White Paper

Cisco usNIC Performance

Author

Ven Immani (TME, SAVTG)

Contents

Key Findings

- Cisco VIC with usNIC technology achieves 2.13 microseconds ping-pong latency using Open MPI across a Cisco Nexus 3548 Switch.
- Cisco VIC with usNIC technology achieves 1168 MBps of MPI ping-pong throughput.
- Cisco VIC with usNIC technology achieves 2336 MBps of MPI Exchange and MPI SendRecv throughput.

Introduction

With the advent of highly dense multicore compute systems, the need for low-latency network communication between compute systems has become paramount. In addition, the cost of building such compute cluster systems, as well as the cost of managing them, has become a major consideration in the design and deployment of these systems.

This white paper presents a simple introduction to Cisco® user-space network interface card (usNIC) technology and also describes performance results using the Intel MPI Benchmark (IMB).

Cisco usNIC

Cisco usNIC is a low-latency interface on top of Cisco UCS® Virtual Interface Card (VIC) 1225. The interface consists of a set of software libraries that enable a data path bypassing the kernel. Cisco UCS VIC 1225 is a converged network adapter (CNA) that supports Ethernet NICs and Fibre Channel host bus adapters (HBAs).

Cisco usNIC technology is a viable approach, since it enables ultra-low-latency network communication between nodes. As such, it can be deployed for various latency-dependent tasks, high-performance computing (HPC) clusters being one of them.

System under Test Configuration

System Hardware and Software

Network Adapter
Hardware: Cisco UCS VIC 1225 CNA

Firmware: 2.1 (2.127)

Driver software: enic-2.1.1.47

usNIC software:

- kmod-usnic_verbs-1.0.0.72-1
- libusnic_verbs-1.0.0.71-1
- openmpi-cisco-1.6.5cisco1.0.0.71-1

Server System

Hardware: Cisco UCS C220 M3 Rack Server with Intel® E5-2690 processor with 1600-MHz DDR3 memory

Firmware: Cisco UCS C-Series Software, Release 1.5 (2)

Network Switch

Hardware: Cisco Nexus® 3548 Switch

Software: Release 6.0 (2) A1(1a)

BIOS Settings

The following BIOS settings were used for this performance testing:

CPU Configuration

Intel Hyper-Threading technology: **Disabled**

Number of enabled cores: **All**

Execute disable: **Disabled**

Intel VT: **Enabled**

Intel VT-d: **Enabled**

Intel VT-d coherency support: **Enabled**

Intel VT-d ATS support: **Enabled**

CPU performance: **HPC**

Hardware prefetcher: **Enabled**

Adjacent cache line prefetcher: **Enabled**

DCU streamer prefetch: **Enabled**

DCU IP prefetcher: **Enabled**

Power technology: **Custom**

Enhanced Intel SpeedStep® technology: **Enabled**

Intel Turbo Boost technology: **Enabled**

Processor power state C6: **Disabled**

Processor power state C1 enhanced: **Disabled**

Frequency floor override: **Disabled**

P-STATE coordination: **HW_ALL**

Energy performance: **Performance**

Memory Configuration

Select memory RAS: **Maximum performance**

DRAM clock throttling: **Performance**

NUMA: **Enabled**

Low-voltage DDR mode: **Performance mode**

DRAM refresh rate: **2x**

Channel interleaving: **Auto**

Rank interleaving: **Auto**

Patrol scrub: **Disabled**

Demand scrub: **Enabled**

Altitude: **300M**

Other settings were left at default values.

Operating System Settings

Kernel Parameters

The following kernel parameter enables support for the Intel I/O memory management unit (IOMMU): **intel_iommu=on**

The following parameter turns off the Intel CPU idle driver: **intel_idle.max_cstate=0**

The following kernel parameter explicitly disables the CPU C1 and C1E states: **idle=poll**

Note: Using **the idle=poll** kernel parameter can result in increased power consumption. Use with caution.

The above kernel parameters were configured in the**/etc/grub.conf** file.

CPU Governor

The OS CPU governor was configured for "performance." In the file**/etc/sysconfig/cpuspeed**, the "GOVERNOR" variable was set to "performance."

GOVERNOR=performance

Other

The test nodes were also configured to operate in runlevel 3 to avoid unnecessary background processes.

In the file **/etc/inittab**, the following line replaced the OS default:

id:3:initdefault

SELinux was disabled.

