Chelsio Communications

ARASTRA

10GbE MPI Benchmarking Report

- **Chelsio 10GbE RDMA NIC (RNIC)**
- **Arastra DX7124-S Switch**

Executive Summary

IN TESTS conducted at the Chelsio facility, results demonstrate successful interoperability between Chelsio's latest SFP+ 10GbE adapter and Arastra's new 10GbE SFP+ switch. Performance results show that the application can sustain line rate and latency numbers are ~7 µsec.

Chelsio Communications • www.chelsio.com • sales@chelsio.com • +1-408-962-3600

Test Configuration

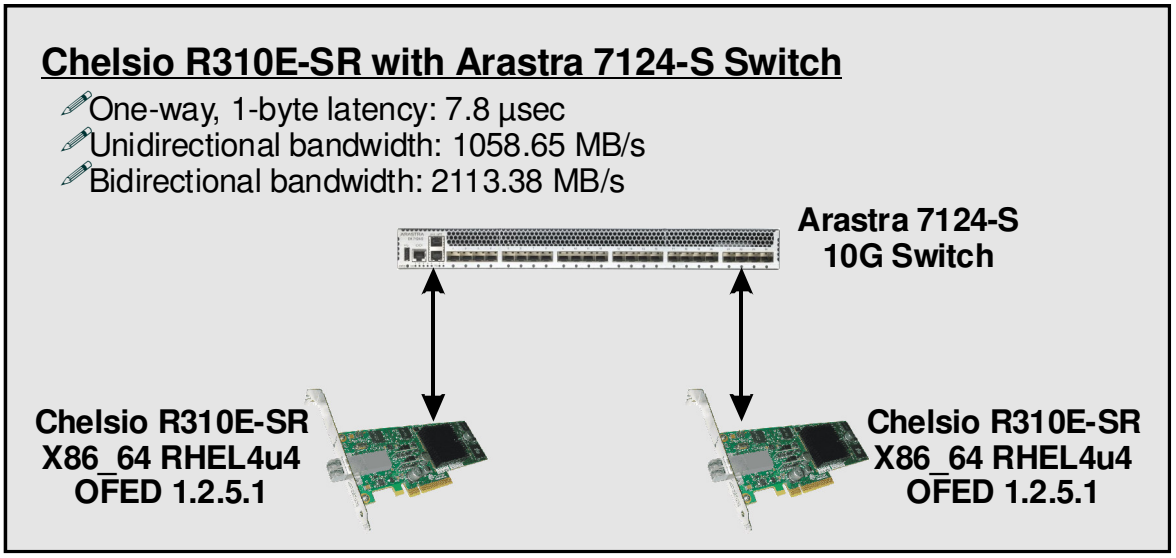
Host Systems

System1:

- CPU: 2 x dual-core Intel Xeon® 5160 @ 3 GHz
- Root Complex: Intel 5000 Chipset
- Memory: 8GB
- OS: Linux 2.6.9-42
- OFED 1.2.5.1
- x86_64

System2:

- CPU: 1 x dual-core Intel Xeon® 3070 @ 2.66 GHz
- Root Complex: Intel 5000 Chipset
- Memory: 8GB
- OS: Linux 2.6.9-42
- OFED 1.2.5.1
- x86_64



