Knights Landing
2nd Generation Intel® Xeon Phi™ products

ALL ABOUT PARALEL PROGRAMMING
Threading, Vectorization, Data Locality
Fortran, C, C++ (plus a little Python)
OpenMPI, MPI, Intel® TBB

New with Knights Landing:
AVX-512,
High Bandwidth Memory (MCDRAM),
Cluster Mode,
Omni-Path
Highly-Parallel Performance to develop on

All the Software Tools & Libraries you need

Support & Training to help you succeed

Leading edge platform capabilities, performance to deliver multi-threaded, vectorized software for today’s HPC workloads!

You can buy your very own Knight Landing development system today!
Why Intel® Xeon Phi™ Processors?
Imagine you could design anything

What is you wanted:

• High performance
• Preserve your investment in code
• Reuse your knowledge of programming languages, tools, techniques

What if the ONLY thing we were willing to “give up” was running non-parallel codes well?

In other words:
what if we assumed ONLY parallel applications would be run?
Imagine you could design anything

What is you wanted:

- **High performance**
- **Preserve your investment in code**
- **Reuse your knowledge of programming languages, tools, techniques**

What if the ONLY thing we were willing to “give up” was running non-parallel codes well?

In other words:
what if we assumed ONLY parallel applications would be run?
If you were plowing a field, which would you rather use… two strong oxen, or 1024 chickens?
Design Question - Best for Computing?

A few powerful vs. Many less powerful.

Diagrams for discussion purposes only, not a precise representation of any product of any company.
Design Question – Our difference: Same programming models, languages, optimizations and tools

A few powerful vs. Many less powerful.

Same programming models, languages, optimizations and tools.
vision
span from few cores to many cores with consistent models, languages, tools, and techniques
# Moderncode: COSMOS

Book Cover Background: Photo of the COSMOS@DiRAC SGI UV2000 based Supercomputer manufactured by SGI, Inc and operated by the Stephen Hawking Centre for Theoretical Cosmology, University of Cambridge. Photo courtesy of Philip Mynott. Book Cover Foreground: 3D visualization of statistical fluctuations in the Cosmic Microwave Background, the remnant of the first measurable light after the Big Bang. CMB data is from the Planck satellite and is the topic of Chapter 10 providing insights into new physics and how the universe evolved. Visualization rendered with Intel’s OSPRay ray tracing open source software by Gregory P. Johnson and Timothy Rowley, Intel Corporation.
Book Cover Background: Photo of the COSMOS@DiRAC SGI UV2000 based Supercomputer manufactured by SGI, Inc and operated by the Stephen Hawking Centre for Theoretical Cosmology, University of Cambridge. Photo courtesy of Philip Mynott. Book Cover Foreground: 3D visualization of statistical fluctuations in the Cosmic Microwave Background, the remnant of the first measurable light after the Big Bang. CMB data is from the Planck satellite and is the topic of Chapter 10 providing insights into new physics and how the universe evolved. Visualization rendered with Intel's OSPRay ray tracing open source software by Gregory P. Johnson and Timothy Rowley, Intel Corporation.
We find that using a simple trapezium rule integrator combined with hand-selected sampling points (to improve accuracy in areas of interest) provides sufficient numerical accuracy to obtain a physically meaningful result, and the reduced space and time requirements of this simplified method give a speed-up of $O(10x)$. 

<table>
<thead>
<tr>
<th>Version</th>
<th>Processor (s)</th>
<th>Coprocessor (s)</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2887.0</td>
<td>-</td>
<td>Original code.</td>
</tr>
<tr>
<td>2</td>
<td>2610.0</td>
<td>-</td>
<td>Loop simplification.</td>
</tr>
<tr>
<td>3</td>
<td>882.0</td>
<td>1991.6</td>
<td>Intel® MKL integration routines and function inlining.</td>
</tr>
<tr>
<td>4</td>
<td>865.9</td>
<td>667.9</td>
<td>Flattened loops and introduced OpenMP threads.</td>
</tr>
<tr>
<td>5</td>
<td>450.6</td>
<td>655.0</td>
<td>Loop reordering and manual nested threading.</td>
</tr>
<tr>
<td>6</td>
<td>385.6</td>
<td>49.5</td>
<td>Blocked version of the loop (for cache).</td>
</tr>
<tr>
<td>7</td>
<td>46.9</td>
<td>34.5</td>
<td>Numerical integration routine (Trapezium Rule).</td>
</tr>
<tr>
<td>8</td>
<td>37.4</td>
<td>37.7</td>
<td>Reduction with DGEMM.</td>
</tr>
<tr>
<td>9</td>
<td>35.1</td>
<td>34.5</td>
<td>Data alignment (for vectorization).</td>
</tr>
<tr>
<td>10</td>
<td>34.3</td>
<td>26.6</td>
<td>Tuning of software prefetching distances.</td>
</tr>
</tbody>
</table>
Knights Landing
2nd Generation Intel® Xeon Phi™

