OPENFABRICS ALLIANCE

12th ANNUAL WORKSHOP 2016

EXTENDING RDMA FOR ALTERNATE FABRICS

Ira Weiny, Network Software Engineer

Intel Corporation

April 5th, 2016

INTEL OMNI-PATH ® ARCHITECTURE (OPA) VERBS COMPATIBILITY

NEW FABRIC, LEGACY APPLICATIONS

It turns out it is possible to support new fabric features and legacy applications without changing the APIs

But…

We can do better

NEW FABRIC FEATURES

- **Per VL MTU**
- **Extended MTU**
  - 8K, “10K”
- **New advanced QoS support**
  - Preemption
  - More advanced buffering
  - Advanced topology support
  - Extended SLs
- **Management**
  - 2K MADs

LEGACY VERBS APPLICATIONS

3 “classes” of Verbs applications

1. **Use rdmacm**
   - Path Record query is hidden
   - Query for Path Records but treat data as opaque
2. **Query for Path Records but interpret data from the Path Record**
3. **Don’t query for Path Records**

LEGACY VERBS APPLICATIONS

1. **Use rdmacm**

- Subclass:
  - Query for Path Records but treat data as opaque (IPoIB, SRP, etc)
- These are the most flexible application
- Already leverage existing infrastructure which abstracts fabric details

LEGACY VERBS APPLICATIONS

2. **Query for Path Records but interpret data from the Path Record**
   - Caution must be used

LEGACY VERBS APPLICATIONS

3. **Don’t query for Path Records**
    - Make assumptions about the fabric which may not be true
    - Technically these are not IB compliant!!!
    - Work with configuration constraints on both IB and OPA

EXISTING VERBS INTERFACE

OPA leverages existing verbs fields by emulating an InfiniBand device

- **Per VL MTU**
  - SL obtained from Path Record
  - SL to VL details kept specific to the hfi1 driver layer
- **Extended MTU**
  - Obtained from Path Record
  - MTU enums are a "natural extension" of the IBTA defined values
- **New QoS support in hardware**
  - SL Obtained from Path Record
  - SL to VL and VL to SL details are contained in the driver
  - SLs are preserved end to end through new mapping tables
  - Extended SLs
    - Verbs is limited to the original 16

MANAGEMENT

- **New scalable MADs**
  - 2K in size
  - Aggregates
- **SMP class version (different name space for OPA)**
  - Different configuration requirements
    - OpenSM will not work
  - Applications no longer require direct access to the fabric
- **InfiniBand GSI MADs are still supported**
  - Still 256 bytes (same class versions)
  - CM (rdmacm)
  - SA
    - Local SA cache of SA data through ibacm
    - Full rdmacm support

VERIFIED ULPS

… and the following work without modifications

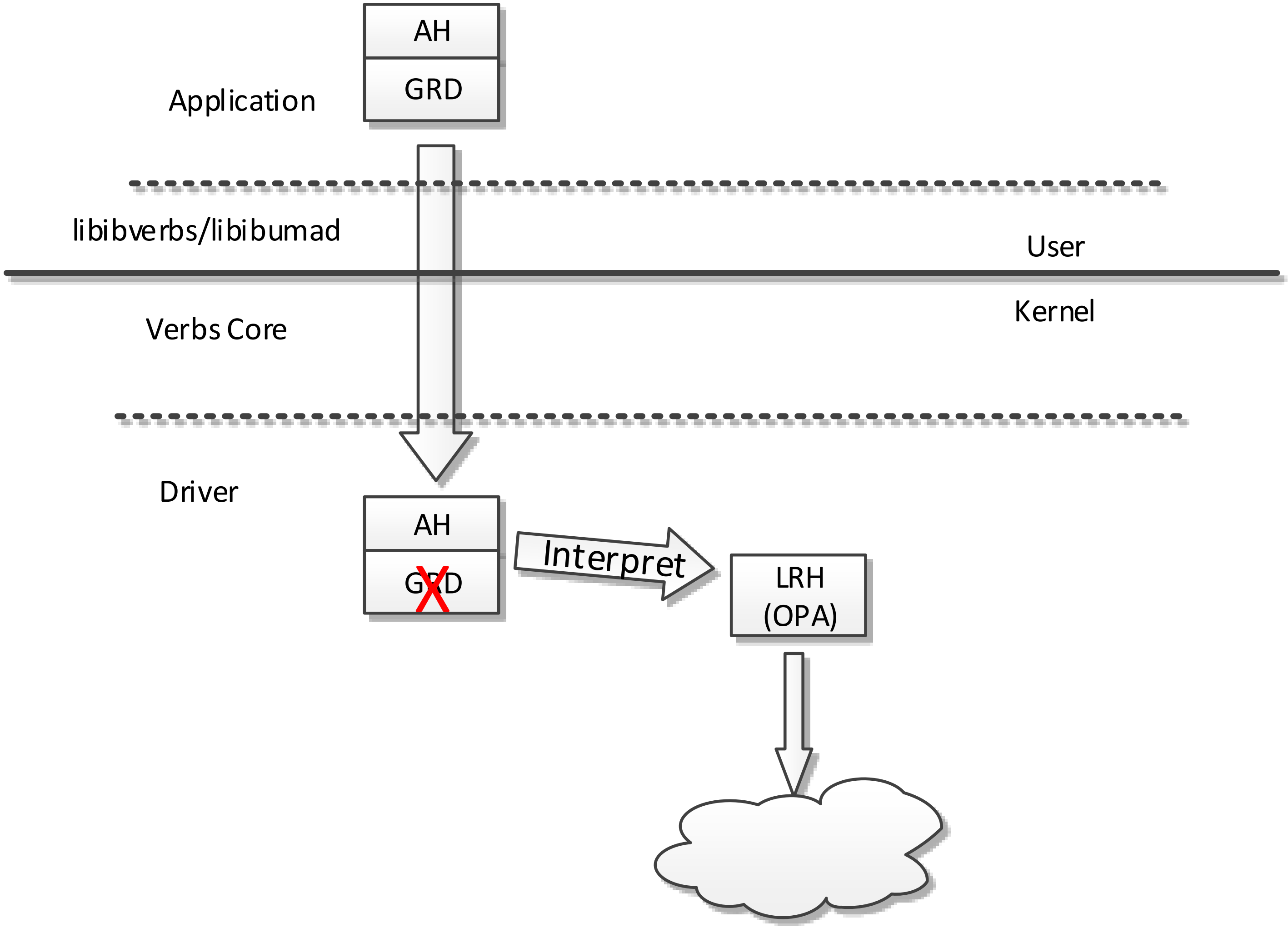
- IPoIB
- SRP
- iSER
- NFSoRDMA
- Lustre
- perftest benchmarks
- MPIs