In the **file/etc/selinux/config**, the following line replaced the OS default:

SELINUX=disabled

Networking Settings

Adapter Configuration

The following vNIC adapter configuration was used:

MTU 9000

Number of VFs instances = 16

Interrupt coalescing timer = 0

The vNIC is directly configured from the onboard Cisco Integrated Management Controller (IMC).

Network Switch Configuration

The Cisco Nexus 3548 was configured with the following settings:

Flow control: **On**

No Drop mode: **Enabled**

Pause: **Enabled**

Network MTU: **9216**

WARP mode: **Enabled**

To enable flow control on the Nexus 3548:

```
configure terminal
interface ethernet 1/1-48
flowcontrol receive on
flowcontrol send on
exit
```

To enable No Drop mode and Pause:

```
configure terminal
class-map type network-qos class1
match qos-group 1
policy-map type network-qos my_network_policy
class type network-qos class1
pause no-drop
system qos
service-policy type network-qos my_network_policy
show running ipqos

configure terminal
class-map type qos class1
match cos 2
policy-map type qos my_qos_policy
class type qos class1
set qos-group 1
system qos
```

```
service-policy type qos input my_qos_policy
```

To enable MTU 9216:

```
configure terminal
policy-map type network-qos jumbo
class type network-qos class-default
mtu 9216
system qos
service-policy type network-qos jumbo
```

To enable WARP mode:

```
configure terminal
hardware profile forwarding-mode warp
copy running-config startup-config
reload
```

Note 1: The above configuration is specific to the system under test presented here and may not be directly applicable to all use cases. Please consult your local network administrator or refer to the Cisco Nexus 3548 Command Reference (see the link in the appendix) for more information.

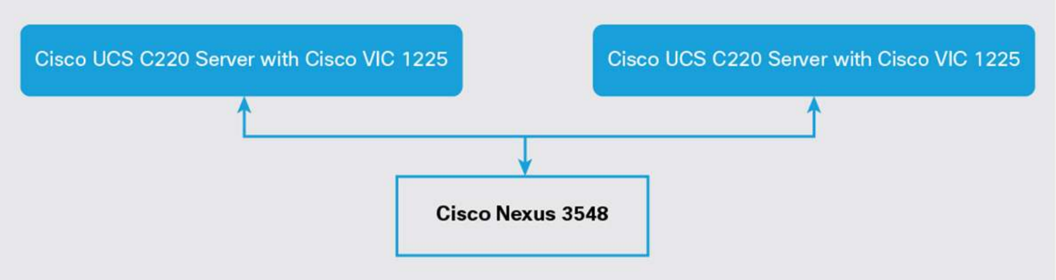
Note 2: The above configuration also specifies send/recv flow control and No Drop mode enabled. This prevents the switch from dropping packets on the network with a combined use of port buffer management and network pause. These settings are specific to the test. In some application instances, it may be optimal not to enable send/recv flow control and 'no drop' mode.

Network Topology

Two nodes were connected to a single Cisco Nexus 3548 Switch. The Cisco Nexus 3548 is an ultra-low-latency-capable switch from Cisco that is well suited for low-latency network messaging. For more details, refer to the product page link for the switch in the appendix.

Figure 1 shows the network topology that was used.

Figure 1. Network Topology


Cisco UCS C220 Server with Cisco VIC 1225
Cisco UCS C220 Server with Cisco VIC 1225
Cisco Nexus 3548

Benchmarking Software

NetPIPE version 3.7.1 was used for testing point-to-point latency and for a throughput comparison between Cisco usNIC and Kernel TCP.

Intel MPI Benchmarks (IMB) version 3.2.4 was used for testing. Refer to the links in the appendix for more information about this software.

