Intel® Xeon Phi™ Processor Codenamed Knights Landing (KNL)

KNL Architecture Overview

ISA

Intel® Xeon® Processor Binary-Compatible (w/Broadwell)

On-package memory

Up to 16GB, ~460 GB/s STREAM at launch

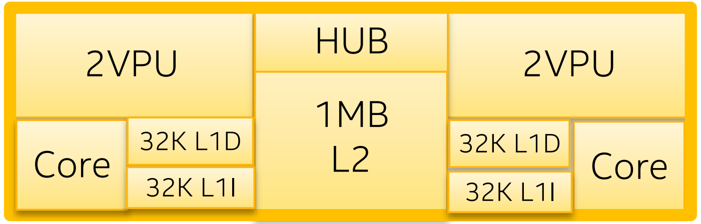
Platform Memory

Up to 384GB (6ch DDR4-2400 MHz)

Fixed Bottlenecks

✓ 2D Mesh Architecture
✓ Out-of-Order Cores
✓ 3X single-thread vs. KNC

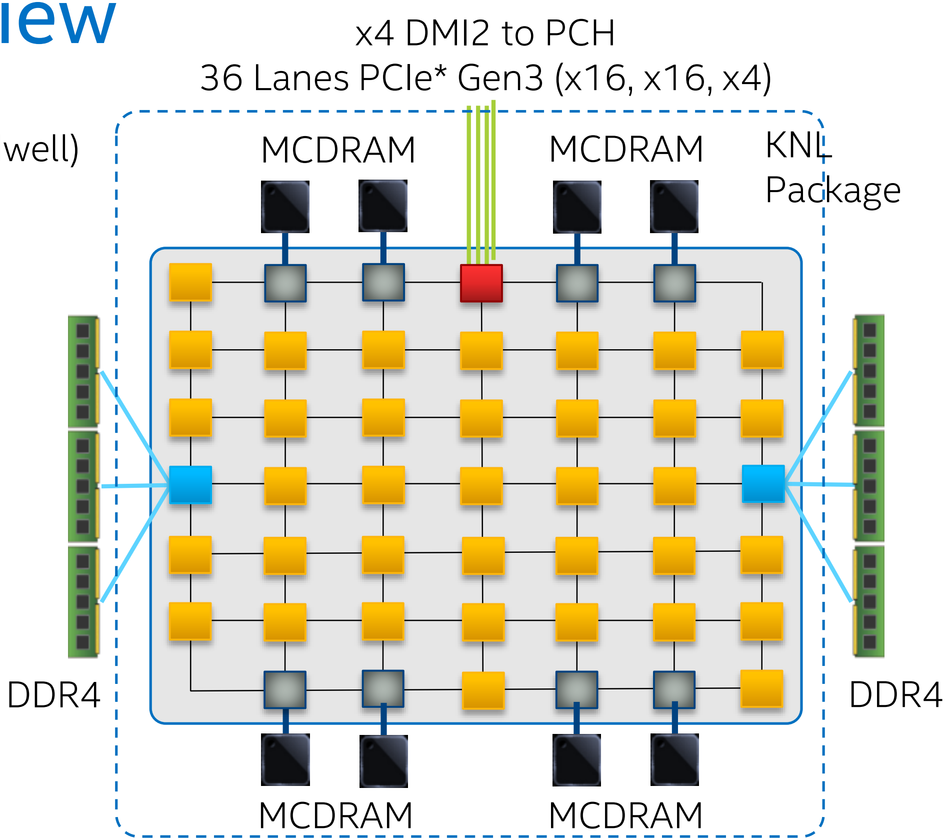
TILE:
(up to 36)



2VPU | HUB | 2VPU

Core | 32K L1D | 32K L1I | 1MB L2 | 32K L1D | 32K L1I | Core

Enhanced Intel® Atom™ cores based on Silvermont Microarchitecture



🟧 Tile

⬛ EDC (embedded DRAM controller)

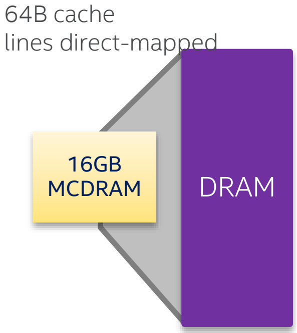
🟦 IMC (integrated memory controller)

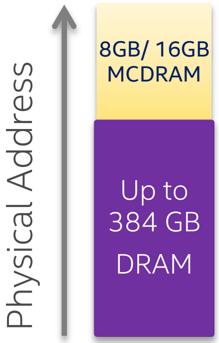
🟥 IIO (integrated I/O controller)

