NwTool
void MHPCCallback::OnSetHPCMode(MI32 i32Mode) -> Thread:NwTool ID:13591 Mode:NONE State:STARTING
13591:Entered TIGHT select() for NwTool

OnConnection - FileSwitch0
Client connected from 10.25.90.31
[S]20190312130441.972    10.25.90.31 10000.0 msgs/s 80.6 Mb/s Min:  11 us Avg:  11 us Max:  17 us Buffered:0 Remain:0
[S]20190312130442.972    10.25.90.31 10000.0 msgs/s 80.6 Mb/s Min:  11 us Avg:  11 us Max:  19 us Buffered:0 Remain:0
[S]20190312130443.972    10.25.90.31 10000.0 msgs/s 80.6 Mb/s Min:  11 us Avg:  11 us Max:  19 us Buffered:0 Remain:0
[S]20190312130444.972    10.25.90.31 10000.0 msgs/s 80.6 Mb/s Min:  11 us Avg:  11 us Max:  20 us Buffered:0 Remain:0
[S]20190312130445.972    10.25.90.31 10000.0 msgs/s 80.6 Mb/s Min:  11 us Avg:  11 us Max:  19 us Buffered:0 Remain:0
[S]20190312130446.972    10.25.90.31 10000.0 msgs/s 80.6 Mb/s Min:  11 us Avg:  11 us Max:  18 us Buffered:0 Remain:0
[S]20190312130447.972    10.25.90.31 10000.0 msgs/s 80.6 Mb/s Min:  11 us Avg:  11 us Max:  15 us Buffered:0 Remain:0
[S]20190312130448.972    10.25.90.31 10000.0 msgs/s 80.6 Mb/s Min:  11 us Avg:  11 us Max:  15 us Buffered:0 Remain:0
```

Machine 1: Acts as a server and transmit packets periodically

Machine 2: Acts as an intermediary server what forward packets to machine 3 and at the same time replies back an acknowledgement to machine 1

Machine 3: Acts as a client of this client server environment and waits for the packets to receive from machine 2.

Measurements:

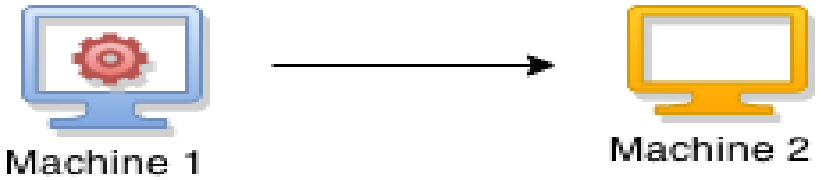
RTT (Round trip time) min, max and average, number of packets flown through the network, number of buffered packets

MCASTNWTOOL

Description:

- This tool is responsible of finding network packet jitter and losses in UDP communication.
- Tool contains two modes, one as a muticaster (Source) and other as a listener (Receiver).
- The tool is capable of publishing stats to standard out and to a file.

Measurements: Jitter, sent and received bytes




```
./MCastNwTool -S -K 1 -i 1 -I 239.1.1.1 -P 42222 -r 100000 -s 1000 -A 2 -t 2
1000
total number of threads 2
On Trace CreateThread B
3 : MCM:21:Listening to multicast 239.1.1.1:42222 [][INet]. iface:(null) receiver:0x67e240 cb:0x667f60 fd:3
On Trace Thread was successfully entered to event loop: 13851:13851:<MReceiver_42222>
On Trace MHPC::Register() -> MReceiver_42222
On Trace void MHPCCallback::OnSetHPCMode(MI32 i32Mode) -> Thread:MReceiver_42222 ID:13851 Mode:NONE
State:STARTING
3 : MCM:21:Listening to multicast 239.1.1.1:42223 [][INet]. iface:(null) receiver:0x7fe750000940 cb:0x65b5a0 fd:12
On Trace Thread was successfully entered to event loop: 13851:13852:<MReceiver_42223>
On Trace MHPC::Register() -> MReceiver_42223
On Trace void MHPCCallback::OnSetHPCMode(MI32 i32Mode) -> Thread:MReceiver_42223 ID:13852 Mode:NONE
State:STARTING
MReceiver_42222 - Msg Rate 23242/s Avg Jitter 1 us Packet Lose 0/s BandWidth 22697KB/S
MReceiver_42223 - Msg Rate 23256/s Avg Jitter 1 us Packet Lose 0/s BandWidth 22710KB/S
```

DTOOL

Description:

Measures disk performance of a system.