ALL ABOUT PARALLEL PROGRAMMING
Threading, Vectorization, Data Locality
Fortran, C, C++ (plus a little Python)
OpenMPI, MPI, TBB

New with Knights Landing:
AVX-512,
High Bandwidth Memory (MCDRAM),
Cluster Mode,
Omni-Path
INTEL® XEON PHI™
PROCESSOR
HIGH PERFORMANCE
PROGRAMMING
KNIGHTS LANDING EDITION
Jim Jeffers | James Reinders | Avinash Sodani
2nd Generation Intel® Xeon Phi™ Products

Codename:

Knights Landing
36 Tiles connected by 2D Mesh Interconnect
36 Tiles connected by 2D Mesh Interconnect
**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

<table>
<thead>
<tr>
<th>Product</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>KNL Self-Boot (Baseline)</td>
<td></td>
</tr>
<tr>
<td>KNL Self-Boot w/ Fabric</td>
<td>Fabric Integrated</td>
</tr>
<tr>
<td>KNL Card</td>
<td>PCIe-Card</td>
</tr>
</tbody>
</table>

Potential future options subject to change without notice.
All timelines, features, products and dates are preliminary forecasts and subject to change without further notification.
Measurements on a pre-production Knights Landing (A0) processor. Results subject to change on production parts.

** Early A0 processor runs, estimated SPEC (not fully compliant nor submitted runs)

++ Code and measurement done by colfaxresearch.com

Data courtesy of Intel Corporation
I say… "Wow!"
512-bit vectors via AVX-512 means: High performance and Binary compatibility

**AVX-512, High Bandwidth Memory, Cluster Mode, Omni-Path**

**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:** Caching/Home Agent. Distributed Tag Directory to keep L2s coherent. MESIF protocol. 2D-Mesh connections for Tile
Support of vectors under 512-bits in size means:

**Binary compatibility**

**AVX-512, High Bandwidth Memory, Cluster Mode, Omni-Path**

**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**
16MB on package DRAM means:

Performance with options

AVX-512, **High Bandwidth Memory**, Cluster Mode, Omni-Path

Memory Modes

**Three Modes. Selected at boot**

---

**Cache Mode**
- 16GB MCDRAM
- DDR

**Flat Mode**
- 16GB MCDRAM
- DDR

**Hybrid Mode**
- 8 or 12GB MCDRAM

- Physical Address

---

- 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
AVX-512, High Bandwidth Memory, Cluster Mode, Omni-Path

Standard "NUMA" node recognition by BIOS, OS, and applications.

Systems will eventually have this as a "norm."

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 two methods:

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

Flat MCDRAM with existing NUMA support in Legacy OS
AVX-512, **High Bandwidth Memory**, Cluster Mode, Omni-Path

### C/C++

“high bandwidth” malloc or new

### Fortran

“high bandwidth” allocatables

#### Flat MCDRAM SW Usage: Code Snippets

<table>
<thead>
<tr>
<th>C/C++</th>
<th>Intel Fortran</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Allocate into DDR</strong></td>
<td>Allocate into MCDRAM</td>
</tr>
</tbody>
</table>
| float *fv;  
fv = (float *)malloc(sizeof(float)*100); | c Declare arrays to be dynamic REAL, ALLOCATABLE :: A(:)  
!DEC$ ATTRIBUTES, FASTMEM :: A  
NSIZE=1024 |
| **Allocate into MCDRAM** | c allocate array ‘A’ from MCDRAM  
ALLOCATE (A(1:NSIZE)) |
| float *fv;  
fv = (float *)hbw_malloc(sizeof(float) * 100); | |
Mesh interconnect means:
Higher Performance with options

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
All-to-all to use all cores together uniformly for instance – OpenMP across all of KNL

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.