Performance Results

Point-to-Point Benchmarks

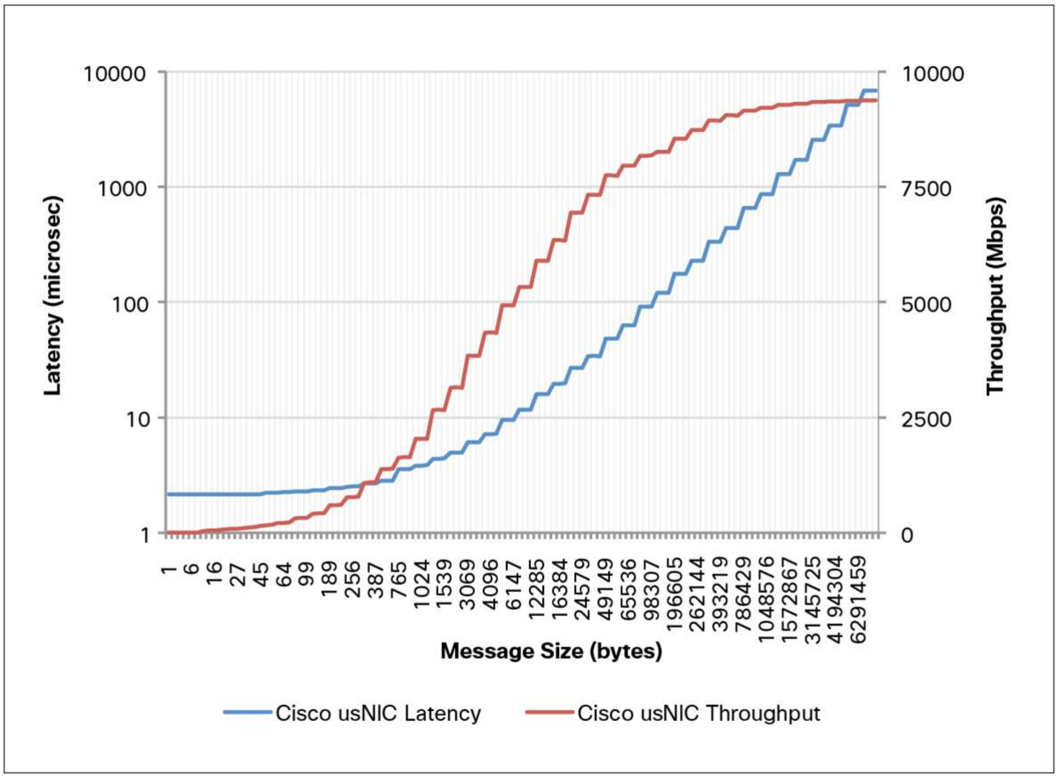
NetPIPE

NetPIPE performs a simple ping-pong test between two nodes and reports half-round-trip (HRT) latencies in microseconds and throughput in MBps for a range of message sizes. Figure 2 shows a graph of the test results.

The following mpi command was used to run the NetPIPE test:

```
hwloc-bind socket:0.core:0 \

/opt/cisco/openmpi/bin/mpirun \

--host n1,n2 \

--mca btl usnic, self, sm \

--bind-to-core \

/opt/NPmpi
```

Figure 2. NetPIPE Latency and Throughput



Intel MPI Benchmarks

IMB runs a set of MPI tests between two nodes and reports latencies (HRT) and throughput in MBps for a range of messages at sizes between and including 2^0 and 2^22.

The following tests were run:

PingPong, PingPing

Sendrecv, Exchange

Allreduce, Reduce, Reduce_scatter

Gather, Gatherv, Scatter, Scatterv

Allgather, Allgatherv, Alltoall, Alltoallv, Bcast

For more information on the benchmarks, refer to the IMB user guide link in the appendix.

The following mpi command was used to run the IMB tests:

```
hwloc-bind socket:0.core:0 \

/opt/cisco/openmpi/bin/mpirun

--host n1,n2 \

--mca btl usnic,self,sm \

--bind-to-core \

/opt/IMB-MPI1 -iter 10000, 10000, 10000
```

Figures 3, 4, and 5 present IMB ping-pong latencies and throughput for a range of message sizes using Cisco usNIC. The message sizes are split up into small, medium, and large to allow a closer look at the trend.