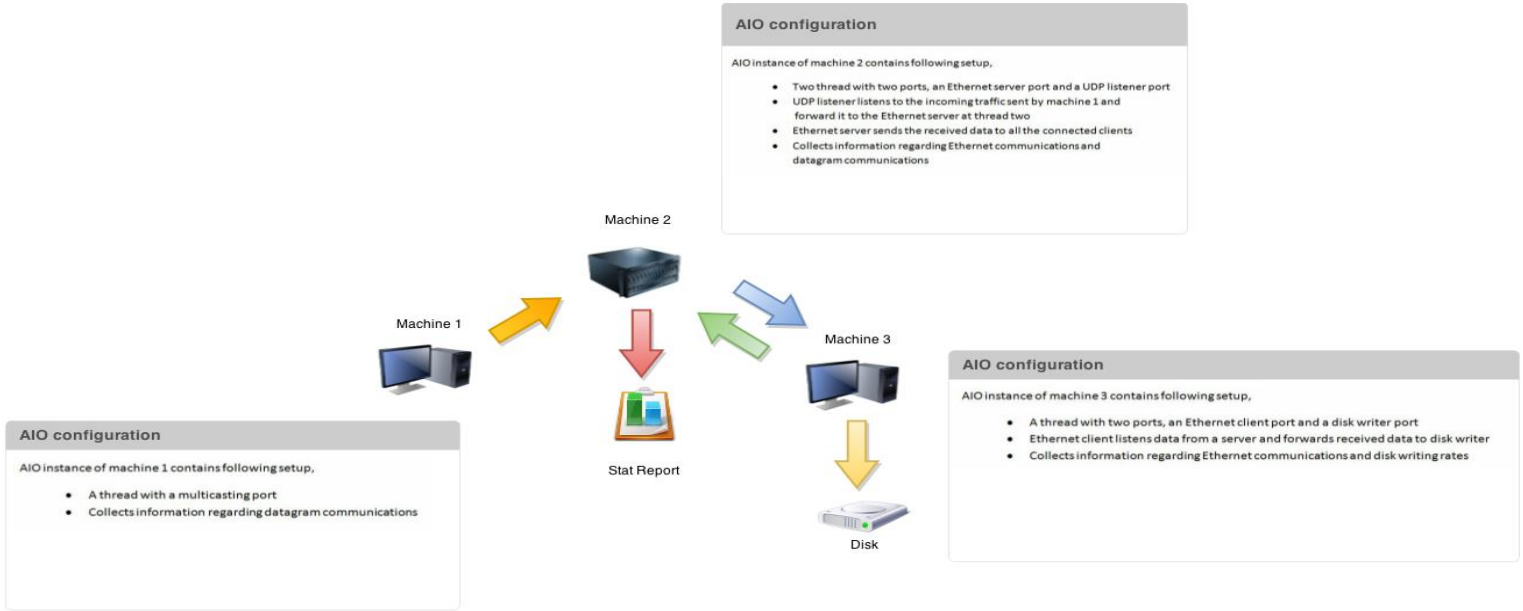
Test types: Read performance, Write performance.

