
12th ANNUAL WORKSHOP 2016

# OPENFABRICS INTERFACES: PAST, PRESENT, AND FUTURE

Sean Hefty

**OFIWG Co-Chair**

**[ April 5th, 2016 ]**

# OFIWG: *develop … interfaces aligned with … application needs*

## Open Source

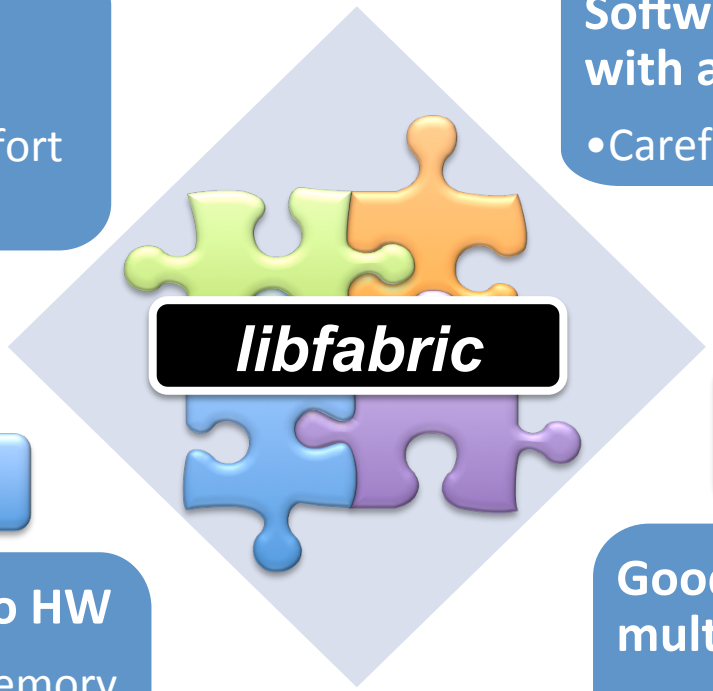
**Expand open source community**

- Inclusive development effort
- App and HW developers

## Application-Centric

**Software interfaces aligned with application requirements**

- Careful analysis of requirement




## Scalable

**Optimized SW path to HW**

- Minimize cache and memory footprint
- Reduce instruction count
- Minimize memory accesses

## Implementation Agnostic

**Good impedance match with multiple fabric hardware**

- InfiniBand*, iWarp, RoCE, Ethernet, UDP offload, Intel®, Cray*, IBM*, others

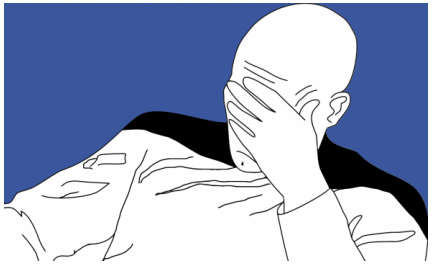
* Other names and brands may be claimed as the property of others