Figure 3. IMB PingPong Small Message

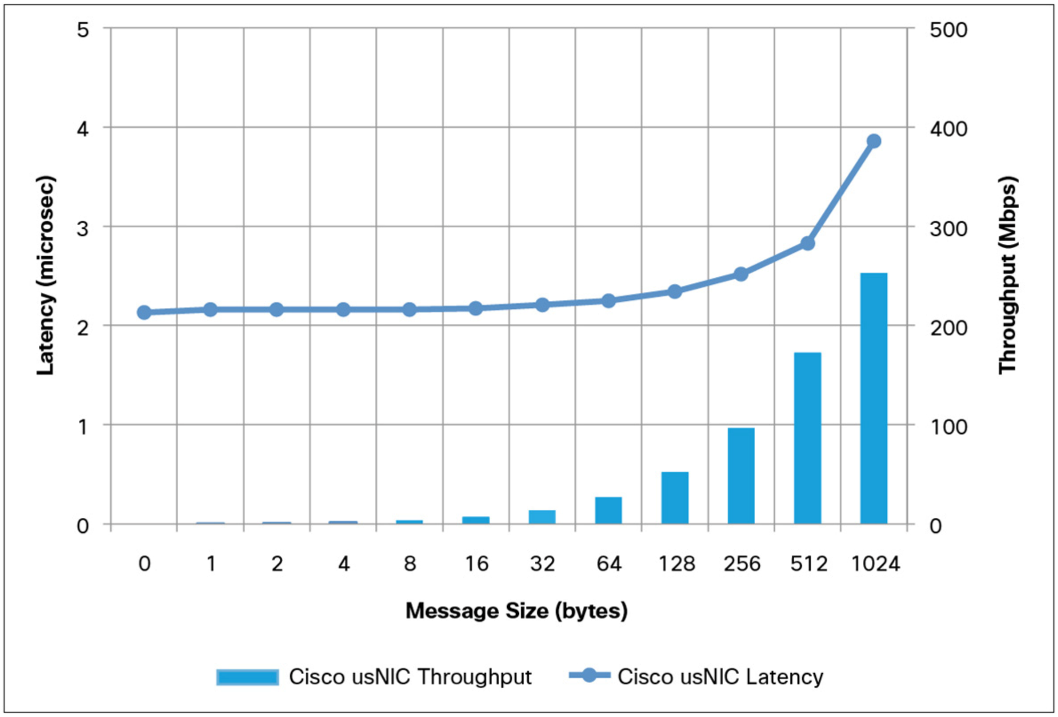


Figure 4. IMB PingPong Medium Message

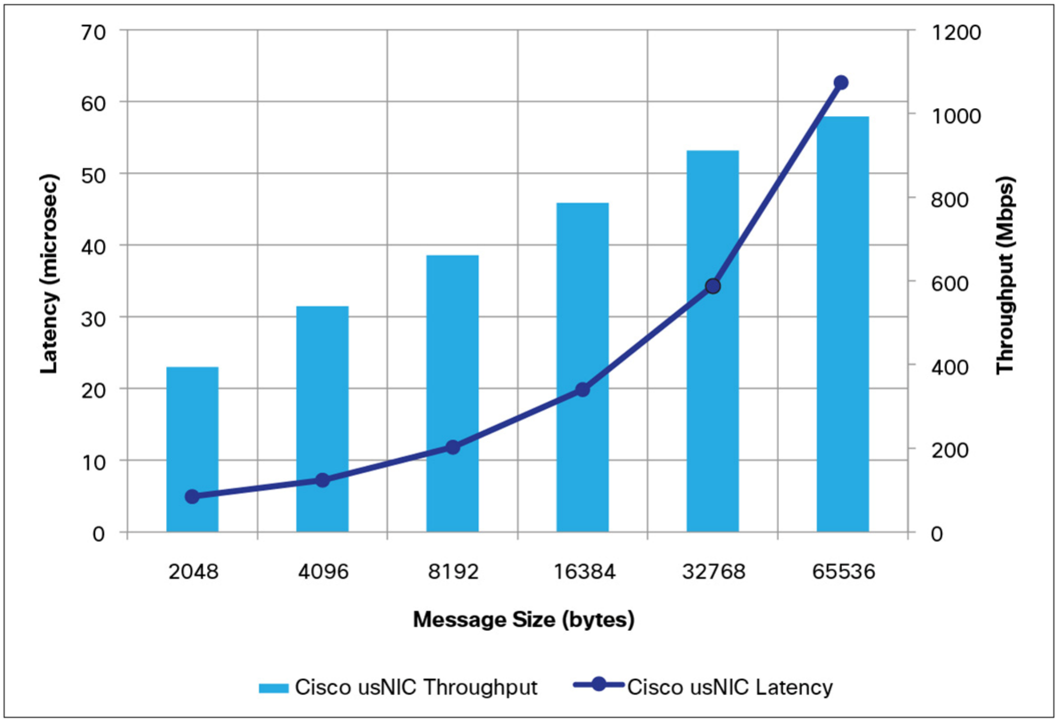
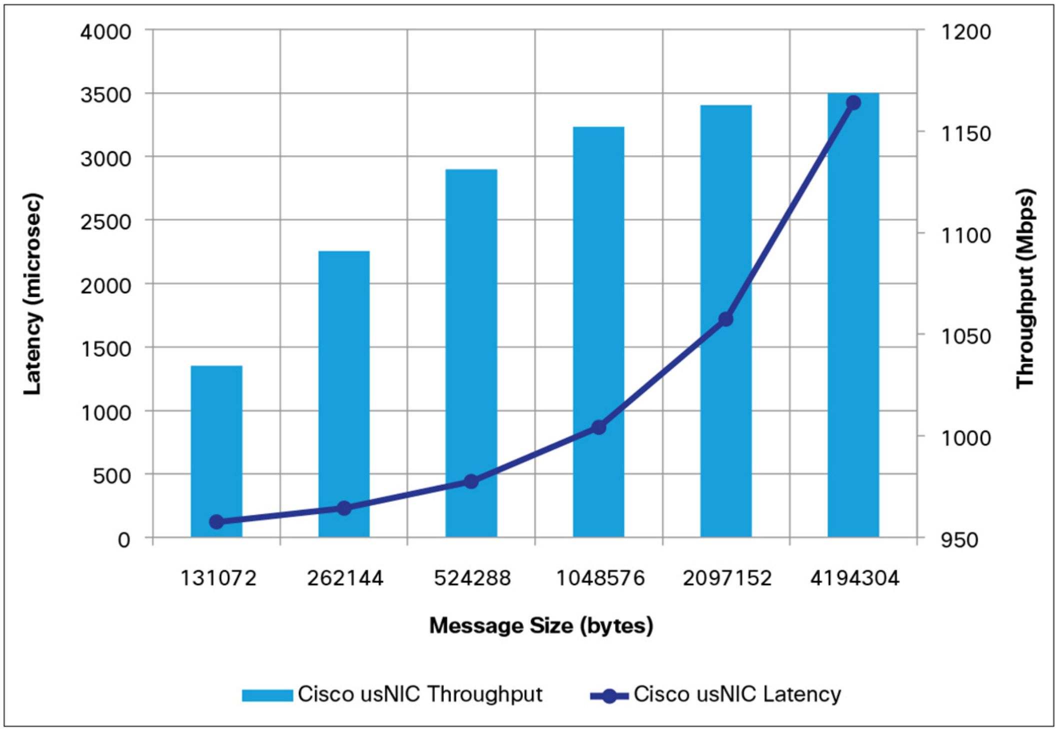