Measurements: Min, max and average read and write latencies, number of bytes read or written in a second

```
baseline@itic21|centos7|3.10.0-957.5.1.el7:0:tools $ ./DTool -W -f /x01/baseline/test -t 5 -r 10000 -s 4096 -i 1 -P 1 A
System Tick = 100 usecs
z_FileName = /x01/baseline/test
i_Rate = 10000
i64_BlockSize = 4096
i_SyncInterval = 0
i64_TickInterval = 100
i_TicksPerIO = 1
i_NumOfIOsPerSlot = 1
Time            | EOF  | Max (us)| Min (us)     | Avg (us)      | Exceed(count) | MB/S          | File          Freq Distribution
051501.484      |0     |  69   |  5           |  6.9764       |  -            |  19.5295      |  /x01/baseline/test     Avg - 6.395152(90%) - 11.030303(10%)
051502.484      |0     |  36   |  5           |  7.0100       |  -            |  39.0593      |  /x01/baseline/test     Avg - 6.409759(90%) - 11.062791(10%)
051503.484      |0     |  34   |  5           |  6.9712       |  -            |  39.0680      |  /x01/baseline/test     Avg - 6.432798(90%) - 10.871393(10%)
051504.484      |0     |  51   |  6           |  7.9821       |  -            |  39.0668      |  /x01/baseline/test     Avg - 7.296496(90%) - 12.703470(10%)
051505.484      |0     |  58   |  6           |  8.0424       |  -            |  39.0650      |  /x01/baseline/test     Avg - 7.367125(90%) - 12.372869(10%)
051506.484      |0     |  40   |  6           |  7.9206       |  -            |  39.0661      |  /x01/baseline/test     Avg - 7.280347(90%) - 12.022963(10%)
051507.484      |0     |  56   |  6           |  8.0021       |  -            |  39.0649      |  /x01/baseline/test     Avg - 7.306933(90%) - 12.311735(10%)
051508.485      |0     |  32   |  6           |  7.9964       |  -            |  39.0671      |  /x01/baseline/test     Avg - 7.317802(90%) - 12.185796(10%)
051509.485      |0     |  30   |  6           |  7.9060       |  -            |  39.0654      |  /x01/baseline/test     Avg - 7.292722(90%) - 11.952888(10%)
051510.485      |0     |  31   |  6           |  7.9312       |  -            |  39.0662      |  /x01/baseline/test     Avg - 7.286440(90%) - 12.090909(10%)
051511.485      |0     |  29   |  6           |  7.9505       |  -            |  39.0659      |  /x01/baseline/test     Avg - 7.309910(90%) - 12.013216(10%)
051512.485      |0     |  30   |  6           |  7.9734       |  -            |  39.0658      |  /x01/baseline/test     Avg - 7.255909(90%) - 12.340893(10%)
051513.485      |0     |  30   |  6           |  7.8666       |  -            |  39.0651      |  /x01/baseline/test     Avg - 7.258670(90%) - 11.860712(10%)
051514.485      |0     |  35   |  6           |  7.9029       |  -            |  39.0652      |  /x01/baseline/test     Avg - 7.283477(90%) - 11.993916(10%)
051515.485      |0     |  35   |  6           |  7.8704       |  -            |  39.0647      |  /x01/baseline/test     Avg - 7.280794(90%) - 11.854926(10%)
051516.485      |0     |  30   |  6           |  7.8753       |  -            |  39.0661      |  /x01/baseline/test     Avg - 7.269944(90%) - 11.880427(10%)
051517.485      |0     |  35   |  6           |  7.9014       |  -            |  39.0646      |  /x01/baseline/test     Avg - 7.273408(90%) - 12.041762(10%)
051518.486      |0     |  205  |  6           |  7.8923       |  -            |  39.0616      |  /x01/baseline/test     Avg - 7.243259(90%) - 12.294163(10%)
051519.486      |0     |  39   |  5           |  6.9757       |  -            |  39.0650      |  /x01/baseline/test     Avg - 6.396796(90%) - 10.991270(10%)
051520.487      |0     |  43   |  5           |  7.7587       |  -            |  39.0169      |  /x01/baseline/test     Avg - 7.147125(90%) - 12.401203(10%)
051521.487      |0     |  30   |  6           |  7.8848       |  -            |  39.0652      |  /x01/baseline/test     Avg - 7.273177(90%) - 11.917236(10%)
051522.487      |0     |  33   |  6           |  8.0038       |  -            |  39.0684      |  /x01/baseline/test     Avg - 7.278633(90%) - 12.363955(10%)
051523.487      |0     |  31   |  6           |  7.9009       |  -            |  39.0650      |  /x01/baseline/test     Avg - 7.283592(90%) - 11.977947(10%)
```

ALL-IN-ONE (AIO) TOOL

- AIO (All in One) tool is an application simulation tool, which enables baseline testing in an application environment without installing the real system.
