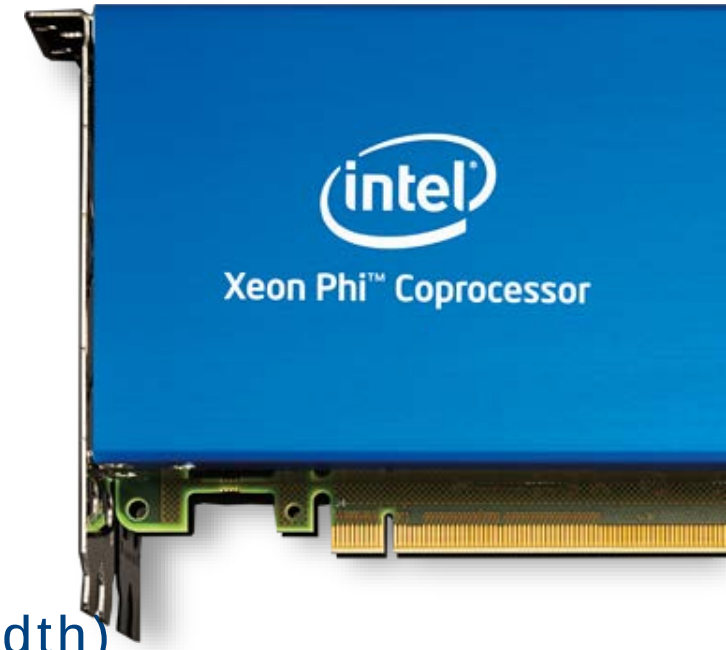
Running Native Lustre* Client inside Intel® Xeon Phi™ coprocessor

Dmitry Eremin, Zhiqi Tao and Gabriele Paciucci

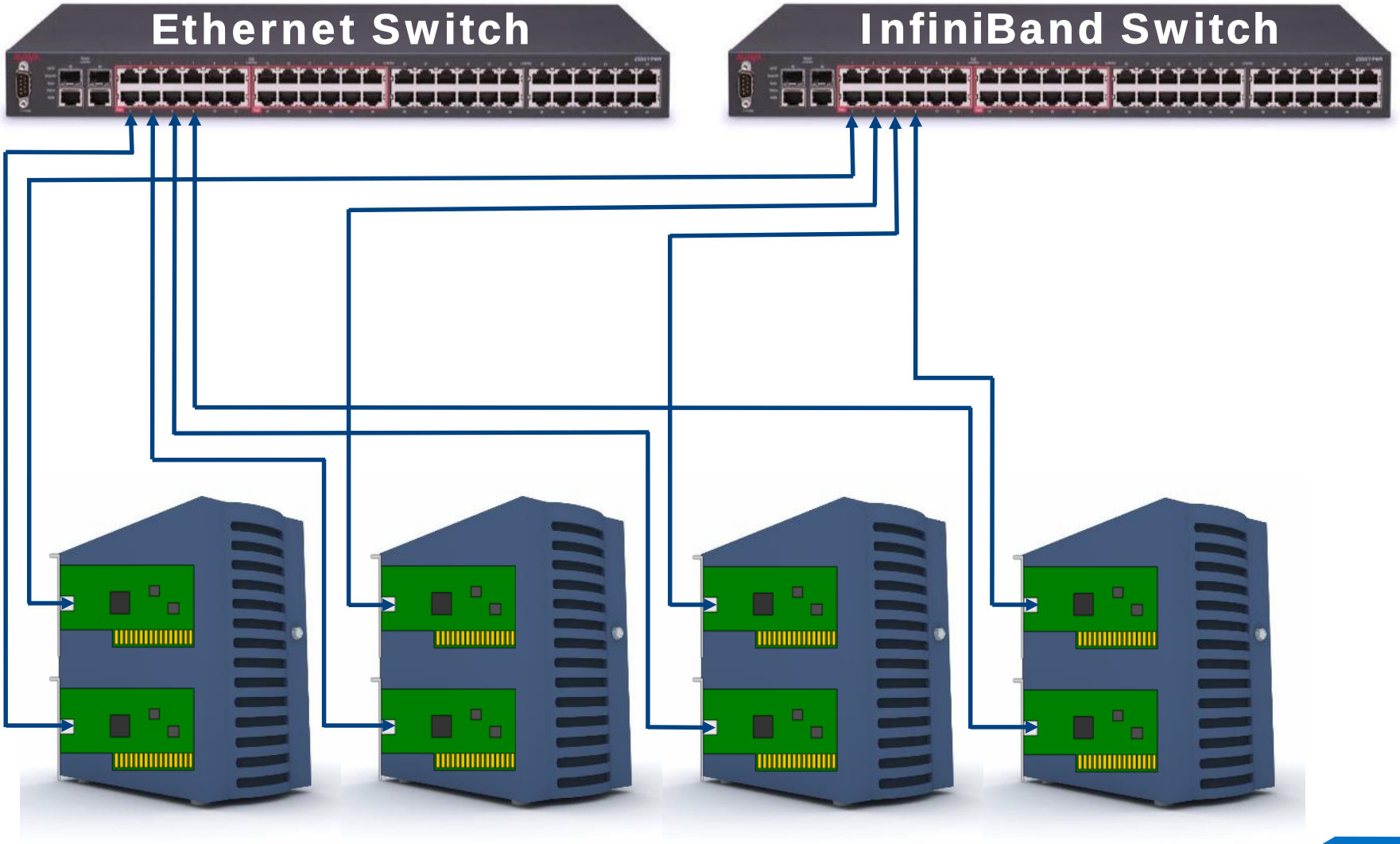
08 April 2014

What is the Intel® Xeon Phi™ coprocessor?

