Open Fabrics Workshop 2013

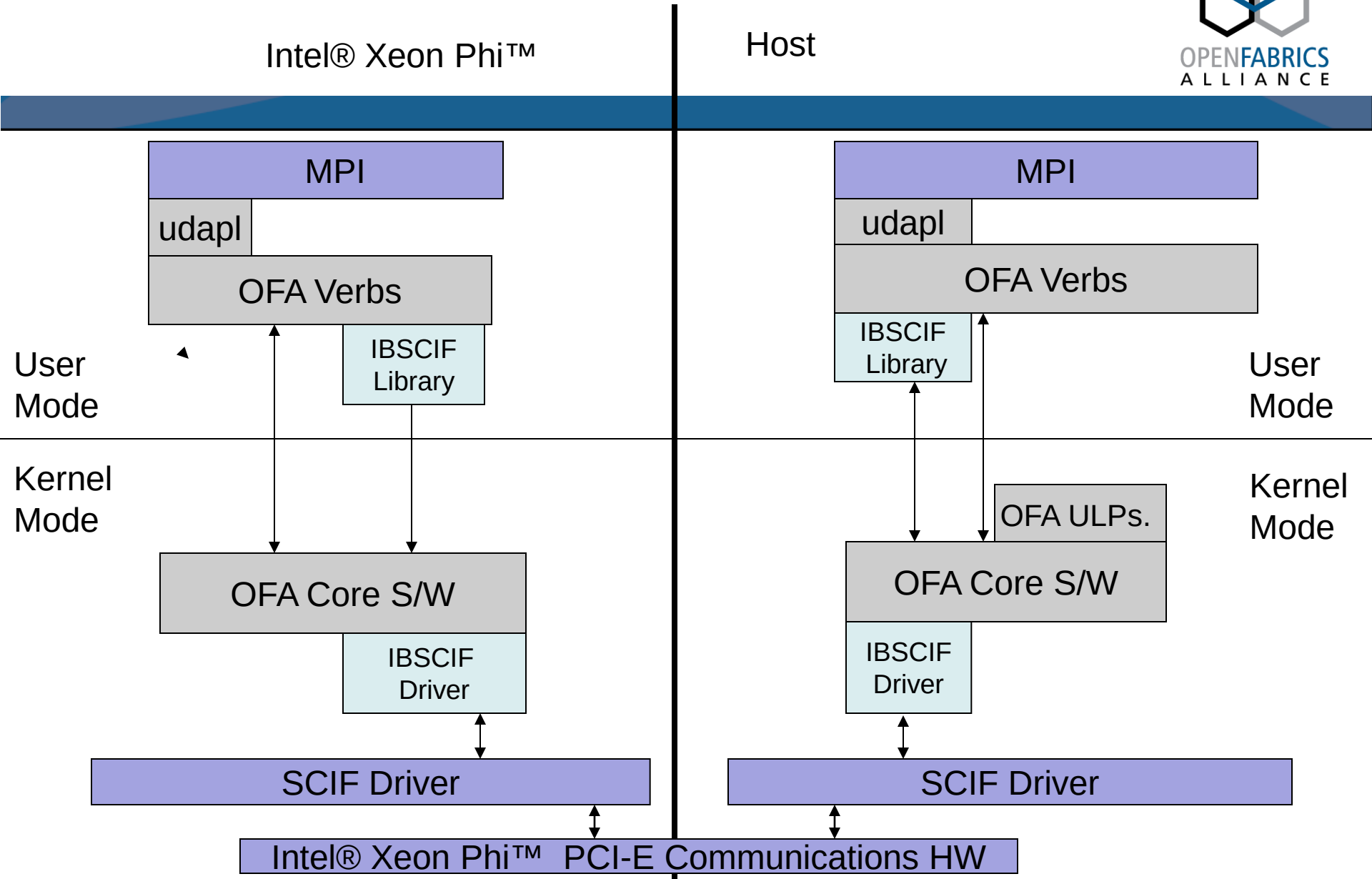
OFS Software for the Intel® Xeon Phi ™

Bob Woodruff

Agenda

- Intel® Coprocessor Communication Link (CCL) Software
  - IBSCIF – RDMA from Host to Intel® Xeon Phi™
  - Direct HCA Access from Intel® Xeon Phi™
  - Proxy Access to HCA via the Host
- How to get the S/W
  - Intel® MPSS S/W release
  - OFED-3.5-MIC Branch Release