- This provide an abstract illustration of the hardware performance without the time spent on installation in configuration.
- AIO can be used for run simple test as well as advance tests.
- Users can setup the application depending on their order flow.



ALL-IN-ONE (AIO) TOOL

```
root@ttic21|centos7|3.10.0-957.5.1.el7:0:tticbench # LD_LIBRARY_PATH=$LD_LIBRARY_PATH:lib PATH=$PATH:bin AIO cnf/EthClient.xml @bind=2
 @IP=10.10.10.20 @port=2377 @proto=Eth @size=1600 @rate=60000 @count=1 @stat_file=kkk.log
Millennium Release Installation System (AIO-1.1)
Copyright (C) 2015 Millennium Information Technologies
All rights reserved

XML Version: 1.1
-----------------------------------
       Tool Arrangement
-----------------------------------
++++++++++ Thread ++++++++++++
  Thread :1
  Polling :true
  Pool Size :500000
  CPU ID :2
  CPU priority :-1
  Stat Forwarding :kkk.log
  Stat Forwarding Time:0
  Stat Summary Interval :1
  Stat Port Threading :false
       | +-> Wrapper 1
              | +-> Module 1
                Type: P2PClientList
                Connection type: Eth
                Ip Address: 10.10.10.20
                Rate: 60000
                Port: 2377
                Ping count: 1
                Packet size: 1600
                Reply Mode:
                Enable Echo Mode: FALSE
                Enable Track Origin: FALSE
                Client count: 1
                Start Id: 1
                Delta Rate: 1
                ModuleSignature: m01s
Error: Error: [10.10.10.20:2377] HPCOptions:1. [Option:0] Error: [INet] 10.10.10.20:2377. Connection refused. getpeername() failed. Tr
ansport endpoint is not connected.
-----------------------------------
AIO configured successfully, instance is starting....
=====================================================
Thread is running.........
Avg: -1.000      Max: -1         Min: -1         Samples: 0
Avg: 1316.867    Max: 17742      Min: -158350    Samples: 56988
Avg: 349.507     Max: 7944       Min: 53         Samples: 59984
Avg: 129.358     Max: 997        Min: 55         Samples: 59382
Avg: 244.822     Max: 5657       Min: 57         Samples: 59614
Avg: 128.809     Max: 1301       Min: 53         Samples: 58825
Avg: 129.182     Max: 1350       Min: 53         Samples: 59145