# OFI APPLICATION REQUIREMENTS

*Give us a **high**-level interface!*

*Give us a **low**-level interface!*

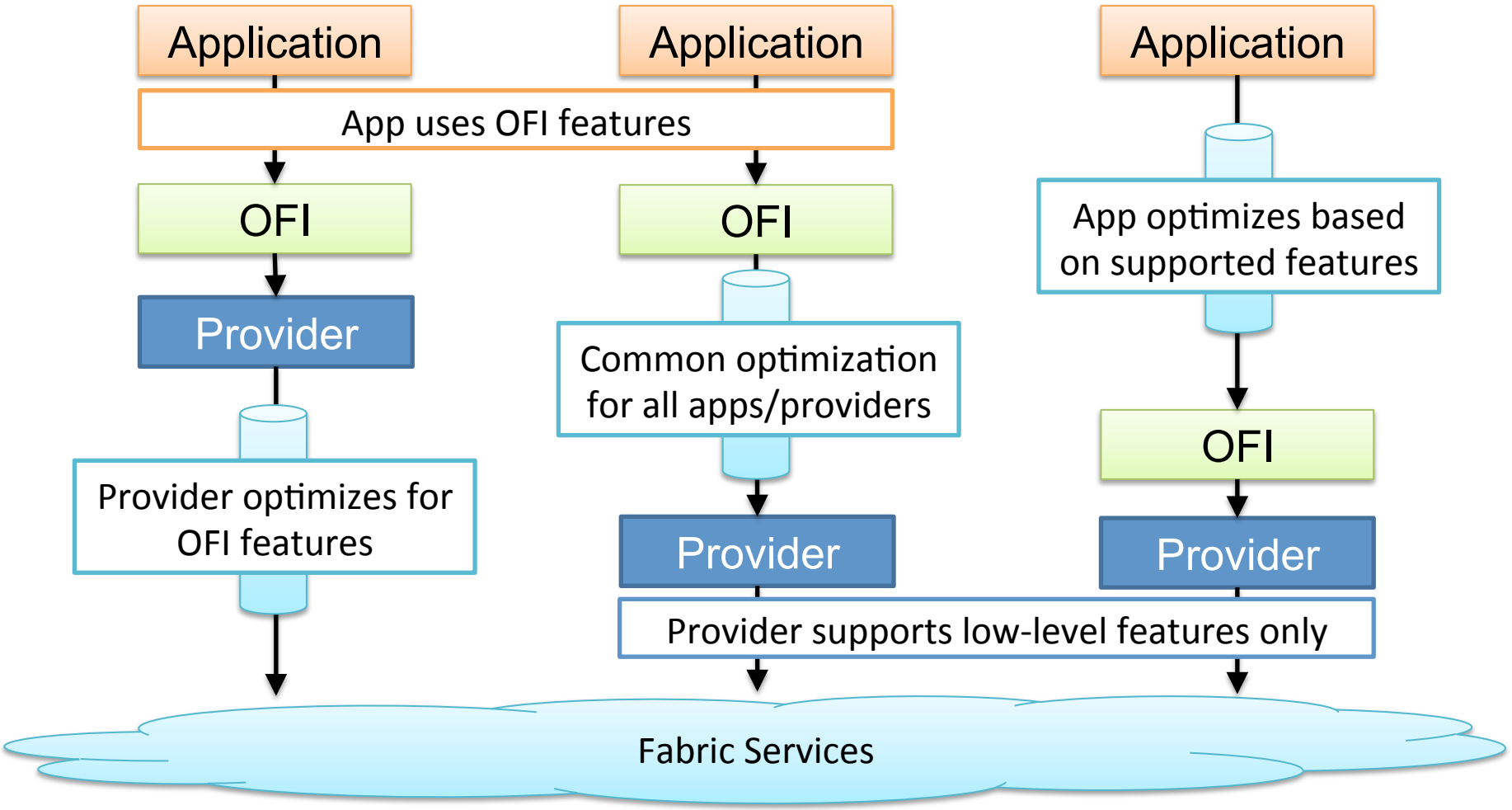
MPI developers



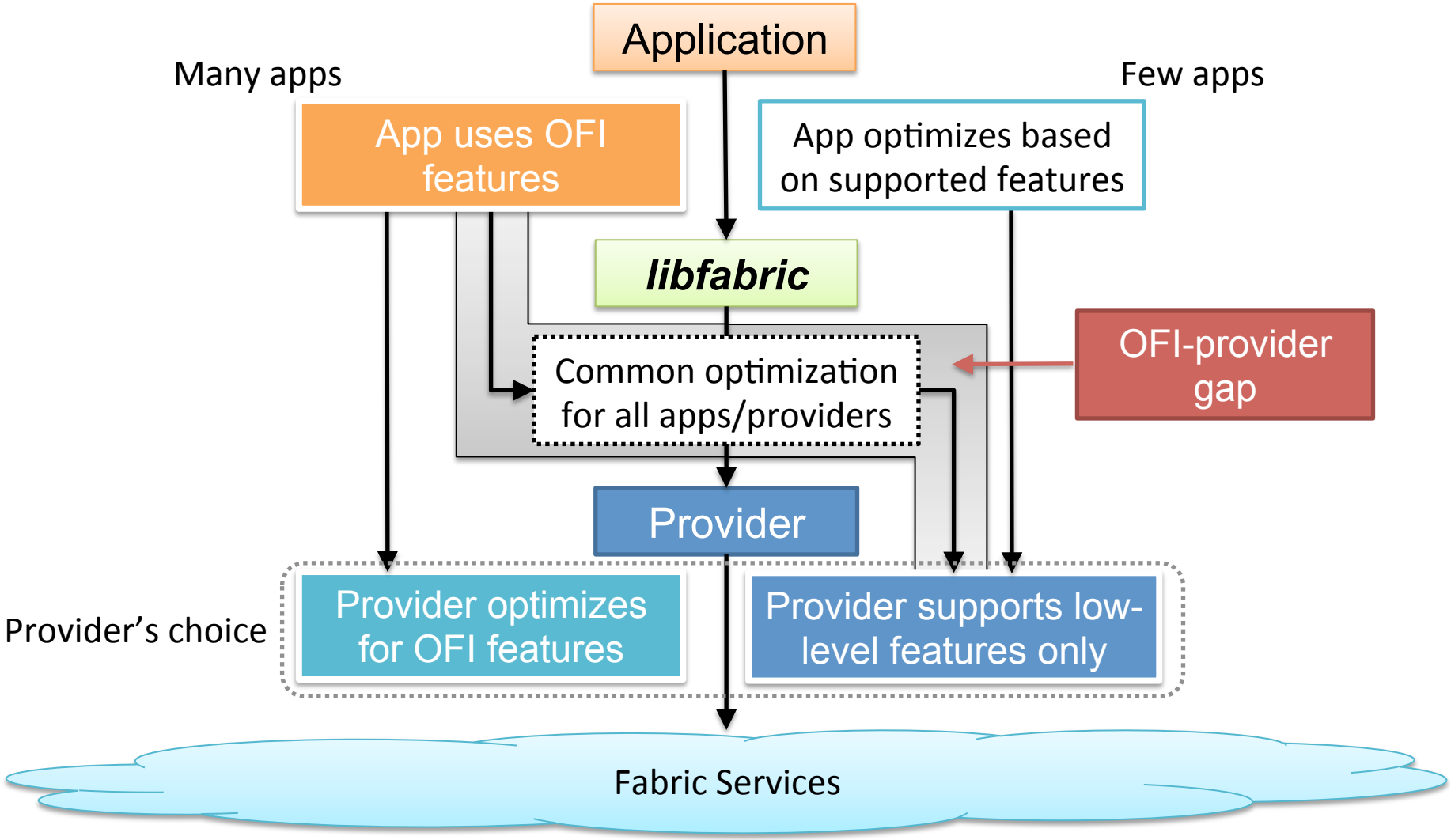
**OFI strives to meet *both* requirements**