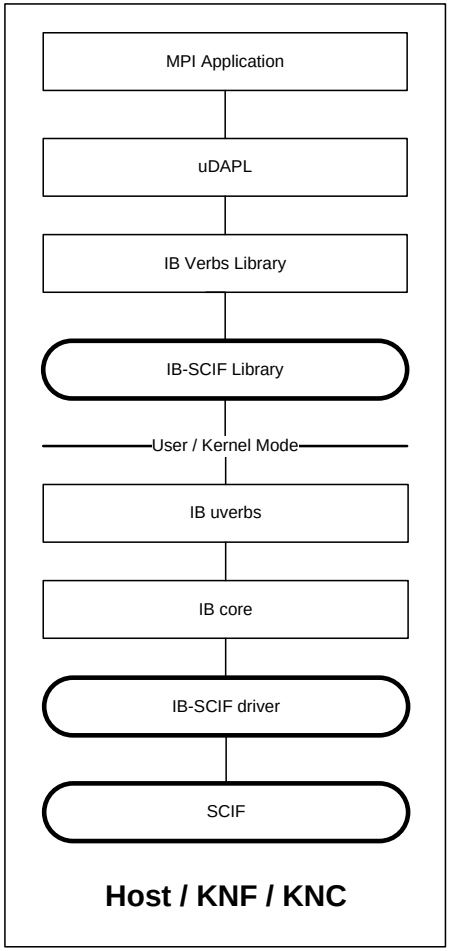
IBSCIF S/W Architecture



RDMA over PCI

(Intel® Xeon™ - Intel® Xeon Phi™)

- Provide standard OFA verbs interfaces for communication between the Intel® Xeon™ and the Intel® Xeon Phi™
- Allows MPI communication between ranks on Intel® Xeon ™ and Intel® Xeon Phi™
- Enables much more standard programming models than proprietary alternatives
  - i.e., MPI, uDAPL, and OFA verbs
- New hardware specific OFA driver (IBSCIF) and Library
  - Plugs into OFA core layer
  - Emulates RDMA QPs in software
  - Uses low level Intel® Xeon Phi™ driver (SCIF) and DMA controllers for Host-Card data transfers