Avg: 244.314     Max: 6817       Min: 57         Samples: 59453
Avg: 127.297     Max: 1298       Min: 53         Samples: 59028
Avg: 128.439     Max: 871        Min: 57         Samples: 59081
Avg: 130.853     Max: 823        Min: 56         Samples: 58351
Avg: 129.422     Max: 629        Min: 54         Samples: 59134
Avg: 240.061     Max: 5654       Min: 54         Samples: 59780
Avg: 130.149     Max: 1381       Min: 54         Samples: 59148
Avg: 128.247     Max: 601        Min: 56         Samples: 59176
Avg: 130.534     Max: 1116       Min: 58         Samples: 58981
Avg: 126.735     Max: 642        Min: 54         Samples: 58979
Avg: 130.390     Max: 1092       Min: 56         Samples: 58606
Avg: 129.307     Max: 719        Min: 52         Samples: 58674
Avg: 129.981     Max: 664        Min: 51         Samples: 58679
Avg: 229.563     Max: 5420       Min: 53         Samples: 59142
Avg: 134.159     Max: 2336       Min: 55         Samples: 58429
Avg: 130.452     Max: 727        Min: 56         Samples: 58768
Avg: 137.994     Max: 905        Min: 56         Samples: 59079
Avg: 129.848     Max: 614        Min: 56         Samples: 59131
Avg: 130.299     Max: 1217       Min: 54         Samples: 58674
Avg: 131.578     Max: 738        Min: 55         Samples: 58917
Avg: 134.405     Max: 1085       Min: 58         Samples: 59008
Avg: 131.431     Max: 740        Min: 54         Samples: 59170
Avg: 133.388     Max: 628        Min: 57         Samples: 59274
```

```
root@ttic20|centos7|3.10.0-957.5.1.el7:0:tticbench # LD_LIBRARY_PATH=$LD_LIBRARY_PATH:lib PATH=$PATH:. ./AIO EthServer.xml @bind=2 @por
t=2377 @proto=Eth
Millennium Release Installation System (AIO-1.1)
Copyright (C) 2015 Millennium Information Technologies
All rights reserved

XML Version: 1.1
-----------------------------------
       Tool Arrangement
-----------------------------------
++++++++++ Thread ++++++++++++
  Thread :1
  Polling :true
  Pool Size :500000
  CPU ID :2
  CPU priority :-1
  Stat Forwarding :EthServer.log
  Stat Forwarding Time:0
  Stat Summary Interval :1
  Stat Port Threading :false
       | +-> Wrapper 1
              | +-> Module 1
                Type: P2PServer
                Connection type: Eth
                Ip Address:
                Rate: 0