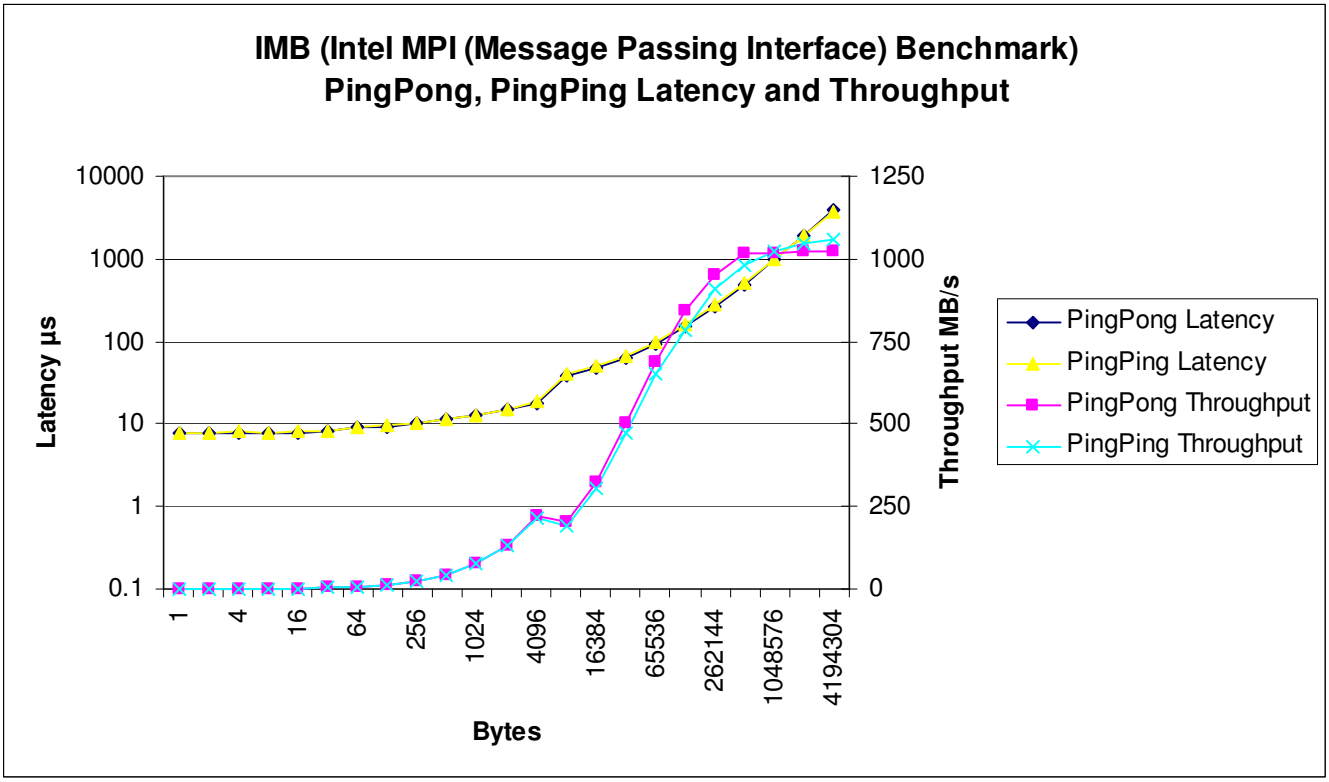
Performance Measurement Software

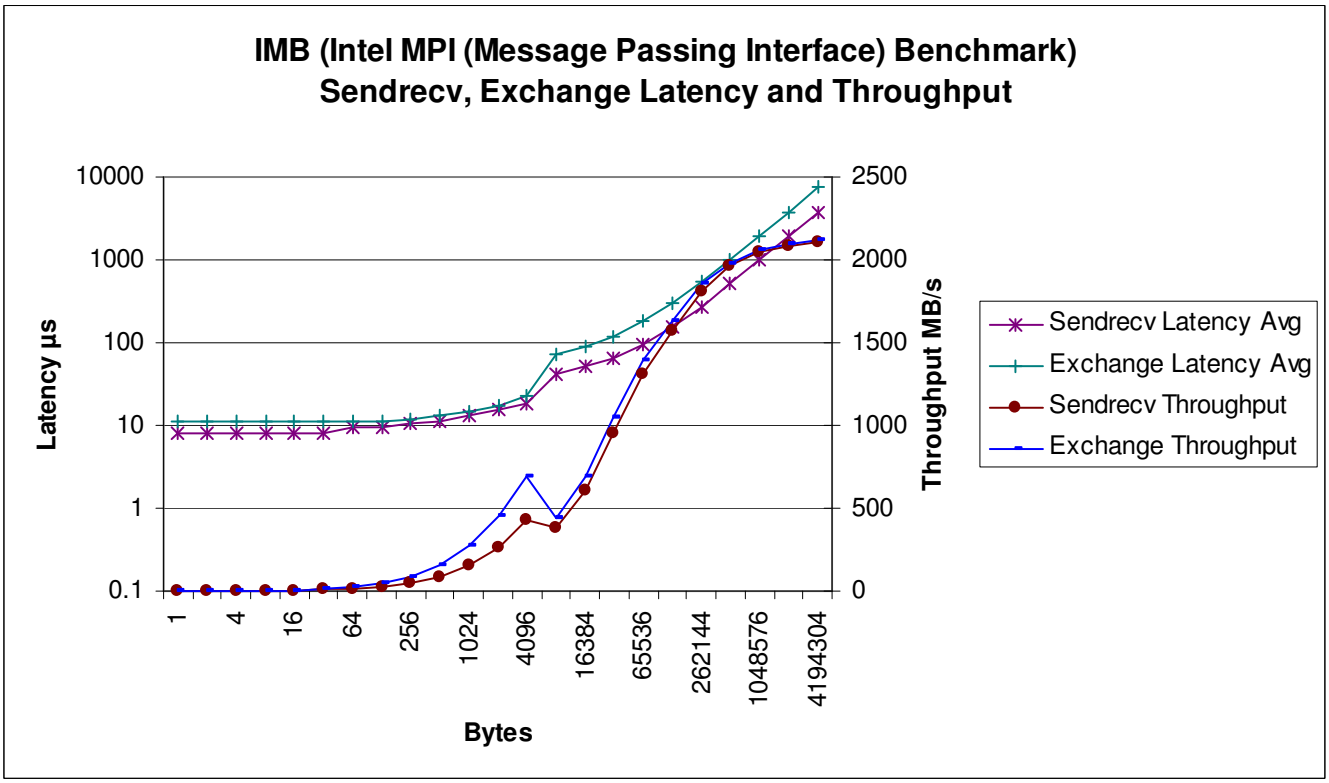
- Test tool: Intel® MPI Benchmark Suite V2.3

