



Intel® Omni-Path Fabric Software Architecture Overview

Todd Rimmer, DCG Architecture

Intel Corporation

November, 2016

Legal Notices and Disclaimers

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.

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>.

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.

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.

No license (express or implied, by estoppel or otherwise) to any intellectual property rights is granted by this document.

Statements in this document that refer to Intel's plans and expectations for the quarter, the year, and the future, are forward-looking statements that involve a number of risks and uncertainties. A detailed discussion of the factors that could affect Intel's results and plans is included in Intel's SEC filings, including the annual report on Form 10-K.

All products, computer systems, dates and figures specified are preliminary based on current expectations, and are subject to change without notice. The products described 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 does not control or audit third-party benchmark data or the web sites referenced in this document. You should visit the referenced web site and confirm whether referenced data are accurate.

Intel, the Intel logo and others are trademarks of Intel Corporation in the U.S. and/or other countries.

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

© 2016 Intel Corporation.

Intel® Scalable System Framework

MODELING & SIMULATION



HPC DATA ANALYTICS



MACHINE LEARNING

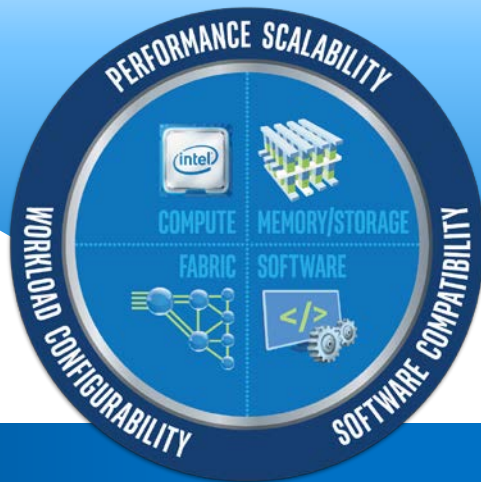


VISUALIZATION



MANY WORKLOADS – ONE FRAMEWORK

**A Flexible
Framework for
Today & Tomorrow**



**Enabling
Breakthrough
System Performance**

Intel® Omni-Path Architecture

Evolutionary Approach, Revolutionary Features, End-to-End Solution

Building on the industry's best technologies

- Highly leverage existing Aries and Intel® True Scale fabric
- Adds innovative new features and capabilities to improve performance, reliability, and QoS
- Re-use of existing OpenFabrics Alliance* software

Robust product offerings and ecosystem

- End-to-end Intel product line
- >100 OEM designs¹
- Strong ecosystem with 70+ Fabric Builders members

HFI Adapters

Single port
x8 and x16



x16
Adapter
(100 Gb/s)

x8 Adapter
(58 Gb/s)

Edge Switches

1U Form Factor
24 and 48 port



48-port
Edge Switch



24-port
Edge Switch

Director Switches

QSFP-based
192 and 768 port



768-port
Director Switch
(20U chassis)



192-port
Director Switch
(7U chassis)

Silicon

OEM custom designs
HFI and Switch ASICs



HFI silicon
Up to 2 ports
(50 GB/s
total b/w)



Switch silicon
up to 48 ports
(1200 GB/s
total b/w)

Software

Open Source
Host Software and
Fabric Manager



Cables

Third Party Vendors
Passive Copper
Active Optical



¹ Source: Intel internal information. Design win count based on OEM and HPC storage vendors who are planning to offer either Intel-branded or custom switch products, along with the total number of OEM platforms that are currently planned to support custom and/or standard Intel® OPA adapters. Design win count as of November 1, 2015 and subject to change without notice based on vendor product plans. *Other names and brands may be claimed as property of others.

Intel® Omni-Path Architecture

Disruptive innovations to knock down the “I/O Wall”


21%

HIGHER PERFORMANCE

Accelerates discovery and innovation

Up to **21% lower latency at scale**, up to **17% higher messaging rate**, and up to **9% higher application performance** than InfiniBand EDR¹


24%

BETTER ECONOMICS

Reduces size of fabric budgets.
Use savings to purchase more compute

up to **24% more compute nodes**
Better price-performance than InfiniBand* EDR reduces fabric spends for a given cluster size. Use savings to get more compute nodes with same total budget²


60%

MORE POWER EFFICIENT

more efficient switches and cards and a reduction in switch count and cables due to the 48-port chip architecture

Up to **60% lower power** than InfiniBand* EDR³



GREATER RESILIENCY

“no compromise” error detection and maintains link continuity with lane failures

No additional latency
penalty for error detection with Packet Integrity Protection⁴

¹ Intel® Xeon® Processor E5-2697A v4 dual-socket servers with 2133 MHz DDR4 memory. Intel® Turbo Boost Technology and Intel® Hyper Threading Technology enabled. BIOS: Early snoop disabled, Cluster on Die disabled, IOU non-posted prefetch disabled, Snoop hold-off timer=9. Red Hat Enterprise Linux Server release 7.2 (Maipo). Intel® OPA testing performed with Intel Corporation Device 24f0 – Series 100 HFI ASIC (B0 silicon). OPA Switch: Series 100 Edge Switch – 48 port (B0 silicon). Intel® OPA host software 10.1 or newer using Open MPI 1.10.x contained within host software package. EDR IB* testing performed with Mellanox EDR ConnectX-4 Single Port Rev 3 MCX455A HCA. Mellanox SB7700 - 36 Port EDR Infiniband switch. EDR tested with MLNX_OFED_Linux-3.2.x. OpenMPI 1.10.x contained within MLNX HPC-X. Message rate claim: Ohio State Micro Benchmarks v. 5.0. osu_mbw_mr, 8 B message (uni-directional), 32 MPI rank pairs. Maximum rank pair communication time used instead of average time, average timing introduced into Ohio State Micro Benchmarks as of v3.9 (2/28/13). Best of default, MXM_TLS=self,rc, and -mca pml yalla tunings. All measurements include one switch hop. Latency claim: HPCC 1.4.3 Random order ring latency using 16 nodes, 32 MPI ranks per node, 512 total MPI ranks. Application claim: GROMACS version 5.0.4 ion_channel benchmark. 16 nodes, 32 MPI ranks per node, 512 total MPI ranks. Intel® MPI Library 2017.0.064. Additional configuration details available upon request. ² Configuration assumes a 750-node cluster, and number of switch chips required is based on a full bisectonal bandwidth (FBB) Fat-Tree configuration. Intel® OPA uses one fully-populated 768-port director switch, and Mellanox EDR solution uses a combination of 648-port director switches and 36-port edge switches. Intel and Mellanox component pricing from www.kernelsoftware.com, with prices as of October 20, 2016. Assumes \$6,200 for a 2-socket Intel® Xeon® processor based compute node. ³ Configuration assumes a 750-node cluster, and number of switch chips required is based on a full bisectonal bandwidth (FBB) Fat-Tree configuration. Intel® OPA uses one fully-populated 768-port director switch, and Mellanox EDR solution uses a combination of director switches and edge switches. Mellanox power data based on Mellanox CS7500 Director Switch, Mellanox SB7700/SB7790 Edge switch, and Mellanox ConnectX-4 VPI adapter card installation documentation posted on www.mellanox.com as of November 1, 2015. Intel OPA power data based on product briefs posted on www.intel.com as of November 16, 2015. ⁴ A CRC check is performed on every Link Transfer Packet (LTP, 1056-bits) transmitted through a switch hop as defined by the Intel® OPA wire protocol, so stated switch latencies always include error detection by definition.