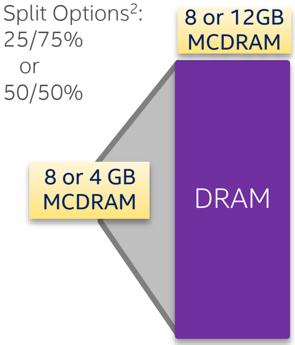
Integrated On-Package Memory Usage Models

Model configurable at boot time and software exposed through NUMA[1]

Platform Memory (DDR4) only available for bootable KNL host processor





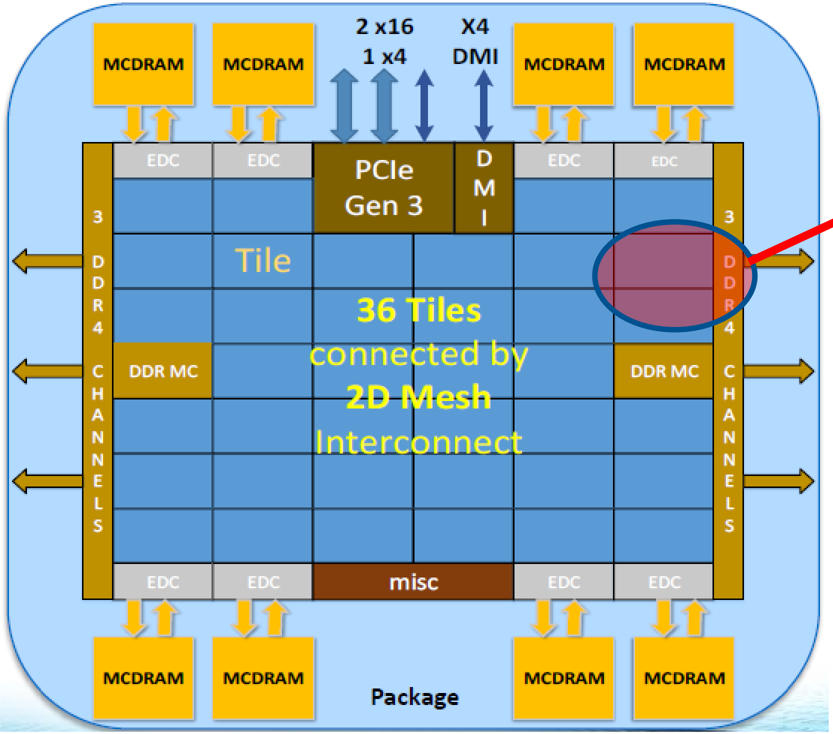


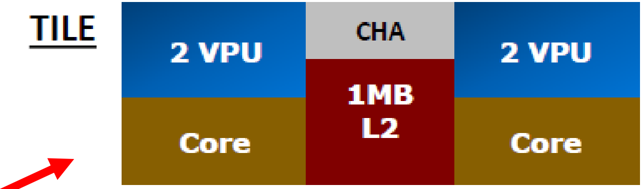
	Cache Model	Flat Model	Hybrid Model
Description	Hardware automatically manages the MCDRAM as a "L3 cache" between CPU and ext DDR memory	Manually manage how the app uses the integrated on-package memory and external DDR for peak perf	Harness the benefits of both Cache and Flat models by segmenting the integrated on-package memory
Usage Model	▪ App and/or data set is very large and will not fit into MCDRAM ▪ Unknown or unstructured memory access behavior	▪ App or portion of an app or data set that can be, or is needed to be "locked" into MCDRAM so it doesn't get flushed out	▪ Need to "lock" in a relatively small portion of an app or data set via the Flat model ▪ Remaining MCDRAM can then be configured as Cache

[1] NUMA = non-uniform memory access

[2] As projected based on early product definition

Knights Landing Overview





Chip: Up To 36 tiles interconnected by 2D Mesh

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

Memory: Up To 16GB on-package MCDRAM + up to 6 channels of DDR4-2400 (up to 384GB)