Figure 5. IMB PingPong Large Message



The rest of the tables present consolidated performance information from various IMB performance tests.

Figure 6. PingPong, PingPing

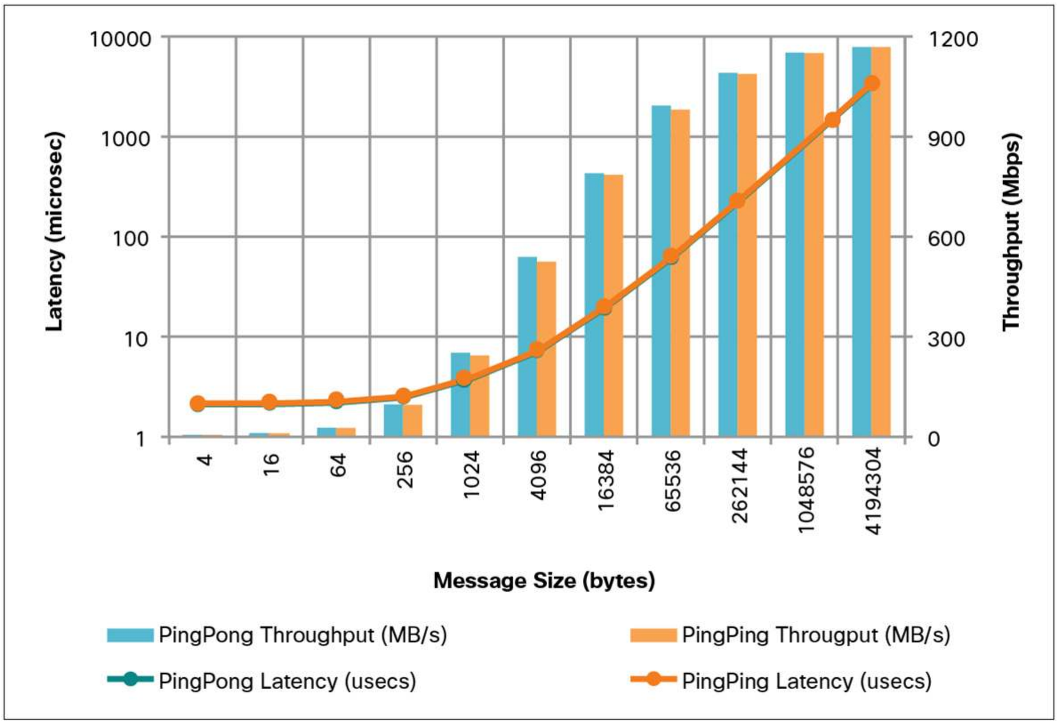


Figure 7. SendRecv, Exchange

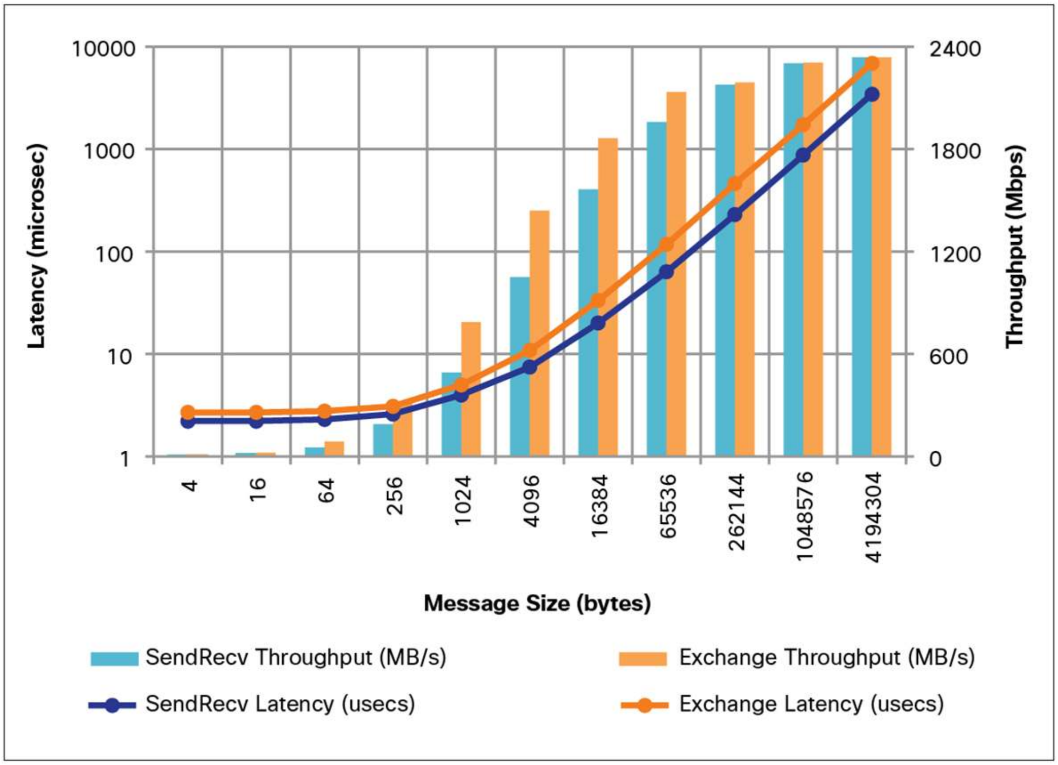


Figure 8. Allreduce, Reduce, Reduce_scatter