                Port: 2377
                Ping count: 1
                Packet size: 200
                Reply Mode: FULL
                Enable Echo Mode: true
                Enable Track Origin: FALSE
                ModuleSignature: T1
-----------------------------------
AIO configured successfully, instance is starting....
=====================================================
Thread is running.........
[ EthS: New connection from: 10.10.10.21 ]
Recv rate: 1600.000      Bytes: 172134400        Samples: 107584
Sent rate: 1599.000      Bytes: 162193800        Samples: 101372
Recv rate: 1600.000      Bytes: 94603200         Samples: 59127
Sent rate: 1600.000      Bytes: 94603200         Samples: 59127
Recv rate: 1600.000      Bytes: 95652800         Samples: 59783
Sent rate: 1600.000      Bytes: 95521600         Samples: 59701
Recv rate: 1600.000      Bytes: 94206400         Samples: 58879
Sent rate: 1600.000      Bytes: 94206400         Samples: 58879
Recv rate: 1600.000      Bytes: 94286400         Samples: 58929
Sent rate: 1600.000      Bytes: 94286400         Samples: 58929
Recv rate: 1600.000      Bytes: 95188800         Samples: 59493
Sent rate: 1600.000      Bytes: 95067200         Samples: 59417
Recv rate: 1600.000      Bytes: 94601600         Samples: 59126
Sent rate: 1600.000      Bytes: 94601600         Samples: 59126
Recv rate: 1600.000      Bytes: 94324800         Samples: 58953
Sent rate: 1600.000      Bytes: 94324800         Samples: 58953
Recv rate: 1600.000      Bytes: 93539200         Samples: 58462
Sent rate: 1600.000      Bytes: 93539200         Samples: 58462
Recv rate: 1600.000      Bytes: 94414400         Samples: 59009
Sent rate: 1600.000      Bytes: 94414400         Samples: 59009
Recv rate: 1600.000      Bytes: 95713600         Samples: 59821
Sent rate: 1600.000      Bytes: 95582400         Samples: 59739
Recv rate: 1600.000      Bytes: 94766400         Samples: 59229
Sent rate: 1600.000      Bytes: 94766400         Samples: 59229
Recv rate: 1600.000      Bytes: 94579200         Samples: 59112
Sent rate: 1600.000      Bytes: 94579200         Samples: 59112
Recv rate: 1600.000      Bytes: 94312000         Samples: 58945
Sent rate: 1600.000      Bytes: 94312000         Samples: 58945