OPA EXTENDED LIDS

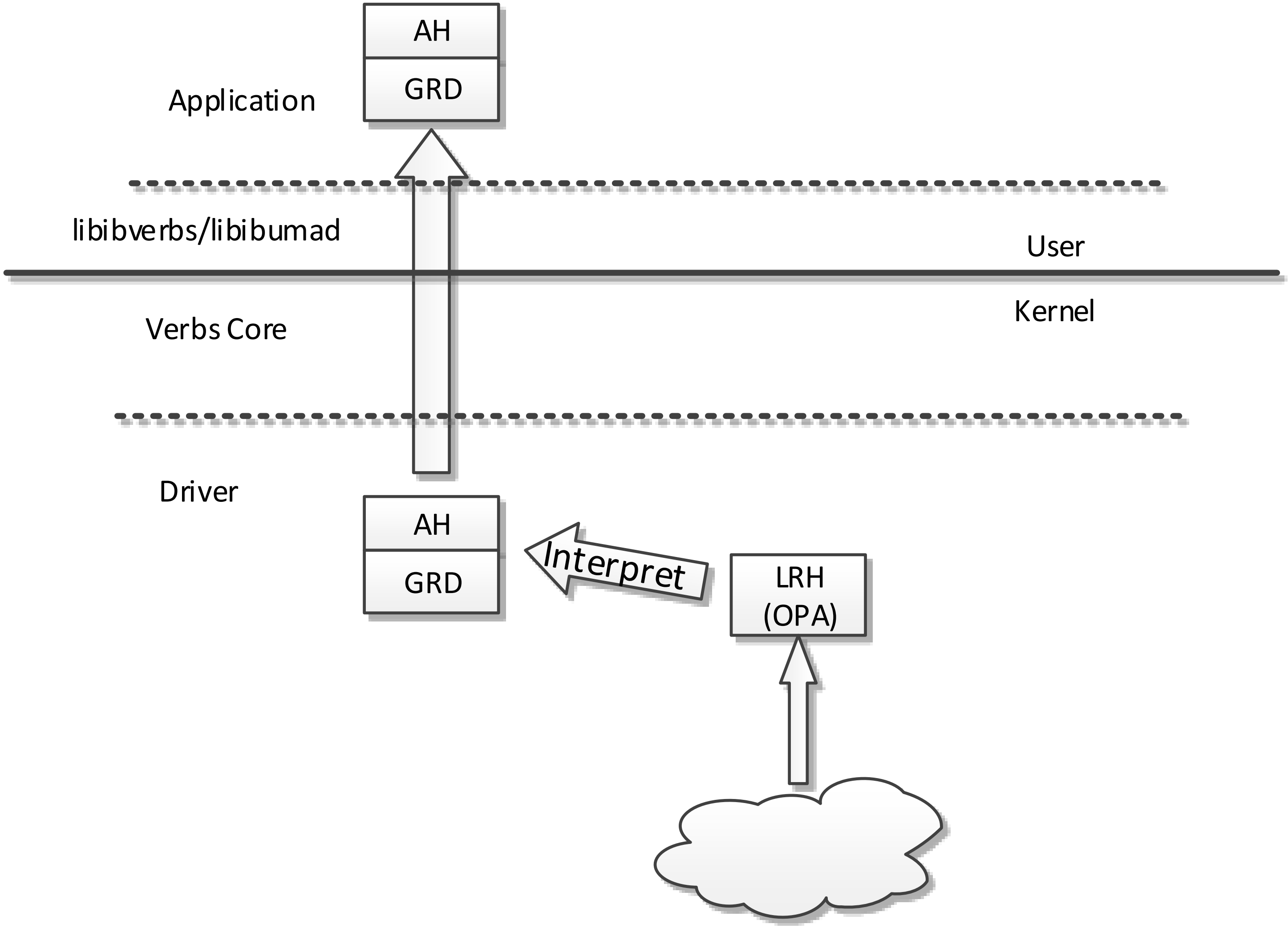
SUPPORTING 32 BIT LIDS

- **OPA LIDs can be up to 32 bits**
  - Rdma stack only supports 16bit LIDs
- **Requirements**
  - Support extended LIDs only on fabrics which require them
  - Minimal to no application changes
- **Architecture**
  - Leverage existing alternate addressing schemes to pass LID data through the verbs stack
  - Keep changes for OPA within OPA specific code as much as possible

APPLICATION SEND



APPLICATION RECV



OPA VERBS APPLICATIONS

- **Most applications already work with OPA**
  - Especially if they work with RoCE
  - A simple audit may be needed
  - Some applications may need to be called with alternate input
    - ib_write_bw for example requires a GID index (similar to RoCE)
- **Some standard ibv_* calls have limitations**
  - ibv_query_port can't return a valid LID or SM LID
  - Applications are not required to use those
  - We require new verbs
- **SM LID is available via sysfs**
  - Is only required for management applications
  - The ibacm daemon is the official SA cache for OPA
- **RoCE Applications have the similar limitations**

GOING FORWARD

VERBS LIMITATIONS

- **MTU is IB specific**
  - Affects RoCE, Usnic, and now OPA
- **Heck pretty much every Verb and data structure is IB specific**
- **Link Layer is required by software for functionality**
  - Bring the new kernel immutable data to user space
  - Remove necessity for Link Layer as an application input

- **Take a more object oriented approach to the interface**
  - Borrow ideas from libfabric/rdmacm
  - Use opaque data structures
  - Use more generic data structures
- **New interfaces which are not InfiniBand specific**

FUTURE MANAGEMENT

- **Remove management from applications**