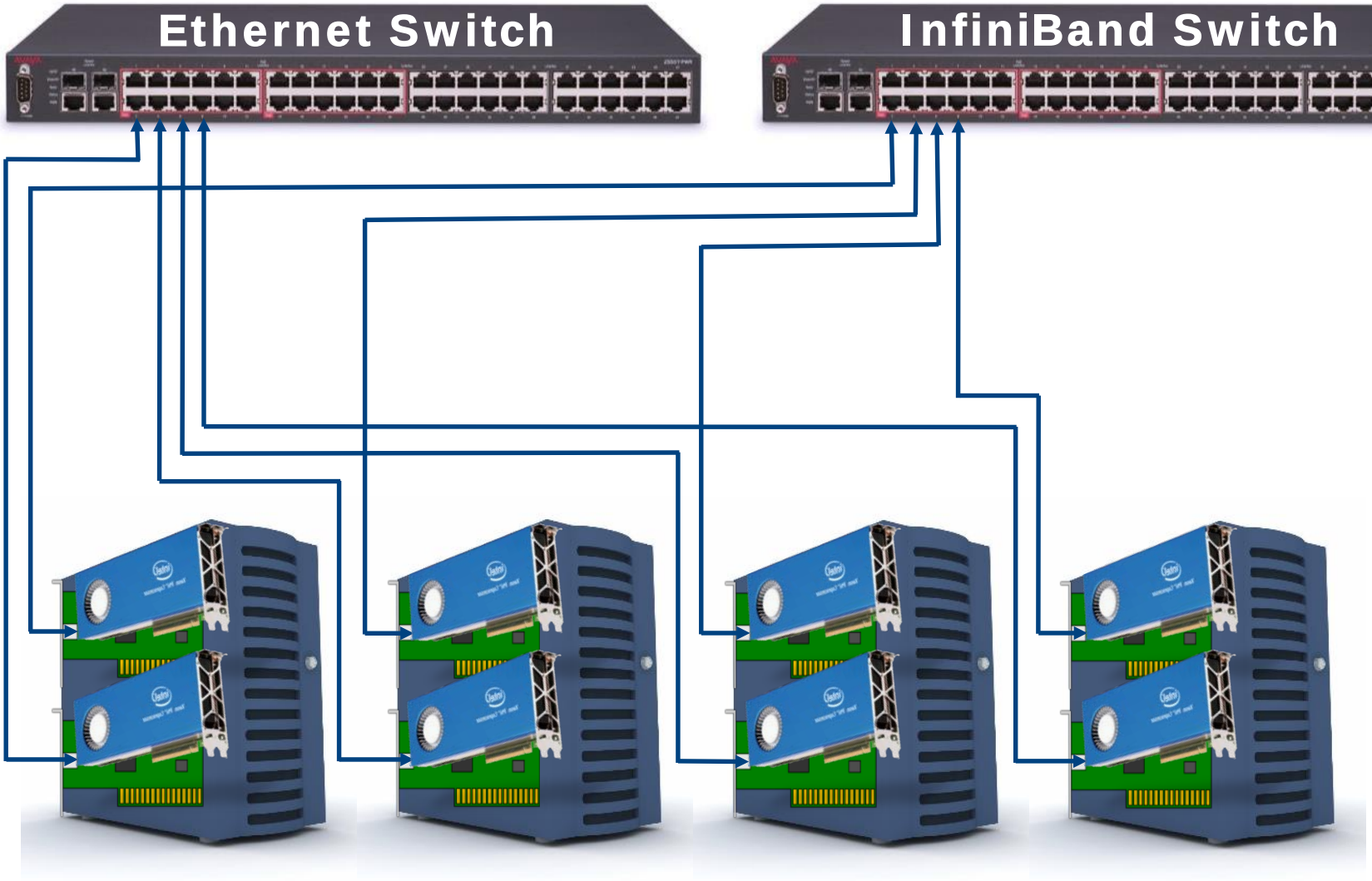
- Up to 61 Cores and 244 Threads per coprocessor
- 8 GB of memory
- Gen2x16 PCI Express: up to 8 GB/s
- 8 memory controllers supporting
  up to 16 GDDR5 channels delivering
  up to 5.5 GT/s (up to 352 GB/s bandwidth)
- http://intel.com/xeonphi