New Intel® OPA Fabric Features: Fine-grained Control Improves Resiliency and Optimizes Traffic Movement



Traffic Flow Optimization

- Optimizes Quality of Service (QoS) in mixed traffic environments, such as storage and MPI
- Transmission of lower-priority packets can be paused so higher priority packets can be transmitted

- Ensures high priority traffic is not delayed → Faster time to solution
- Deterministic latency → Lowers run-to-run timing inconsistencies



Packet Integrity Protection

- Allows for rapid and transparent recovery of transmission errors on an Intel® OPA link without additional latency
- Resends 1056-bit bundle w/errors only instead of entire packet (based on MTU size)

- Fixes happen at the link level rather than end-to-end level
- Much lower latency than Forward Error Correction (FEC) defined in the InfiniBand* specification¹



Dynamic Lane Scaling

- Maintain link continuity in the event of a failure of one of more physical lanes
- Operates with the remaining lanes until the failure can be corrected at a later time

- Enables a workload to continue to completion. **Note:** InfiniBand will shut down the entire link in the event of a physical lane failure

¹ Lower latency based on the use of InfiniBand with Forward Error Correction (FEC) Mode A or C in the public presentation titled "Option to Bypass Error Marking (supporting comment #205)," authored by Adeel Ran (Intel) and Oran Sela (Mellanox), January 2013. Mode A modeled to add as much as 140ns latency above baseline, and Mode C can add up to 90ns latency above baseline. Link: www.ieee802.org/3/bj/public/jan13/ran_3bj_01a_0113.pdf

Intel® Omni-Path Architecture Rapid Ramp



It's been quite a year since launch at Supercomputing 2015



Major wins across the globe:



50% MSS of 100Gb systems on the June '16 Top500 list¹

Expect dramatic increase for November '16 list next week

Gained significant momentum in the market²

¹ Source Top500.org; **Other names and brands may be claimed as the property of others

² As measured by 100Gbps High Perf. Fabrics revenue

Intel® Omni-Path Architecture: Software Components and Usage Model

Element Management Stack

- Runs on embedded Intel Atom processor included in managed switches
- “Traditional System Mgmt” – e.g. Signal integrity, Thermal monitoring, voltage monitoring, etc.

Host Software Stack

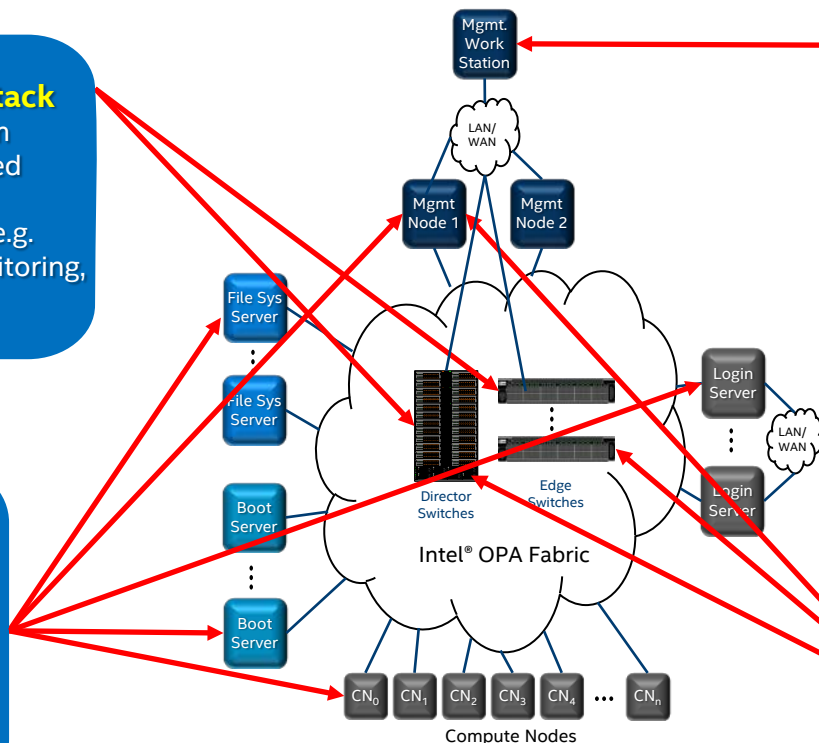
- Runs on all Intel® OPA-connected host nodes
- High performance, highly scalable MPI implementation via PSM and extensive set of upper layer protocols
- Boot over Fabric

Fabric Management GUI

- Runs on workstation with a local screen/keyboard
- Provides interactive GUI access to Fabric Management TCO features (configuration, monitoring, diagnostics, element management drill down)

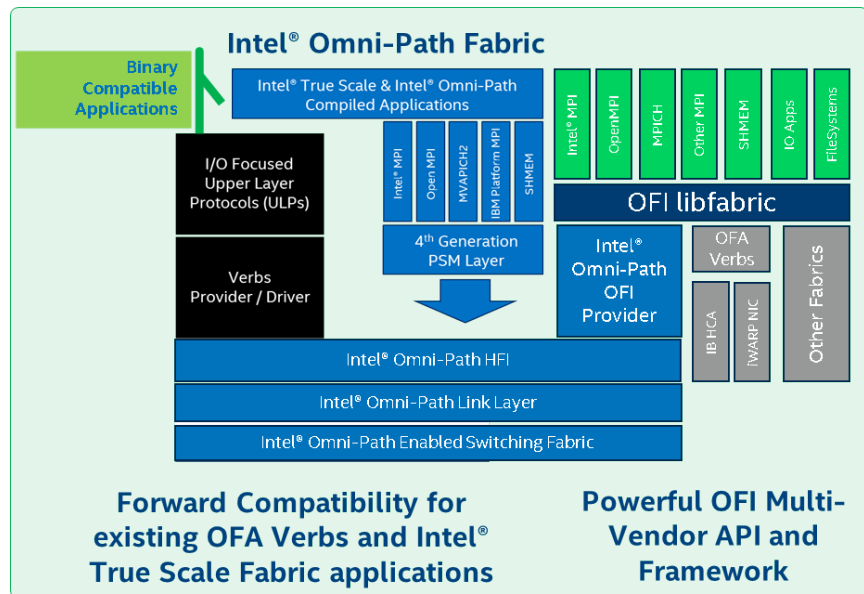
Fabric Management Stack

- Runs on OPA-connected management nodes or switch embedded Atom processor
- Initializes, configures and monitors the fabric routing, QoS, security, and performance
- Includes toolkit for TCO functions: configuration, monitoring, diags, and repair



Intel® Omni-Path Architecture

Optimized host implementation



Host Strategy: Leverage OpenFabrics Alliance* (OFA)

- OpenFabrics Alliance compliant: Off-the-shelf application compatibility
- Provides an extensive set of mature upper layer protocols
- Integrates 4th generation proven, scalable PSM capability for HPC
- OpenFabrics Interface (OFI) API aligned with application requirements

Access: Open Source Key Elements

- Host software stack via OFA
- Intel® Omni-Path FastFabric Tools, Fabric Manager, and GUI

