OPENFABRICS ALLIANCE

14th ANNUAL WORKSHOP 2018

INTEL® OMNI-PATH ARCHITECTURE AND NVIDIA GPU SUPPORT

Ravindra Babu Ganapathi

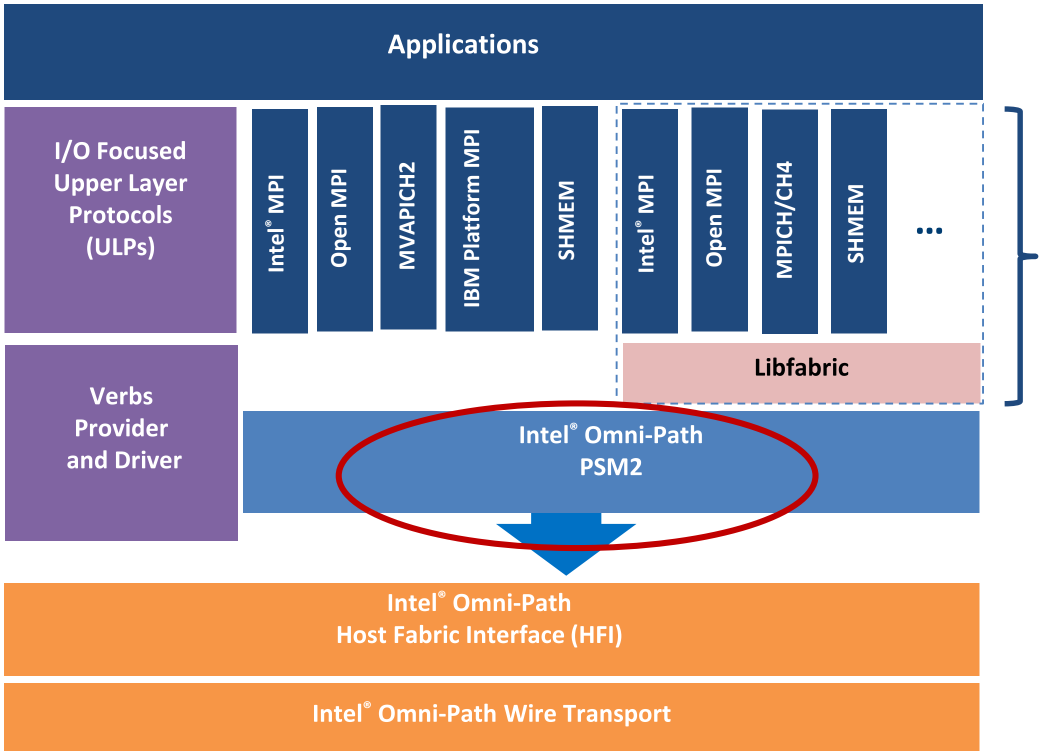
Intel Corporation

[April, 2018]

