Knights Landing (KNL): 2nd Generation Intel® Xeon Phi™ Processor

Avinash Sodani
KNL Chief Architect
Senior Principal Engineer, Intel Corp.

Legal

Knights Landing: Next Intel® Xeon Phi™ Processor

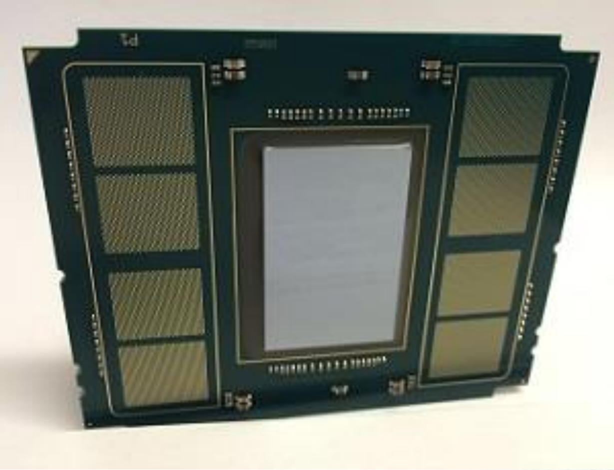
Intel® Many-Core Processor targeted for HPC and Supercomputing

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**



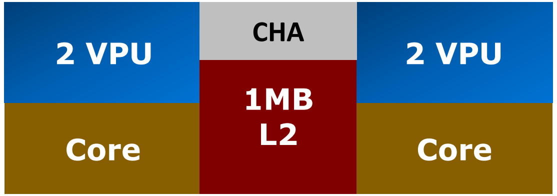
Three products

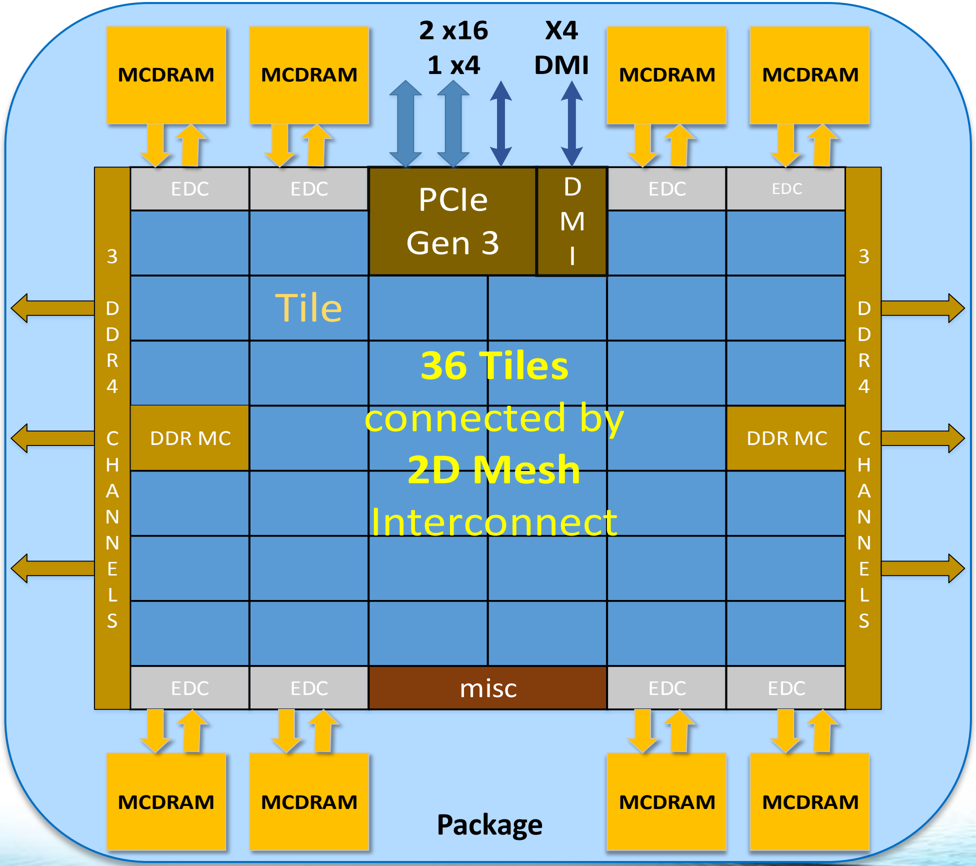
KNL Self-Boot	KNL Self-Boot w/ Fabric	KNL Card
(Baseline)	(Fabric Integrated)	(PCIe-Card)

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

TILE





Omni-path not shown

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 Peak Perf: 3+TF DP and 6+TF SP Flops

Scalar Perf: ~3x over Knights Corner

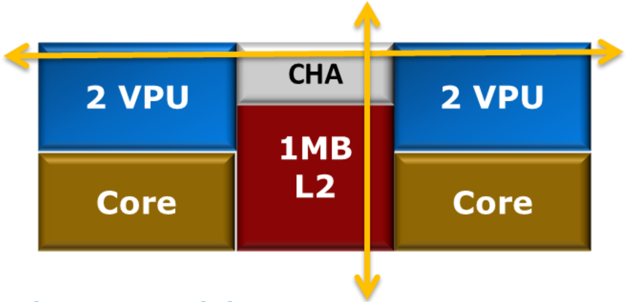
Streams Triad (GB/s): MCDRAM : 400+; DDR: 90+

Source Intel: All products, computer systems, dates and figures specified are preliminary based on current expectations, and are subject to change without notice. KNL data are preliminary based on current expectations and are subject to change without notice. 1Binary Compatible with Intel Xeon processors using Haswell Instruction Set (except TSX). 2Bandwidth numbers are based on STREAM-like memory access pattern when MCDRAM used as flat memory. Results have been estimated based on internal Intel analysis and are provided for informational purposes only. Any difference in system hardware or software design or configuration may affect actual performance.