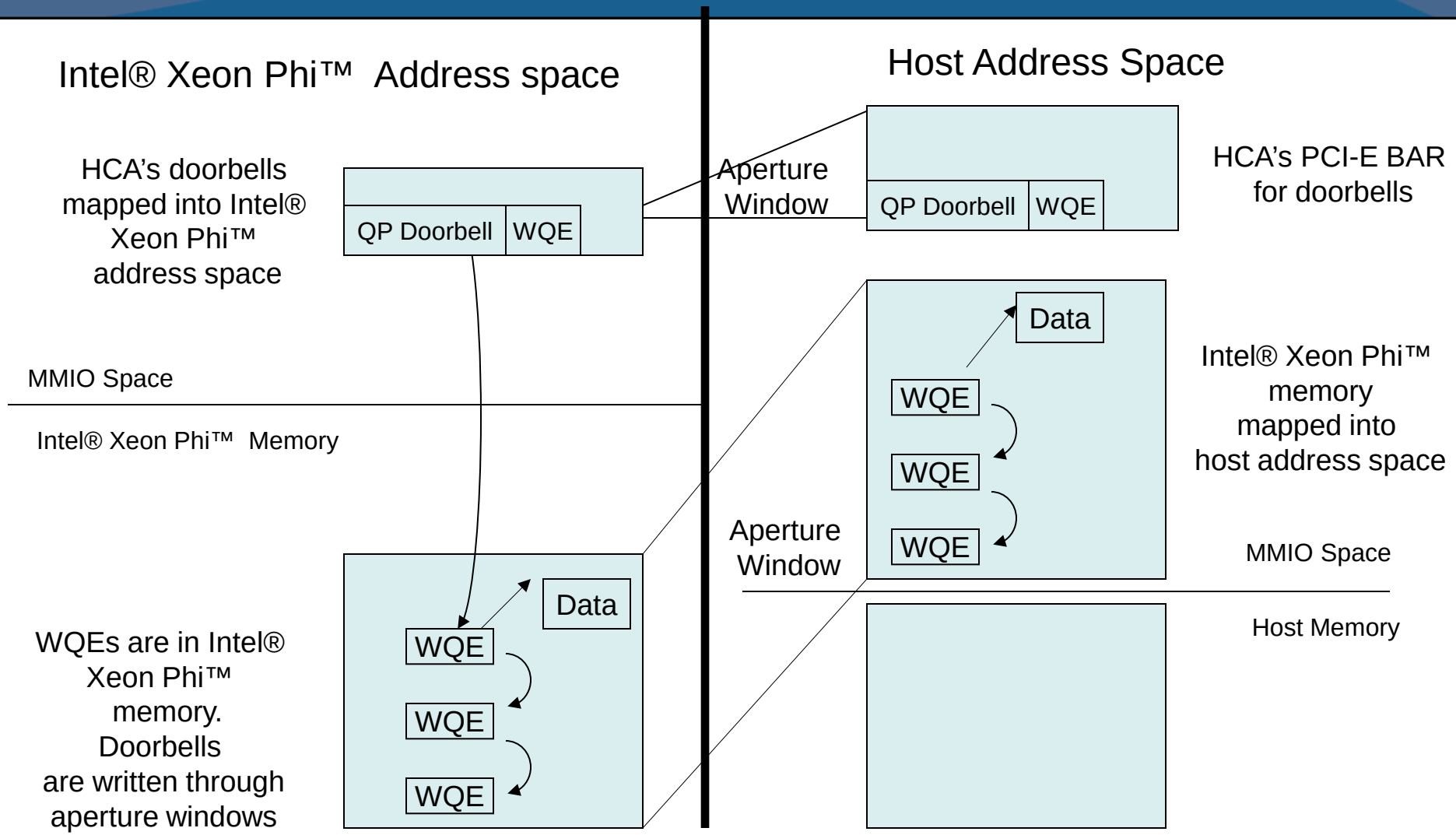


Direct Access to HCA
from Intel® Xeon Phi™

- Goal of Direct HCA Access
  - Provide standard OFA interfaces from a Intel® Xeon Phi™ to an InfiniBand HCA
  - Allows MPI ranks on Intel® Xeon Phi™ to communicate directly to other nodes in the fabric via the HCA
  - Allows heterogeneous MPI with MPI ranks both on Intel® Xeon™ and Intel® Xeon Phi™
  - Enables much more standard programming models (i.e. MPI) than proprietary alternatives
- Make programming the accelerator look more like a normal computer