INTEL® OMNI-PATH ARCHITECTURE HPC

DESIGN FOCUS ARCHITECTED FOR YOUR MPI APPLICATION

Designed for Performance at Extreme Scale

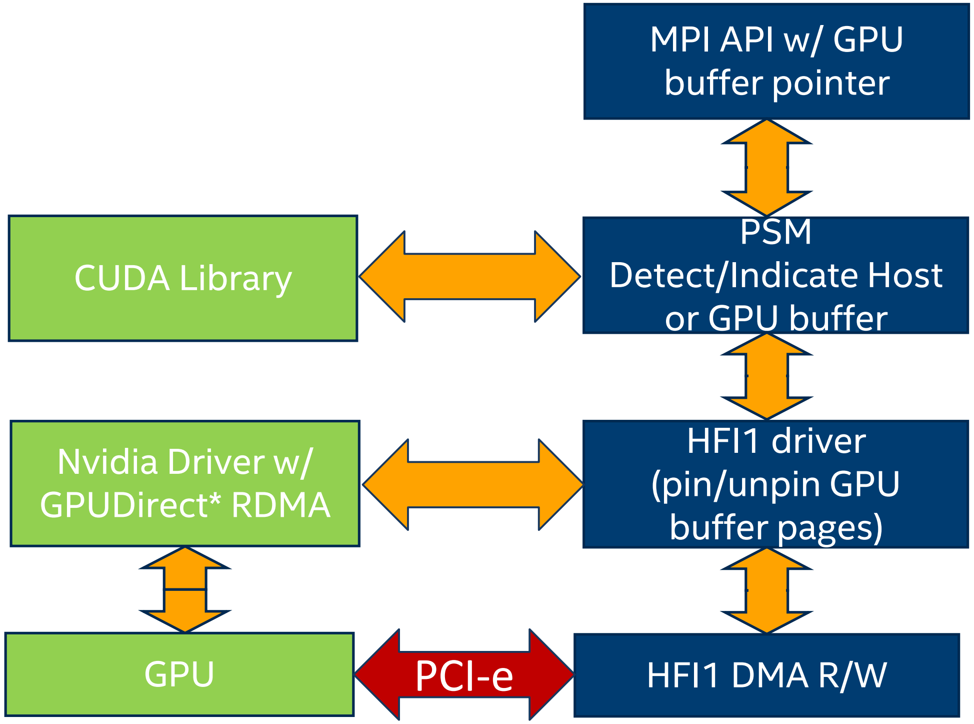

Applications
I/O Focused Upper Layer Protocols (ULPs)
Intel® MPI
Open MPI
MVAPICH2
IBM Platform MPI
SHMEM
Intel® MPI
Open MPI
MPICH/CH4
SHMEM
...
Libfabric
Verbs Provider and Driver
Intel® Omni-Path PSM2
Intel® Omni-Path Host Fabric Interface (HFI)
Intel® Omni-Path Wire Transport

Emerging OpenFabrics Interface (OFI) enabled HPC middleware

PSM2

- PSM2 API is a low level high performing communication interface
- Semantics matches with that of compute middleware such as MPI and/or OFI
- PSM2 EP maps to HW context, Each EP associated with matched queue
- API's are optimized for both latency and bandwidth
- PIO/Eager for small message latency
- DMA/Expected for optimal Bandwidth with large message size
- Intel MPI, Open MPI and MVAPICH2 use PSM2 transport for Omni-Path Fabric
- PSM2 Programmer's Guide available @ https://www.intel.com/content/dam/support/us/en/documents/network-and-i-o/fabric-products/Intel_PSM2_PG_H76473_v8_0.pdf

SIMPLISTIC HIGH-LEVEL VIEW OF OPA + GPU STACK



MPI API w/ GPU buffer pointer

PSM Detect/Indicate Host or GPU buffer

CUDA Library

HFI1 driver (pin/unpin GPU buffer pages)