Typical Cluster Configuration



Typical Cluster Configuration with Intel® Xeon Phi™ coprocessors

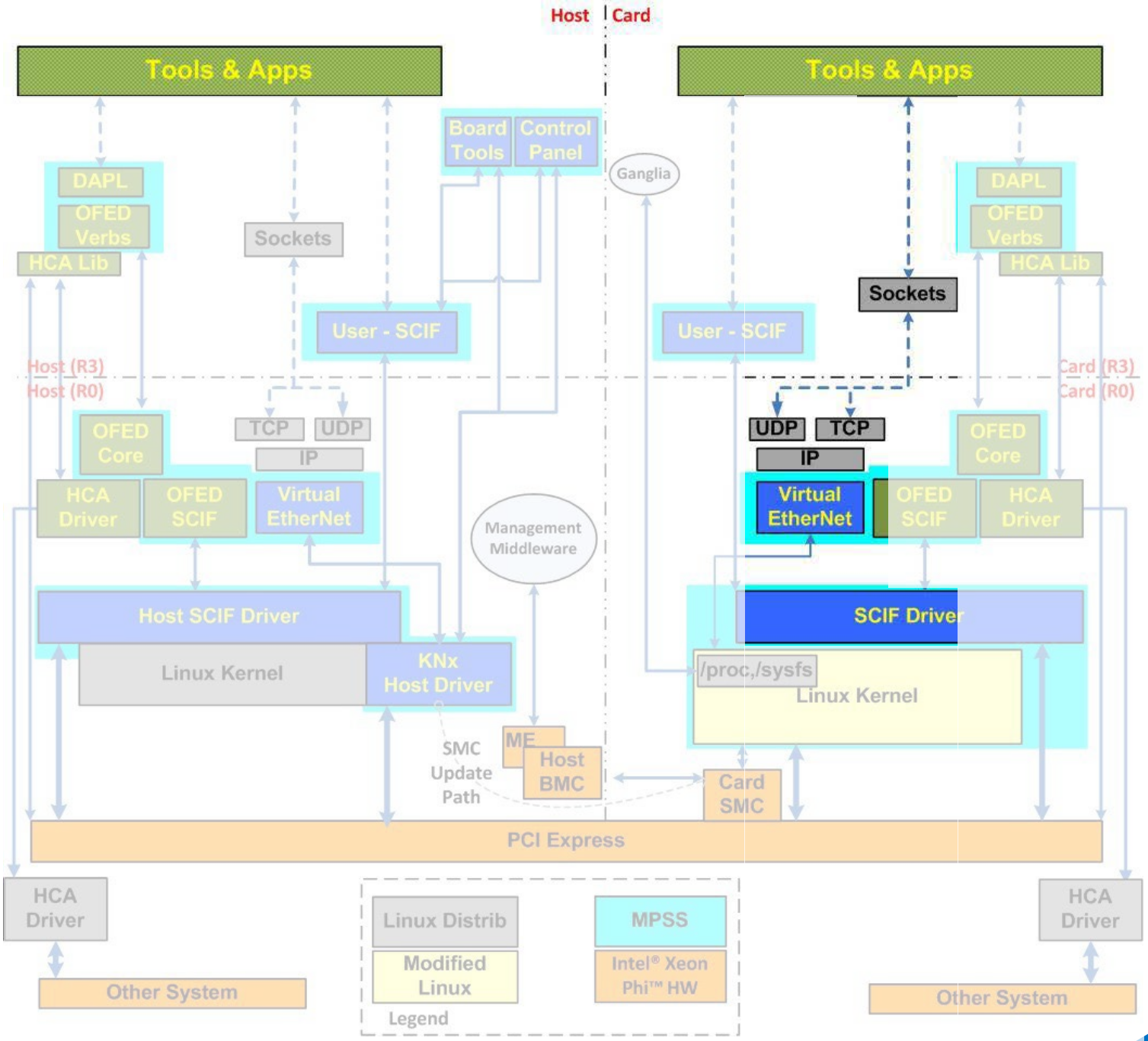


Intel® Xeon Phi™ coprocessor Software Architecture

- Intel® Xeon Phi coprocessors, based on Intel® Many Integrated Core (Intel® **MIC**) Architecture

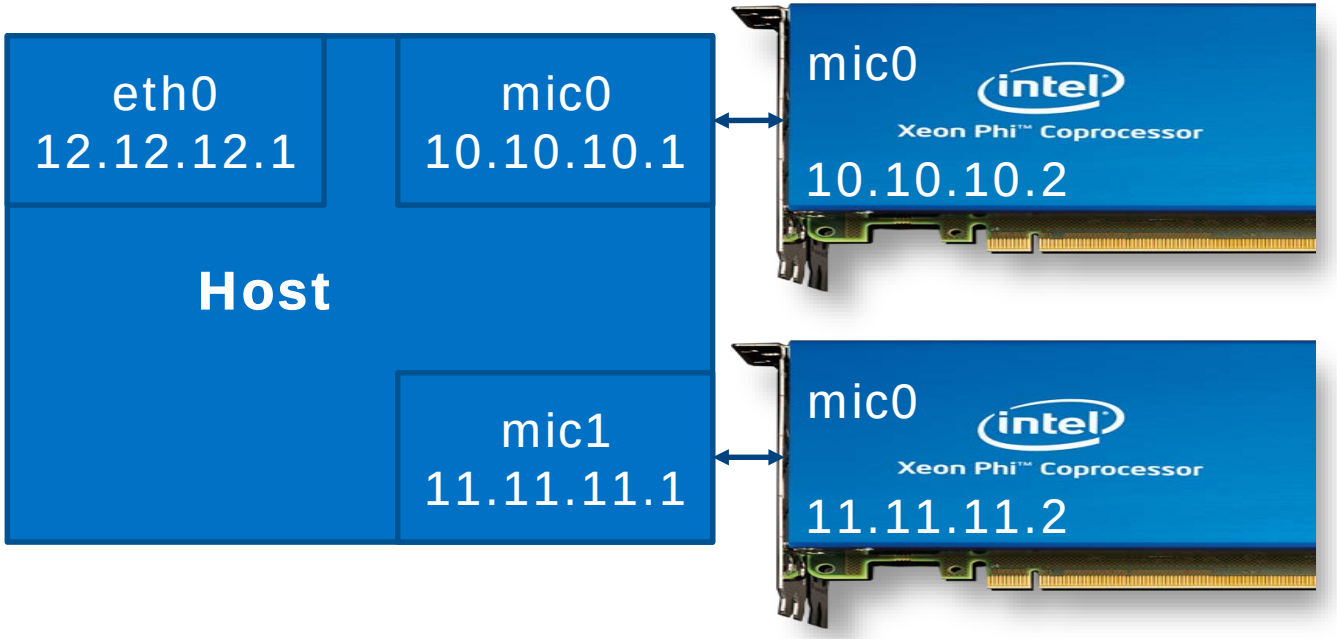
- Intel® MIC Software Stack (Intel® **MPSS**)

- The Symmetric Communication Interface (**SCIF**) API is the communication backbone between the host processors and the Intel® Xeon Phi coprocessors in a heterogeneous computing environment


Host
Card
Tools & Apps
Tools & Apps
Board Tools
Control Panel
Ganglia
DAPL
OFED Verbs
HCA Lib
Sockets
User - SCIF
DAPL
OFED Verbs
HCA Lib
Sockets
User - SCIF
Host (R3)
Host (R0)
Card (R3)
Card (R0)
OFED Core
TCP
UDP
IP
HCA Driver
OFED SCIF
Virtual EtherNet
UDP
TCP
IP
OFED Core
Virtual EtherNet
OFED SCIF
HCA Driver
Management Middleware
Host SCIF Driver
SCIF Driver
Linux Kernel
KNx Host Driver
/proc,/sysfs
Linux Kernel
SMC Update Path
ME
Host BMC
Card SMC
PCI Express
HCA Driver
Other System
Linux Distrib
Modified Linux
MPSS
Intel® Xeon Phi™ HW
Legend
HCA Driver
Other System

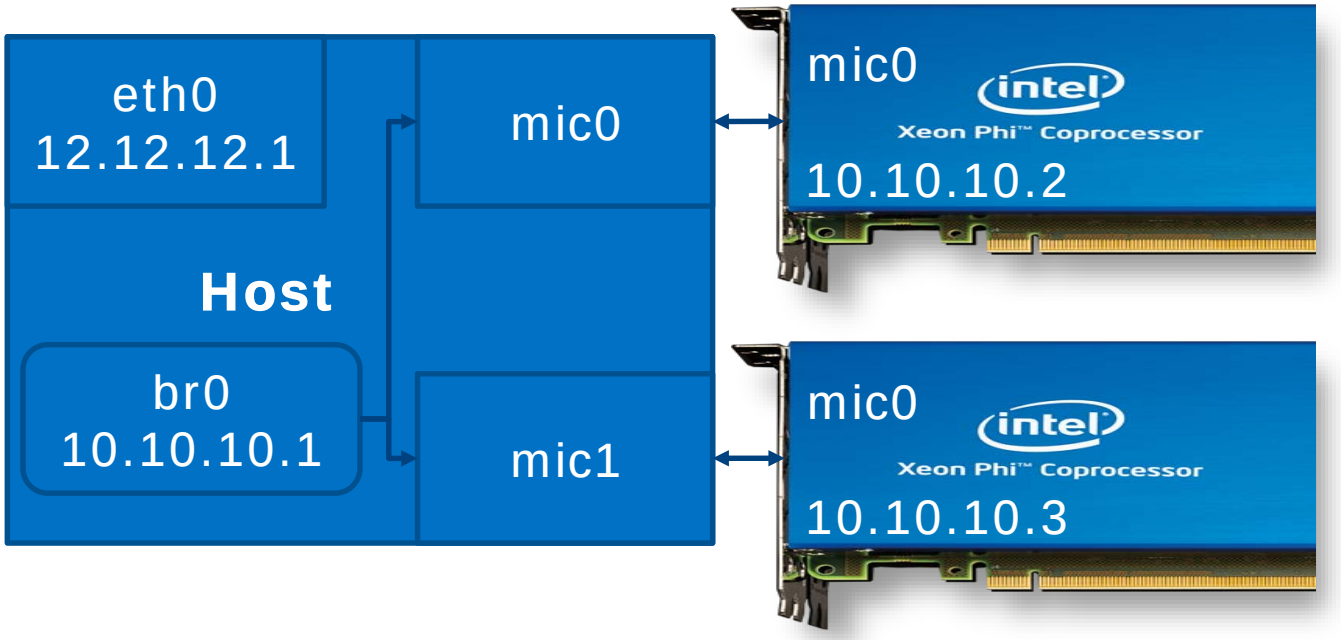
Intel® Xeon Phi™ coprocessors Static Pair Topology