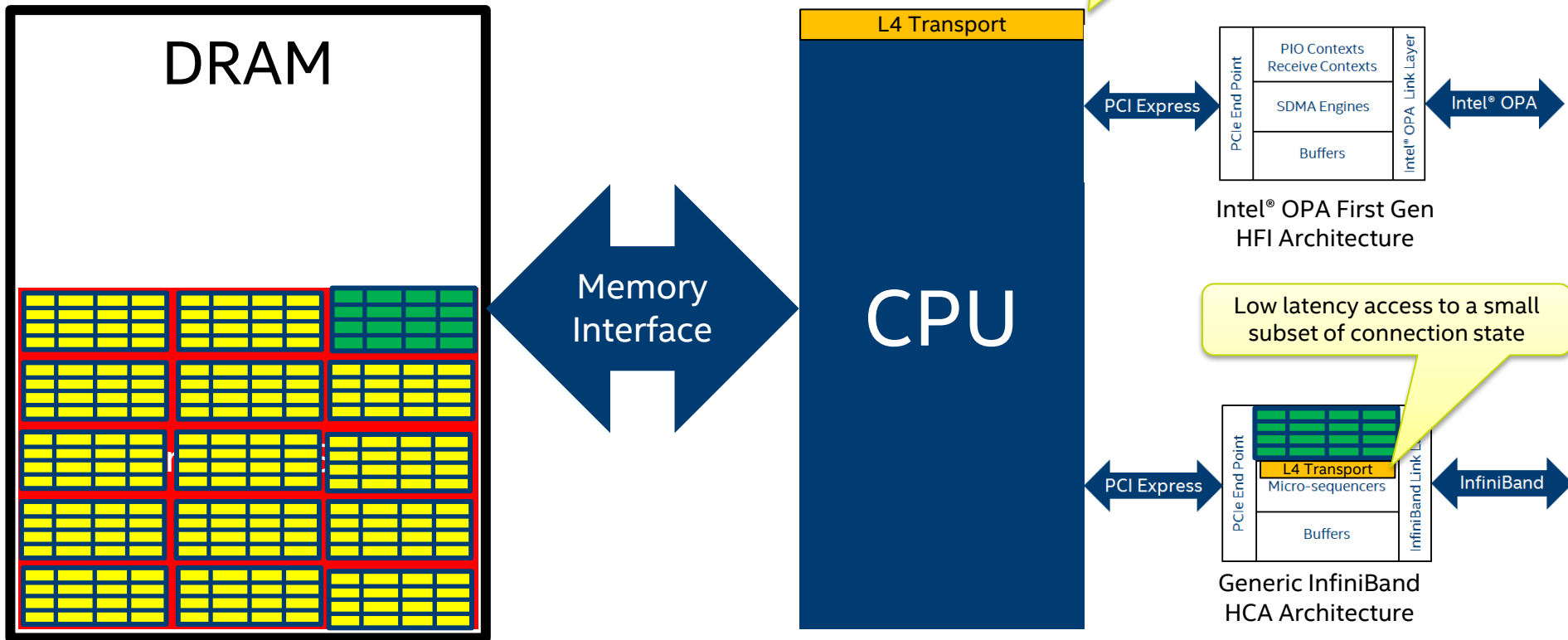
Channels: Integrate into Linux* Distributions

- Intel® Omni-Path Architecture support included in standard distributions
 - Starting with RHEL 7.3 and SLES 12sp2
- Delta distribution of OFA stack atop Linux distributions as needed

Maintains existing HPC fabric software approach

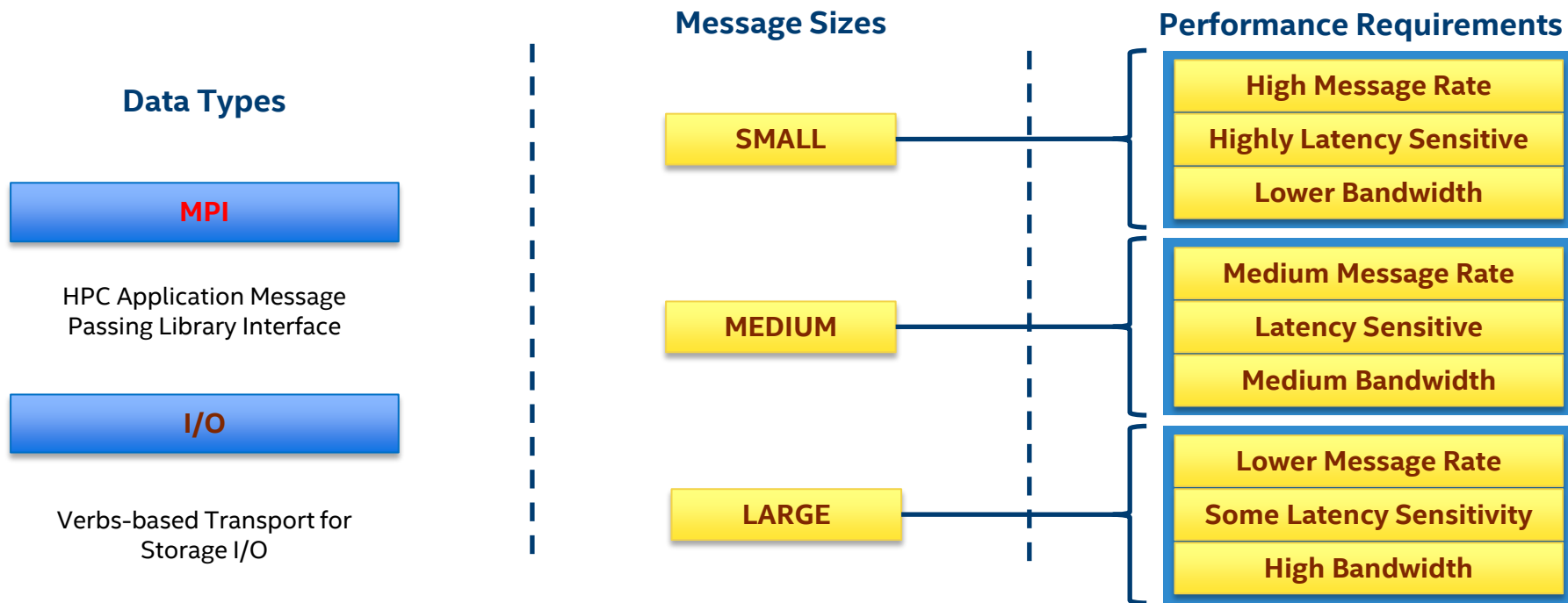
Onload vs. Offload

InfiniBand HCA vs. Intel® OPA HFI



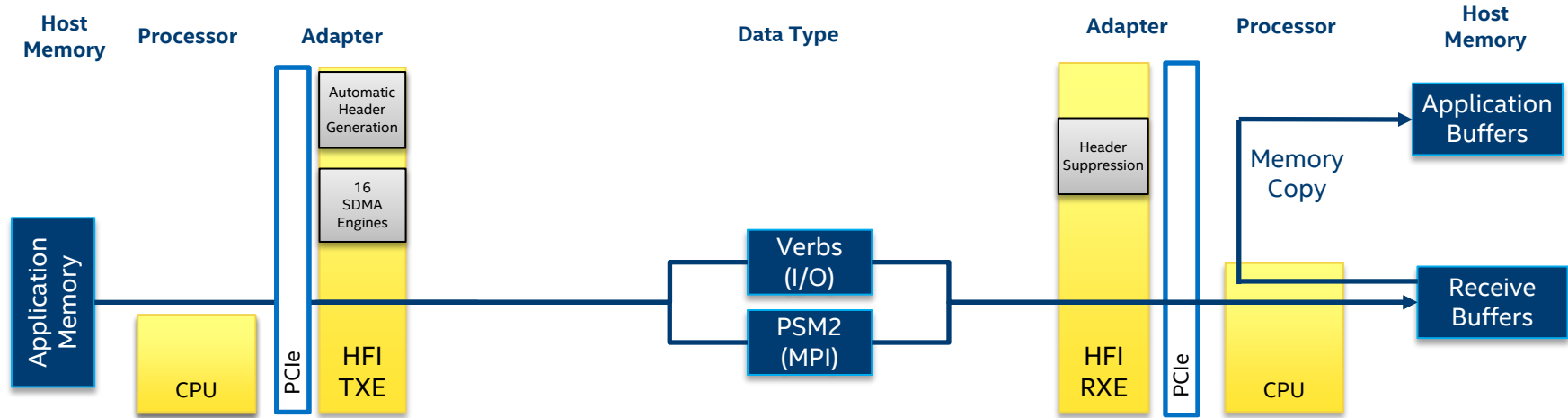
Choosing the Right Data Transfer Approach