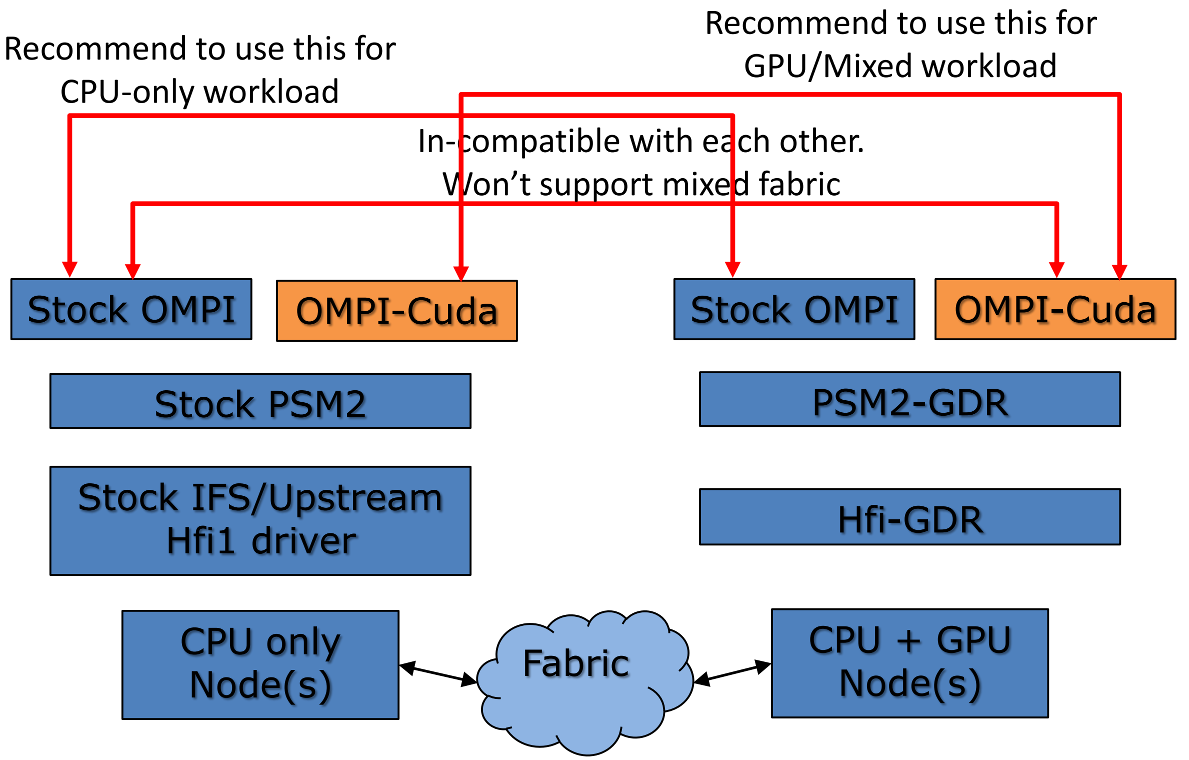
Nvidia Driver w/ GPUDirect* RDMA

GPU

PCI-e

HFI1 DMA R/W

OPEN MPI STACK VIEW

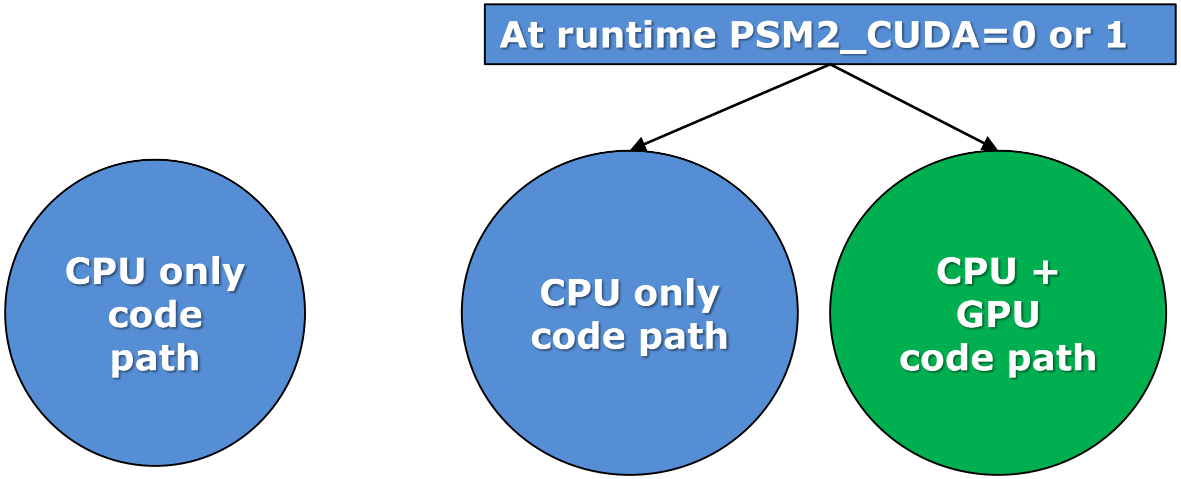


Recommend to use this for CPU-only workload

Recommend to use this for GPU/Mixed workload

In-compatible with each other.
Won't support mixed fabric

PSM2 CODE STRUCTURE



Single Source in GITHUB to build CPU or CPU + GPU binary

RUNNING MPI WORKLOADS

- **PSM2 level runtime ENV vars to mpirun**
  - **GPU Workloads**
    - Set PSM2_CUDA=1
      - Enable cuda code path at runtime
    - Set PSM2_GPUDIRECT=1
      - Enable GPUDirect* RDMA technology
    - PSM2_GDRCPY=1 by default when PSM2_GPUDIRECT=1
      - Enables low latency transfers for small messages
  - **CPU-only workloads:** default values, no need to set the variables
    - PSM2_CUDA=0
    - PSM2_GPUDIRECT=0
- **Example using CUDA-aware Open MPI**
  - mpirun -np 2 --map-by ppr:1:node -host host1,host2 -x PSM2_CUDA=1 -x PSM2_GPUDIRECT=1 -x HFI_UNIT=1 ./osu_latency -d cuda D D

PSM2 GPU PLATFORM SPECIFIC TUNING

- **Defaults are expected to be optimal in most cases**
- **PSM2_GDR_COPY_SEND_THRESH (32 bytes)**