- Every Intel® Xeon Phi coprocessor card is assigned to a separate subnet known only to the host
- Host is a router for any network communications



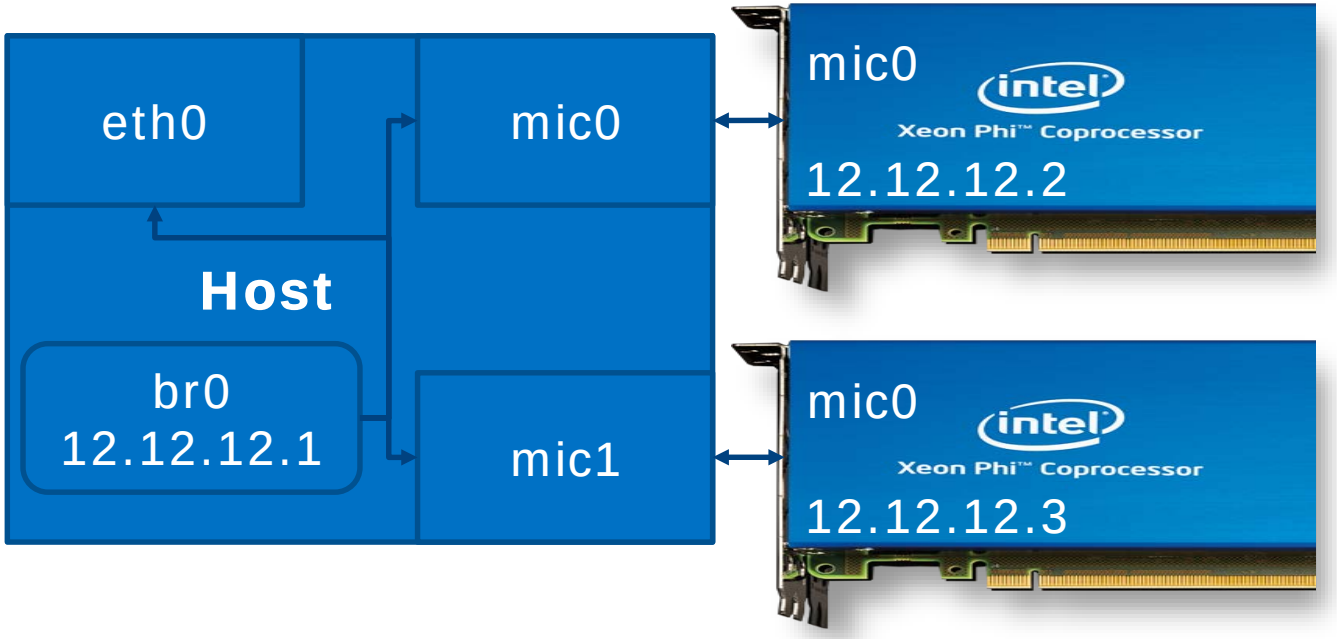
Intel® Xeon Phi™ coprocessors Internal Bridge Topology

- Bridging together multiple Intel® Xeon Phi coprocessor card virtual network interfaces, on the same host
- Host is a router for external network communications



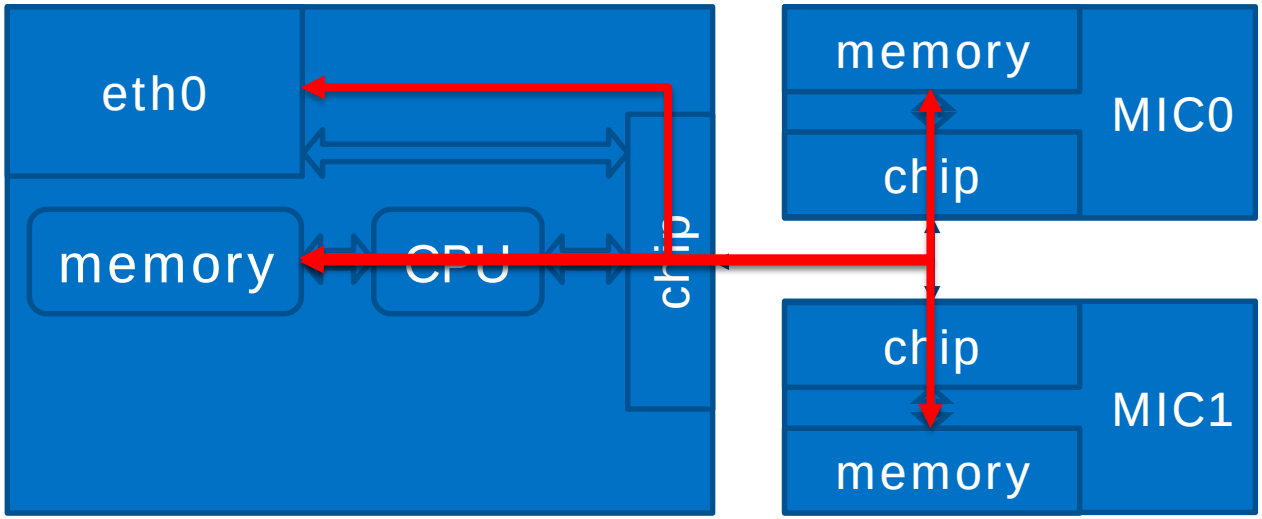
Intel® Xeon Phi™ coprocessors External Bridge Topology