MPI and Storage Traffic Performance Needs



Multimodal Data Acceleration

Highest performance small message transfer



Host Driven Send

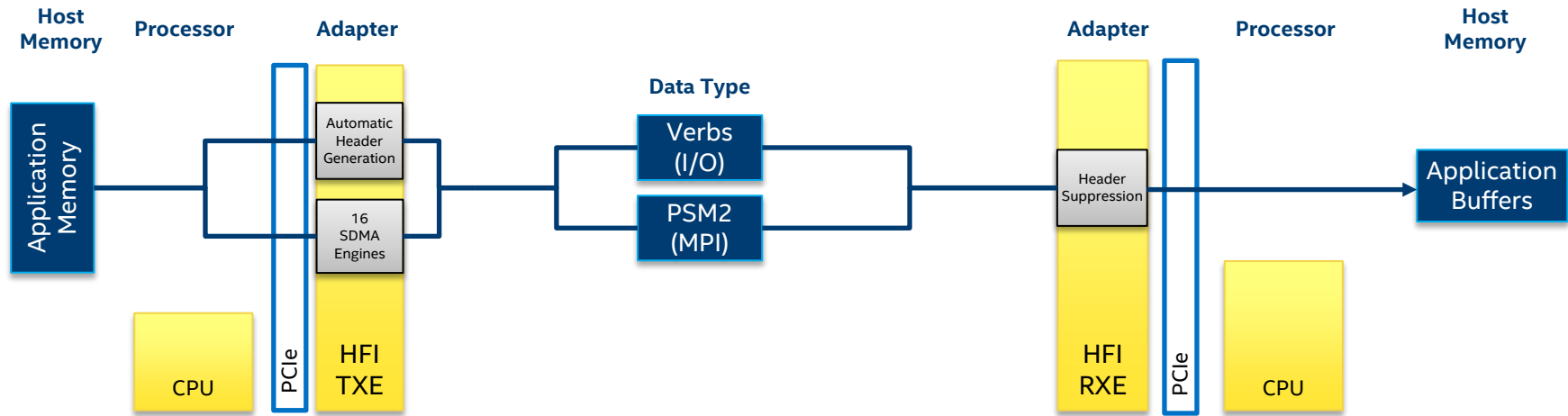
- ❖ Optimizes latency and message rate for high priority messages
- ❖ Transfer time lower than memory handle exchange, memory registration

Receive Buffer Placement

- ❖ Data placed in receive buffers
- ❖ Buffers copied to application buffer

Multimodal Data Acceleration

Lowest overhead RDMA-based large message transfer



Send DMA (SDMA) Engine

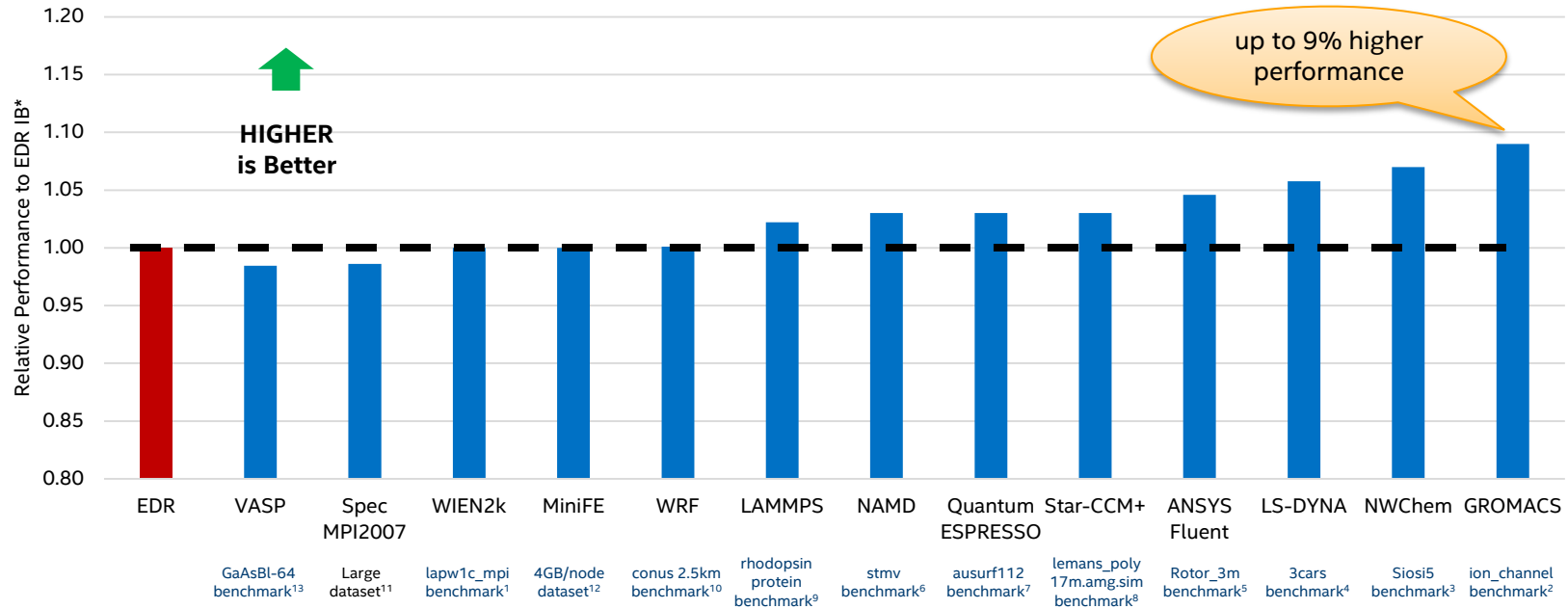
- ❖ Stateless offloads on send side
- ❖ DMA setup required

Direct Data Placement

- ❖ Direct data placement on receive side
- ❖ Eliminates memory copy

Intel® Omni-Path Architecture (Intel® OPA)

Application Performance - Intel® MPI - 16 Nodes



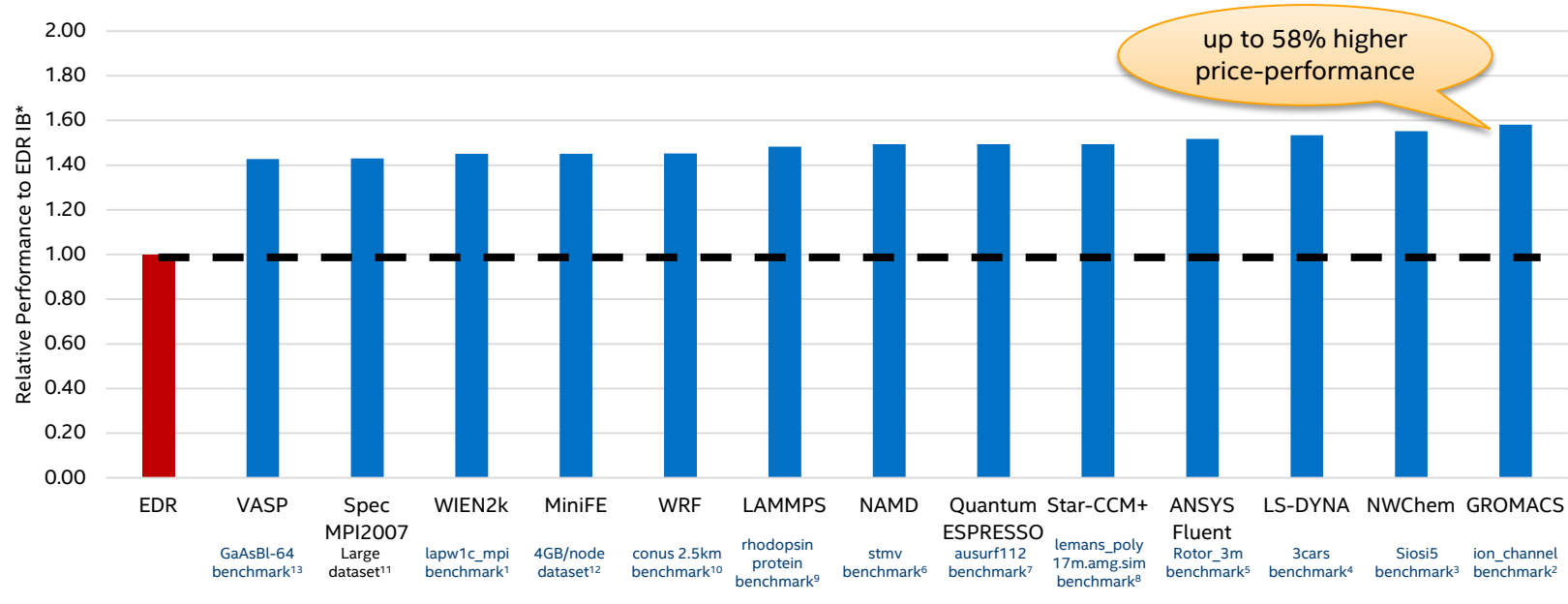
*Spec MPI2007 results estimates until published