  - Send side threshold for GDR Copy, above this limit uses GPUDirect technology
- **PSM2_GDR_COPY_RECV_THRESH (64000 bytes)**
  - Send side threshold for GDR Copy, above this uses GPUDirect technology
- **PSM2_GPUDIRECT_SEND_THRESH (30000 bytes)**
  - Above this threshold switch to 2MB window pipeline sends through the host
- **PSM2_GPUDIRECT_RECV_THRESH (UINT_MAX)**
  - Above this threshold switch to 2MB window pipeline receives through the host
  - Default assumes both OPA and GPU are on the same CPU socket
  - Set this variable when both OPA and GPU are connected to different sockets

PSM2 NUMA AWARENESS

- **PSM2 Device Selection algorithm**
  - Combination of first and best fit algorithms
    - Find all active OPA devices (units) in system.
    - If only one device found then return and use this device for all communication
  - Scan for OPA devices that are on same NUMA node(root complex)
    - Uniformly distribute the process among the OPA devices found
    - If no devices are found in current NUMA node, then select OPA device from remote NUMA node.

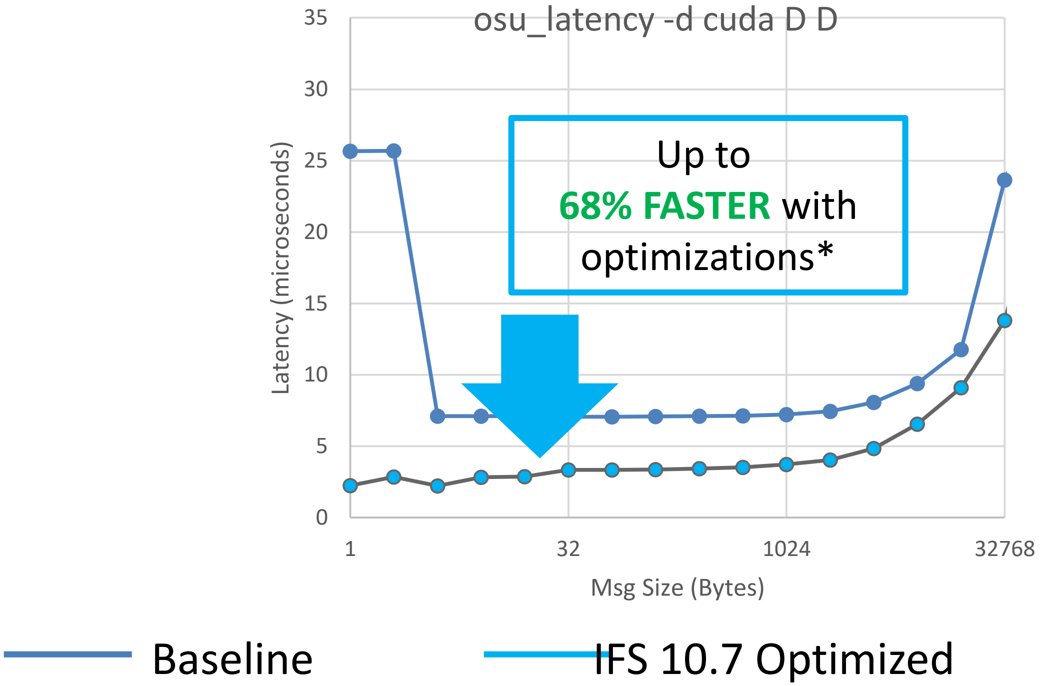
Ravindra Babu Ganapathi ; Aravind Gopalakrishnan ; Russell W. McGuire,