# OFI SOFTWARE DEVELOPMENT STRATEGIES

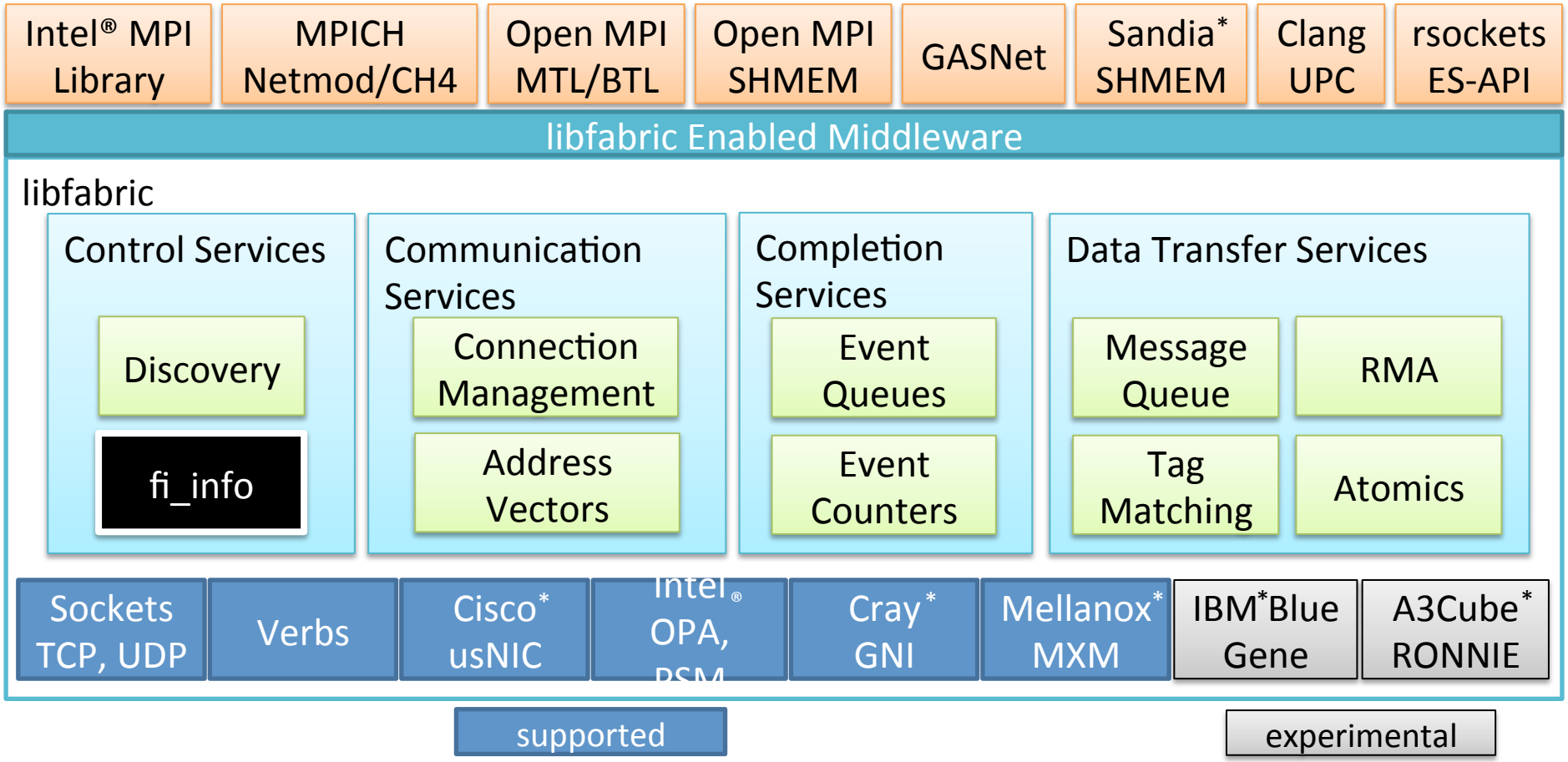
## One Size Does Not Fit All



# OFI DEVELOPMENT STATUS

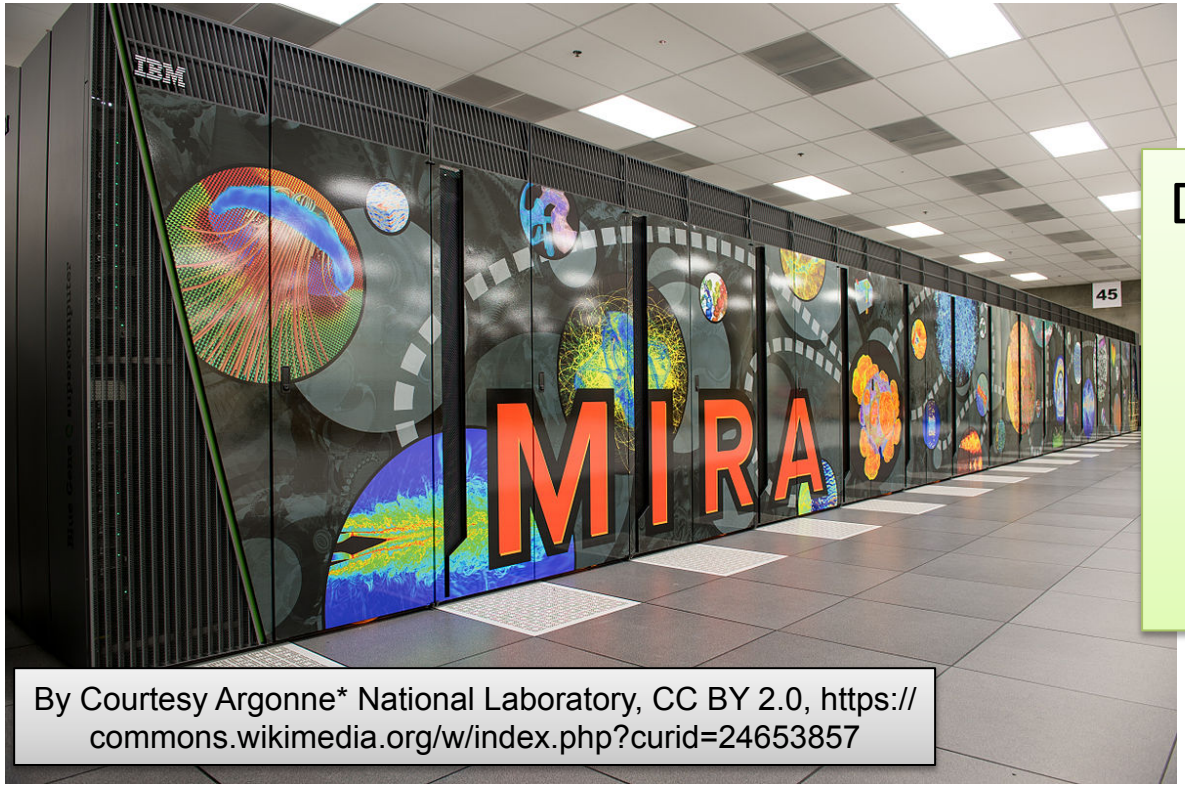


# OFI LIBFABRIC COMMUNITY




**Because of the OFI-provider gap, not all apps work with all providers**

# LIBFABRIC