To Exascale And Beyond:

Intel's Scalable System Framework and OpenSHMEM

James Dinan
Extreme Scale Software Pathfinding Team

OpenSHMEM Workshop
August 2016

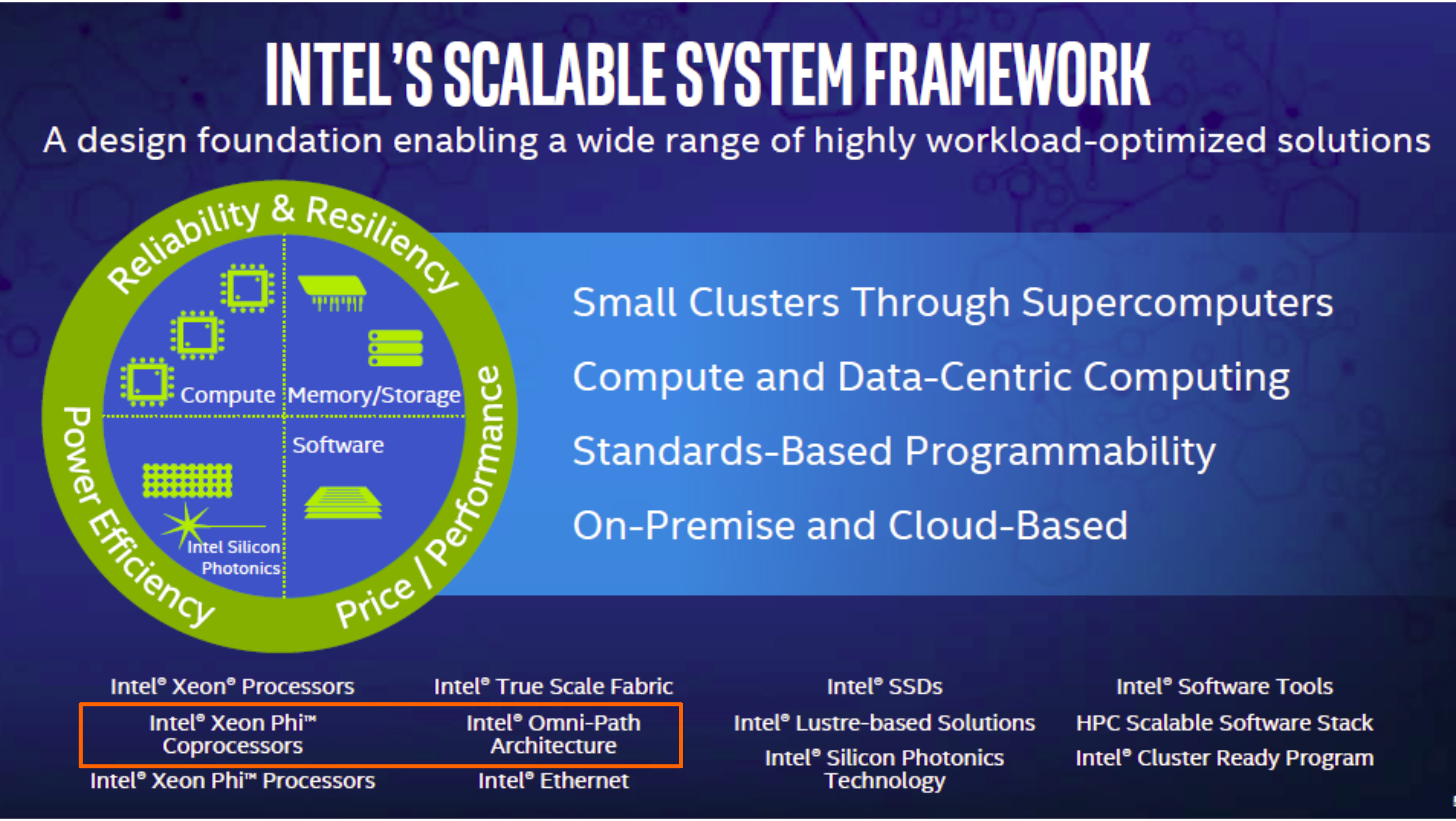
Legal Notices and Disclaimers

Intel technologies' features and benefits depend on system configuration and may require enabled hardware, software or service activation. Learn more at intel.com, or from the OEM or retailer.

No computer system can be absolutely secure.

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.

Intel, the Intel logo, Xeon and Xeon Phi 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.



INTEL'S SCALABLE SYSTEM FRAMEWORK

A design foundation enabling a wide range of highly workload-optimized solutions

Small Clusters Through Supercomputers

Compute and Data-Centric Computing

Standards-Based Programmability

On-Premise and Cloud-Based

Intel® Xeon® Processors
Intel® Xeon Phi™ Coprocessors
Intel® Xeon Phi™ Processors

Intel® True Scale Fabric
Intel® Omni-Path Architecture
Intel® Ethernet

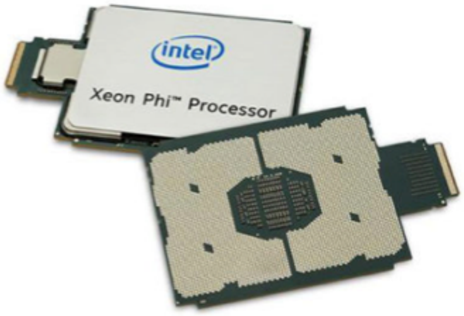
Intel® SSDs
Intel® Lustre-based Solutions
Intel® Silicon Photonics Technology