**see following slide for system configurations

No Intel® OPA or EDR specific optimizations applied to any workloads except LS-DYNA and ANSYS Fluent: Intel® OPA HFI driver parameter: eager_buffer_size=8388608
WIEN2k comparison is for 8 nodes because EDR IB measurements did not scale above 8 nodes*

Intel® Omni-Path Architecture (Intel® OPA)

Application Performance Per Fabric Dollar* - Intel® MPI - 16 Nodes



*Spec MPI2007 results estimates until published

**see following slide for system configurations

*All pricing data obtained from www.kernelsoftware.com May 4, 2016. All cluster configurations estimated via internal Intel configuration tool. 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. **Fabric hardware assumes one edge switch, 16 network adapters and 16 cables.**

*No Intel® OPA or EDR specific optimizations applied to any workloads except LS-DYNA and ANSYS Fluent: Intel® OPA HFI driver parameter: eager_buffer_size=8388608
WIEN2k comparison is for 8 nodes because EDR IB* measurements did not scale above 8 nodes*

System & Software Configuration [include with previous slides]

Common configuration for bullets 1-11 unless otherwise specified: Intel® Xeon® Processor E5-2697A v4 dual socket servers. 64 GB DDR4 memory per node, 2133 MHz. RHEL 7.2. BIOS settings: Snoop hold-off timer = 9, Early snoop disabled, Cluster on die disabled, IOU Non-posted prefetch disabled. Intel® Omni-Path Architecture (Intel® OPA): Intel Fabric Suite 10.0.1.0.50. Intel Corporation Device 24f0 – Series 100 HFI ASIC (Production silicon). OPA Switch: Series 100 Edge Switch – 48 port (Production silicon). EDR Infiniband: MLNX_OFED_LINUX-3.2-2.0.0.0 (OFED-3.2-2.0.0). Mellanox EDR ConnectX-4 Single Port Rev 3 MCX455A HCA. Mellanox SB7700 – 36 Port EDR Infiniband switch.

1. WIEN2k version 14.2. <http://www.wien2k.at/>. http://www.wien2k.at/reg_user/benchmark/. Run command: "mpirun ... lapw1c_mpi lapw1.def". Intel Fortran Compiler 17.0.0 20160517. Compile flags: -FR -mp1 -w -prec_div -pc80 -pad -ip -DINTEL_VML -traceback -assume buffered_io -DFTW3 -l/opt/intel/compilers_and_libraries_2017.0.064/linux/mkl/include/fftw/ -DParallel.shm:tmi fabric used for Intel® OPA and shm:dapl fabric used for EDR IB®.
 2. GROMACS version 5.0.4. Intel Composer XE 2015.1.133. Intel MPI 5.1.3. FFTW-3.3.4. ~/bin/cmake .. -DGMX_BUILD_OWN_FFTW=OFF -DREGRESSIONTEST_DOWNLOAD=OFF -DCMAKE_C_COMPILER=icc -DCMAKE_CXX_COMPILER=icpc -DCMAKE_INSTALL_PREFIX=~/gromacs-5.0.4-installed. Intel® OPA MPI parameters: I_MPI_FABRICS=shm:tmi, EDR MPI parameters: I_MPI_FABRICS=shm:dapl
 3. NWChem release 6.6. Binary: nwchem_comex-mpi-pr_mkl with MPI-PR run over MPI-1. Workload: sios3 and sios5. Intel® MPI Library 2017.0.064. 2 ranks per node, 1 rank for computation and 1 rank for communication. shm:tmi fabric for Intel® OPA and shm:dapl fabric for EDR, all default settings. Intel Fabric Suite 10.2.0.0.153. http://www.nwchem-sw.org/index.php/Main_Page
 4. LS-DYNA MPP R8.1.0 dynamic link. Intel Fortran Compiler 13.1 AVX2. Intel® OPA - Intel MPI 2017 Library Beta Release Candidate 1. mpi.2017.0.0.BETA.U1.RC1.x86_64.wv20.20160512.143008. MPI parameters: I_MPI_FABRICS=shm:tmi. HFI driver parameter: eager_buffer_size=8388608. EDR MPI parameters: I_MPI_FABRICS=shm:ofa.
 5. ANSYS Fluent v17.0, Rotor_3m benchmark. Intel® MPI Library 5.0.3 as included with Fluent 17.0 distribution, and libpsm_infinipath.so.1 added to the Fluent syslib library path for PSM/PSM2 compatibility. Intel® OPA MPI parameters: -pib.infinipath, EDR MPI parameters: -pib.dapl
 6. NAMD: Intel Composer XE 2015.1.133. NAMD V2.11, Charm 6.7.0, FFTW 3.3.4. Intel MPI 5.1.3. Intel® OPA MPI parameters: I_MPI_FABRICS=shm:tmi, EDR MPI parameters: I_MPI_FABRICS=shm:dapl
 7. Quantum Espresso version 5.3.0. Intel Compiler 2016 Update 2. ELPA 2015.11.001 (<http://elpa.mpcdf.mpg.de/elpa-tar-archive>). Minor patch set for QE to accommodate latest ELPA. Most optimal NPOOL, NDIAG, and NTG settings reported for both OPA and EDR. Intel® OPA MPI parameters: I_MPI_FABRICS=shm:tmi, EDR MPI parameters: I_MPI_FABRICS=shm:dapl
 8. CD-adapco STAR-CCM+® version 11.04.010. Workload: lemans_poly_17m.amg.sim benchmark. Intel® MPI version 5.0.3.048. 32 ranks per node. OPA command: \$ /starccm+ -ldlibpath /STAR-CCM+11.04.010/mpi/intel/5.0.3.048/linux-x86_64/lib64 -ldpreload /usr/lib64/psm2-compat/libpsm_infinipath.so.1 -mpi intel -mppflags "-env I_MPI_DEBUG 5 -env I_MPI_FABRICS shm:tmi -env I_MPI_TMI_PROVIDER psm" -power -rsh ssh -np 512 -machinefile hosts -benchmark:"-nps 512,256,128,64,32 -nits 20 -preits 40 -tag lemans_opa_n16" lemans_poly_17m.amg.sim. EDR command: \$ /starccm+ -mpi intel -mppflags "-env I_MPI_DEBUG 5" -power -rsh ssh -np 512 -machinefile hosts -benchmark:"-nps 512,256,128,64,32 -nits 20 -preits 40 -tag lemans_edr_n16" lemans_poly_17m.amg.sim
 9. LAMMPS (Large-scale Atomic/Molecular Massively Parallel Simulator) Feb 16, 2016 stable version release. MPI: Intel® MPI Library 5.1 Update 3 for Linux. Workload: Rhodopsin protein benchmark. Number of time steps=100, warm up time steps=10 (not timed) Number of copies of the simulation box in each dimension: 8x8x4 and problem size: 8x8x4x32k = 8,192k atoms Intel® OPA: MPI parameters: I_MPI_FABRICS=shm:tmi, I_MPI_PIN_DOMAIN=core EDR: MPI parameters: I_MPI_FABRICS=shm:dapl, I_MPI_PIN_DOMAIN=core
 10. WRF version 3.5.1, Intel Composer XE 2015.1.133. Intel MPI 5.1.3. NetCDF version 4.4.2. FCBASEOPTS=-w -ftz -align all -fno-alias -fp-model precise. CFLAGS_LOCAL = -w -O3 -ip. Intel® OPA MPI parameters: I_MPI_FABRICS=shm:tmi, EDR MPI parameters: I_MPI_FABRICS=shm:dapl
 11. Spec MPI 2007: 16 nodes, 32 MPI ranks/node. SPEC MPI2007, Large suite, <https://www.spec.org/mpi/>. *Intel Internal measurements marked estimates until published. Intel MPI 5.1.3. Intel® OPA MPI parameters: I_MPI_FABRICS=shm:tmi, EDR MPI parameters: I_MPI_FABRICS=shm:dapl
- **Common configuration for bullets 12-13:** Intel® Xeon® Processor E5-2697 v4 dual socket servers. 128 GB DDR4 memory per node, 2400 MHz. RHEL 6.5. BIOS settings: Snoop hold-off timer = 9. Intel® OPA: Intel Fabric Suite 10.0.1.0.50. Intel Corporation Device 24f0 – Series 100 HFI ASIC (Production silicon). OPA Switch: Series 100 Edge Switch – 48 port (Production silicon). IOU Non-posted prefetch disabled. 2). Mellanox EDR based on internal measurements: Mellanox EDR ConnectX-4 Single Port Rev 3 MCX455A HCA. Mellanox SB7700 – 36 Port EDR Infiniband switch. IOU Non-posted prefetch enabled.
12. MiniFE 2.0, Intel compiler 16.0.2. Intel® MPI Library version 5.1.3. Build settings: -O3 -xCORE-AVX2 -DMINIFE_CSR_MATRIX -DMINIFE_GLOBAL_ORDINAL="long long int", mpirun -bootstrap ssh -env OMP_NUM_THREADS 1 - perhost 36 miniFE.nx=200 ny=200 nz=200, 200x200x200 grid using 36 MPI ranks pinned to 36 cores per node. Intel® OPA MPI parameters: I_MPI_FABRICS=shm:tmi, EDR MPI parameters: I_MPI_FABRICS=shm:dapl. Intel® Turbo Mode technology and Intel® Hyper threading technology disabled.
 13. VASP (developer branch). MKL: 11.3 Update 3 Product build 20160413. Compiler: 2016u3. Intel MPI-2017 Build 20160718. elpa-2016.05.002. Intel® OPA MPI parameters: I_MPI_FABRICS=shm:tmi, EDR MPI parameters: I_MPI_FABRICS=shm:dapl, I_MPI_PLATFORM=BDW, I_MPI_DAPL_PROVIDER=ofa-v2-mlx5_0-1u, I_MPI_DAPL_DIRECT_COPY_THRESHOLD=331072. Intel® Turbo Mode technology disabled. Intel Hyper Threading technology enabled.