SCALABILITY




By Courtesy Argonne* National Laboratory, CC BY 2.0, https://commons.wikimedia.org/w/index.php?curid=24653857

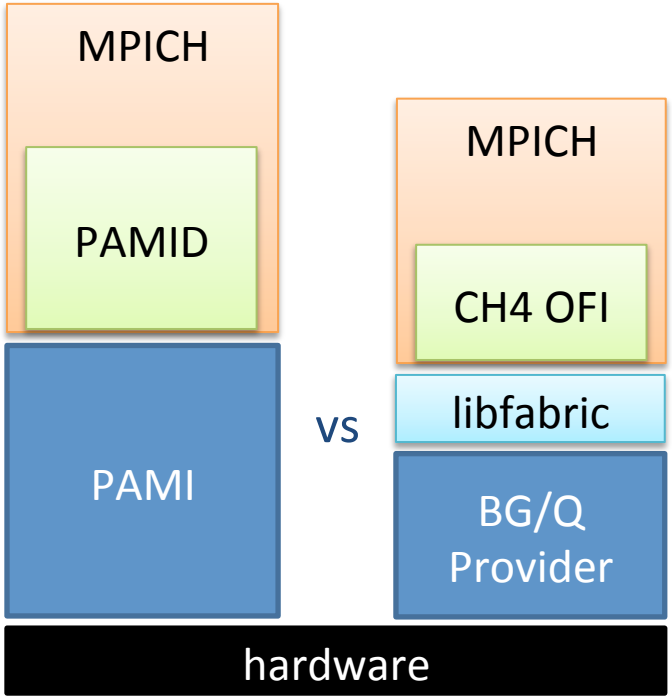
Developed to evaluate the Aurora software stack at scale and assist applications in the transition from Mira to Aurora