- Bridge the Intel® Xeon Phi coprocessor virtual connections to a physical Ethernet device
- Independent network communications
- Maximum MTU is 9000

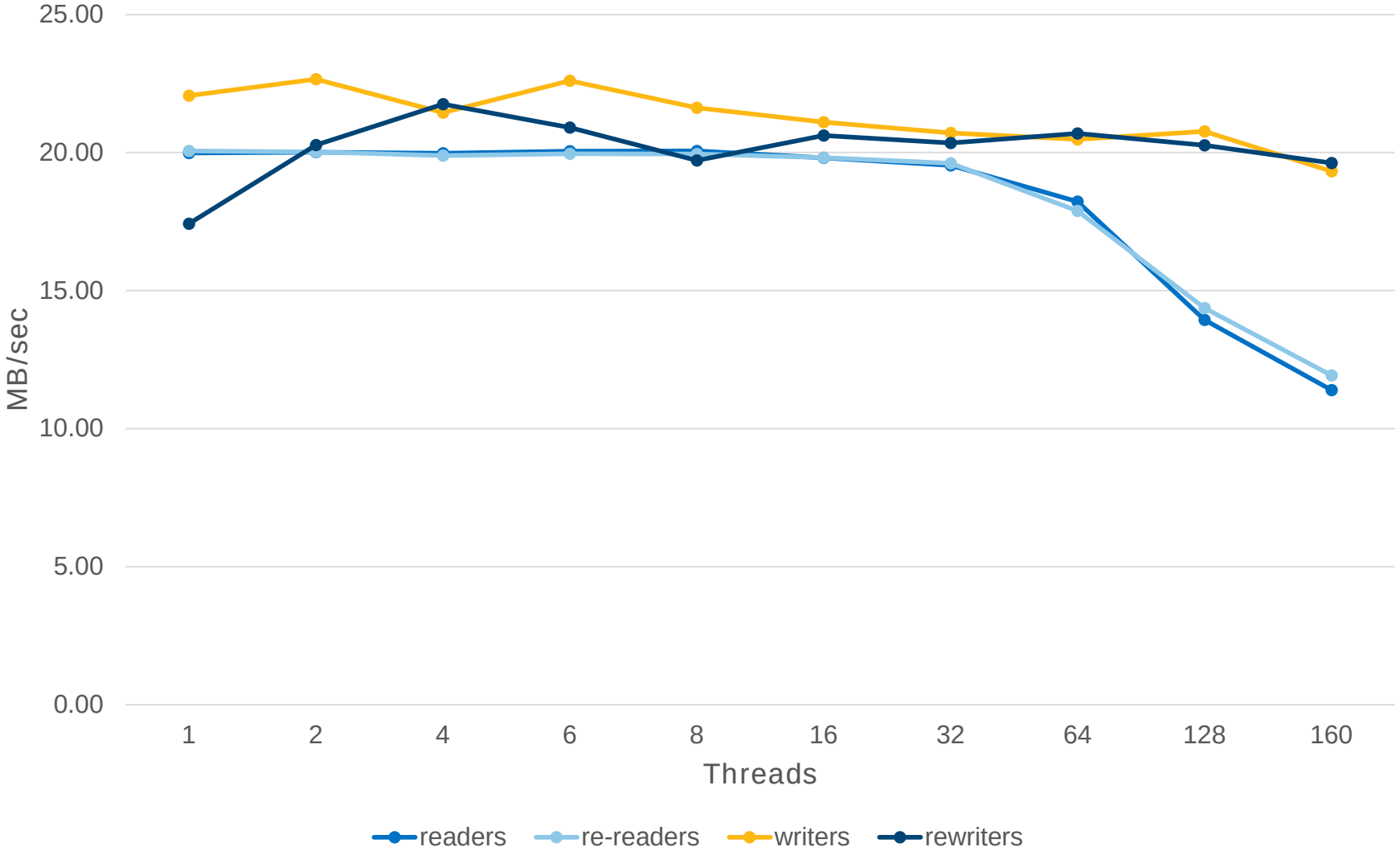


Data transfer over virtual Ethernet

- All data passed through host memory
- Intensive data transfer cause intensive host usage