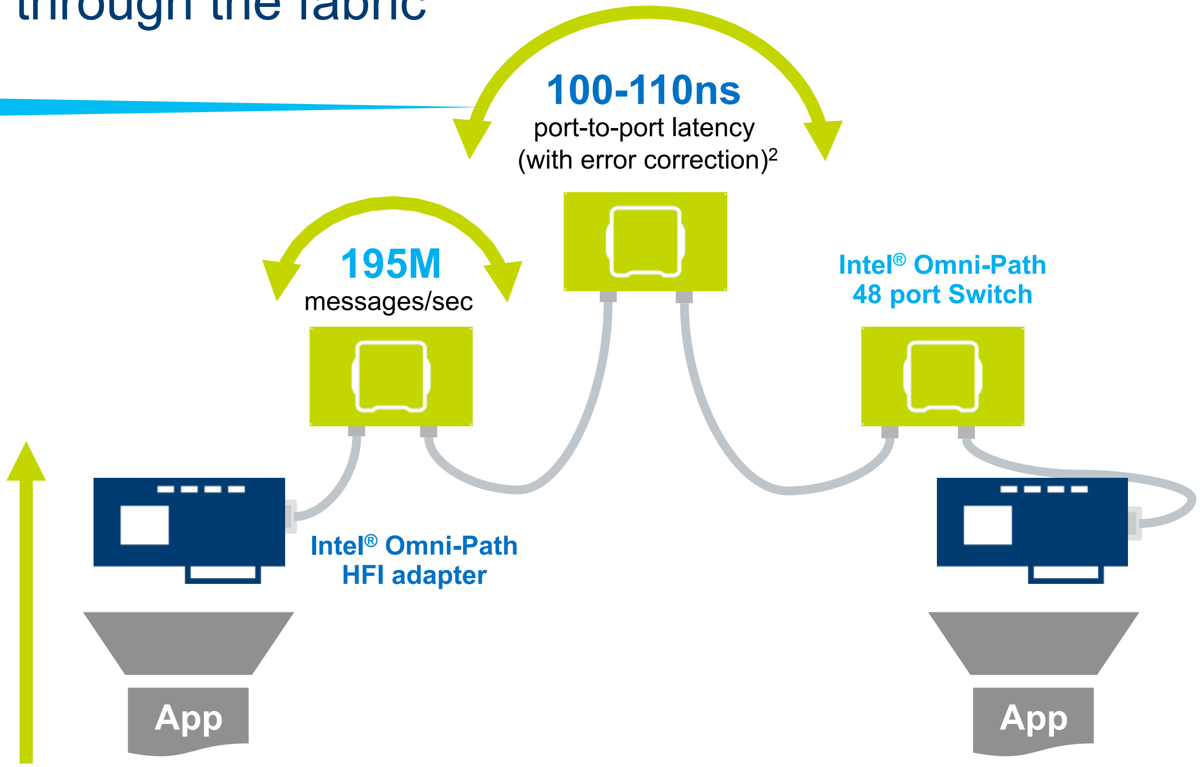
Intel® Software Tools
HPC Scalable Software Stack
Intel® Cluster Ready Program

Intel® Omni-Path Architecture

Accelerating data movement through the fabric

Port latency
(includes
error detection!)





160 M
messages/sec
(MPI message rate uni-directional)

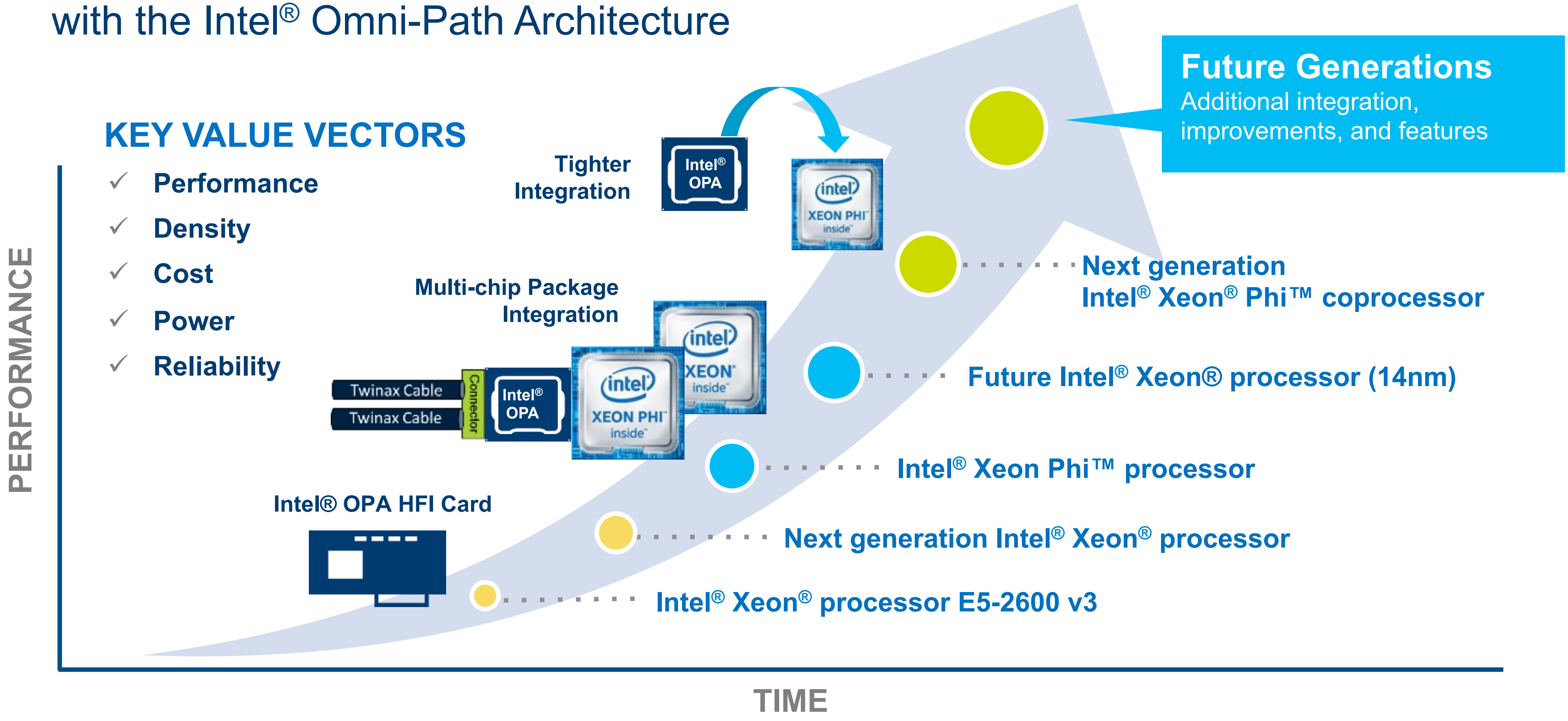
Software and workloads used in performance tests may have been optimized for performance only on Intel microprocessors. Performance tests, such as SYSmark and MobileMark, are measured using specific computer systems, components, software, operations and functions. Any change to any of those factors may 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.