Native provider implementation that directly uses the Blue Gene/Q hardware and network interfaces for communication

# LIBFABRIC SCALABILITY

PAMI and libfabric performance

*Completely subjective* software stack comparison

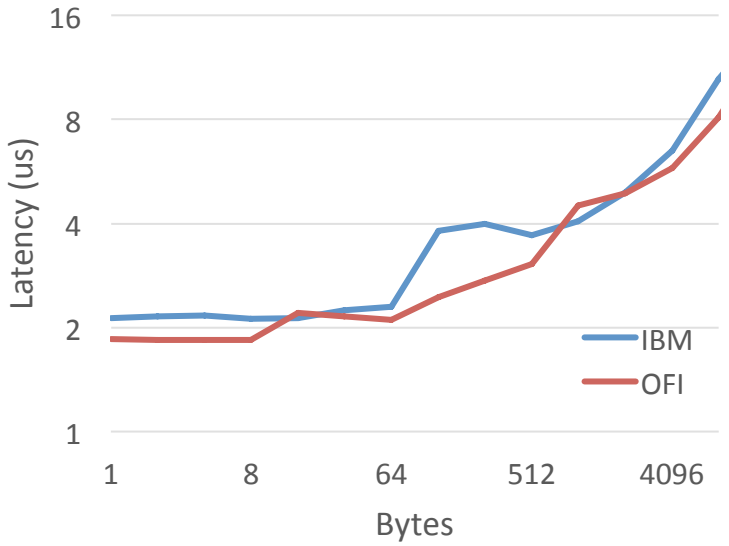


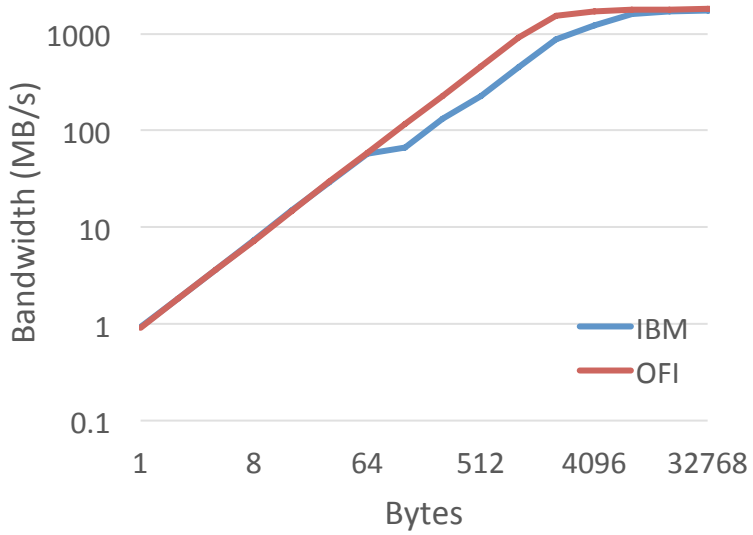

32 nodes on ALCF Vesta machine

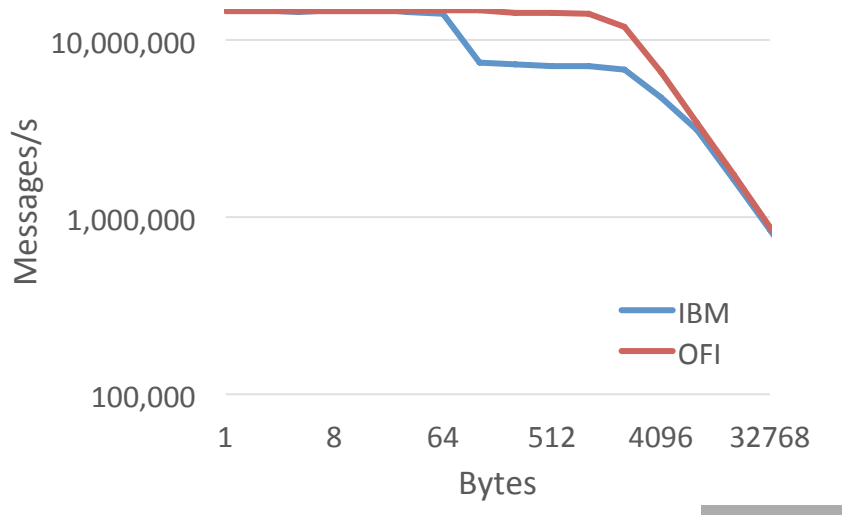
- **IBM MPICH / PAMI**
  - IBM XL C compiler for BG, v12.1
  - Optimized for single-threaded latency
  - …/comm/xl.legacy.ndebug/bin/mpicc
  - v1r2m2
- **MPICH / CH4 / libfabric**
  - gcc 4.4.7
  - global locks, inline, direct, etc.
  - *Provider not optimized for performance*