KNL Tile:

2 Cores, each with 2 VPU

1M L2 shared between two Cores



Core: Changed from Knights Corner (KNC) to KNL. Based on 2-wide OoO Silvermont™ Microarchitecture, but with *many* changes for HPC.

4 thread/core. Deeper OoO. Better RAS. Higher bandwidth. Larger TLBs.

2 VPU: 2x AVX512 units. 32SP/16DP per unit. X87, SSE, AVX1, AVX2 and EMU

L2: 1MB 16-way. 1 Line Read and ½ Line Write per cycle. Coherent across all Tiles

CHA: **C**aching/**H**ome **A**gent. Distributed Tag Directory to keep L2s coherent. MESIF protocol. 2D-Mesh connections for Tile

Many Trailblazing Improvements in KNL

Improvements	What/Why
Self Boot Processor	No PCIe bottleneck
Binary Compatibility with Xeon	Runs all legacy software. No recompilation.
New Core: Atom™ based	~3x higher ST performance over KNC
Improved Vector density	3+ TFLOPS (DP) peak per chip
New AVX 512 ISA	New 512-bit Vector ISA with Masks
Scatter/Gather Engine	Hardware support for gather and scatter
New memory technology: MCDRAM + DDR	Large High Bandwidth Memory → MCDRAM Huge bulk memory → DDR
New on-die interconnect: Mesh	High BW connection between cores and memory
Integrated Fabric: Omni-Path	Better scalability to large systems. Lower Cost

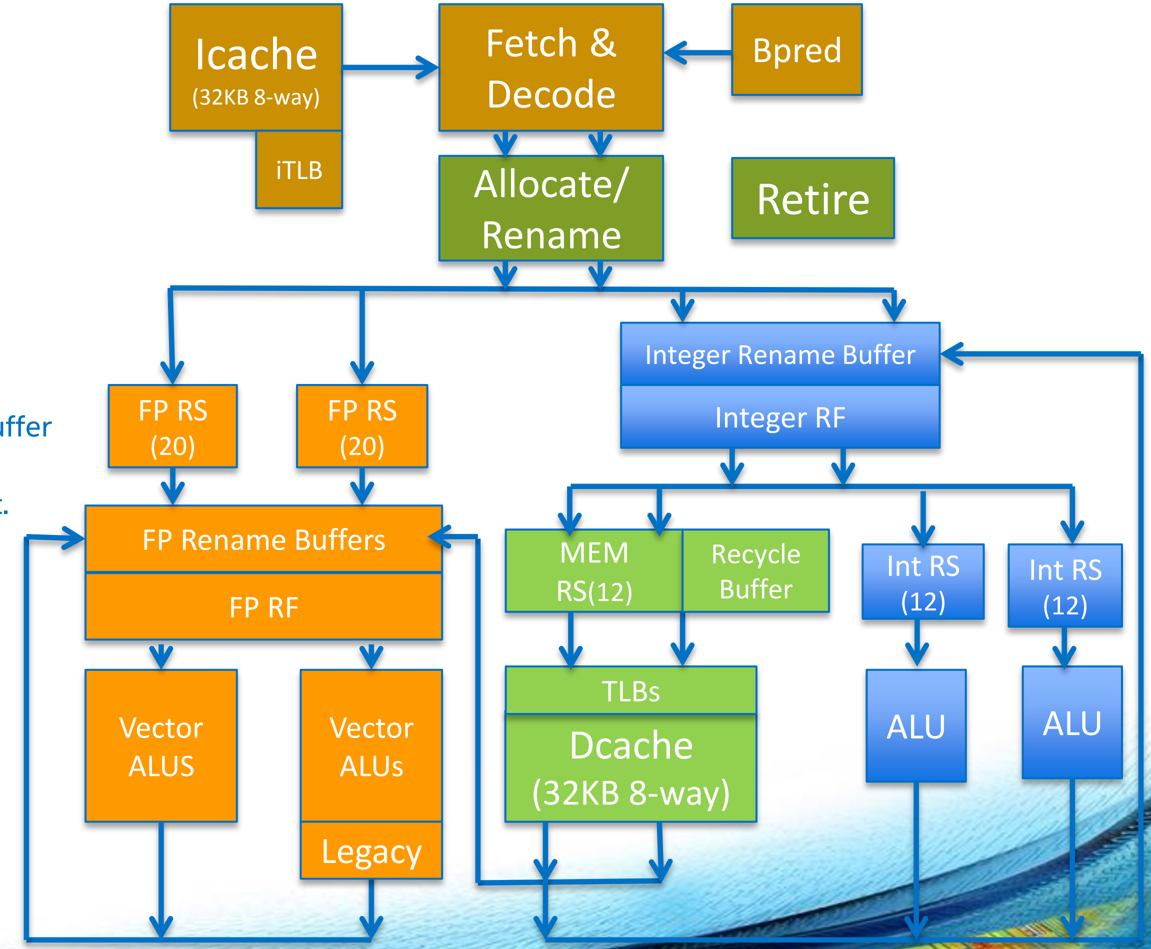
Core & VPU

- Out-of-order core w/ 4 SMT threads
- VPU tightly integrated with core pipeline

- 2-wide Decode/Rename/Retire
- ROB-based renaming. 72-entry ROB & Rename Buffers
- Up to 6-wide at execution
- Int and FP RS OoO.
- MEM RS inorder with OoO completion. Recycle Buffer holds memory ops waiting for completion.
- Int and Mem RS hold source data. FP RS does not.