CPU-Fabric Integration

with the Intel® Omni-Path Architecture

KEY VALUE VECTORS

- ✓ Performance
- ✓ Density
- ✓ Cost
- ✓ Power
- ✓ Reliability



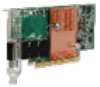
PERFORMANCE
TIME
Intel® OPA HFI Card
Twinax Cable
Twinax Cable
Connector
Intel® OPA
Multi-chip Package Integration
XEON PHI inside
XEON inside
Tighter Integration
Intel® OPA
XEON PHI inside
Intel® Xeon® processor E5-2600 v3
Next generation Intel® Xeon® processor
Intel® Xeon Phi™ processor
Future Intel® Xeon® processor (14nm)
Next generation Intel® Xeon® Phi™ coprocessor
Future Generations
Additional integration, improvements, and features

Intel® Omni-Path Architecture

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

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

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 designs[1]
- Strong ecosystem with 70+ Fabric Builders members

[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 Network Layers

Layer 1 – Physical Layer

- Leverages existing Ethernet and InfiniBand* PHY standards

Layer 1.5 – Link Transfer Protocol

- Provides reliable delivery of Layer 2 packets, flow control, and link control across a single link

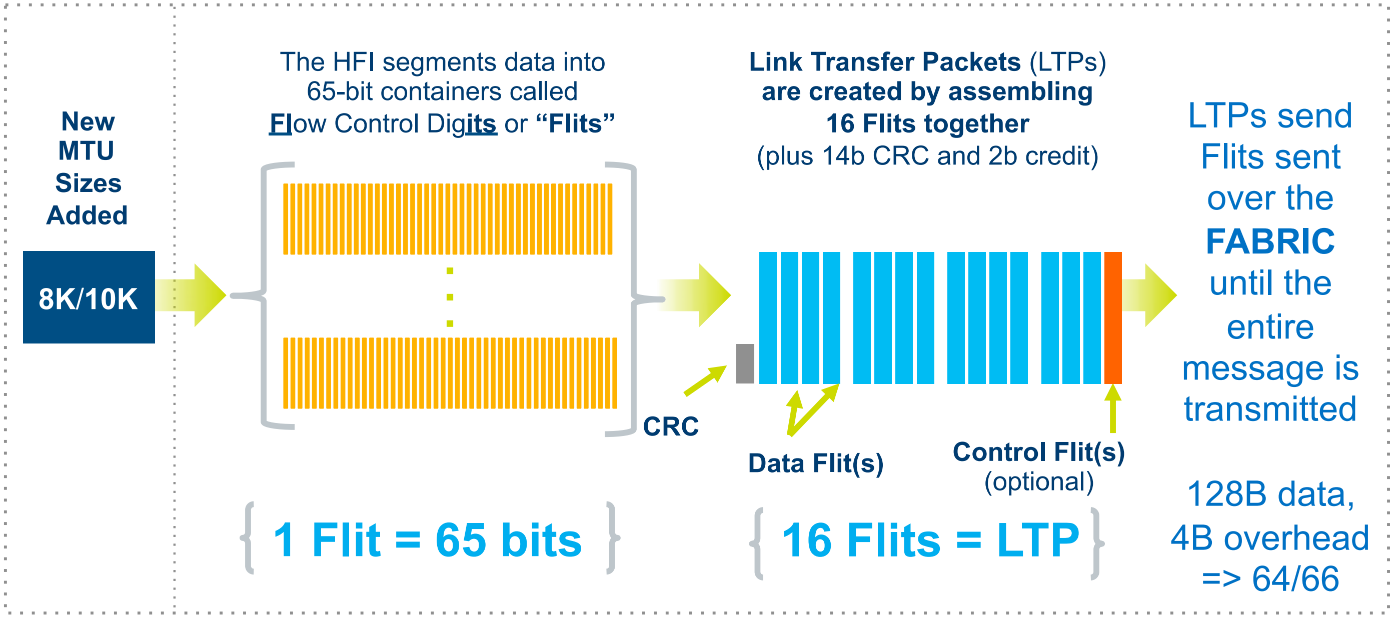
Layer 2 – Data Link Layer

- Provides fabric addressing, switching, resource allocation, and partitioning support

Layers 4-7 – Transport to Application Layers

- Provide interfaces between software libraries and HFIs

Intel® Omni-Path Architecture Link Transfer Layer



New MTU Sizes Added

8K/10K

The HFI segments data into 65-bit containers called Flow Control Digits or **"Flits"**

Link Transfer Packets (LTPs) **are created by assembling 16 Flits together** (plus 14b CRC and 2b credit)

LTPs send Flits sent over the **FABRIC** until the entire message is transmitted

128B data, 4B overhead => 64/66

CRC: Cyclic Redundancy Check

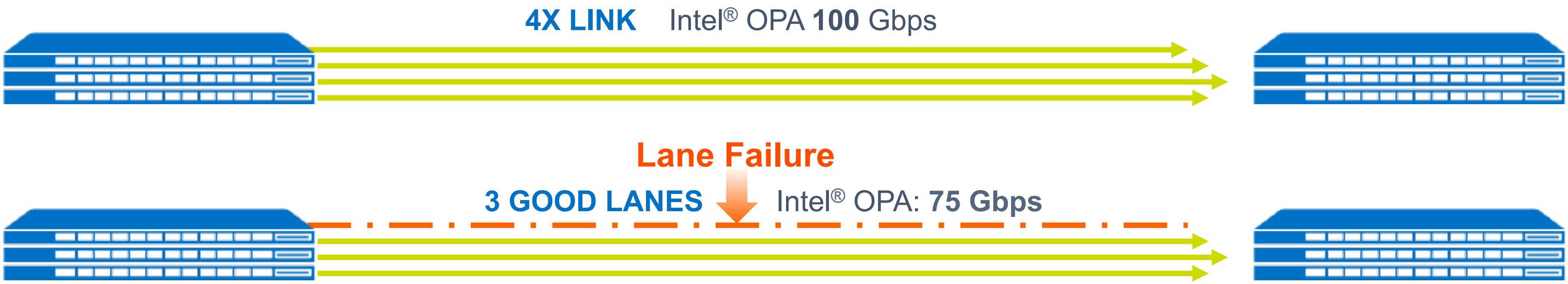
PiP (Packet Integrity Protection) – Link error detection/correction in units of LTPs

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

	Description	Benefits
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* specification[1]
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 Adee 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 Fabric Link Level Innovation:
Dynamic Lane Scaling (DLS) **Traffic Protection**



User Setting (per Fabric):

- Set maximum degrade option allowable
  - 4x – Any lane failure would cause link reset or take down
  - 3x – Still operates at degraded bandwidth (75 Gbps)
  - 2x – Still operates at degraded bandwidth (50 Gbps)
  - 1x – Still operates at degraded bandwidth (25 Gbps)

Link Recovery:

- PIP is used to recover link without reset – An Intel® OPA innovation

Intel® OPA still passing data at reduced bandwidth with link recovery via PIP

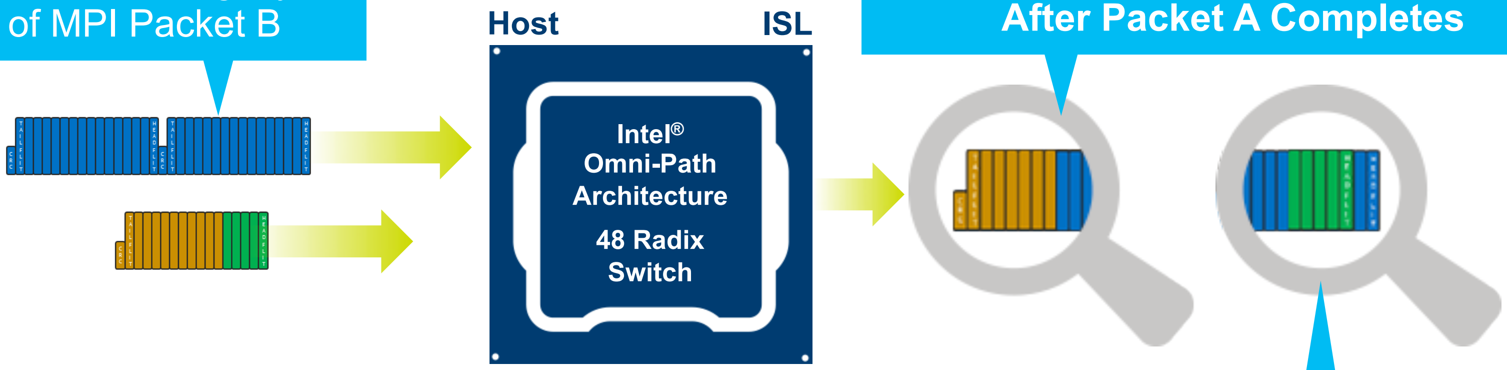
Intel technologies' features and benefits depend on system configuration and may require enabled hardware, software or service activation. Performance varies depending on system configuration. *Other names and brands may be claimed as property of others.

Intel® Omni-Path Fabric Link Level Innovation:

Traffic Flow Optimization (TFO) - **Preemption**

Packet A Arrives Slightly Ahead of MPI Packet B

Same Priority Packet C **Transmits After Packet A Completes**



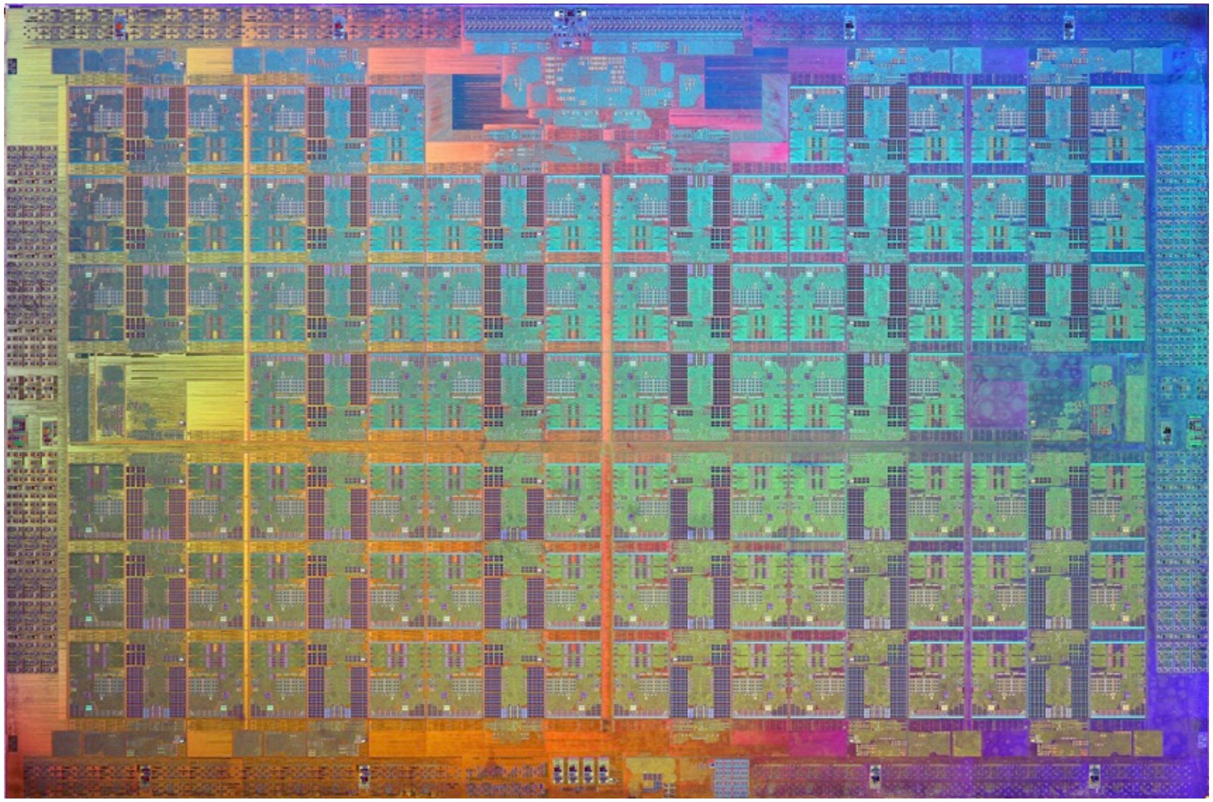
Suspends Packet A to Send **High Priority Packet B** then Resumes A

Packet A (VL0) Storage Traffic	🟦	**Large Packet with Lower Priority**
Packet B (VL0) MPI Traffic	🟩	**Packet B is on a Higher Priority VL**
Packet C (VL0) MPI Traffic	🟧	**Other Traffic on a Lower Priority VL**

VL = Virtual Lane (Each Lane Has a Different Priority)

Intel technologies' features and benefits depend on system configuration and may require enabled hardware, software or service activation. Performance varies depending on system configuration.

Knights Landing: Next Intel® Xeon Phi™ Processor



First **self-boot** Intel® Xeon Phi™ processor that is **binary compatible** with main line IA. Boots standard OS.