Recv rate: 1600.000      Bytes: 94616000         Samples: 59135
Sent rate: 1600.000      Bytes: 94616000         Samples: 59135
Recv rate: 1600.000      Bytes: 93750400         Samples: 58594
Sent rate: 1600.000      Bytes: 93750400         Samples: 58594
Recv rate: 1600.000      Bytes: 93860800         Samples: 58663
```

RDMA VS NON-RDMA EVALUATION

RDMA Categories:

- Infiniband (IB)
- Remote Direct Memory Access over Converged Ethernet (RoCE)

Non RDMA Categories:

- RAW Ethernet
- Solarflare Openonload
- Mellanox Voltaire Message Accelerator (VMA)
- Intel Omnipath
- Exablaze Exasock
- Myricom DBLRUN

DATA COLLECTION – VARIABLES AND OUTPUT

Two main variables:

- Packet Sizes - 100/500/800/1000 Bytes
- Message Rates - 100/ 500/ 1,000/ 5,000/ 10,000/ 25,000/ 50,000/ 100,000/ 250,000/ 500,000/ 750,000/ 1,000,000

Main Outputs/Results:

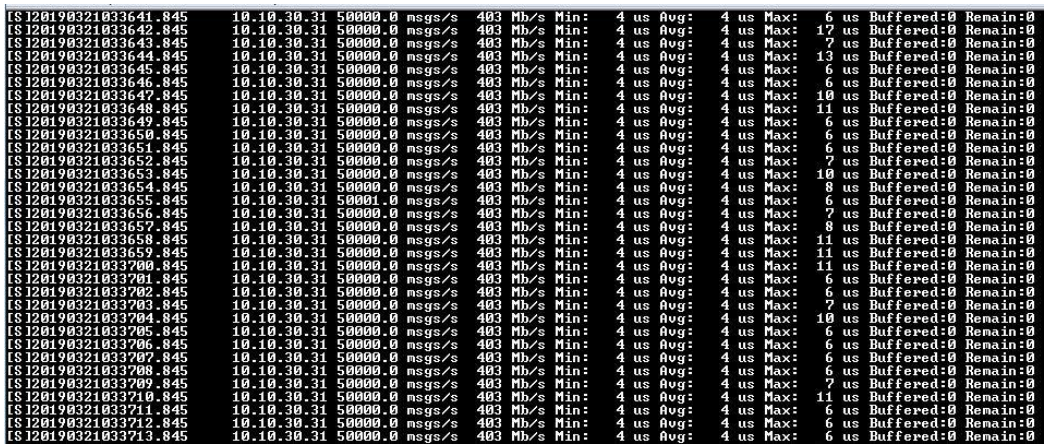
- Minimum Two-Way Delay per second (min)
- Average Two-Way Delay per second (avg)
- Maximum Two-Way Delay per second (max)
- Percentiles to filter max (90%, 95%, 99.5%)

DATA COLLECTION – OUTPUT AND RESULTS

Example of output Log Entry:

[S]20150729075828.815 100.1 msgs/s 104 Kb/s

Min: 2us Avg: 2us Max:13us Buffered:0 Remain:0

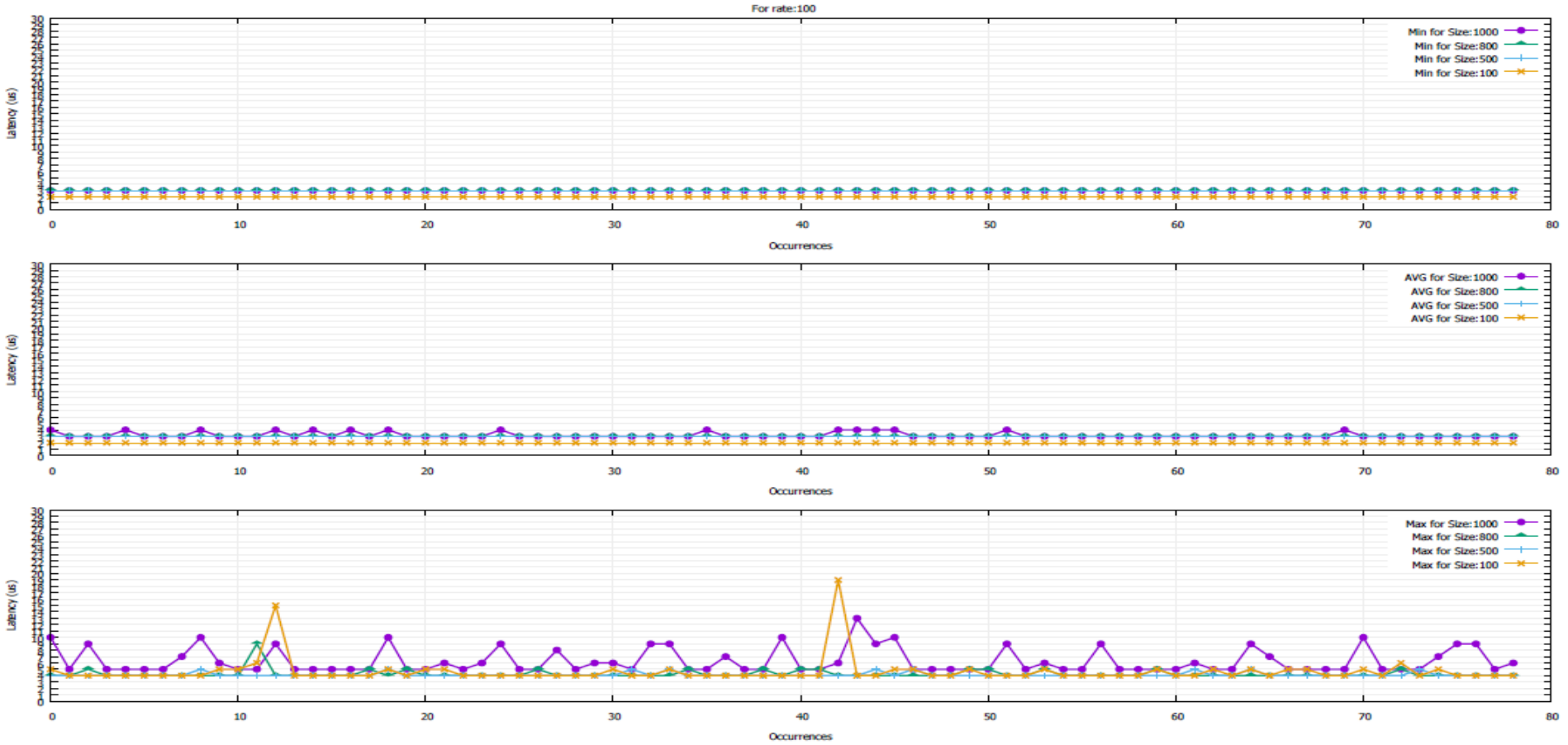


Simplified Tabulation:

A. *Infiniband*

Test Tool	Rate (pps)	Size (Bytes)	min(µs)	avg(µs)	max(µs)	max 90%	max 95%	max 99.5%
NWTool	100	100	2.00	2.00	3.00	3.00	3.00	3.00
NWTool	100	500	2.00	2.66	5.00	4.00	4.00	5.00
NWTool	100	800	3.00	3.00	5.00	4.00	4.00	5.00
NWTool	100	1000	3.00	3.00	10.00	9.00	9.00	10.00
NWTool	500	100	2.00	2.00	5.00	3.00	3.00	5.00
NWTool	500	500	2.00	2.85	6.00	4.00	5.00	6.00
NWTool	500	800	3.00	3.00	4.00	4.00	4.00	4.00
NWTool	500	1000	3.00	3.00	11.00	10.00	11.00	11.00
NWTool	5000	100	2.00	2.00	8.00	5.00	7.00	8.00

PERFORMANCE COMPARISON GRAPHS – AN EXAMPLE



DISCUSSION OF RESULTS (SUMMERY)

Expected latency boundaries :

- Minimum Level of Average RT Latency per Second (**min of avg**) =5us
- Maximum Level of Average RT Latency per Second (**max of avg**) =7us
- Minimum Level of Maximum RT Latency per Second (**min of max**)=75us
- Maximum Level of Maximum RT Latency per Second (**max of max**)=100us

Successful Categories in the full range of tests:

- **IB100G-EDR**