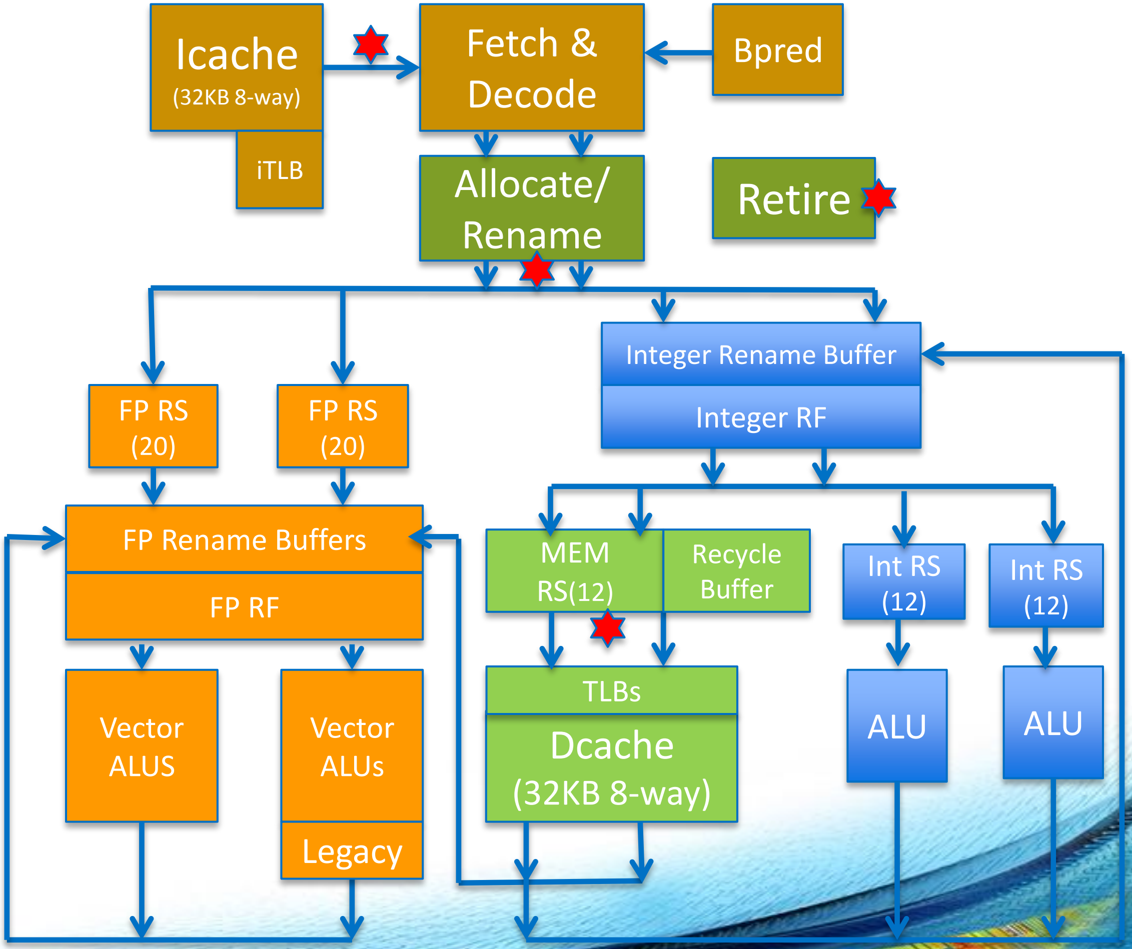
- 2x 64B Load & 1 64B Store ports in Dcache.
- 1st level uTLB: 64 entries
- 2nd level dTLB: 256 4K, 128 2M, 16 1G pages

- L1 Prefetcher (IPP) and L2 Prefetcher.
- 46/48 PA/VA bits
- Fast unaligned and cache-line split support.
- Fast Gather/Scatter support

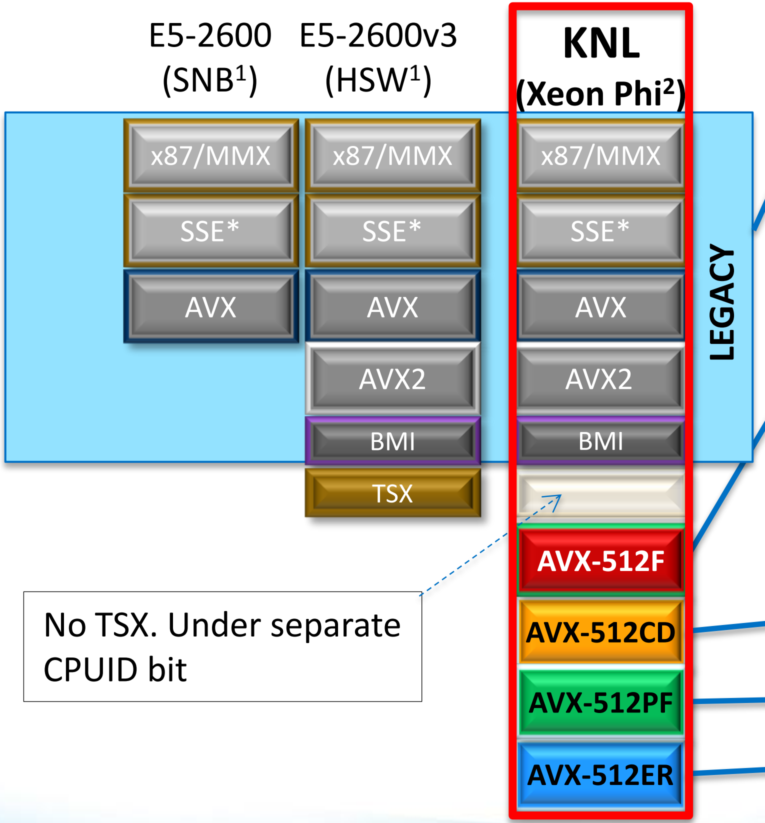


Threading

- 4 Threads per core. Simultaneous Multithreading.
- Core resources **shared** or **dynamically repartitioned** between active threads
  - ROB, Rename Buffers, RS: Dynamically partitioned
  - Caches, TLBs: Shared
  - E.g., 1 thread active → uses full resources of the core
- Several Thread Selection points in the pipeline. (★)
  - Maximize throughput while being fair.
  - Account for available resources, stalls and forward progress