**Typical Read L2 miss**
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: Sub-NUMA Clustering (SNC)

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.

1) L2 miss, 2) Directory access, 3) Memory access, 4) Data return
Integrated Fabric means Lower Cost, power; Higher Density; Plus the Advancements of Omni-Path in latency, bandwidth & scaling

**Integrated Fabric**

**Lower Cost, power; Higher Density; Plus the Advancements of Omni-Path in latency, bandwidth & scaling**

---

**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
Intel® Parallel Studio XE 2017 Beta

Submitted by RAVI (Intel) on March 28, 2016

Contents

- How to enroll in the Beta program
- What’s New in the 2017 Beta
- Frequently Asked Questions
- Beta duration and schedule
- Support
- Beta webinars
- Beta Release Notes
- Known issues
- Next steps

How to enroll in the Beta program

Complete the pre-beta survey at registration link


BETA for “2017” Product – NOW
Many factors impact achieving good vectorization for our applications. The Vectorization Advisor directly analyzes an application and provides feedback on the extent of current vectorization and on possible steps to achieve more effective vectorization. Vectorization Advisor works with any compiler although some features in the Intel® compilers will increase the effectiveness of advice from the Vectorization Advisor tool. It is like having an expert sitting next to us who never tires of digging into an application to analyze what is really happening.

The Vectorization Advisor is one of the two major workflows (feature sets) available in the Intel® Advisor “2016” and later versions. The Intel Advisor also includes a thread prototyping feature set which can be useful for analysis of scaling for threads. In this chapter, we focus on using the Vectorization Advisor to help us maximize our vectorization performance.

How close is my application to maximum performance? Insight into this is helped by a “roofline model” analysis, in the Advisor Roofline Report section.
Can recommend:

- AoS to SoA
- AoSoA
- Use of SDLT
- Use of MCDRAM
Mask Utilization and FLOPS Profiler

Mask-aware:
- FLOPs Report
- Vector Efficiency
- Memory Access Pattern,
  (coming soon): Roofline Analysis Graph

Vectorization efficiency and FLOP/s in Survey Report and Loop Analytics.
...supplements AI-based analysis with a dynamic FLOP/s profile and peak FLOPs and memory sub-system throughput levels providing enlightening “bounds and bottlenecks” analysis for complex workloads.
Intel® Distribution for Python*

- Faster NumPy/SciPy performance powered by Intel® Math Kernel Library and Intel® Threading Building Blocks and Intel® Data Analytics Acceleration Library
- ~3X speedups on single thread
- Easy installation
- Python 2.7 & 3.5
- Windows & Linux & OS X

Some Performance Benchmarks
Intel® Distribution for Python*

Python* Performance Boost on Select Numerical Functions
Intel® Distribution for Python (Technical Preview) vs. Ubuntu Python

<table>
<thead>
<tr>
<th>Function</th>
<th>Speedup</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVD</td>
<td>1</td>
</tr>
<tr>
<td>Cholesky</td>
<td>22X</td>
</tr>
<tr>
<td>QR</td>
<td>61X</td>
</tr>
<tr>
<td>Inv</td>
<td>70X</td>
</tr>
<tr>
<td>DGEMM</td>
<td>97X</td>
</tr>
</tbody>
</table>

Configuration Info: - Versions: Intel® Distribution for Python 2.7.10 Technical Preview 1 (Aug 03, 2015), Ubuntu* built Python*: Python 2.7.10, NumPy 1.9.2 built with gcc 4.8.4; Hardware: Intel® Xeon® CPU E5-2668 v3 @ 2.10GHz (2 sockets, 16 cores each, HT=OFH), 64GB or RAM & LIMMS of 667MHz; Operating System: Ubuntu 14.04 LTS.

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 brands and names are the property of their respective owners. Benchmark Source: Intel Corporation

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.
#ModernCode

Optimizations for Intel® Xeon® and Intel® Xeon Phi™ products share the same:

- Languages
- Directives
- Libraries
- Tools
XeonPhiDeveloper.com

You can buy your very own Knight Landing development system today!

Highly-Parallel Performance to develop on

All the Software Tools & Libraries you need

Support & Training to help you succeed

Leading edge platform capabilities, performance to deliver multi-threaded, vectorized software for today’s HPC workloads!
Q&A

Thank you!