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.



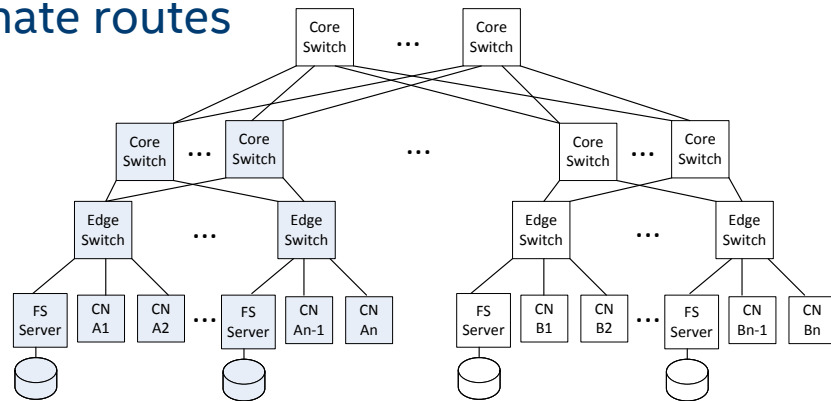
Intel® OPA Management Software

	LEGEND					
		Intel Updates	Intel Provides (new)	3rd Party	Existing	
Management Interfaces	Unified Management (console / tools)					
Managers	Intel® OPA Fabric Manager	Cluster Manager & Provisioning	System Manager	Device Manager	Job Manager	
Transport Layer	Intel® OPA	TCP/UDP/IP				
Network	Intel® OPA	IPoFabric	Ethernet	Ethernet	IPoFabric	Ethernet
Management Agents	FM Agents	PXE Boot	BMC ME	Device Agents	Job Manager Daemon	
Managed Items	Interconnect	OS Images	Motherboards	IO Devices	Job Processes	

- Intel® OPA leverages existing stacks for each type of management
- Assorted 3rd party unified management consoles
- Intel® OPA provides a scalable centralized fabric management stack

Advanced Switching Features

- Route Balancing per Device Group
 - Scalable unit and cross sub-cluster storage vs compute optimization
- Self discovering FatTree
 - Flexibility in cable placement, option to handle HFIs at core
- Medium Grained Adaptive Routing
 - Can specify any and all potential alternate routes
- Dispersive Routing
 - Maximized dispersion and resiliency
- Congestion Control




Fabric Diagnostic and Debug Features

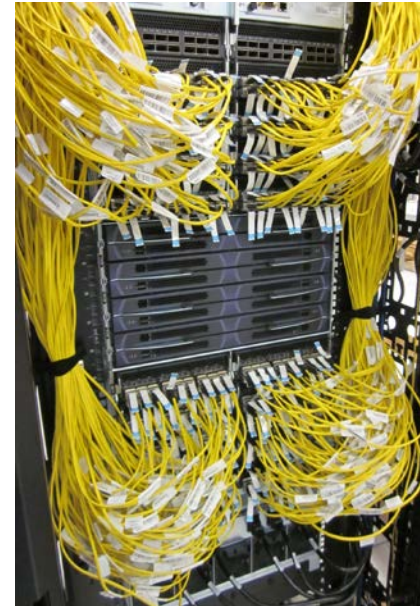
In a typical cluster the majority of fabric FRUs are cables

- Managing cable FRUs is one of the biggest sysadmin challenges



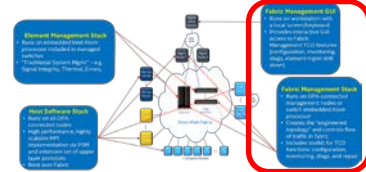
Cable FRU management Intel® OPA Fabric innovations:

- Link Quality Indicator – “5 bars” instantaneous view of link quality
 - In every HW port, monitored by FM, FastFabric Tools, FM GUI 
- Topology verification – Are cables in correct places?
 - FM can warn or quarantine incorrect links
 - FastFabric online and offline topology analysis
- Port type information
 - QSFP/Standard, Fixed/Backplane, Variable, Disconnected, ...
- QSFP CableInfo – shows all key cable /transceiver info
 - Vendor, model, length, technology, date, etc.
 - Fully integrated into FM, FastFabric tools, FM GUI
- Link Down Reason
 - LinkDownReason and NeighborLinkDownReason – most recent reason link went down