MPI Process and Network Device Affinitization for Optimal HPC Application Performance, High-Performance Interconnects (HOTI), 2017

OPEN MPI ENABLING

- **Open MPI handles GPU Buffers when built with CUDA Support**
- **Converter flag added specific to PSM2 MTL**
  - Indicates PSM2 support for GPUDirect* to OPAL layer
  - Flag allows OPAL layer to skip CUDA convertor set up phase
  - Facilitates to bypass CUDA transfers in OPAL for contiguous MPI data-types
  - PSM2 automatically handles all GPU buffers
- **PSM2 handles all pt2pt and blocking collectives**
- **Open MPI continues to handle non-contiguous MPI data-types**
  - Pack/Unpack datatypes into contiguous memory before transfers
- **Open MPI Upstream info**
  - Enabled Open MPI branches v2.x, v3.0.x, v3.1.x to support OPA + GPU
  - Version 2.1.3 released with this feature (released 03/15/18)
  - Upcoming versions v3.1.0, v3.0.1 will also have the feature (currently both are release candidates)

GPU BUFFER TRANSFER LATENCY - UPCOMING INTEL® OPA OPTIMIZATIONS NVIDIA* CORPORATION TESLA P100*

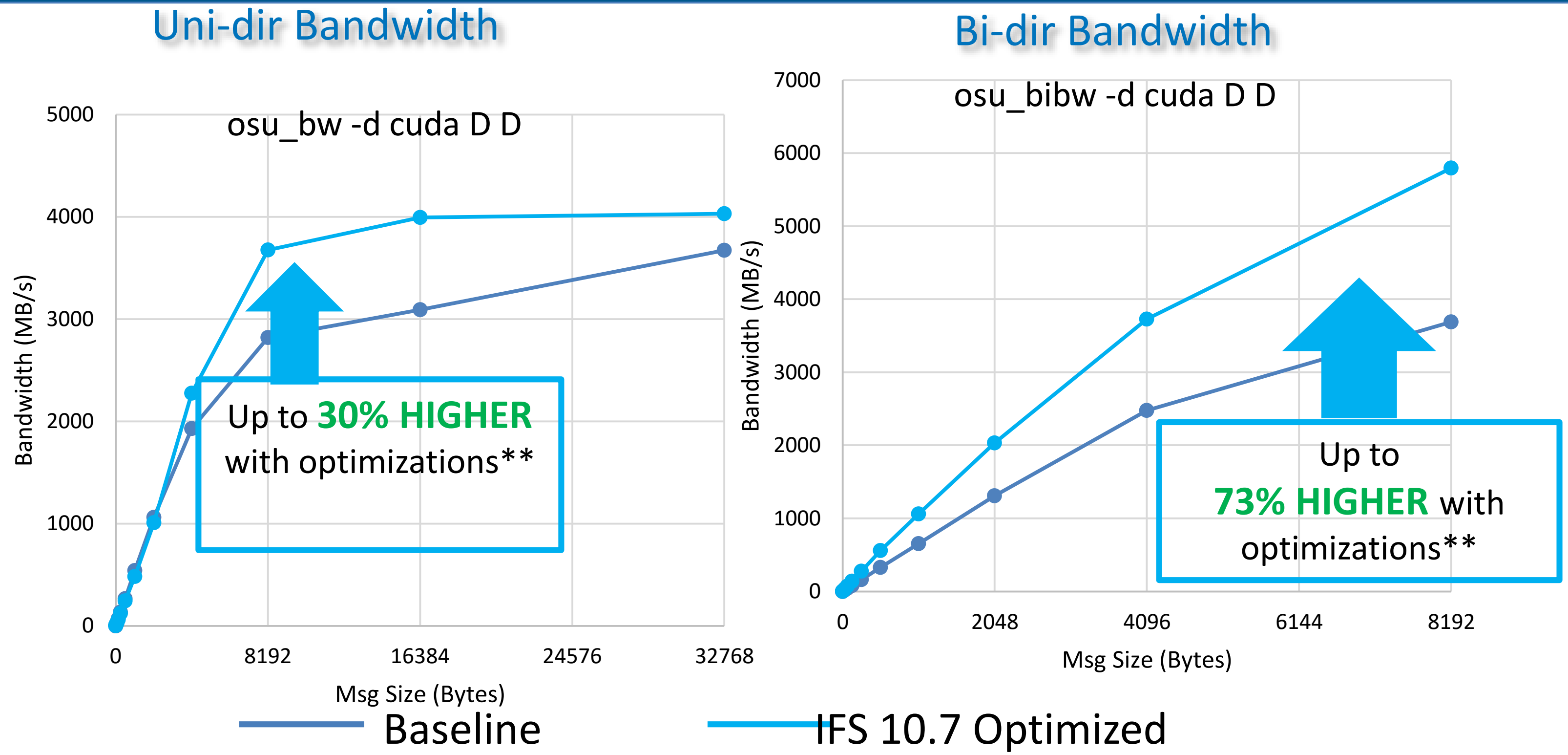


Up to **68% FASTER** with optimizations*

Intel® Xeon® processor E5-2699 v4, SLES 12.3 4.4.73-5-default, 0xb00001b microcode. Intel Turbo Boost Technology enabled. Dual socket servers connected back to back with no switch hop. NVIDIA* P100 and Intel® OPA HFI both connected to second CPU socket. 64GB DDR4 memory per node, 2133MHz.
OSU Microbenchmarks version 5.3.2 Open MPI 2.1.2-cuda-hfi as packaged with IFS 10.7.
* 68% higher claim based on 4 byte latency ** 30% higher claim based on 8KB uni-directional bandwidth. 73% higher claim based on 64B bi-directional bandwidth.
Optimized performance: mpirun -np 2 --map-by ppr:1:node -host host1,host2 -x PSM2_CUDA=1 -x PSM2_GPUDIRECT=1 -x HFI_UNIT=1 ./osu_latency -d cuda D D
Baseline performance: same as above but with "-x PSM2_GDRCPY=0 (off)"

Performance estimates were obtained prior to implementation of recent software patches and firmware updates intended to address exploits referred to as "Spectre" and "Meltdown." Implementation of these updates may make these results inapplicable to your device or system. Software and workloads used in performance tests may have been optimized for performance only on Intel microprocessors. Performance tests, such as SYSmark and MobileMark, are measured using specific computer systems, components, software, operations and functions. Any change to any of those factors may cause the results to vary. You should consult other information and performance tests to assist you in fully evaluating your contemplated purchases, including the performance of that product when combined with other products. For more complete information visit http://www.intel.com/performance.