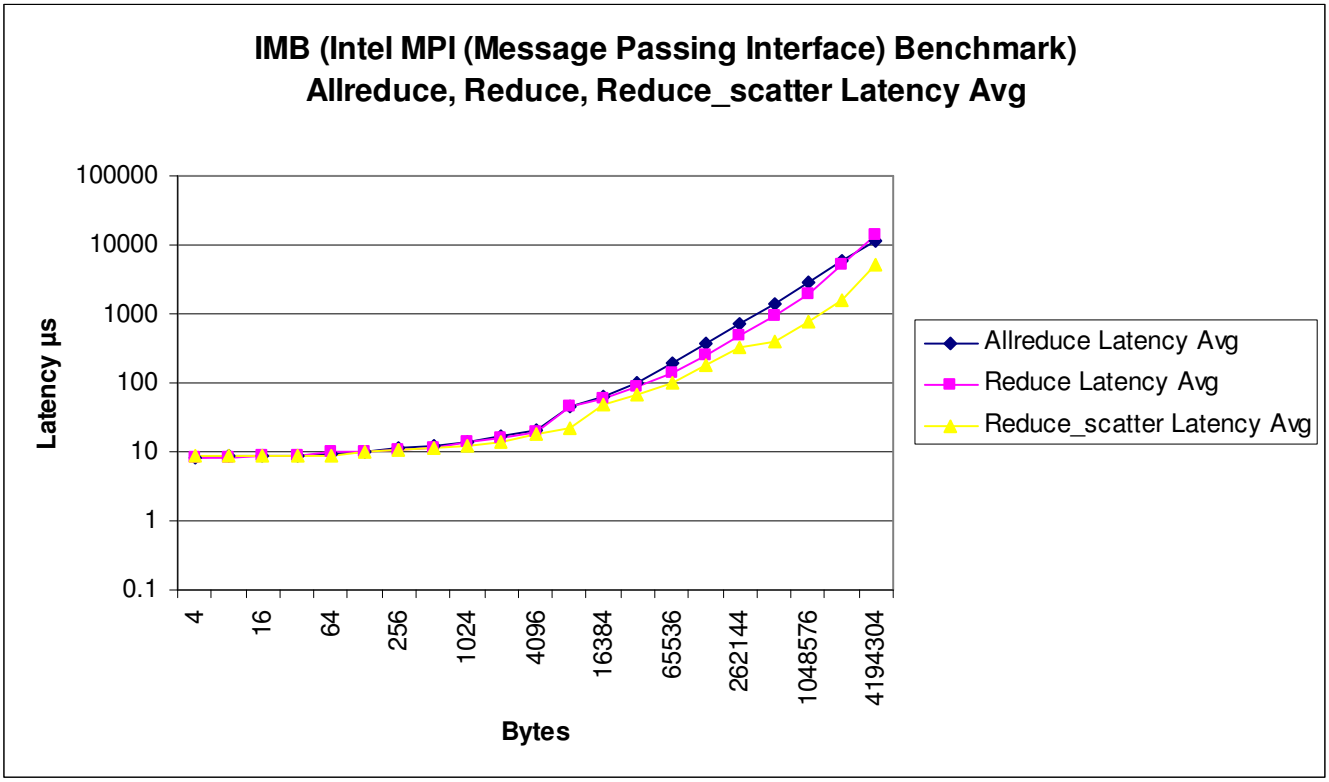
Results

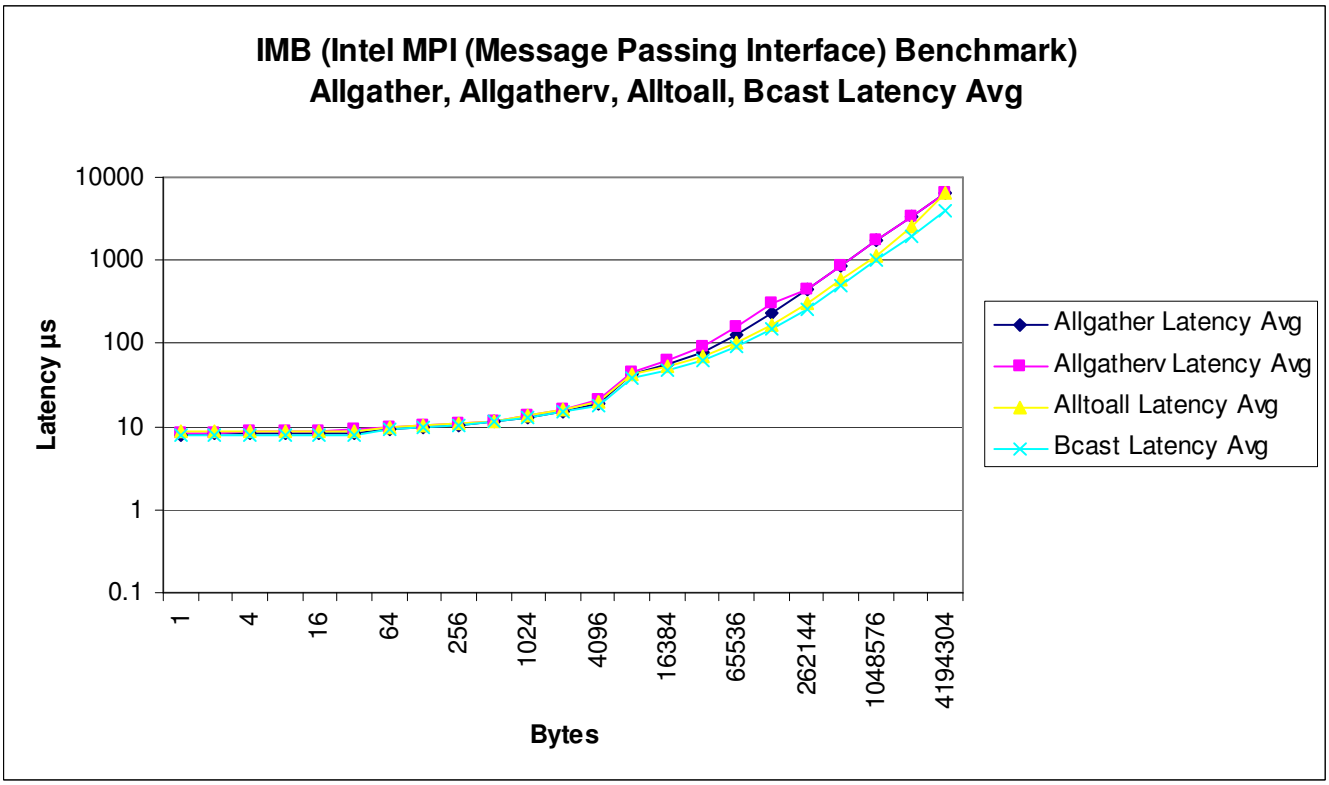
	Latency (µsec)	*Throughput (MB/s)*
Ping-Pong	7.7	1025.13
Ping-Ping	7.8	1058.65
SendRecv	8.0	2109.43
Exchange	11.1	2113.38
Allreduce	8.4	
Reduce	8.4	
Reduce Scatter	8.7	
Allgather	8.0	
Allgatherv	8.4	
Alltoall	8.5	
Bcast	7.9	