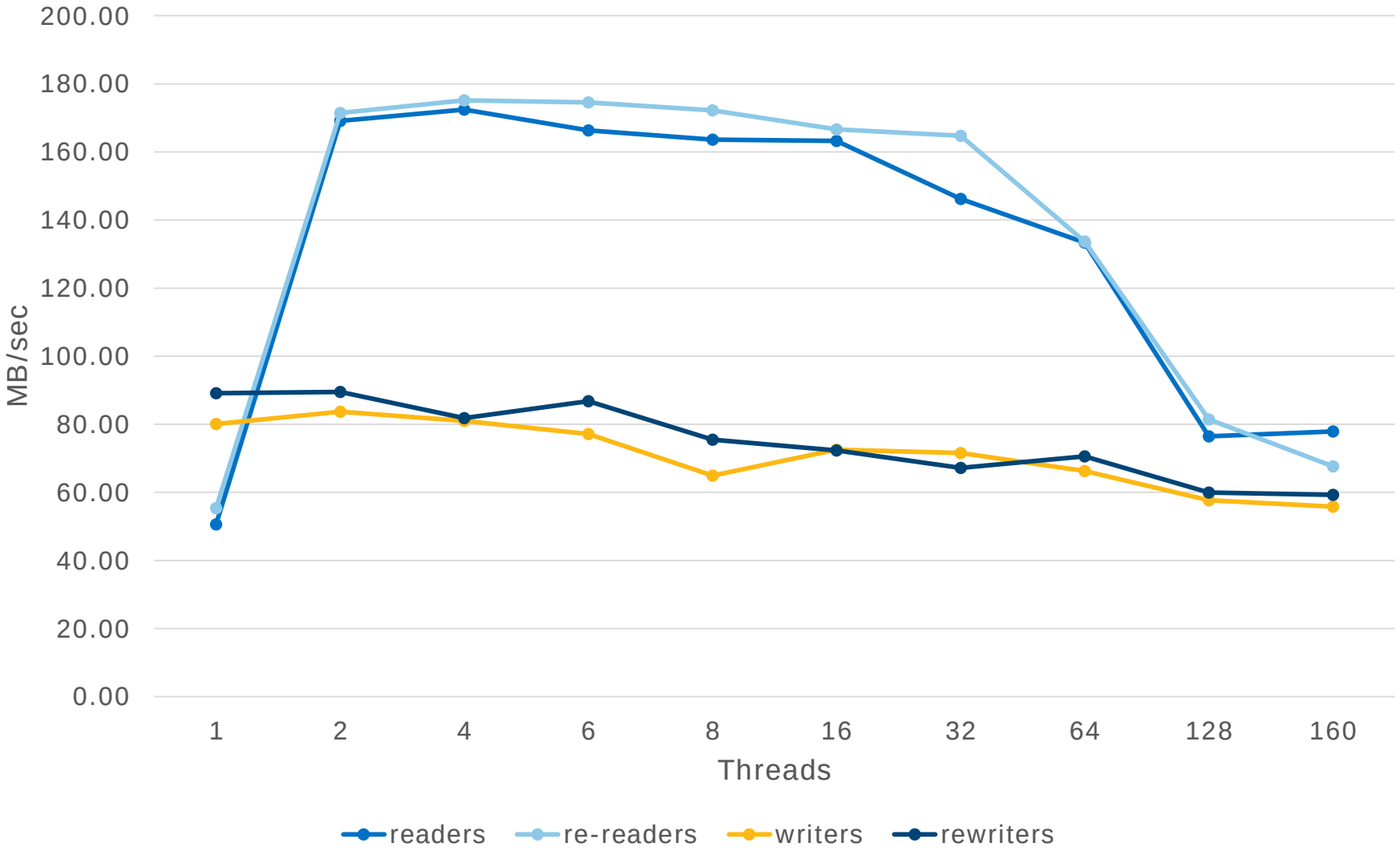


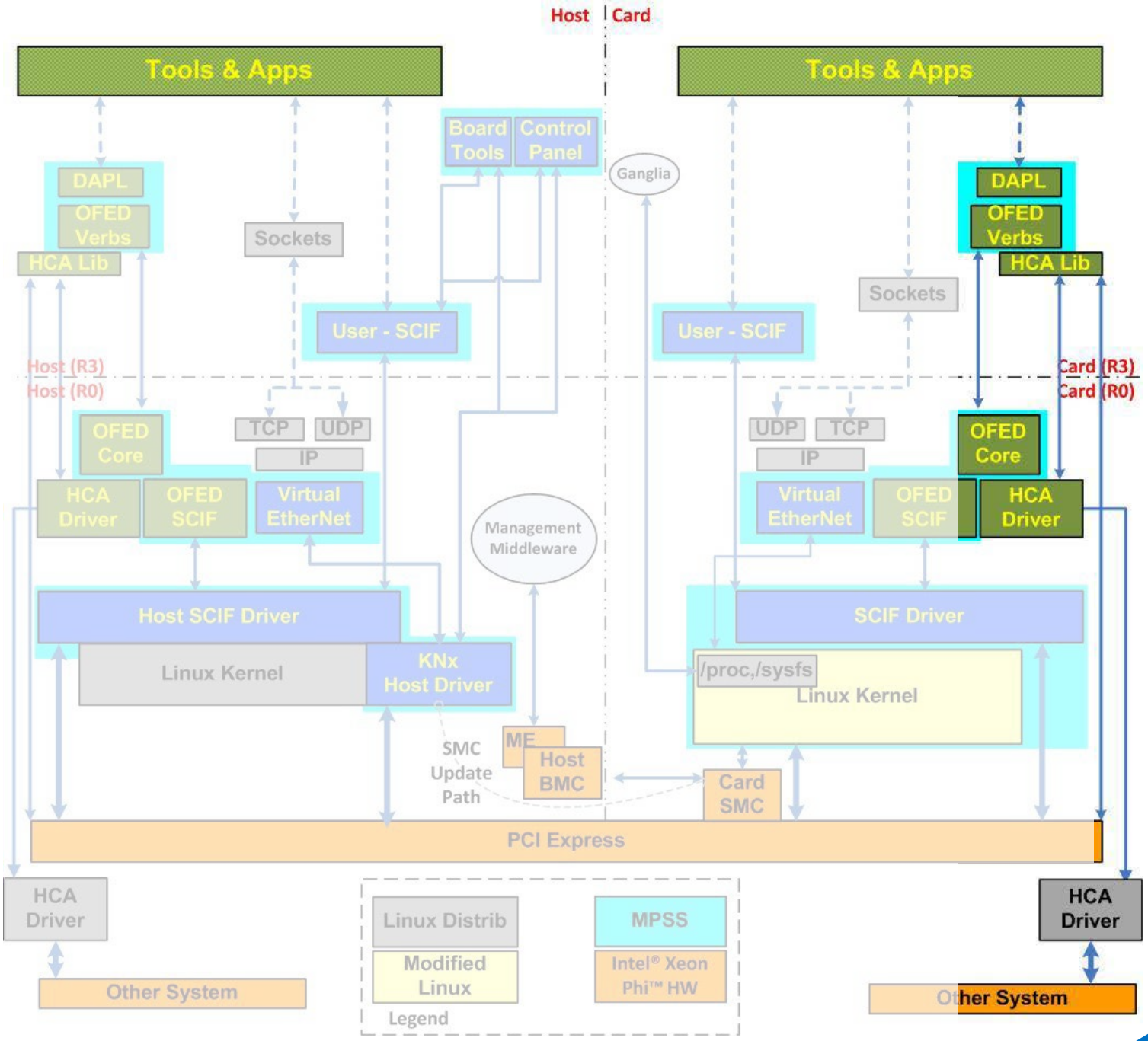
The NFS throughput over virtual Ethernet (MTU is 1500)



Results have been estimated based on internal Intel analysis and are provided for informational purposes only. Any difference in system hardware or software design or configuration may affect actual performance.