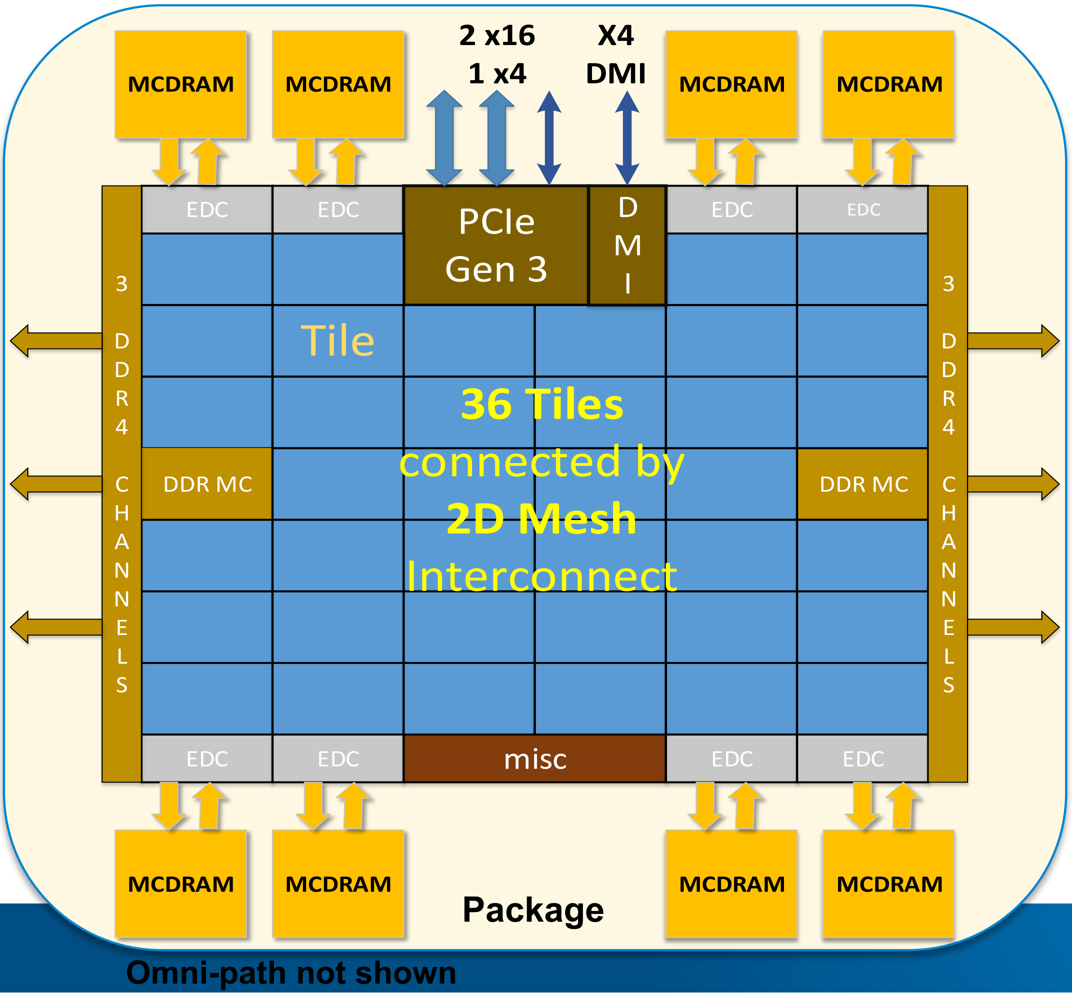
Significant improvement in scalar and **vector** performance

Integration of **Memory on package**: innovative memory architecture for high bandwidth and high capacity

Integration of **Fabric on package**

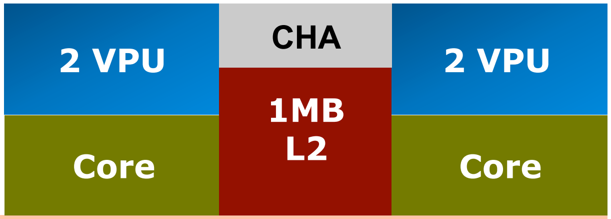
Potential future options subject to change without notice.
All timeframes, features, products and dates are preliminary forecasts and subject to change without further notification.

Knights Landing Overview



Omni-path not shown

TILE



Chip: 36 Tiles interconnected by **2D Mesh**

Tile: 2 Cores + 2 VPU/core + 1 MB L2

Memory: MCDRAM: 16 GB on-package; High BW

DDR4: 6 channels @ 2400 up to 384GB

IO: 36 lanes PCIe* Gen3. 4 lanes of DMI for chipset

Node: 1-Socket only

Fabric: Omni-Path on-package (not shown)

Vector[1]: up to 2 TF/s Linpack/DGEMM; 4.6 TF/s SGEMM

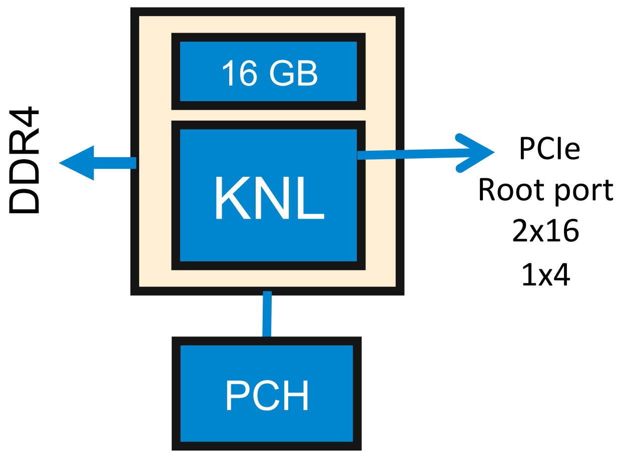
Streams Triad[1]: MCDRAM up to 490 GB/s; DDR4 90 GB/s

Scalar[2]: Up to ~3x over current Intel® Xeon Phi™ co-processor 7120 (“Knights Corner”)

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. Configurations:

1. Intel Xeon Phi processor 7250 (16GB, 1.4 GHz, 68-cores) running LINPACK (score 2000 GFLOPS), DGEMM (score 2070 GFLOPS), SGEMM (4605 GFLOPS), STREAM (DDR4 = 90 GB/s and MCDRAM = 490 GB/s), 96 GB DDR4-2133 memory, BIOS R00.RC085, Cluster Mode = Quad, MCDRAM Flat or Cache, RHEL* 7.0, MPSP 1.2.2, Intel MKL 11.3.2, Intel MPI 5.1.2, DGEMM 20K x 20K, LINPACK 100K x 100K size
2. Intel estimates based on estimated 1-user SPECint*_rate_base2006 comparing configuration 1 to Intel Xeon Phi co-processor 7120A hosted on 2x Intel Xeon processor E5-2697 v3.