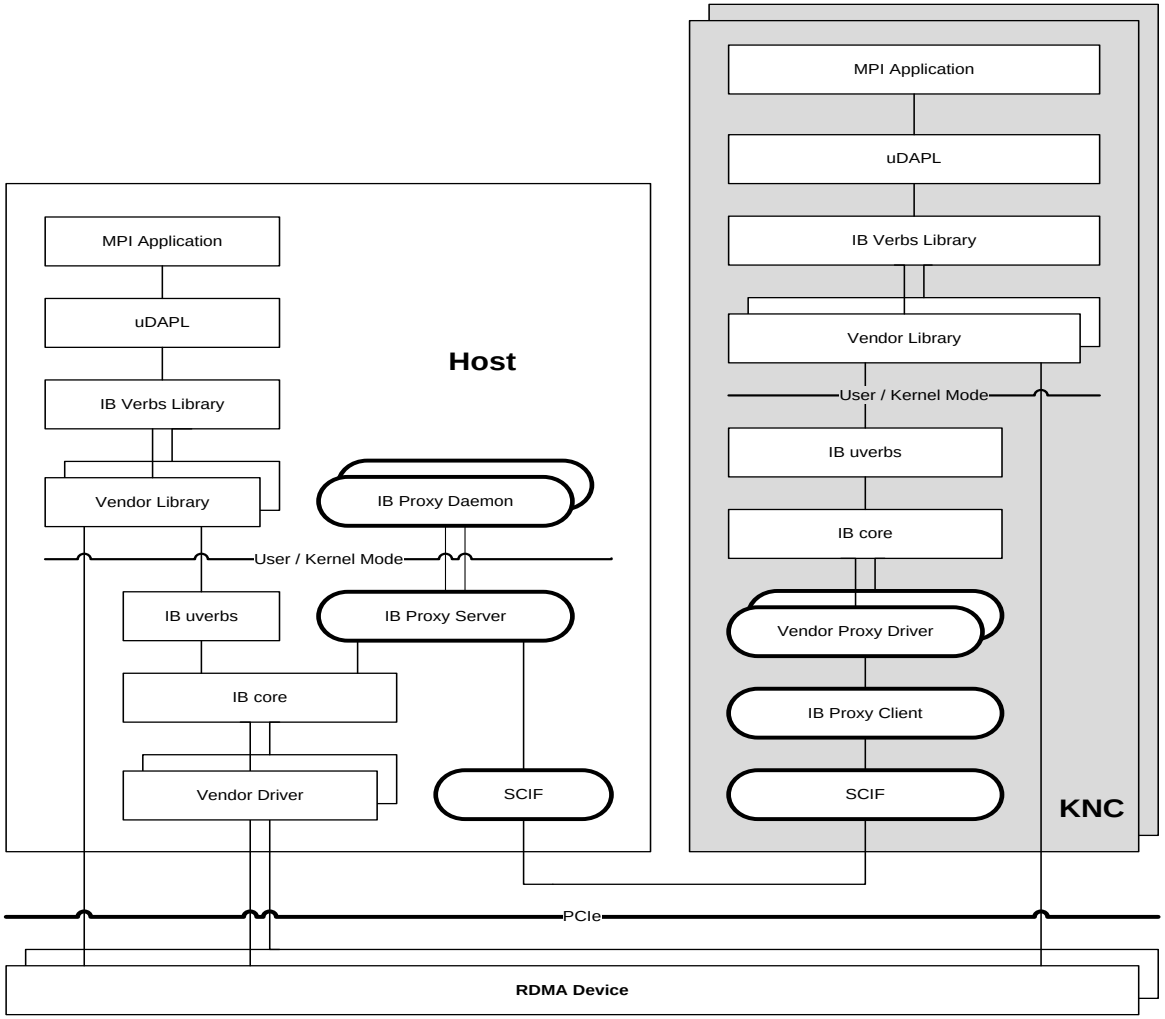
Memory Mapping HCA and Intel® Xeon Phi™ Memory

The memory mapping magic that makes Direct HCA access possible



Intel® Xeon Phi™ CCL Direct Software Architecture

The software magic that makes Direct HCA access possible



Intel® Xeon Phi™ CCL Direct S/W Components

- Vendor Proxy Driver – a HCA specific driver that presents itself to the OFA core layer on the card as if it were the real HCA driver.
  - i.e., presents itself as a mlx4 driver
  - A unique Vendor Proxy Driver is needed for each different HCA type.
- IB Proxy Client Driver – driver on the KNC card that sends requests for priviledged operations to the host CCL-Direct Proxy server for processing.
  - E.g. QP allocation, Memory Registration
- IB Proxy Server Driver/Daemon – host software that calls down to the host OFA core layer to allocate the resources, QPs etc., and map the QP doorbells to the KNC address space.
- All other OFA S/W on KNC is the same S/W from OFA
  - i.e., OFA core, ib verbs library, etc. with one or 2 small changes