GPU BUFFER TRANSFER BANDWIDTH - UPCOMING INTEL® OPA OPTIMIZATIONS NVIDIA* CORPORATION TESLA P100*



Uni-dir Bandwidth

Bi-dir Bandwidth

Up to 30% HIGHER with optimizations**

Up to 73% HIGHER with optimizations**

Intel® Xeon® processor E5-2699 v4, SLES 12.3 4.4.73-5-default, 0xb00001b microcode. Intel Turbo Boost Technology enabled. Dual socket servers connected back to back with no switch hop. NVIDIA* P100 and Intel® OPA HFI both connected to second CPU socket. 64GB DDR4 memory per node, 2133MHz.

OSU Microbenchmarks version 5.3.2 Open MPI 2.1.2-cuda-hfi as packaged with IFS 10.7.

* 68% higher claim based on 4 byte latency ** 30% higher claim based on 8KB uni-directional bandwidth. 73% higher claim based on 64B bi-directional bandwidth.

Optimized performance: mpirun -np 2 --map-by ppr:1:node -host host1,host2 -x PSM2_CUDA=1 -x PSM2_GPUDIRECT=1 -x HFI_UNIT=1 ./osu_latency -d cuda D D

Baseline performance: same as above but with “-x PSM2_GDRCPY=0 (off)”

Performance estimates were obtained prior to implementation of recent software patches and firmware updates intended to address exploits referred to as "Spectre" and "Meltdown." Implementation of these updates may make these results inapplicable to your device or system. Software and workloads used in performance tests may have been optimized for performance only on Intel microprocessors. Performance tests, such as SYSmark and MobileMark, are measured using specific computer systems, components, software, operations and functions. Any change to any of those factors may cause the results to vary. You should consult other information and performance tests to assist you in fully evaluating your contemplated purchases, including the performance of that product when combined with other products. For more complete information visit http://www.intel.com/performance.

NOTICES AND DISCLAIMERS

INFORMATION IN THIS DOCUMENT IS PROVIDED IN CONNECTION WITH INTEL® PRODUCTS. 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. Intel products are not intended for use in medical, life-saving, life-sustaining, critical control or safety systems, or in nuclear facility applications.

Intel products 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.

Intel may make changes to dates, specifications, product descriptions, and plans referenced in this document at any time, without notice.

This document may contain information on products in the design phase of development. The information herein is subject to change without notice. Do not finalize a design with this information.

Intel processors of the same SKU may vary in frequency or power as a result of natural variability in the production process.

Performance estimates were obtained prior to implementation of recent software patches and firmware updates intended to address exploits referred to as "Spectre" and "Meltdown." Implementation of these updates may make these results inapplicable to your device or system.

Designers must not rely on the absence or characteristics of any features or instructions marked "reserved" or "undefined." Intel reserves these for future definition and shall have no responsibility whatsoever for conflicts or incompatibilities arising from future changes to them.

Intel Corporation or its subsidiaries in the United States and other countries may have patents or pending patent applications, trademarks, copyrights, or other intellectual property rights that relate to the presented subject matter. The furnishing of documents and other materials and information does not provide any license, express or implied, by estoppel or otherwise, to any such patents, trademarks, copyrights, or other intellectual property rights.

Some features may require you to purchase additional software, services or external hardware.

Performance tests and ratings are measured using specific computer systems and/or components and reflect the approximate performance of Intel products as measured by those tests. Any difference in system hardware or software design or configuration may affect actual performance. Buyers should consult other sources of information to evaluate the performance of systems or components they are considering purchasing. For more information on performance tests and on the performance of Intel products, visit Intel Performance Benchmark Limitations

Intel, the Intel logo are trademarks or registered trademarks of Intel Corporation or its subsidiaries in the United States and other countries.

Other names and brands may be claimed as the property of others.

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

14th ANNUAL WORKSHOP 2018

THANK YOU

Ravindra Babu Ganapathi

Intel Corporation

intel®