Knights Landing Products

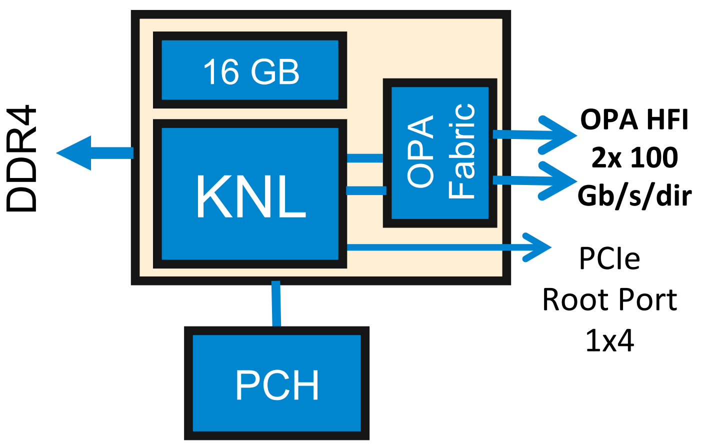


KNL

DDR Channels: 6

MCDRAM: up to 16 GB

Gen3 PCIe (Root port): 36 lanes



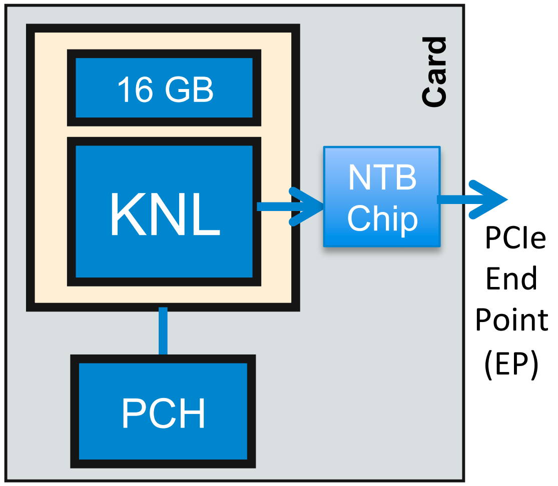
KNL with Omni-Path

DDR Channels: 6

MCDRAM: up to 16 GB

Gen3 PCIe (Root port): 4 lanes

Intel® Omni-Path Arch.: 200 Gb/s/dir



KNL Card

No DDR Channels

MCDRAM: up to 16 GB

Gen3 PCIe (End point): 16 lanes

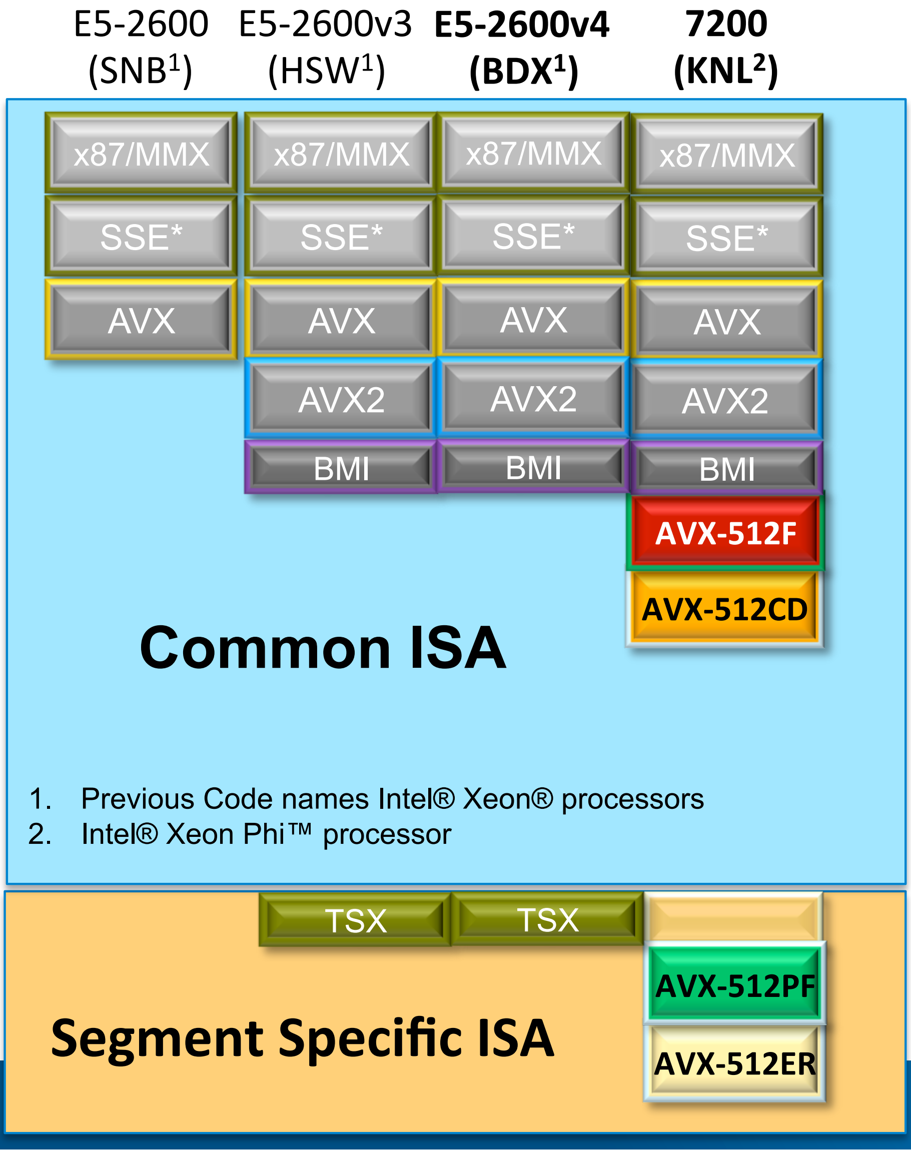
NTB Chip to create PCIe EP

Self Boot Socket

PCIe Card

Potential future options subject to change without notice. Codenames.
All timeframes, features, products and dates are preliminary forecasts and subject to change without further notification.