# LIBFABRIC SCALABILITY

OSU* MPI Performance Tests v5.0



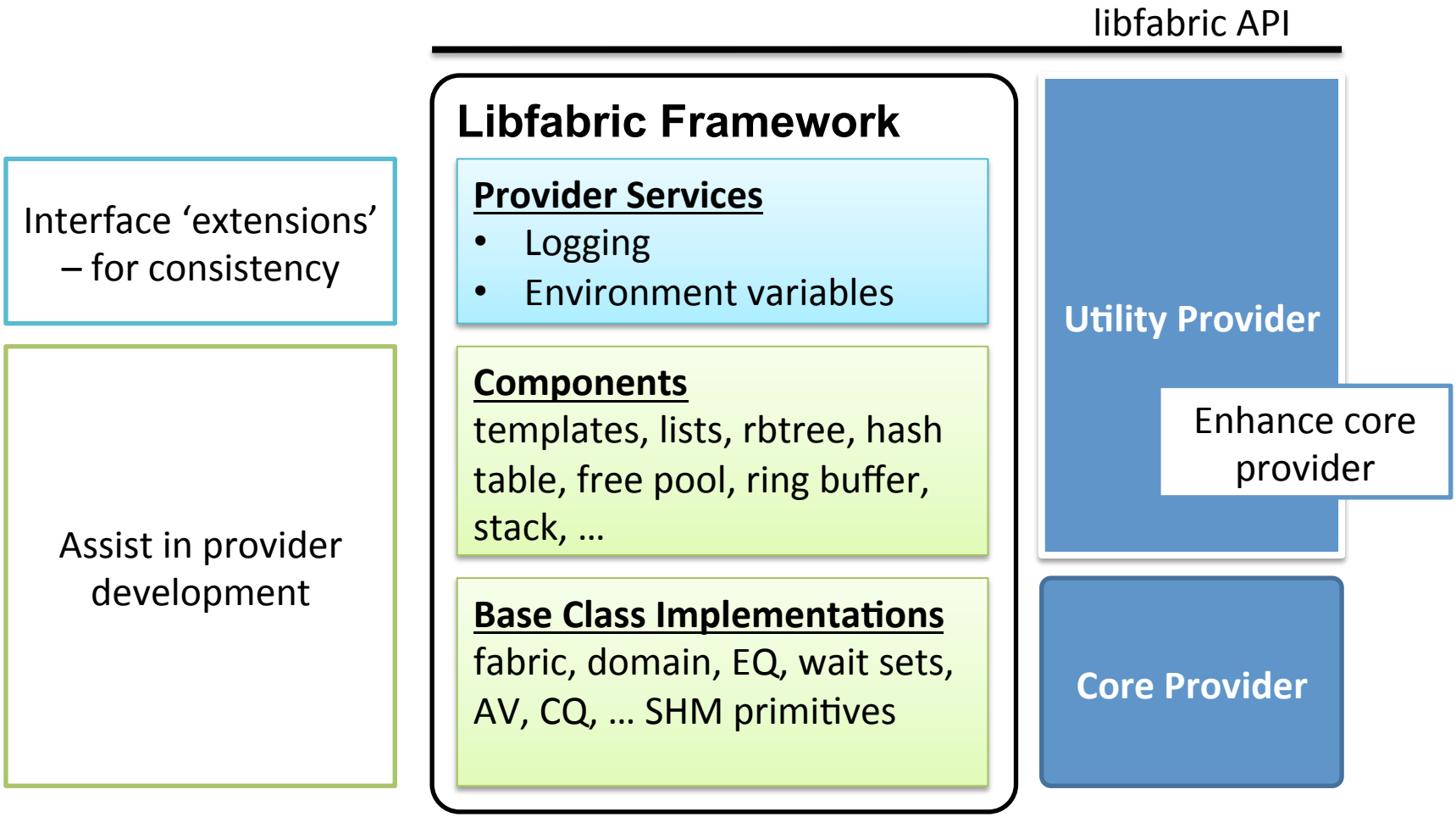

MPI scale out testing:

- cpi – 1M ranks,
- ISx benchmark – 0.5M ranks

Tests document performance of components on a particular test, in specific systems. Differences in hardware, software, or configuration will affect actual performance. Consult other sources of information to evaluate performance as you consider your purchase. For more complete information about performance and benchmark results, visit http://www.intel.com/performance.