KNL ISA



KNL implements all legacy instructions

- Legacy binary runs w/o recompilation
- KNC binary requires recompilation

KNL introduces AVX-512 Extensions

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

Conflict Detection: Improves Vectorization

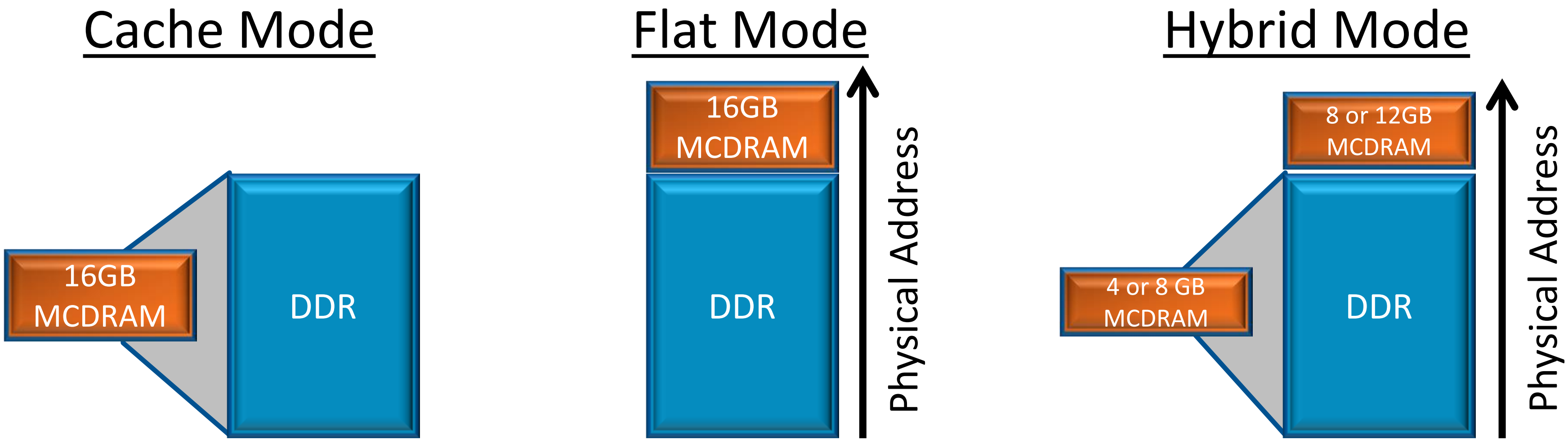
Prefetch: Gather and Scatter Prefetch

Exponential and Reciprocal Instructions

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

Memory Modes

Three Modes. Selected at boot



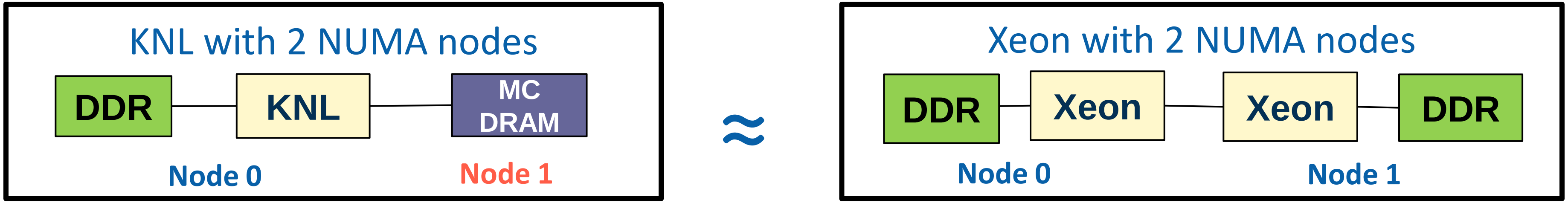
- SW-Transparent, Mem-side cache
- Direct mapped. 64B lines.
- Tags part of line
- Covers whole DDR range

- MCDRAM as regular memory
- SW-Managed
- Same address space

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

Flat MCDRAM: SW Architecture

MCDRAM exposed as a separate NUMA node



Memory allocated in DDR by default → Keeps non-critical data out of MCDRAM.

Apps explicitly allocate critical data in MCDRAM. Using <u>two</u> methods:

- **"Fast Malloc"** functions in High BW library (https://github.com/memkind)
  - Built on top to existing *libnuma* API
- **"FASTMEM"** Compiler Annotation for Intel Fortran

Flat MCDRAM with existing NUMA support in Legacy OS

Flat MCDRAM SW Usage: Code Snippets

C/C++ (*https://github.com/memkind)

Allocate into DDR

```c
float   *fv;
fv = (float *)malloc(sizeof(float)*100);
```

Allocate into MCDRAM

```c
float   *fv;
fv = (float *)hbw_malloc(sizeof(float) * 100);
```