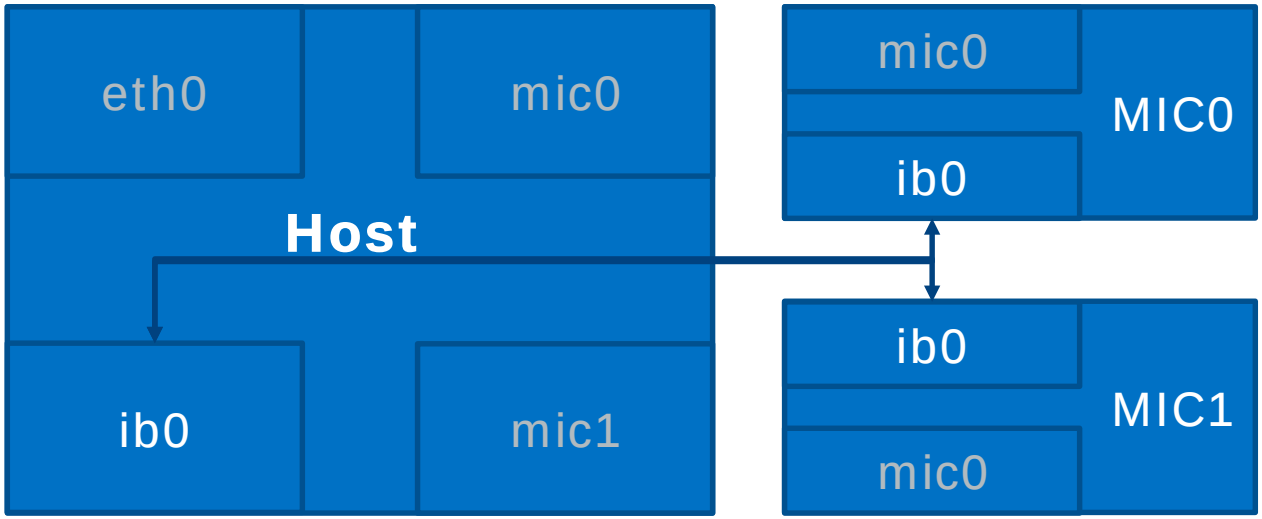
The NFS throughput over virtual Ethernet (MTU is 64512)




Host
Card
Tools & Apps
Tools & Apps
Board Tools
Control Panel
Ganglia
DAPL
OFED Verbs
HCA Lib
Sockets
User - SCIF
DAPL
OFED Verbs
HCA Lib
Sockets
User - SCIF
Host (R3)
Host (R0)
Card (R3)
Card (R0)
OFED Core
TCP
UDP
IP
HCA Driver
OFED SCIF
Virtual EtherNet
UDP
TCP
IP
OFED Core
Virtual EtherNet
OFED SCIF
HCA Driver
Management Middleware
Host SCIF Driver
Linux Kernel
KNx Host Driver
SCIF Driver
/proc,/sysfs
Linux Kernel
SMC Update Path
ME
Host BMC
Card SMC
PCI Express
HCA Driver
Other System
Linux Distrib
Modified Linux
MPSS
Intel® Xeon Phi™ HW
Legend
HCA Driver
Other System

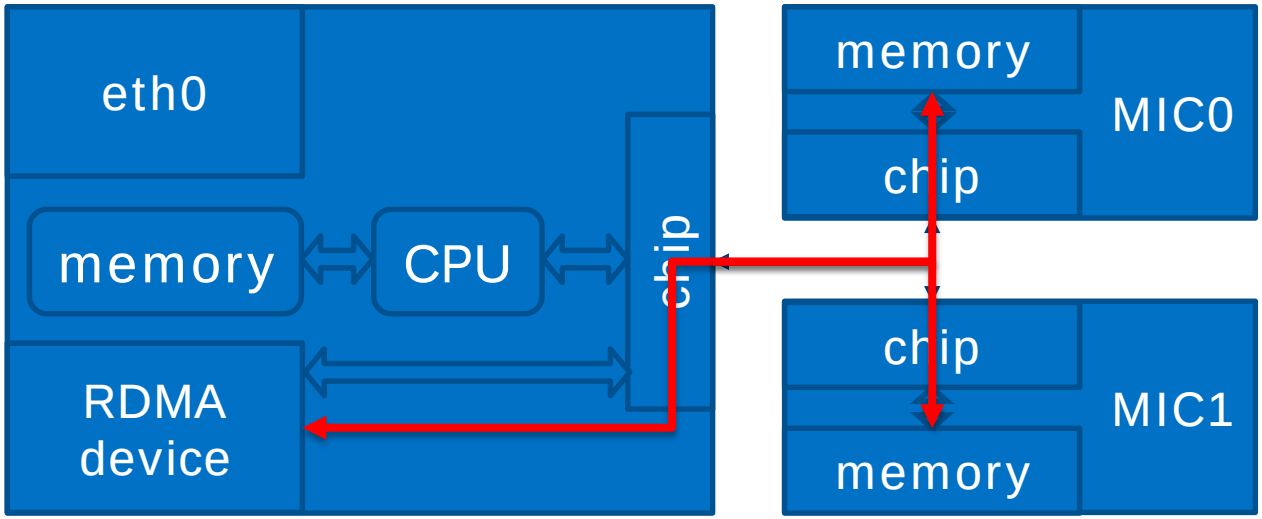
The virtual IB interface in Intel® Xeon Phi™ coprocessor

- Become available in Intel® MPSS starting with v3.1
- Required to have the OFED version 1.5.4.1 InfiniBand drivers installed
- Supports Intel® True Scale and Mellanox Fabrics



Data transfer over virtual IB

- RDMA transfer passed directly to RDMA device
- RDMA device hardware is shared between Linux-based host and Intel® Xeon Phi coprocessor applications



The Coprocessor Operating System (coprocessor OS)

- Based on a standard Linux kernel source code
- Coprocessor OS is a minimal:
  - Busybox minimal shell environment
  - Linux Standard Base (LSB) Core libraries
- Can be extended with loadable kernel modules (LKMs)

The Lustre* 2.4 Client for Intel® Xeon Phi™ coprocessor

- Download the SOURCE package from "MPSS 2.1 release for Linux" section
- Unpack it and then unpack the **package-full_src-k1om.tar.bz2** file in any location
- Execute the following commands in this place:

```
# export PATH=/usr/linux-k1om-4.7/bin:$PATH
# make defconfig-miclinux
# make -C card/kernel ARCH=k1om modules_prepare
# sh autogen.sh
# ./configure --with-linux=<unpacked_path>/card/kernel \
    --without-o2ib \
    --host=x86_64-k1om-linux --build=x86_64-pc-linux
# make rpms
```

The Lustre* 2.5 Client for Intel® Xeon Phi™ coprocessor

- Download the SOURCE, mpss-3.x-k1om.tar and OS specific files from "MPSS 3.x release for Linux" section
- Unpack from them “kernel-dev-*.rpm”, “ofed-driver-*-devel-*.rpm” and “linux-*.tar.bz2” files
- Prepare Intel® MPSS sources for Lustre* build:

```
# rpm2cpio kernel-dev-*.rpm | cpio –idm
# rpm2cpio ofed-driver-*-devel-*.rpm | cpio –idm
# tar xjvf linux-*.tar.bz2 && cd linux-*
# cp -f ../boot/config-* .config
# cp -f ../boot/Module.symvers-* Module.symvers
# . /opt/mpss/3.x/environment-setup-k1om-mpss-linux
# make ARCH=k1om silentoldconfig modules_prepare
```