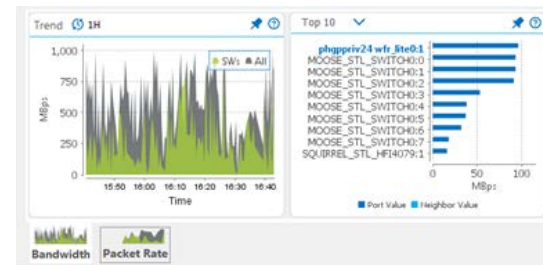
Fabric Diagnostic and Debug Features

Fabric utilization and performance monitoring is critical to fabric operations



Intel® OPA Fabric Innovations

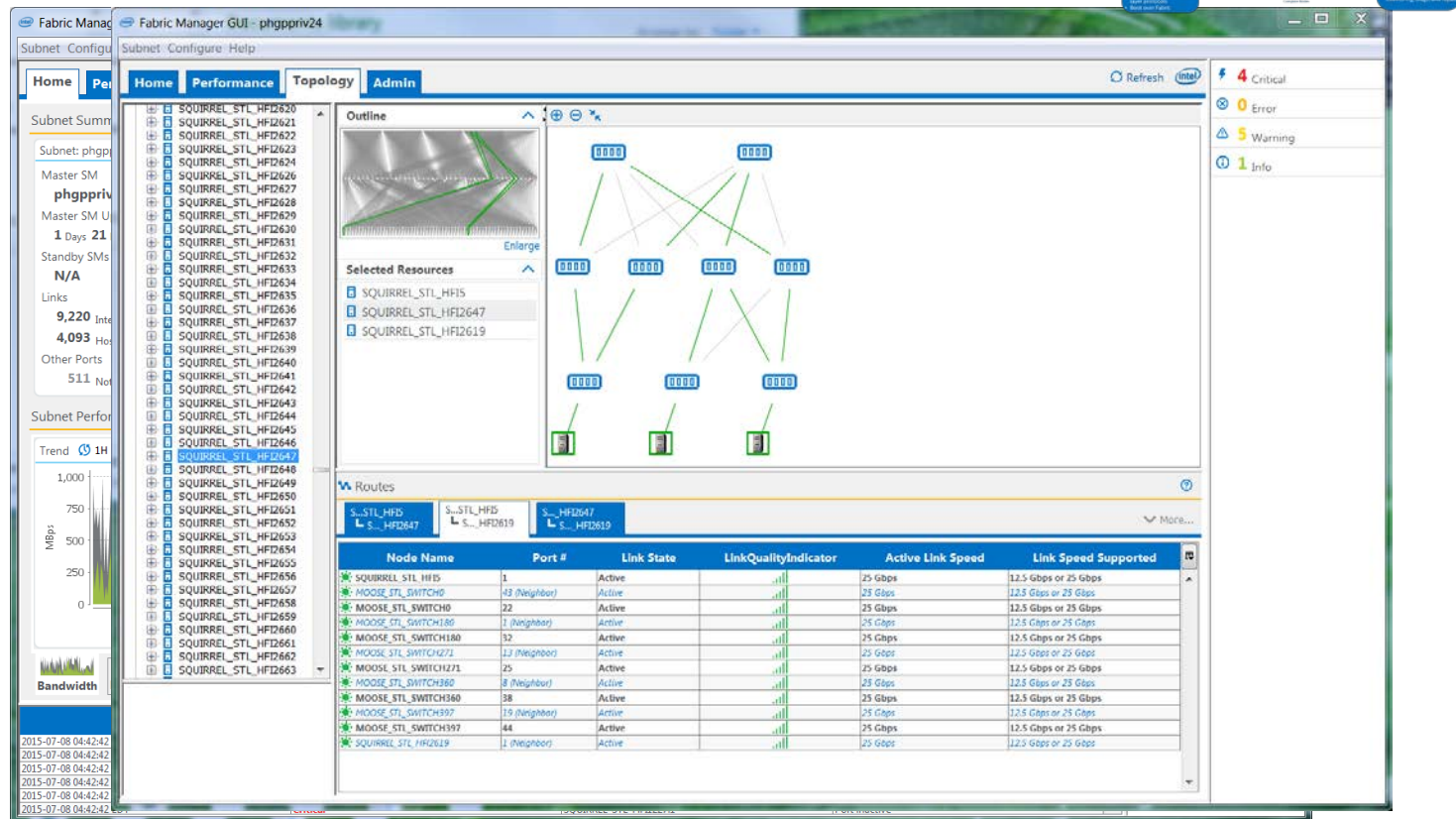
- FM's monitors fabric and maintains history of fabric performance and errors
 - Over 130 performance counters per port
 - Including utilization, packet rate and congestion per VL
 - 64-bit counters (many decades to rollover)
- PM Device Groups allow performance analysis for a specific set of devices
 - Storage vs compute, etc.
- PM/PA database sync – PM data retained during FM failover
- Statistics available via FastFabric CLI, FM GUI, PA APIs



Omni-Path scalable FM GUI

Advanced
User
Interface

Exposes New
OPA Features



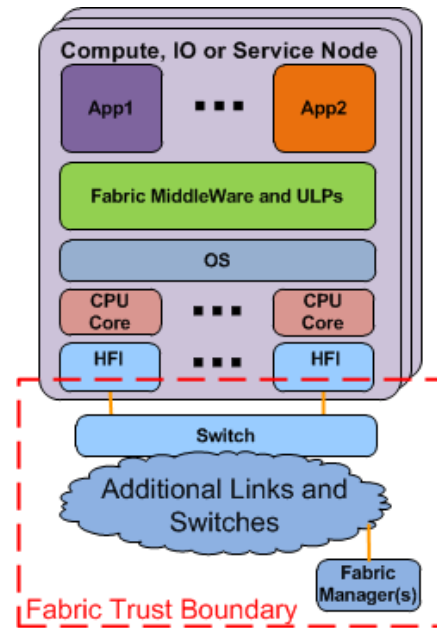
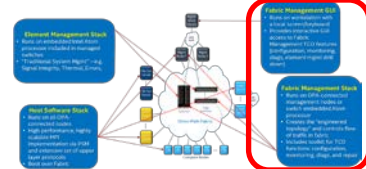
Fabric Security



Even friendly users/apps make mistakes
Need fabric management access secured

Intel® OPA Fabric Innovations

- SMA/PMA protection
 - Specific enablement of Mgmt. Nodes by switches
 - SMA and PMA protocols strictly protected
 - SMA pacing at non-mgmt. hosts limits SMA denial of service attacks
- Host spoofing prevention
 - SLID verification at neighbor switch
 - NodeType, NodeGUID and PortGUID verification (hardware assisted)
 - Topology verification by SM
 - Catches unexpected devices or attempts to spoof other devices
- Cluster information restricted
 - SA limits data available to non-mgmt. nodes
 - SSL security for FM GUI connection to FM



Virtual Fabrics Overview

Unifying Concept for Security and QoS

- Allow multiple applications on cluster at once
- Allow sysadmin to control degree of isolation
- Composed of application, devicegroup and policy

Applications

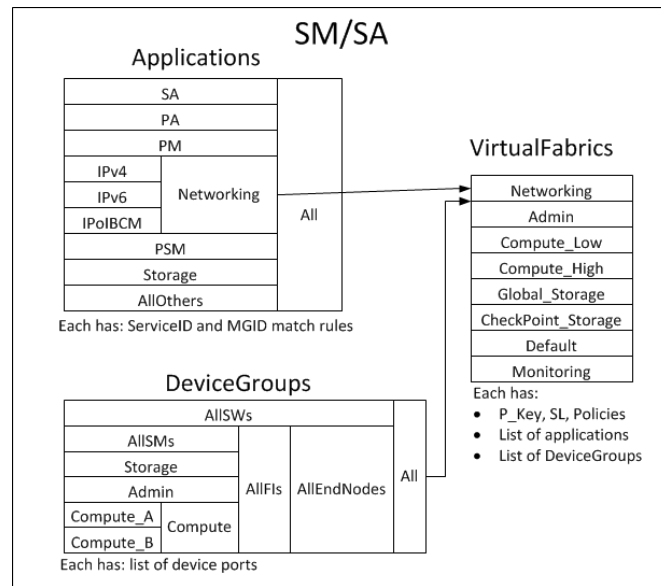
- Identified by PathRecord ServiceID or Multicast MGID