Intel Fortran

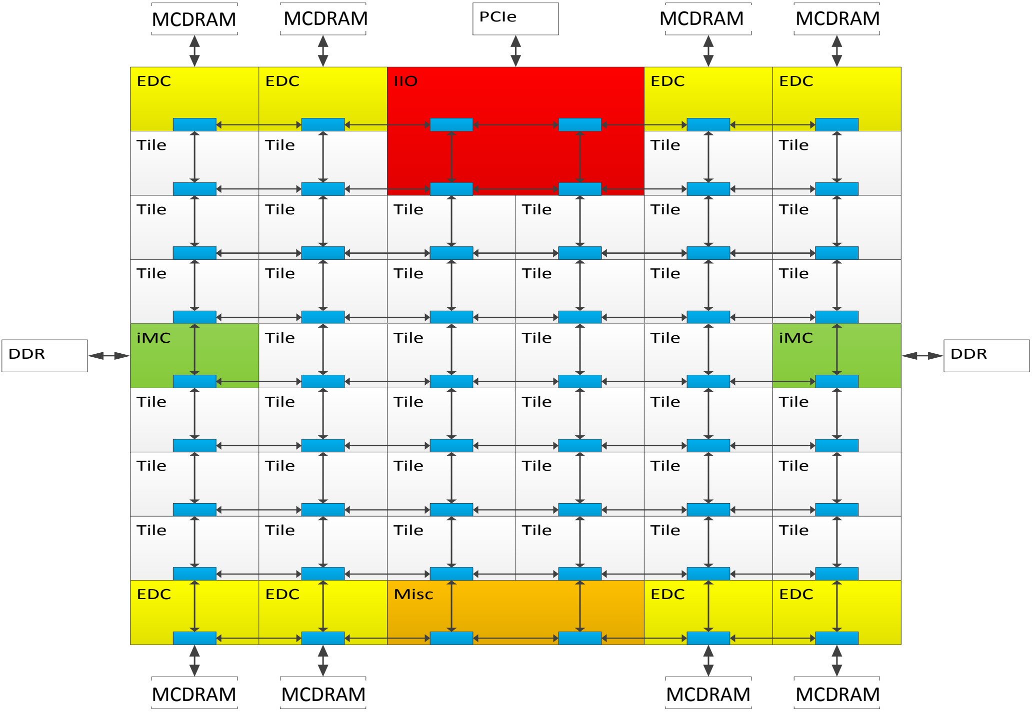
Allocate into MCDRAM

```fortran
c      Declare arrays to be dynamic
       REAL, ALLOCATABLE :: A(:)

!DEC$ ATTRIBUTES, FASTMEM :: A

       NSIZE=1024
c      allocate array 'A' from MCDRAM
c
       ALLOCATE (A(1:NSIZE))
```

KNL Mesh Interconnect



Mesh of Rings

- Every row and column is a (half) ring
- YX routing: Go in Y → Turn → Go in X
- Messages arbitrate at injection and on turn