  - opensm-libs required by openmpi… ☹
  - Leverage ibacm/librdmacm

- **Enhance management for scalability**
  - Enhance MAD timeout mechanisms
    - RMPP total transfer
  - More efficient processing of queues
    - Threads per device
  - General clean up
    - New tracing mechanism

- **Enhance management interfaces to be fabric agnostic**

CALL TO ACTION

- **Have to emulate a verbs device now**
  - Why?!?!?!?!?!
  - USNIC, Gave up and are now a libfabric user…
- **Take a more object oriented approach**
  - What is a QP?
  - What is an address (address handle)?
- **User space has libfabric, what does the kernel have?**

Apps should be agnostic to the fabric…
But the interface has to be agnostic first.

Make a “verbs easy” button

LEGAL DISCLAIMERS

INFORMATION IN THIS DOCUMENT IS PROVIDED IN CONNECTION WITH INTEL PRODUCTS. NO LICENSE, EXPRESS OR IMPLIED, BY ESTOPPEL OR OTHERWISE, TO ANY INTELLECTUAL PROPERTY RIGHTS IS GRANTED BY THIS DOCUMENT. EXCEPT AS PROVIDED IN INTEL'S TERMS AND CONDITIONS OF SALE FOR SUCH PRODUCTS, INTEL ASSUMES NO LIABILITY WHATSOEVER AND INTEL DISCLAIMS ANY EXPRESS OR IMPLIED WARRANTY, RELATING TO SALE AND/OR USE OF INTEL PRODUCTS INCLUDING LIABILITY OR WARRANTIES RELATING TO FITNESS FOR A PARTICULAR PURPOSE, MERCHANTABILITY, OR INFRINGEMENT OF ANY PATENT, COPYRIGHT OR OTHER INTELLECTUAL PROPERTY RIGHT.

All information provided here is subject to change without notice. Contact your Intel representative to obtain the latest Intel product specifications and roadmaps.

Any forecasts of goods and services needed for Intel's operations are provided for discussion purposes only. Intel will have no liability to make any purchase in connection with forecasts published in this document.