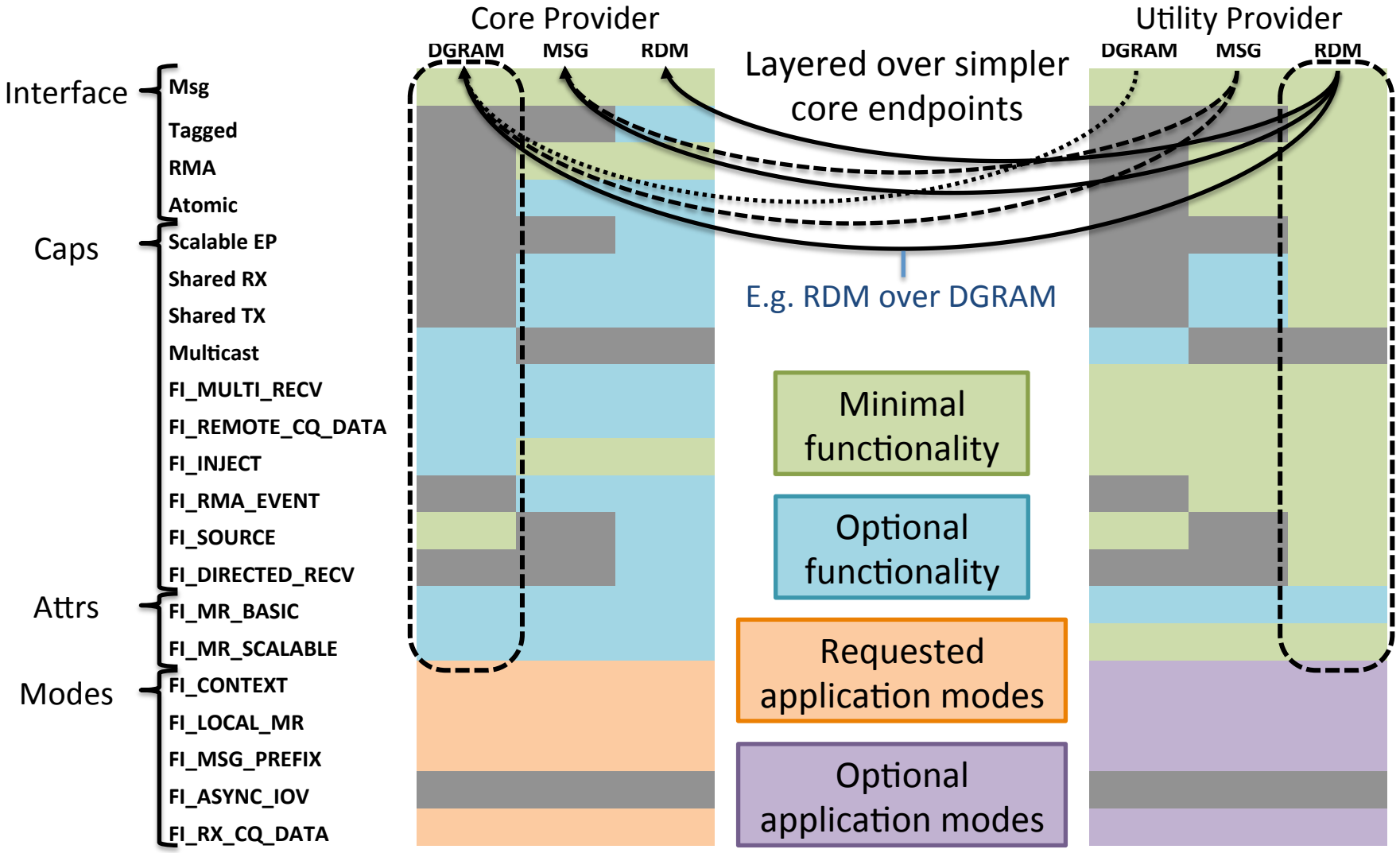
* Other names and brands may be claimed as the property of others