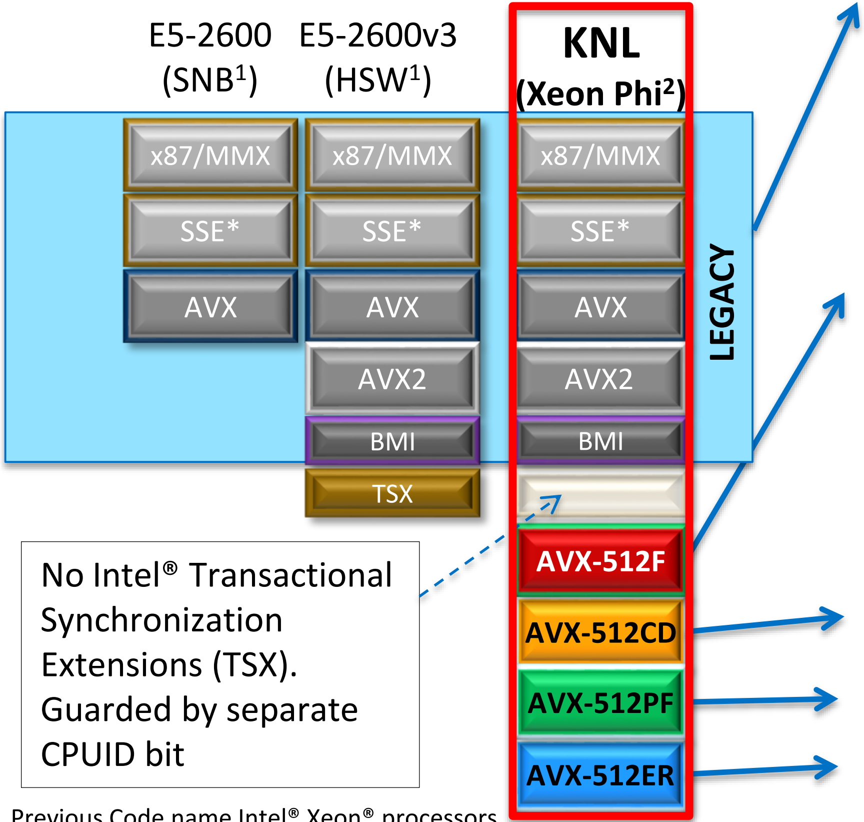
IO: 36 lanes PCIe Gen3 + 4 lanes DMI for chipset

Node: 1-socket only

Fabric: Omni-path in package (not shown)

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. 1. Binary Compatible with Intel Xeon processors using Haswell Instructions Set (except TSX), 2 Bandwidth 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 & software design or configuration may affect actual performance.

KNL ISA



No Intel® Transactional Synchronization Extensions (TSX). Guarded by separate CPUID bit

KNL implements all legacy instructions

- Existing binaries run w/o recompilation
- KNC binaries require 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