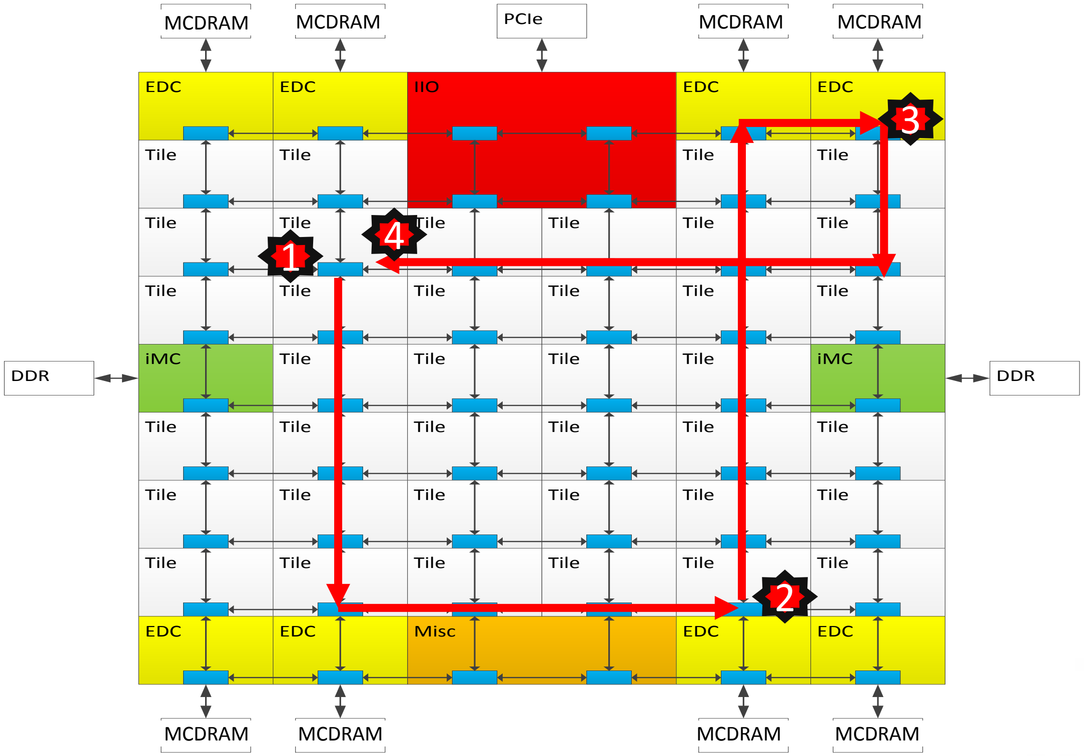
Cache Coherent Interconnect

- MESIF protocol (F = Forward)
- Distributed directory to filter snoops

Three Cluster Modes

(1) All-to-All (2) Quadrant (3) Sub-NUMA Clustering

Cluster Mode: All-to-All



Address uniformly hashed across all distributed directories

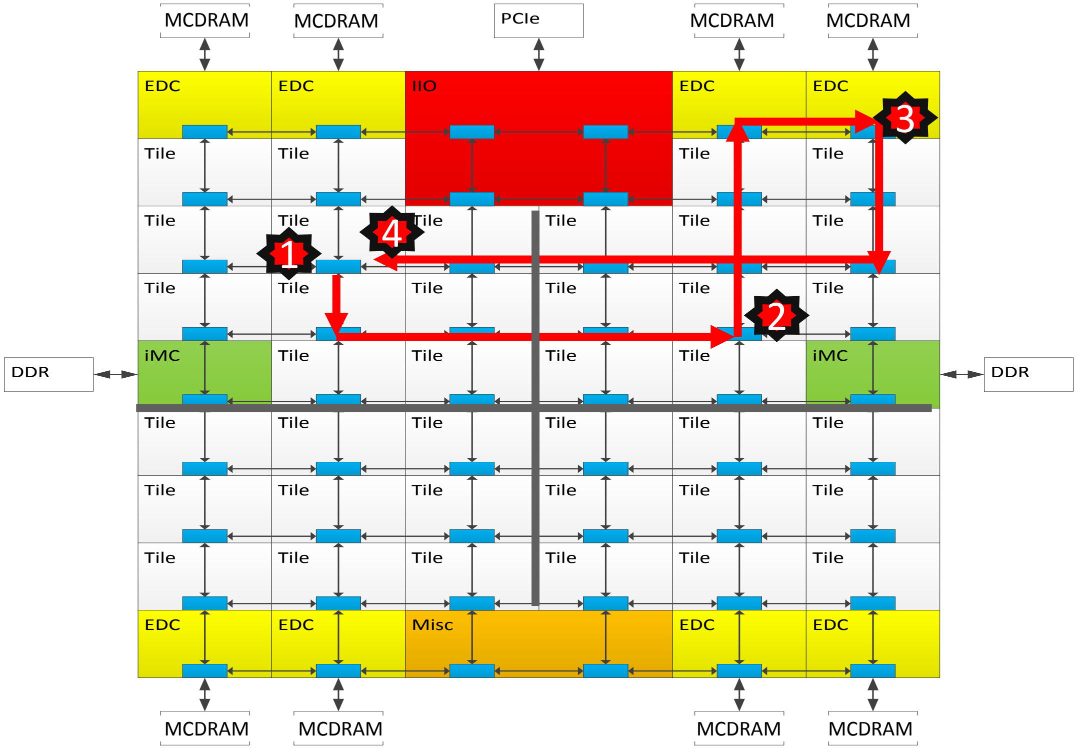
No affinity between Tile, Directory and Memory

Most general mode. Lower performance than other modes.

<u>Typical Read L2 miss</u>

1. L2 miss encountered
2. Send request to the distributed directory
3. Miss in the directory. Forward to memory
4. Memory sends the data to the requestor

Cluster Mode: Quadrant



Chip divided into four virtual Quadrants

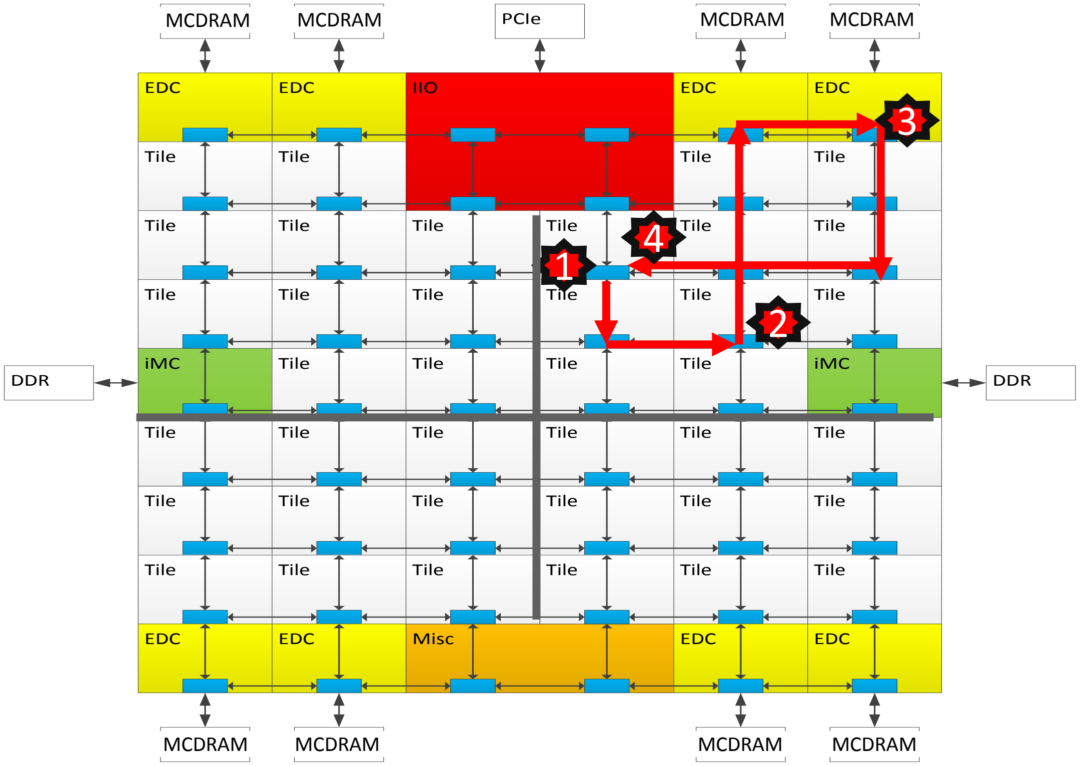
Address hashed to a Directory in the same quadrant as the Memory

Affinity between the Directory and Memory

Lower latency and higher BW than all-to-all. SW Transparent.

1) L2 miss, 2) Directory access, 3) Memory access, 4) Data return

Cluster Mode: Sub-NUMA Clustering (SNC)



1) L2 miss, 2) Directory access, 3) Memory access, 4) Data return

Each Quadrant (Cluster) exposed as a separate NUMA domain to OS.

Looks analogous to 4-Socket Xeon

Affinity between Tile, Directory and Memory

Local communication. Lowest latency of all modes.

SW needs to NUMA optimize to get benefit.