The Lustre* 2.5 Client for Intel® Xeon Phi™ coprocessor (cont)

- Build Lustre* sources for Intel® Xeon Phi coprocessor with Intel® MPSS sources with virtual IB support:

```
# . /opt/mpss/3.x/environment-setup-k1om-mpss-linux
# sh autogen.sh
# ./configure --with-linux=<path>/linux-2.6.38+mpss3.x \
    --with-o2ib=<path>/usr/src/ofed-driver-*.el6.x86_64 \
    --host=k1om-mpss-linux --build=x86_64-pc-linux
# make rpms
```

Install & Configure the Lustre* Client

- Only two Lustre* RPMs should be installed on host:
  - **lustre-client-mic-<version>.x86_64.rpm**
  - **lustre-client-mic-modules-<version>.x86_64.rpm**

- ssh mic0 "echo 'options lnet networks=\"o2ib0(ib0)\"' > /etc/modprobe.d/lustre.conf"

Host configuration in /etc/modprobe.d/lustre.conf

options lnet networks="o2ib0(ib0)"

Mounting on Host:

mount –t lustre 8.8.8.8@o2ib:/lustrefs /mnt/lustrefs

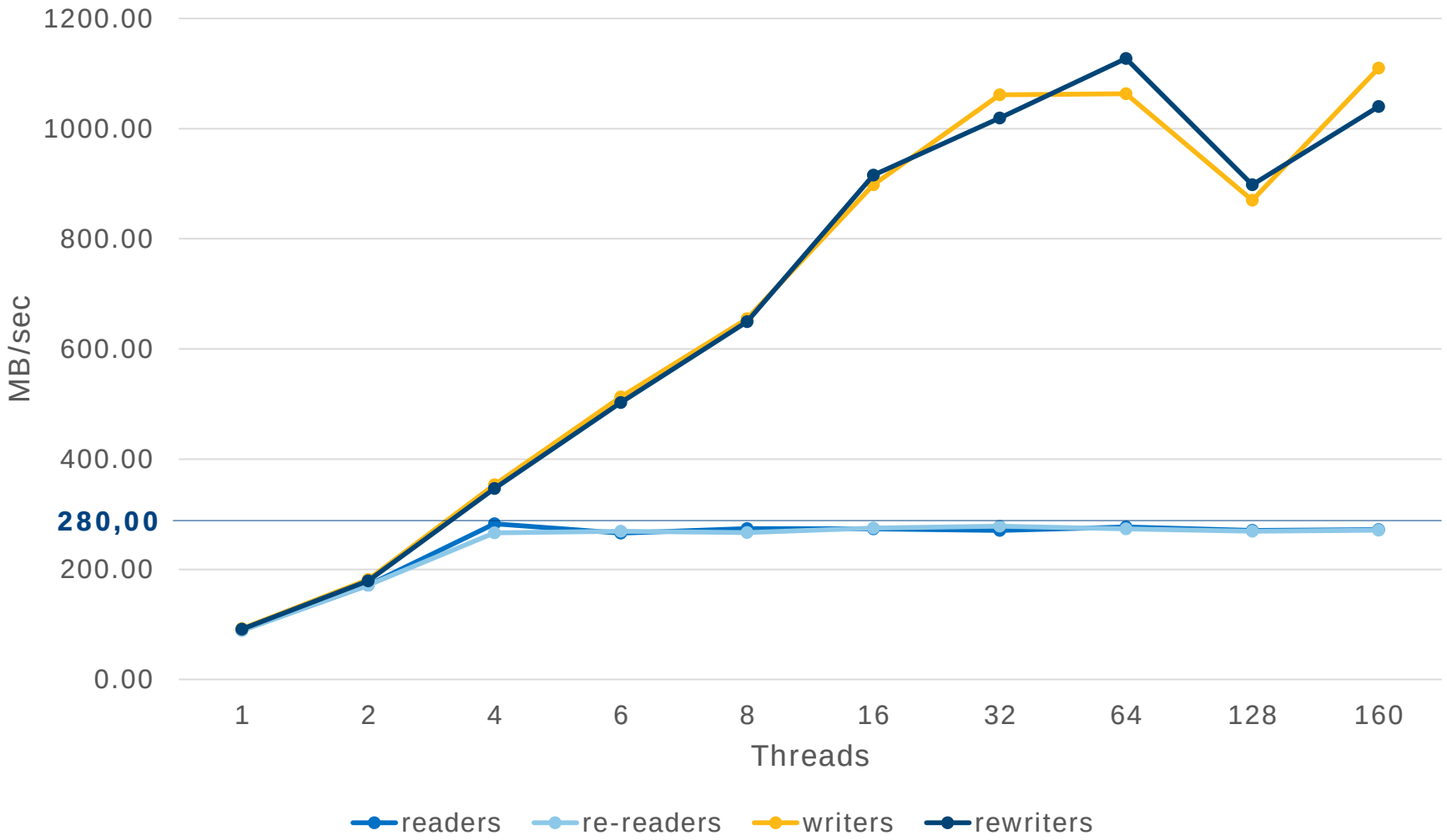
Xeon Phi configuration in /etc/modprobe.d/lustre.conf

options lnet networks="o2ib0(ib0)"

Mounting on Xeon Phi:

mount.lustre 8.8.8.8@o2ib:/lustrefs /mnt/lustrefs

The Lustre* throughput over virtual IB

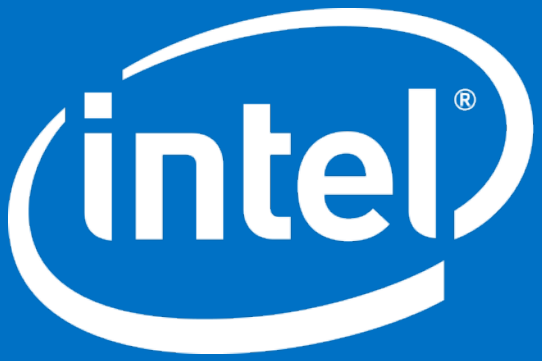


Results have been estimated based on internal Intel analysis and are provided for informational purposes only. Any difference in system hardware or software design or configuration may affect actual performance.

Server and storage infrastructure,
networking and hardware support was supplied by
Intel's High Performance computing Labs in Swindon (UK).

Special thanks to **Jamie Wilcox** and **Adam Roe**.

Questions?


intel