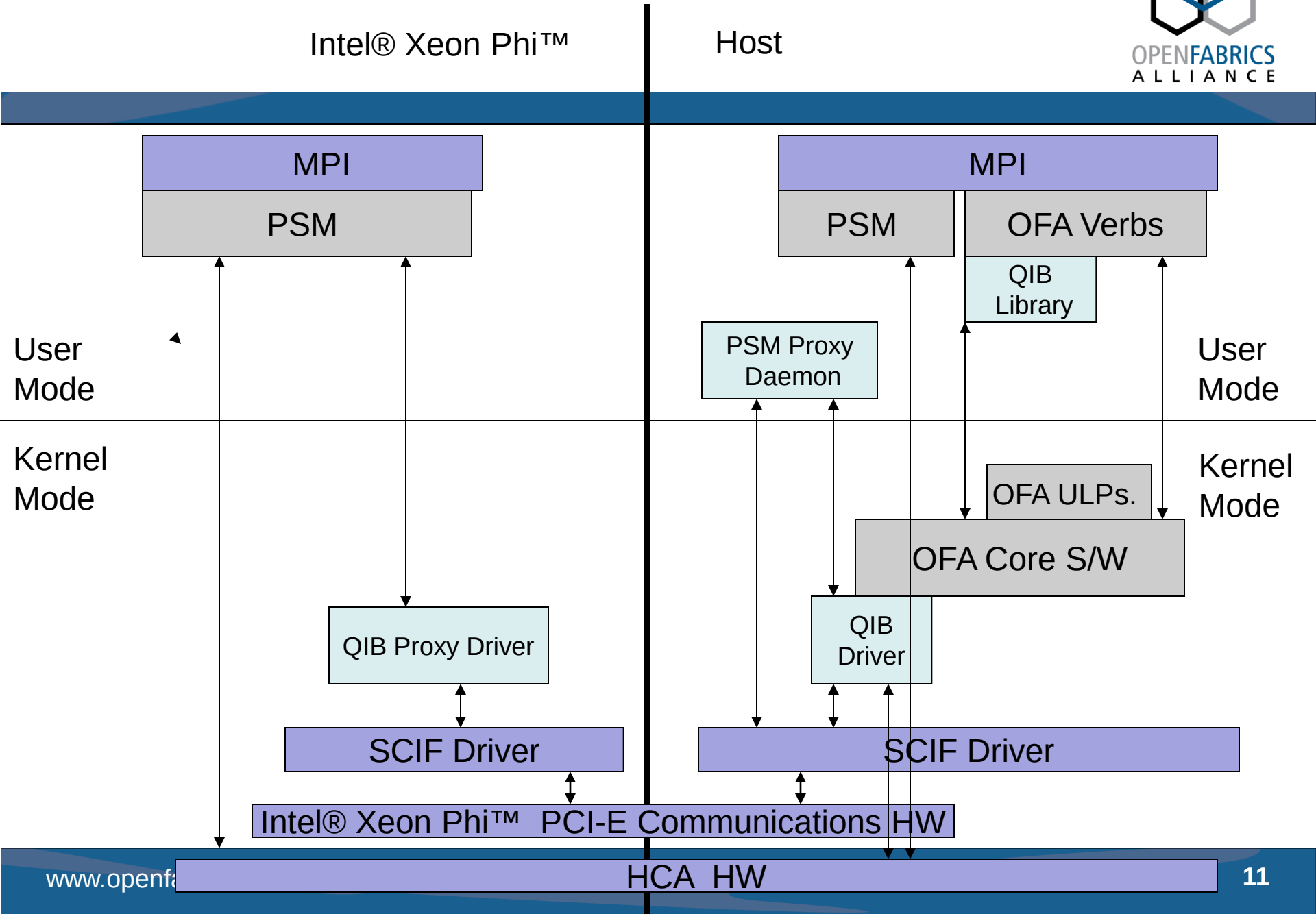
Intel® Xeon Phi™ CCL Direct S/W Components

- Once the QP is allocated and mapped into the user process of the MPI rank on the Intel® Xeon Phi™, access to the HCA happens directly without any host interaction.
- Data transfers to/from the HCA to/from the Intel® Xeon Phi ™ happen directly using PCI-E peer-to-peer transactions.
- Completion interrupts are received on the host and the completion event is proxied back to the card via the IB Proxy Client/Server.