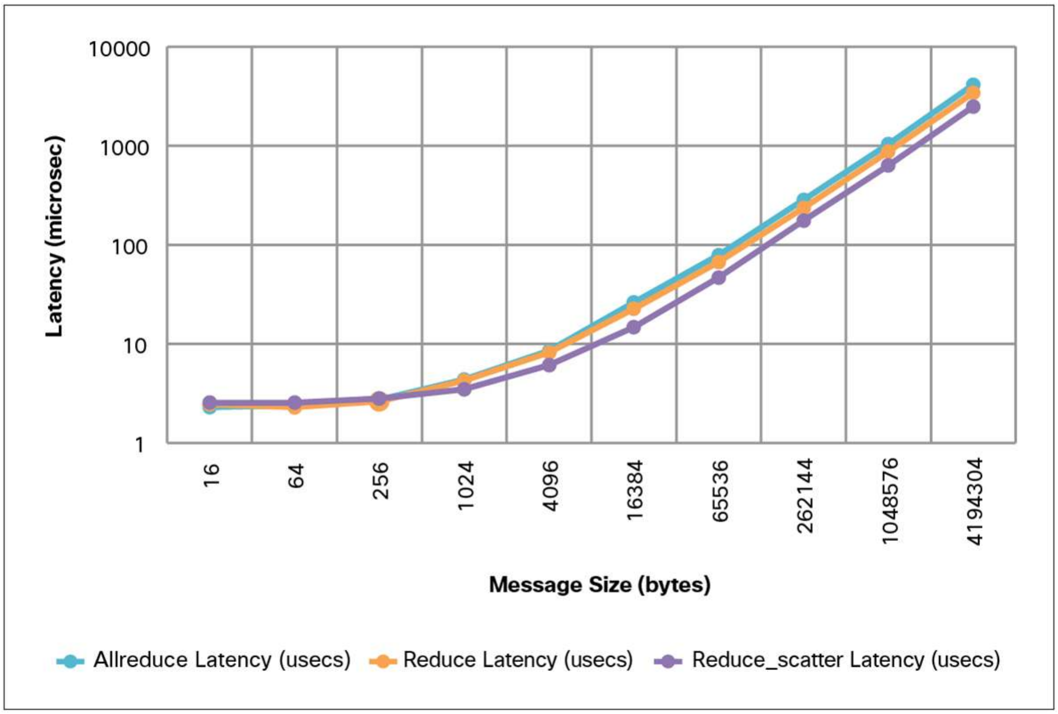


Figure 9. Gather, Gatherv, Scatter, Scatterv

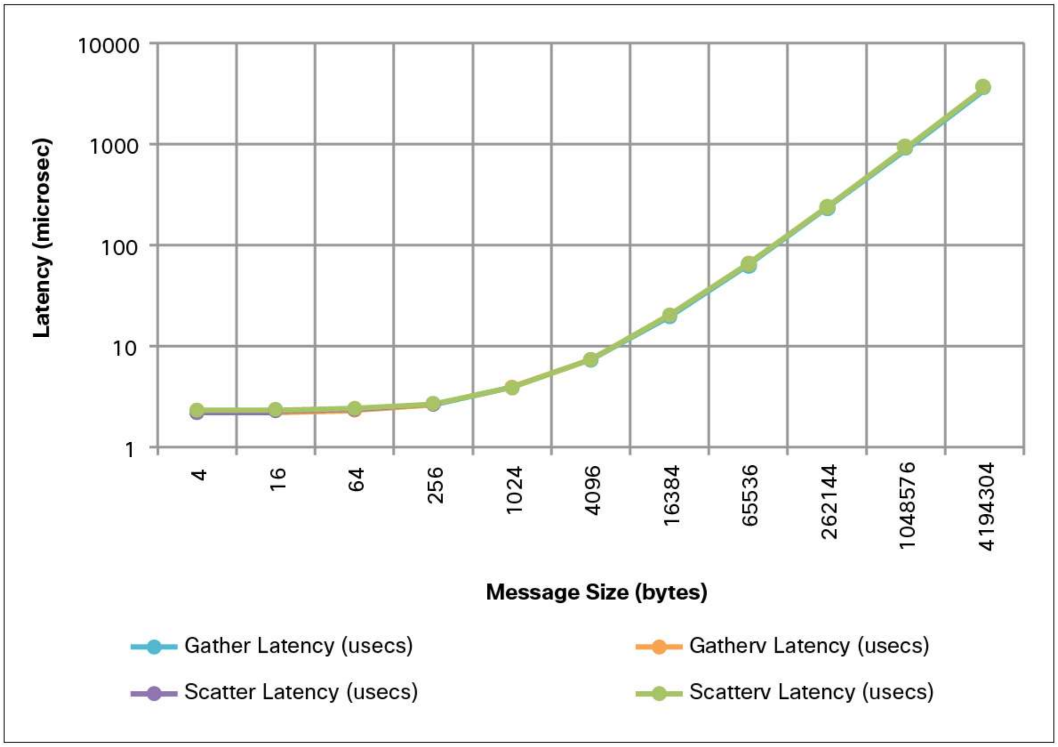
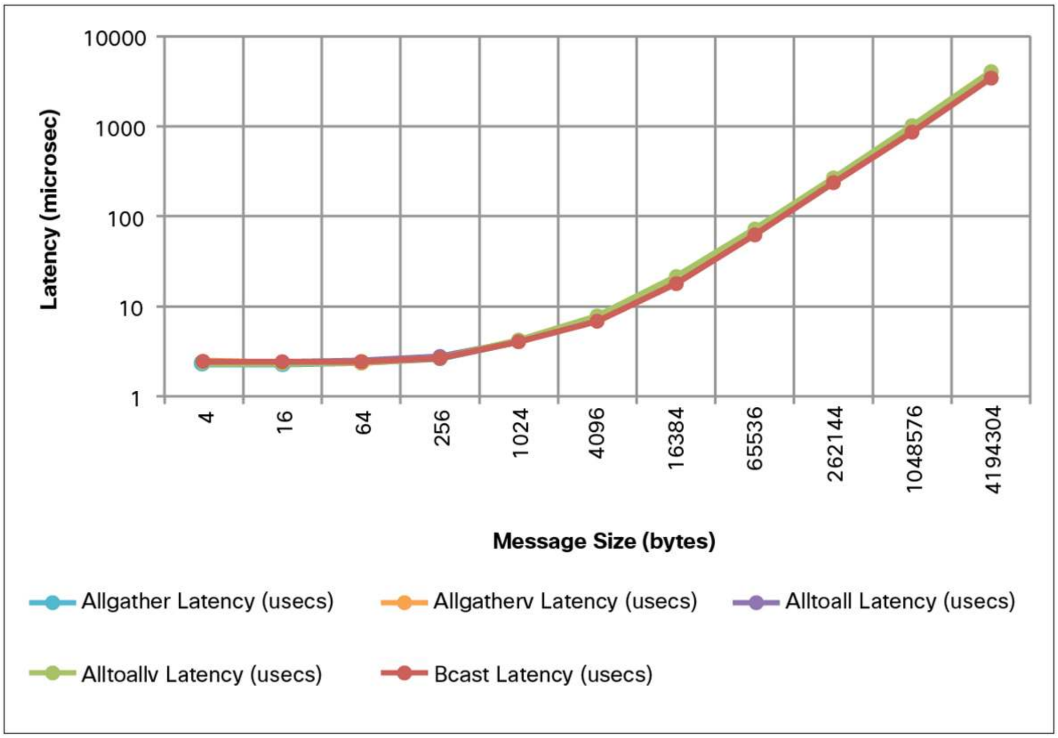