10/18/07


IMB (Intel MPI (Message Passing Interface) Benchmark)
PingPong, PingPing Latency and Throughput
Latency µs
Throughput MB/s
Bytes
10000
1000
100
10
1
0.1
1250
1000
750
500
250
0
1
4
16
64
256
1024
4096
16384
65536
262144
1048576
4194304
PingPong Latency
PingPing Latency
PingPong Throughput
PingPing Throughput


IMB (Intel MPI (Message Passing Interface) Benchmark)
Sendrecv, Exchange Latency and Throughput
Latency µs
Throughput MB/s
Bytes
10000
1000
100
10
1
0.1
2500
2000
1500
1000
500
0
1
4
16
64
256
1024
4096
16384
65536
262144
1048576
4194304
Sendrecv Latency Avg
Exchange Latency Avg
Sendrecv Throughput
Exchange Throughput


IMB (Intel MPI (Message Passing Interface) Benchmark)
Allreduce, Reduce, Reduce_scatter Latency Avg
Latency µs
100000
10000
1000
100
10
1
0.1
4
16
64
256
1024
4096
16384
65536
262144
1048576
4194304
Bytes
Allreduce Latency Avg
Reduce Latency Avg
Reduce_scatter Latency Avg


IMB (Intel MPI (Message Passing Interface) Benchmark)
Allgather, Allgatherv, Alltoall, Bcast Latency Avg
Latency µs
10000
1000
100
10
1
0.1
1
4
16
64
256
1024
4096
16384
65536
262144
1048576
4194304
Bytes
Allgather Latency Avg
Allgatherv Latency Avg
Alltoall Latency Avg
Bcast Latency Avg

Test Notes

- **PingPing:** Measures the startup (latency) and throughput of a single message sent between two processes, with each node issuing the commands concurrently.
- **Sendrecv:** Creates a chain of sending and receiving between upstream and downstream neighbors.
- **Exchange:** Creates a chain of sending and receiving between upstream and downstream neighbors, with each node issuing the send/receive commands concurrently.
- **Allreduce:** Measures the time to perform the allreduce collective operator where each process in a group owns a vector whose elements are combined with the corresponding elements owned by other processes, producing a new vector of the same length.
- **Reduce:** Measures the time to perform the allreduce collective operator where each process in a group owns a vector whose elements are combined with the corresponding elements owned by other processes, producing a new vector of the same length, changing the root of the process cyclically.
- **Reduce_scatter:** Measures the time to perform the allreduce collective operator, split between all processes, where each process in a group owns a vector whose elements are combined with the corresponding elements owned by other processes, producing a new vector of the same length.
- **Allgather:** Measures the time that it sends X bytes and gathers X*processes bytes.
- **Allgatherv:** Measures the time that it sends X bytes and gathers X*(#processes) bytes, and sees if MPI produces more overhead due to the more complicated situation.
- **Alltoall:** Measures the time that it takes for every process to input X*(#processes) bytes (X for each process) and receives X*(#processes) bytes (X for each process).
- **Bcast:** Measures the time that it takes a root process to broadcast X bytes to all while the root of the operation is changed cyclically.

MMVIII-0304