Intel® Xeon Phi™ CCL Direct Limitations

- Currently only supports user space OFA clients,
  - E.g., MPI
- Not all OFA core services are available on the Intel® Xeon Phi™
  - Currently no support for rdma_cm, umad, sa
- CCL Direct provides best path for low latency.
  - Peak BW is limited on some platform configurations as PCI-E P2P BW is not optimized on some platforms

Direct HCA access for Intel ® InfiniBand HCAs



Intel® HCA Direct for Truescale™ HCA

S/W Components

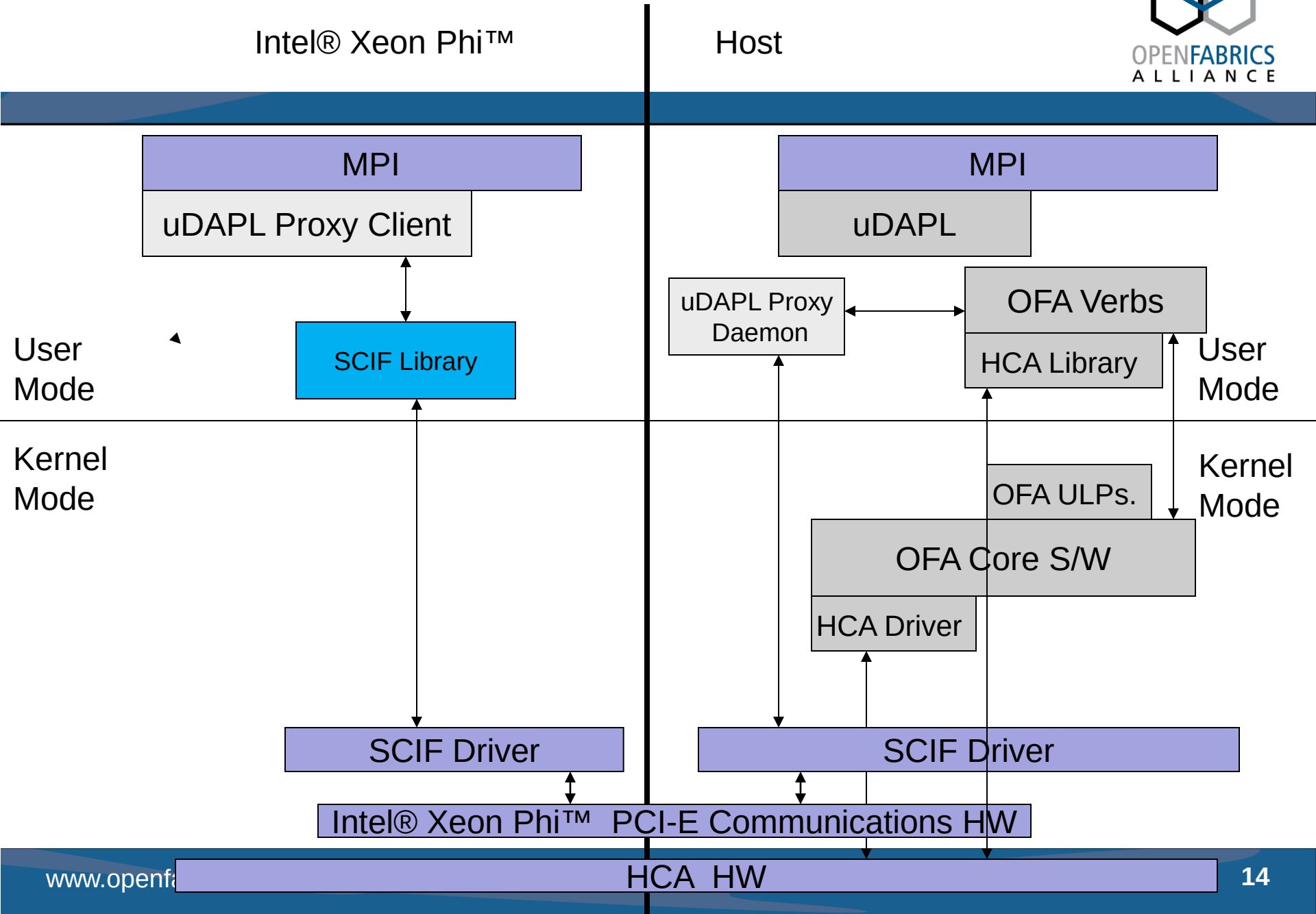
- QIB Proxy Driver
  - Handles privileged operations
  - Mapping HCA resources into KNC address space
  - Proxies calls to the real QIB driver/PSM Proxy Daemon running on the host
- PSM Proxy Daemon
  - Handles proxied calls from the QIB Proxy Driver running on the MIC.
- Once HCA resources are mapped into Intel® Xeon Phi™ address space, access to the HCA are direct via PSM library