Cost reduction scenarios described are intended as examples of how a given Intel- based product, in the specified circumstances and configurations, may affect future costs and provide cost savings. Circumstances will vary. Intel does not guarantee any costs or cost reduction.

Results have been estimated or simulated using internal Intel analysis or architecture simulation or modeling, and provided to you for informational purposes. Any differences in your system hardware, software or configuration may affect your actual performance.

Intel technologies' features and benefits depend on system configuration and may require enabled hardware, software or service activation. Performance varies depending on system configuration. No computer system can be absolutely secure. Check with your system manufacturer or retailer or learn more at intel.com.

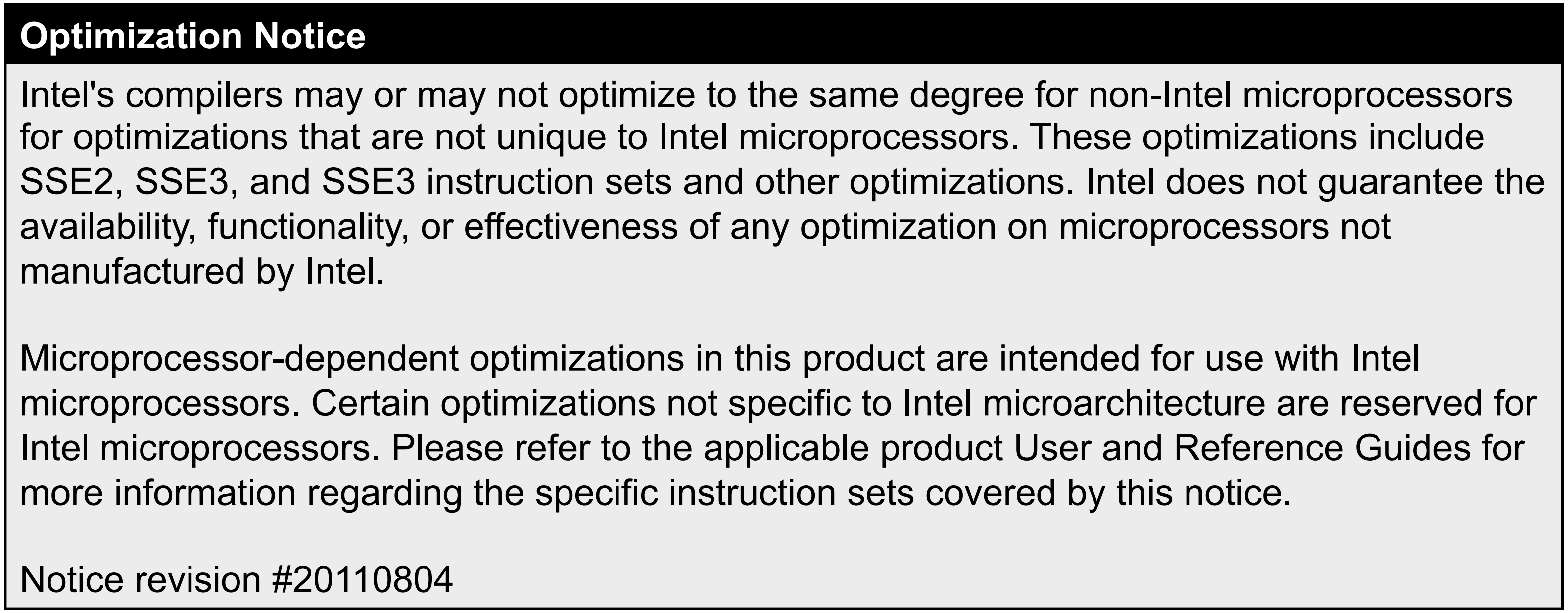
Intel processor numbers are not a measure of performance. Processor numbers differentiate features within each processor family, not across different processor families: Go to: Learn About Intel® Processor Numbers

The products described in this document may contain design defects or errors known as errata which may cause the product to deviate from published specifications. Current characterized errata are available on request.

Copies of documents which have an order number and are referenced in this document, or other Intel literature, may be obtained by calling 1-800-548-4725, or go to: http://www.intel.com/design/literature.htm

Intel, Intel Xeon, Intel Xeon Phi™ are trademarks or registered trademarks of Intel Corporation or its subsidiaries in the United States or other countries.

OPTIMIZATION NOTICE



Optimization Notice

Intel's compilers 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 SSE2, SSE3, and SSE3 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. Certain optimizations not specific to Intel microarchitecture are reserved for Intel microprocessors. Please refer to the applicable product User and Reference Guides for more information regarding the specific instruction sets covered by this notice.

Notice revision #20110804

OPENFABRICS ALLIANCE

12th ANNUAL WORKSHOP 2016

THANK YOU

Ira Weiny, Network Software Engineer

Intel Corporation

intel