KNL with Omni-Path™

Omni-Path™ Fabric integrated *on package*

First product with integrated fabric

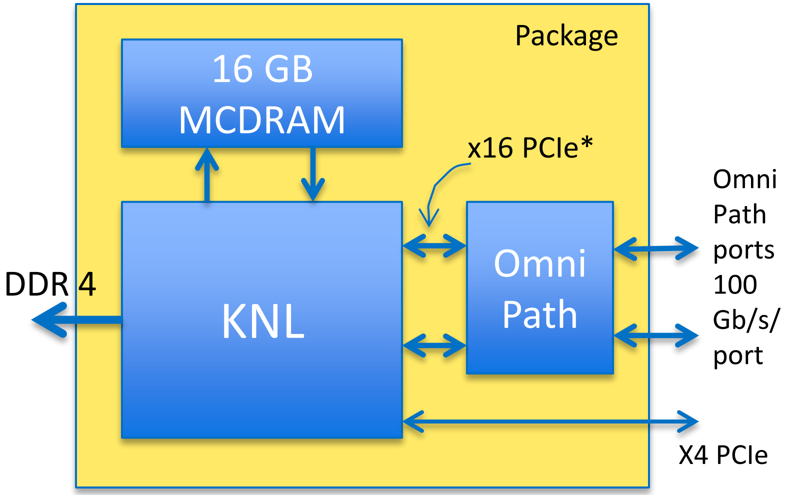
Connected to KNL die via 2 x16 PCIe* ports

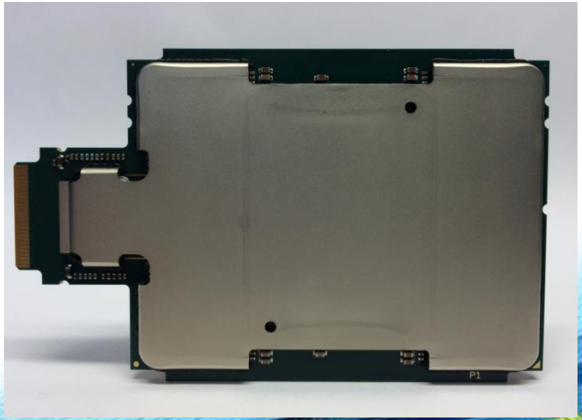
Output: 2 Omni-Path ports

- 25 GB/s/port (bi-dir)

Benefits

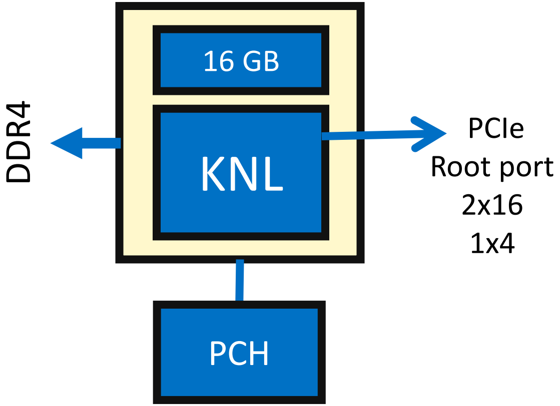
- Lower cost, latency and power
- Higher density and bandwidth
- Higher scalability





*On package connect with PCIe semantics, with MCP optimizations for physical layer