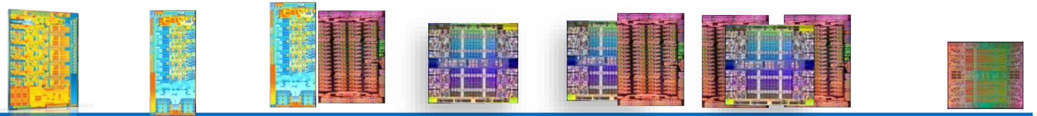
Exponential and Reciprocal Instructions

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

Case Study: The STAC-A2 benchmark

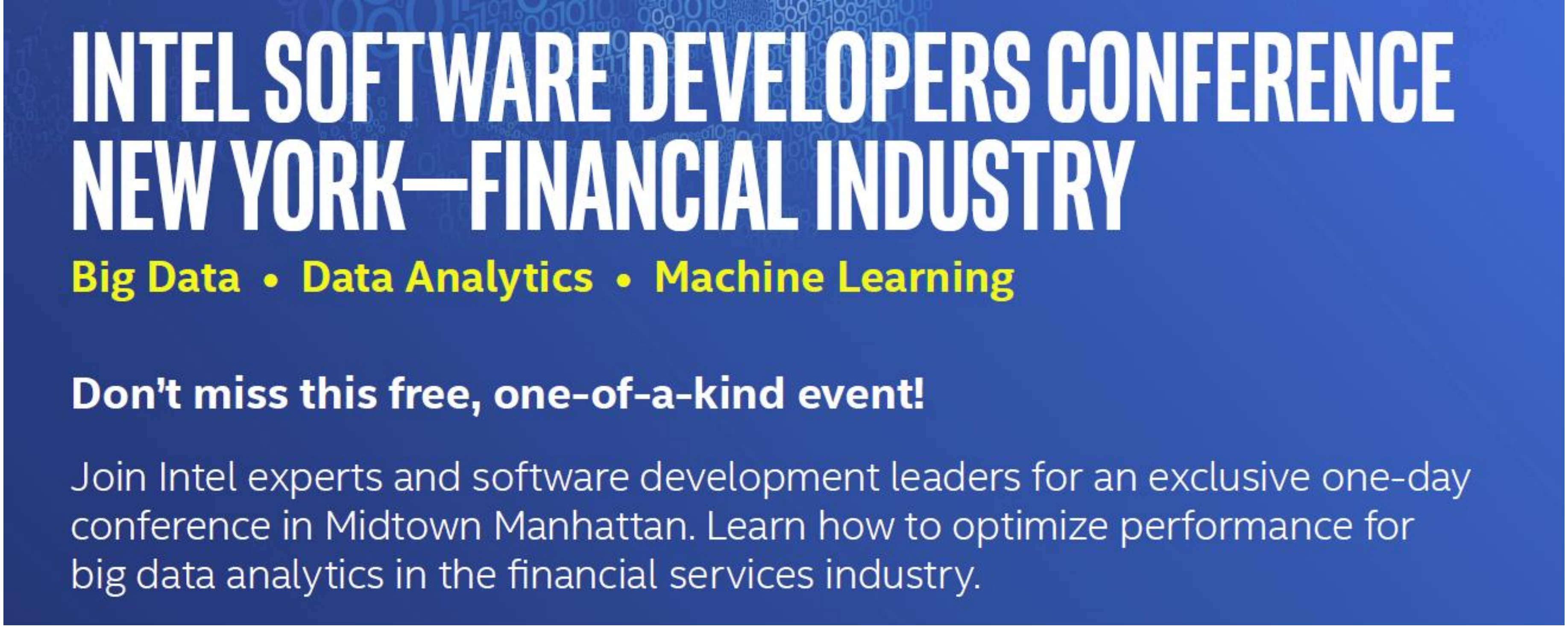
STAC-A2 evaluates Monte Carlo over 5 assets and the Greeks

- 5 assets, 25K path, 252 time steps
- For American-style options using the Heston Model
- Compute Greeks: Theta, Rho, Delta, Gamma, Cross-Gamma, Model Vega, Correlation Vega:



	Intel Xeon processor E5 2690	Intel Xeon processor E5 2697-V2	Intel Xeon E5 2697-V2 + Xeon Phi	Intel Xeon processor E5 2697-V3	Intel Xeon E5 2697-V3+ Xeon Phi	Intel Xeon E5 2697-V3+ 2*Xeon Phi	Intel Xeon processor E5 2699-V4
	2013	2014	2014	2014	2014	2015	2016
cores	16	24	24+61	36	36+61	36+122	44
Threads	32	48	48+244	72	72+244	72+488	88
vectors	256	256	256+512	256	256+512	256+2*512	256
Parallelization	OpenMP	TBB	TBB	TBB	TBB	TBB	TBB
Vectorization	#SIMD	OpenMP	OpenMP	OpenMP	OpenMP	OpenMP	OpenMP
Heterogeneity	N/A	N/A	OpenMP	N/A	OpenMP	TBB	N/A
Greek time	5.8	1.0	0.63	0.81	0.53	0.216	0.371

27x overall improvement



INTEL SOFTWARE DEVELOPERS CONFERENCE NEW YORK—FINANCIAL INDUSTRY

Big Data • Data Analytics • Machine Learning

Don't miss this free, one-of-a-kind event!

Join Intel experts and software development leaders for an exclusive one-day conference in Midtown Manhattan. Learn how to optimize performance for big data analytics in the financial services industry.

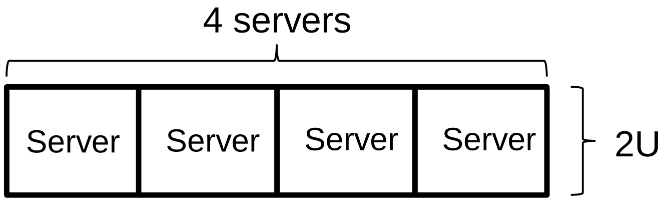
June 23, 2016

Le Parker Meridien, 119 W. 56th St., New York, NY 10019

Register now at intel.ly/1TKRGPT

STAC-A2 on KNL – Background

- Testing out pre-release product. First configuration is below.
- Hardware
  - 1 x Intel Xeon Phi 7250 (Knights Landing)
    - 68 physical cores
    - 272 logical cores
  - 96GB DRAM, 16GB MCDRAM
  - Intel white box, effectively 0.5U



- Software
  - STAC-A2 Pack for Intel Composer XE Rev H
    - Derived from Rev F. Ideal for homogeneous systems
  - Intel Composer XE, Intel Threading Building Blocks
- First STAC-A2 results using just one socket

STAC-A2 on KNL – Results Highlights