# ADDRESSING THE OFI-PROVIDER GAP



# UTILITY PROVIDER

Performance is a primary objective



# MOVING FORWARD

Analyze requests to expand OFI community

Sockets – TCP/UDP

NetworkDirect


FreeBSD

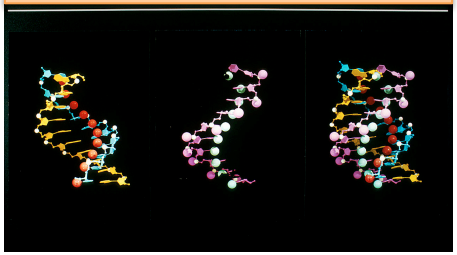
ORACLE

SOLARIS

Windows

Beyond Linux*

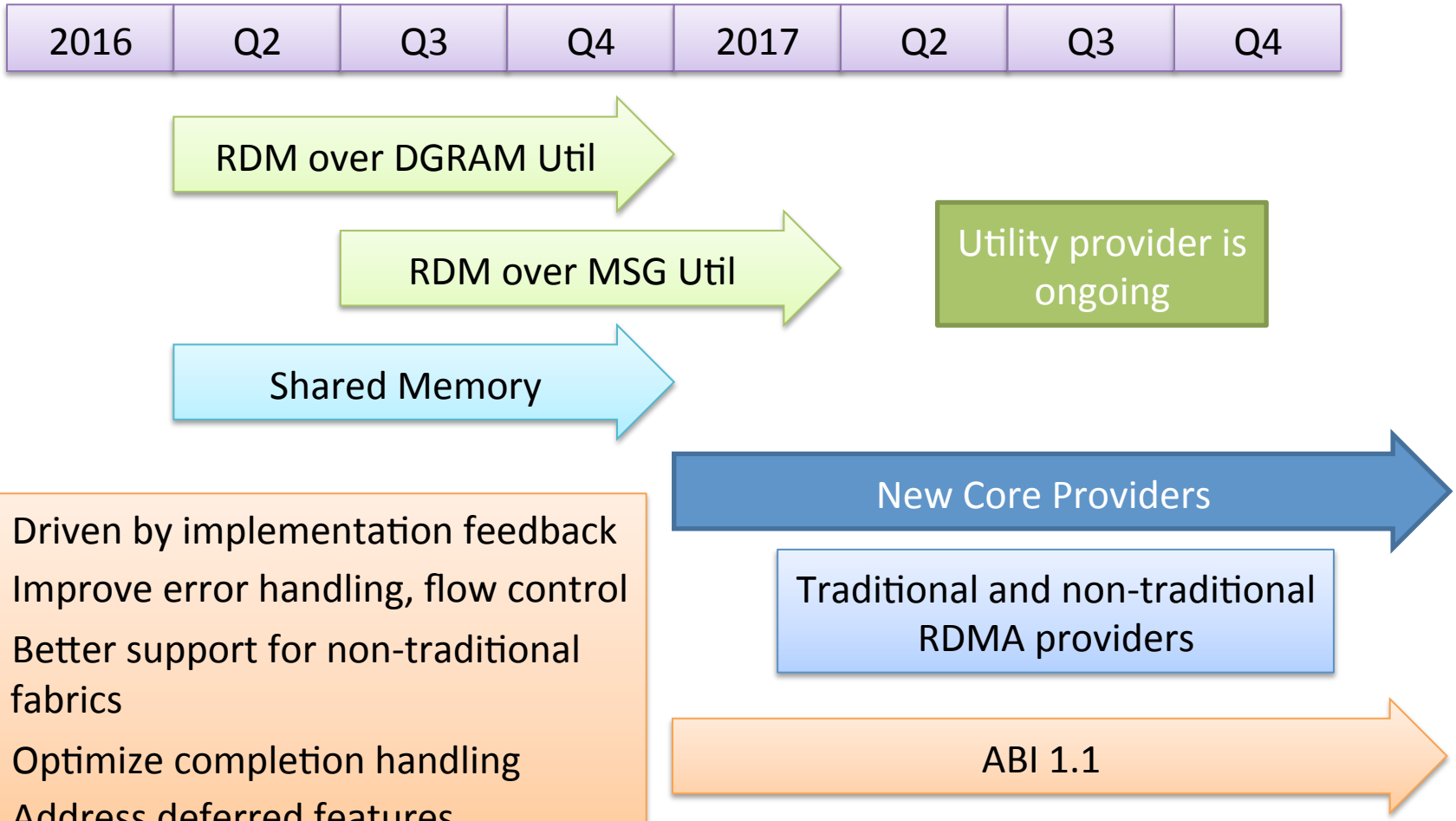
Beyond HPC



Enterprise, Cloud, Storage (NVM)


Stronger engagement with these communities

# TARGET SCHEDULE



# SUMMARY

- **OFIWG development model working well**
- **Interest in OFI and libfabric is high**
- **Growing community**
- **Significant effort being made to simplify the lives of developers**
  - Applications and providers

# LEGAL DISCLAIMER & OPTIMIZATION NOTICE

- **No license (express or implied, by estoppel or otherwise) to any intellectual property rights is granted by this document. Intel disclaims all express and implied warranties, including without limitation, the implied warranties of merchantability, fitness for a particular purpose, and non-infringement, as well as any warranty arising from course of performance, course of dealing, or usage in trade. This document contains information on products, services and/or processes in development. All information provided here is subject to change without notice. Contact your Intel representative to obtain the latest forecast, schedule, specifications and roadmaps. The products and services described may contain defects or errors known as errata which may cause deviations from published specifications. Current characterized errata are available on request.**
- **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.**
- 
- ***Other names and brands may be claimed as the property of others**

**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 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. 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

Sean Hefty

**OFIWG Co-Chair**