Device Group

- Identified by Node Names or other mechanisms

Policies

- QoS settings (SL, BW, TFO, etc), security settings (Pkey)



Virtual Fabrics Address Resolution

Transparently Integrated into OFA mechanisms

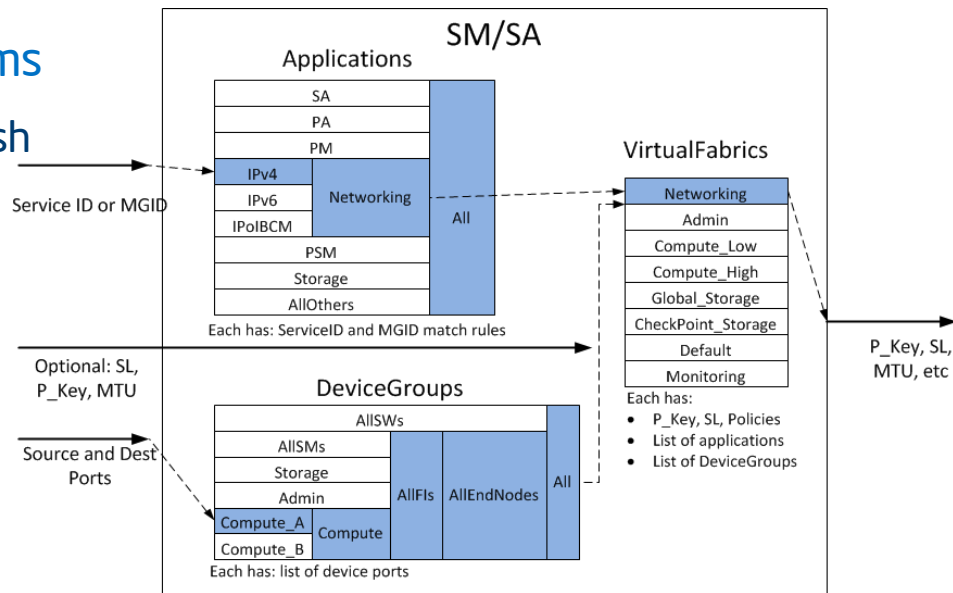
- PathRecord query, RDMA CM connection establish
- Multicast Join, IPoIB

Implemented in FM's SA

- Resolves request, finds matching application, device group, and associated vFabric
- Returns proper SL, PKey, etc

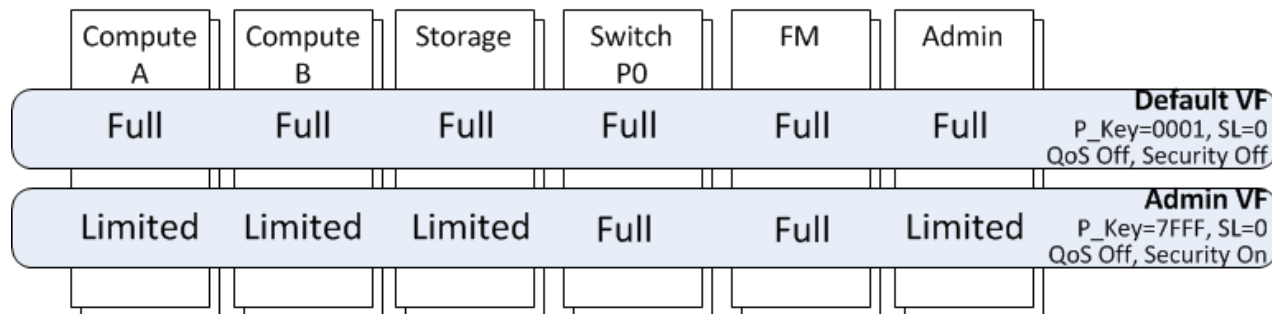
Supports Other Mechanisms

- VF Info queries to get SL, PKey, etc for a given vFabric
- Useful for scalable MPI job launch



Virtual Fabrics Examples

Default Virtual Fabrics Configuration

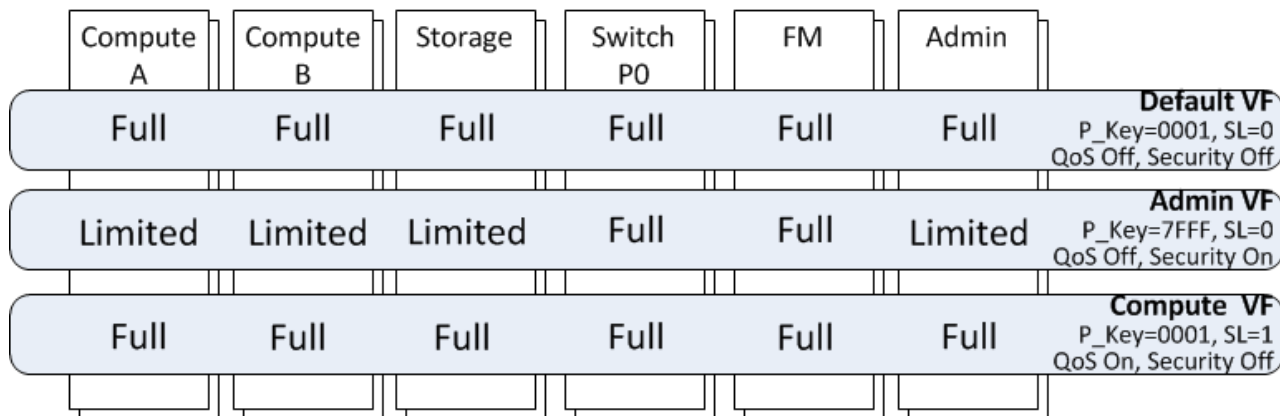


Applications

AllOthers

SA, PA, PM
(SM implicit)

Simple QoS Virtual Fabrics Configuration



Applications

AllOthers

SA, PA, PM
(SM implicit)

Compute

Virtual Fabrics Examples

Simple Security Virtual Fabrics Configuration

Compute A	Compute B	Storage	Switch P0	FM	Admin	
Full	N/A	N/A	N/A	N/A	N/A	Compute_A VF P_Key=0001, SL=0 QoS Off, Security On
Limited	Limited	Limited	Full	Full	Limited	Admin VF P_Key=7FFF, SL=0 QoS Off, Security On
N/A	Full	N/A	N/A	N/A	N/A	Compute_B VF P_Key=0002, SL=0 QoS Off, Security On
Limited	Limited	Full	N/A	Limited	Full	Services VF P_Key=0003, SL=0 QoS Off, Security On

Applications

Compute,
Networking

SA, PA, PM
(SM implicit)

Compute,
Networking

Storage,
AllOthers

Virtual Fabrics Ground Rules

Always get the SL and PKey from the FM

- PathRecord query, Multicast Join, RDMA CM
- VF Info query and related scripts (opagetvf)
- A priori discussion with sysadmin and direct application parameters for SL and PKey

Use scalable PathRecord query mechanisms

- RDMA CM, ibacm, kernel SA query
- Do not hand build your own SA MADs via umad
 - Often secured to root access and will not scale on large fabrics

Be aware 0x7fff/0xffff is admin pkey

- Secured by default, not for use by application traffic
- Only includes SMA, PMA, SA, PA “applications”

INTEL® FABRIC BUILDERS

An ecosystem working together to enable world class solutions based on Intel® Omni-Path Fabric



<https://fabricbuilders.intel.com/>

Last update November 3, 2016

Intel® Omni-Path Architecture Summary

Next generation HPC fabric that builds on Intel® True Scale Fabric

- Full end-to-end solution: Switches, adapters, host software, management software, cabling, silicon
- Optimized CPU, host and fabric architecture that cost effectively scales from entry to extreme deployments
- Comprehensive & mature host software that is compatible with existing Intel True Scale Fabric and Open Fabric Alliance (OFA) APIs
- Many advanced fabric administration features
- An established ecosystem and rapidly growing customer base