Intel ISA



E5-2600 (SNB[1])	E5-2600v3 (HSW[1])	**E5-2600v4 (BDX[1])**	**7200 (KNL[2])**
x87/MMX	x87/MMX	x87/MMX	x87/MMX
SSE*	SSE*	SSE*	SSE*
AVX	AVX	AVX	AVX
	AVX2	AVX2	AVX2
	BMI	BMI	BMI
			AVX-512F
			AVX-512CD

Common ISA

	TSX	TSX	
			AVX-512PF
			AVX-512ER

Segment Specific ISA

1. Previous Code names Intel® Xeon® processors
2. Intel® Xeon Phi™ processor

KNL implements all legacy instructions

AVX-512 Extensions

- 512-bit FP/Integer Vectors
- 32 regs, & 8 mask regs
- Gather/Scatter

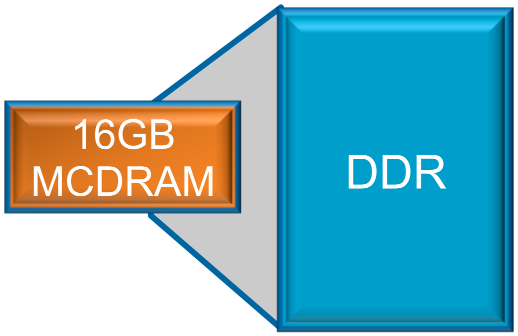
Conflict Detection: Improves Vectorization

Prefetch: Gather and Scatter Prefetch

Exponential and Reciprocal Instructions

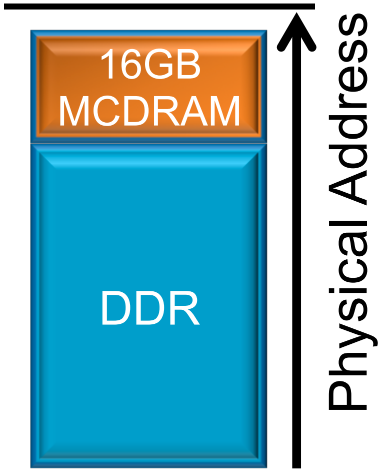
Three Memory Modes, Selected at Boot

Cache Mode



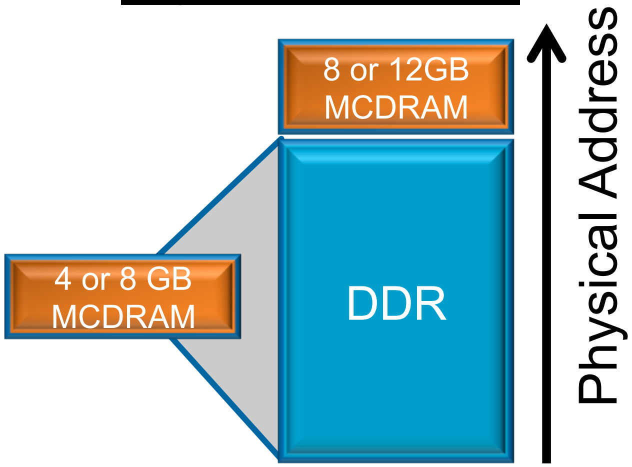
- SW-Transparent, Mem-side cache
- Direct mapped. 64B lines.
- Tags part of line
- Covers whole DDR range

Flat Mode



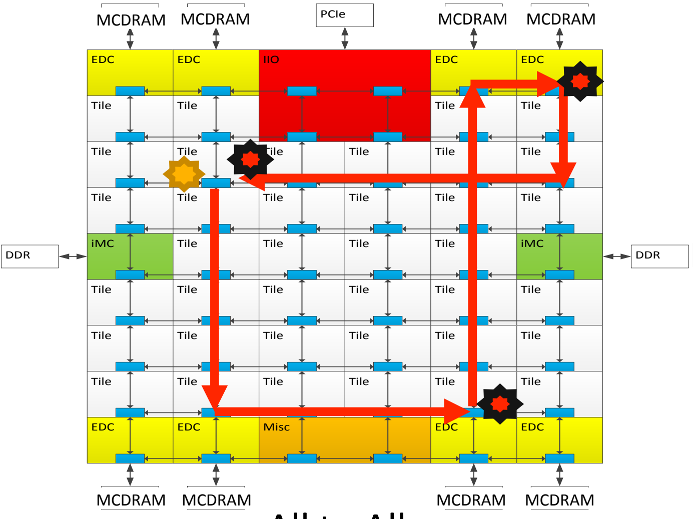
- MCDRAM as regular memory
- SW-Managed (e.g. memkind)
- Same address space