Knights Landing Products

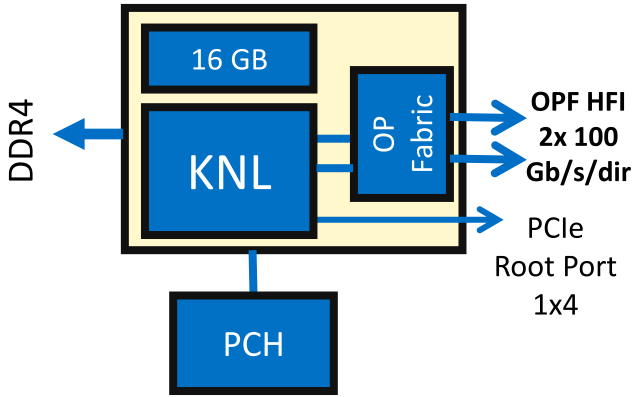


KNL

DDR Channels: 6

MCDRAM: up to 16 GB

Gen3 PCIe (Root port): 36 lanes



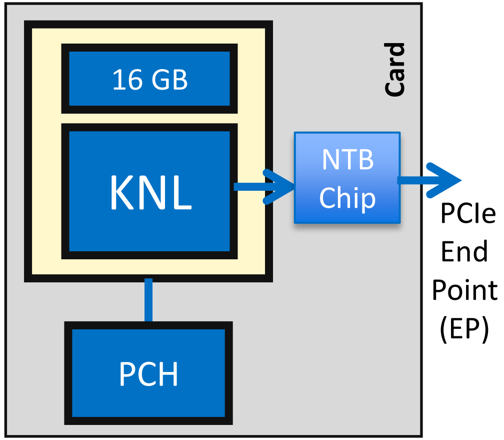
KNL with Omni-Path

DDR Channels: 6

MCDRAM: up to 16 GB

Gen3 PCIe (Root port): 4 lanes

Omni-Path Fabric: 200 Gb/s/dir



KNL Card

No DDR Channels

MCDRAM: up to 16 GB

Gen3 PCIe (End point): 16 lanes

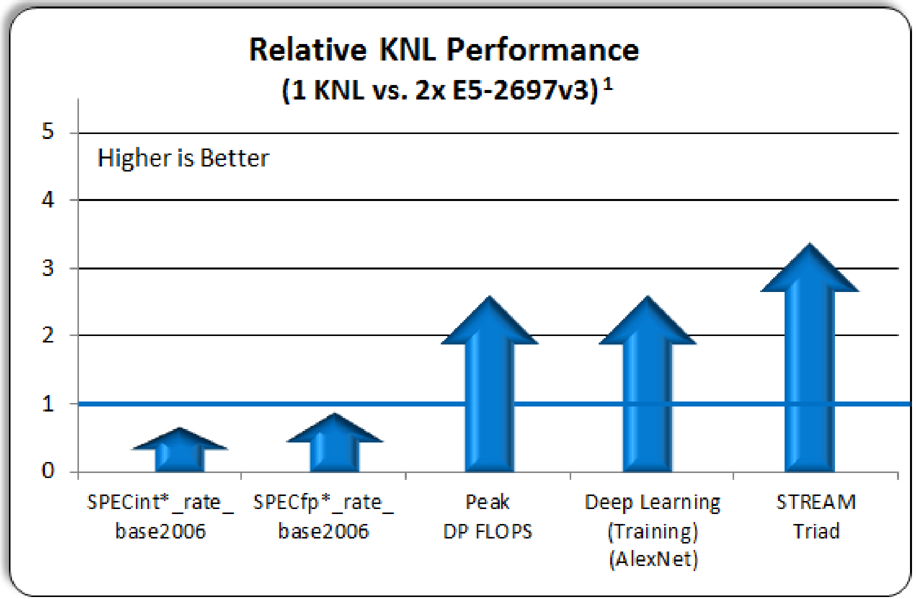
NTB Chip to create PCIe EP

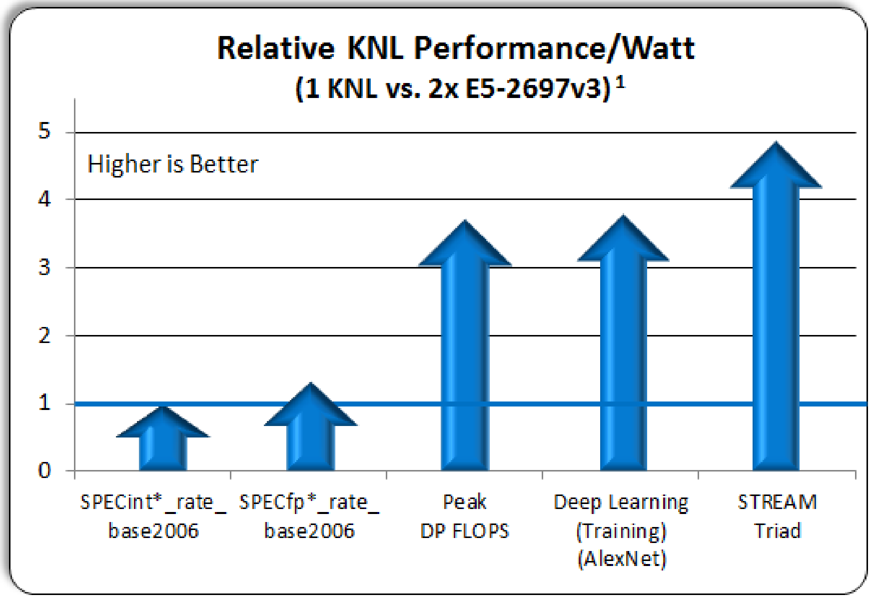
Self Boot Socket

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.

KNL Performance





Significant performance improvement for compute and bandwidth sensitive workloads, while still providing good general purpose throughput performance.

1. Projected KNL Performance (1 socket, 200W CPU TDP) vs. 2 Socket Intel® Xeon® processor E5-2697v3 (2x145W CPU TDP)

Backup

High Bandwidth (HBW) Malloc API

```
HBWMALLOC(3)                    HBWMALLOC                    HBWMALLOC(3)

NAME
       hbwmalloc - The high bandwidth memory interface

SYNOPSIS
       #include <hbwmalloc.h>

       Link with -ljemalloc -lnuma -lmemkind -lpthread

       int hbw_check_available(void);
       void* hbw_malloc(size_t size);
       void* hbw_calloc(size_t nmemb, size_t size);
       void* hbw_realloc (void *ptr, size_t size);
       void hbw_free(void *ptr);
       int hbw_posix_memalign(void **memptr, size_t alignment, size_t size);
       int hbw_posix_memalign_psize(void **memptr, size_t alignment, size_t size, int
pagesize);
       int hbw_get_policy(void);
```

Publicly released at https://github.com/memkind

AVX-512 PF, ER and CD Instructions

- Intel AVX-512 Prefetch Instructions (PFI)
- Intel AVX-512 Exponential and Reciprocal Instructions (ERI)
- Intel AVX-512 Conflict Detection Instructions (CDI)

CPUID	Instructions	Description
AVX512PF	PREFETCHWT1	Prefetch cache line into the L2 cache with intent to write
	VGATHERPF{D,Q}{0,1}PS	Prefetch vector of D/Qword indexes into the L1/L2 cache
	VSCATTERPF{D,Q}{0,1}PS	Prefetch vector of D/Qword indexes into the L1/L2 cache with intent to write
AVX512ER	VEXP2{PS,PD}	Computes approximation of 2^x with maximum relative error of 2^{-23}
	VRCP28{PS,PD}	Computes approximation of reciprocal with max relative error of 2^{-28} before rounding
	VRSQRT28{PS,PD}	Computes approximation of reciprocal square root with max relative error of 2^{-28} before rounding
AVX512CD	VPCONFLICT{D,Q}	Detect duplicate values within a vector and create conflict-free subsets
	VPLZCNT{D,Q}	Count the number of leading zero bits in each element
	VPBROADCASTM{B2Q,W2D}	Broadcast vector mask into vector elements

Glossary

KNL: Knights Landing
KNC: Knights Corner
HSW: Haswell
SNB: Sandy Bridge
OoO: Out-of-Order
ROB: Reorder Buffer
RS: Reservation Stations
VPU: Vector Processing Unit
EMU: Extended Math Unit
TLB: Translation Look-aside Buffer
CHA: Caching/Home Agent
PA/VA: Physical Address/Virtual Address

MCDRAM: “Multi-Channel” DRAM. High BW memory

EDC: Memory Controller for MCDRAM

MESIF: Coherence procotol (Modified, Exclusive, Shared., Invalid, Forward)

NTB: Non-Transparent Bridge

PCH: Chipset

NUMA: Non-Uniform Memory Access


intel