Figure 10. Allgather, Allgatherv, Alltoall, Alltoallv, Bcast



Conclusion

With a small packet MPI ping-pong latency of 2.13 microsec and a maximum ping-pong throughput of 1168 MBps the Cisco usNIC on the Cisco UCS VIC 1225 with the Cisco Nexus 3548 enables a full HPC stack solution that is capable of both low latency and high throughput. Therefore, it is a compelling approach to running HPC tasks with Open MPI on standard Ethernet networks.

Appendix

Cisco UCS VIC 1225 CNA product page:
http://www.cisco.com/en/US/prod/collateral/modules/ps10277/ps12571/data_sheet_c78-708295.html.

Cisco UCS C220 M3 Rack Server product page: http://www.cisco.com/en/US/products/ps12369/index.html.

Cisco Nexus 3548 Switch product page: http://www.cisco.com/en/US/products/ps12581/index.html.

Cisco Nexus 3548 Command Reference:
http://www.cisco.com/en/US/products/ps11541/prod_command_reference_list.html.

NetPIPE homepage: http://www.scl.ameslab.gov/netpipe/.

Intel MPI Benchmarks (IMB): http://software.intel.com/en-us/articles/intel-mpi-benchmarks.

IMB user guide is available at:
http://software.intel.com/sites/products/documentation/hpc/ics/imb/32/IMB_Users_Guide/IMB_Users_Guide.pdf.

Americas Headquarters
Cisco Systems, Inc.
San Jose, CA

Asia Pacific Headquarters
Cisco Systems (USA) Pte. Ltd.
Singapore

Europe Headquarters
Cisco Systems International BV Amsterdam,
The Netherlands

Cisco has more than 200 offices worldwide. Addresses, phone numbers, and fax numbers are listed on the Cisco Website at **www.cisco.com/go/offices.**

Printed in USA C11-729460-00 01/14