Proxy access to HCA via the Host

HCA Proxy uDAPL provider (CCL-Proxy)

- Peer-to-peer BW is limited on some platforms
  - Copying the data from the Intel® Xeon Phi™ and then sending it out the HCA can achieve higher bandwidth in these cases
  - Intel® Xeon Phi™ CCL Proxy handles this case
- CCL-Proxy is a new uDAPL provider
  - Client Library on the Intel® Xeon Phi™
  - Server daemon on the Intel® Xeon™ host
- CCL-Proxy first copies the data from the card to the host and pipelines the data out the HCA

uDAPL Proxy HCA access from Intel® Xeon Phi™



Intel MPI dual-DAPL Provider Support

- The Intel® Xeon Phi™ CCL Direct data path provides low latency and the Intel® Xeon Phi™ CCL Proxy data path provides high bandwidth.
- Intel MPI has a new feature called dual-DAPL.
  - MPI dual-DAPL sends small message down the Intel® Xeon Phi™ CCL Direct path and large messages down the Intel® Xeon Phi™ CCL Proxy path
  - This allows native Intel® Xeon Phi™ MPI applications to get both low latency and high bandwidth
  - This MPI feature is currently an experimental feature in the slipping version of Intel® MPI

OFS S/W for Intel® Xeon Phi™

How do I get the software ?

- All the OFS software for Intel® Xeon Phi™ is open source (Under the OFA BSD+GPL license)
- Intel® MIC Platform Software Stack (MPSS)
  - IBSCIF, CCL-Direct, CCL-Proxy and PSM-Direct
  - Currently shipped as add-on RPMS to the OFED-1.5.4.1 release
  - **http://software.intel.com/en-us/articles/intel-manycore-platform-software-stack-mpss**
- OFED-3.5-MIC Branch Release
  - Currently under development

    http://www.openfabrics.org/downloads/ofed-3.5-mic
  - Technology preview release from OFA that allows people to use the new S/W before it goes upsteam
  - Includes all the standard OFED-3.5 S/W plus the new IBSCIF, CCL-Direct, CCL-Proxy, and PSM-Direct code and support for RHEL EL 6.4

OFS S/W for Intel® Xeon Phi™

Questions ?

Optimization Notice

Optimization Notice

Intel compilers, associated libraries and associated development tools may include or utilize options that optimize for instruction sets that are available in both Intel and non-Intel microprocessors (for example SIMD instruction sets), but do not optimize equally for non-Intel microprocessors. In addition, certain compiler options for Intel compilers, including some that are not specific to Intel micro-architecture, are reserved for Intel microprocessors. For a detailed description of Intel compiler options, including the instruction sets and specific microprocessors they implicate, please refer to the "Intel Compiler User and Reference Guides" under "Compiler Options." Many library routines that are part of Intel compiler products are more highly optimized for Intel microprocessors than for other microprocessors. While the compilers and libraries in Intel compiler products offer optimizations for both Intel and Intel-compatible microprocessors, depending on the options you select, your code and other factors, you likely will get extra performance on Intel microprocessors.

Intel compilers, associated libraries and associated development tools may or may not optimize to the same degree for non-Intel microprocessors for optimizations that are not unique to Intel microprocessors. These optimizations include Intel® Streaming SIMD Extensions 2 (Intel® SSE2), Intel® Streaming SIMD Extensions 3 (Intel® SSE3), and Supplemental Streaming SIMD Extensions 3 (SSSE3) instruction sets and other optimizations. Intel does not guarantee the availability, functionality, or effectiveness of any optimization on microprocessors not manufactured by Intel. Microprocessor-dependent optimizations in this product are intended for use with Intel microprocessors.

While Intel believes our compilers and libraries are excellent choices to assist in obtaining the best performance on Intel and non-Intel microprocessors, Intel recommends that you evaluate other compilers and libraries to determine which best meet your requirements. We hope to win your business by striving to offer the best performance of any compiler or library; please let us know if you find we do not.

Notice revision #20110307


intel®
Software

Backup slides