Hybrid Mode

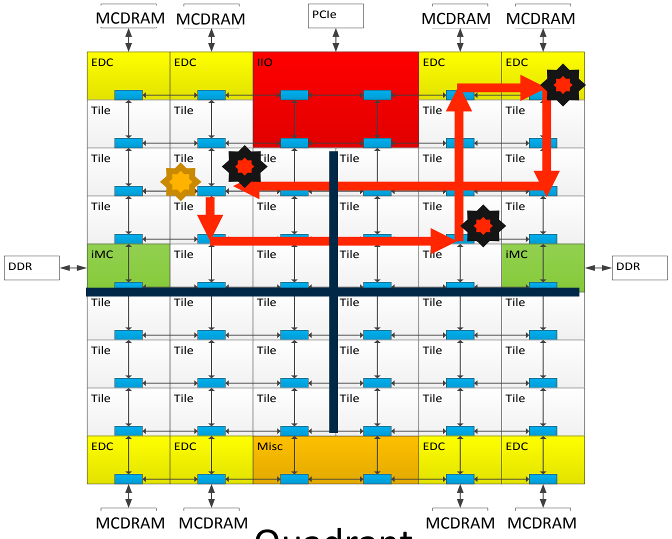


- Part cache, Part memory
- 25% or 50% cache
- Benefits of both

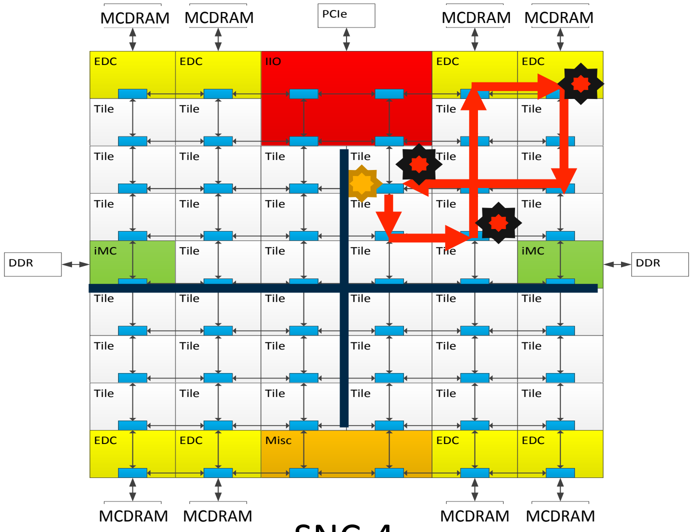
KNL Mesh Interconnect – Mesh of Rings



All to All



Quadrant

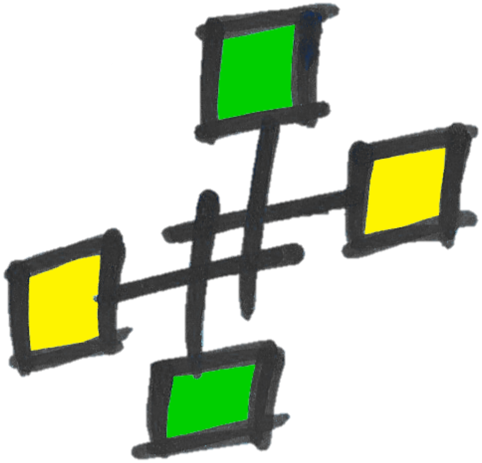


SNC-4

Three Cluster Modes:

1. **All-to-All**: No affinity between Tile, Directory and Memory
2. **Quadrant**: Affinity between Directory and Memory: Default mode. SW transparent
3. **Sub-NUMA Clustering**: Affinity between Tile, Directory, Memory. SW visible

Observations for OpenSHMEM



On-node communication is growing in importance

- OpenSHMEM can and should address both off-node and on-node communication needs
- PGAS can provide coherence alternative

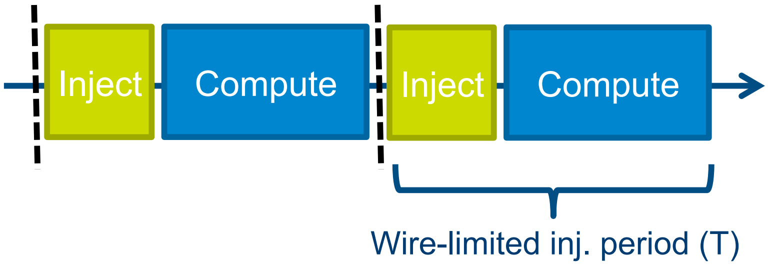
Hybrid processes + threads programming is already important

- Threads have an advantage with sharing on-node resources
- OpenSHMEM is late to the party – let's bring something good!

Heterogeneous memory has arrived

- HBW, DDR, large pages, NUMA, nonvolatile memory, …
- API should enable access to diverse memory technologies and allow users to control data placement and locality

Why Thread Safety Is Not Enough



Saturating the fabric with small messages

- Message rate of 160M/sec
- Processor with 72 cores
- Assuming all cores are sending messages, $T = T_{inject} + T_{compute} = 72/160M = 450ns$

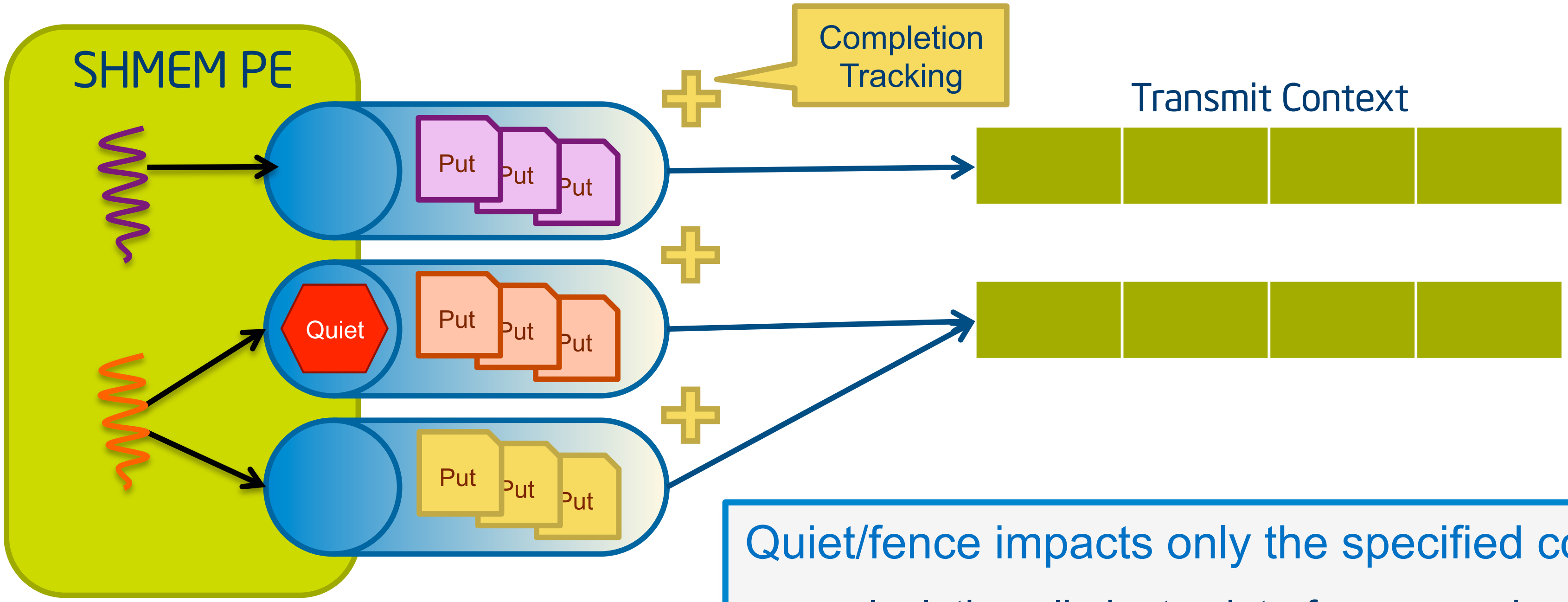
Low injection time is critical for small message workloads

- OpenSHMEM threading extensions must not burden critical paths
  - Taking a mutex
  - Accessing thread-local storage
  - Issuing an atomic operation

Contexts were designed to integrate threads while avoiding these overheads

Opinions expressed are those of the speaker and do not necessarily reflect the views of Intel Corp.

Contexts: You Want Them



Quiet/fence impacts only the specified context

- Isolation eliminates interference and synchronization within the OpenSHMEM runtime

Threads/PEs use contexts to overlap comm./comp.