- **IB56G-FDR**
- **RoCE 100GE**
- **RoCE 40GE**
- **Solarflare OpenOnload 40GE**

DISCUSSION OF RESULTS (SUMMERY) CONT.

Break Even/Melting points of remaining categories (Descending order)

- **RoCE 10GE**: 1,000,000 messages/second at the packet size of 500bytes.
- **Openonload 10GE**: 1,000,000 messages/second at the packet size of 500bytes.
- **MLX Ethernet 40GE**: 750,000 messages/second at the packet size of 100bytes.
- **SFN Ethernet 40GE**: 750,000 messages/second at the packet size of 100bytes.
- **MLX Ethernet 10GE**: 750,000 messages/second at the packet size of 100bytes.
- **IPoIB (EDR)**: 500,000 messages/second at the packet size of 500bytes.
- **IPoIB (FDR)**: 500,000 messages/second at the packet size of 500bytes.
- **SFN Ethernet 10GE**: 500,000 messages/second at the packet size of 100bytes.

RESULTS SUMMERY

Technology/Functional Benefits	**RDMA**	**Openonload**	**Ethernet**	**IPoIB**
Low Latency Benefits	Highest	High	Medium	Lowest
Low Jitter Benefits	Highest	High	Low	Lowest
Scalability Benefits	Highest	High	Low	Lowest
Interoperability Benefits	Lowest	High	Highest	High

- **Scale: Highest >High> >Medium> >Low>Lowest**

FURTHER EVALUATIONS

- **Further testing with more variants**:
  - Intel Omnipath
  - Cisco User Space Network Interface Card (us NIC)/Libfabric and MPI
  - Voltaire Message Accelerator (VMA)
  - Exablaze Exasock
  - Myricom DBLRUN
  - Layer-1 Switches (Metamako/Exablaze/xCelor)
  - White Rabbit
- **Further testing on different hardware platforms (Eg: IBM Power) and operating systems.**

QUESTIONS FOR THE OFA COMMUNITY

- **Network level HCA failover mechanisms for IB**
  - -APM supports only between two ports of the same NIC
- **Teaming Support for IB**
  - -Teaming project not aimed support for IB
- **Active-Active bonding support for IB**
- **Time synchronization between IB only hosts**
  - -PTP support
- **IB packet capturing and monitoring at (sub) nano-second precision**
- **Reliable multicast on Ethernet**
- **IB stack troubleshooting difficulties**
- **Exponential multi-session latency problem of Onload techniques**

15th ANNUAL WORKSHOP 2019

THANK YOU

Sampath Tilakumara – sampath@lseg.com / Indika Prasad Kumara – indikap@lseg.com

Millennium IT Software (LSEG Technology)

March 21, 2019