Usage models and APIs for Intel Knights Landing In-Package Memory

Jeff Hammond
Intel Parallel Computing Lab
(slides content borrowed shamelessly from others)

Legal Disclaimer & Optimization Notice

INFORMATION IN THIS DOCUMENT IS PROVIDED IN CONNECTION WITH INTEL PRODUCTS. NO LICENSE, EXPRESS OR IMPLIED, BY ESTOPPEL OR OTHERWISE, TO ANY INTELLECTUAL PROPERTY RIGHTS IS GRANTED BY THIS DOCUMENT. EXCEPT AS PROVIDED IN INTEL'S TERMS AND CONDITIONS OF SALE FOR SUCH PRODUCTS, INTEL ASSUMES NO LIABILITY WHATSOEVER AND INTEL DISCLAIMS ANY EXPRESS OR IMPLIED WARRANTY, RELATING TO SALE AND/OR USE OF INTEL PRODUCTS INCLUDING LIABILITY OR WARRANTIES RELATING TO FITNESS FOR A PARTICULAR PURPOSE, MERCHANTABILITY, OR INFRINGEMENT OF ANY PATENT, COPYRIGHT OR OTHER INTELLECTUAL PROPERTY RIGHT.

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

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

Extreme Scalability Group Disclaimer

- I work in Intel Labs ~~and therefore don't know anything about Intel products~~ *and leverage my colleagues for accurate information about Intel products.*
- ~~I am not an official spokesman for Intel.~~
- ~~I do not speak for my collaborators, whether they be inside or outside Intel.~~ *I have done my best to capture the expertise of my colleagues.*
- ~~You may or may not be able to reproduce any performance numbers I report.~~ *All performance data comes from the product owners.*
- Hanlon's Razor (blame Jeff's stupidity, not malice).

Hardware details

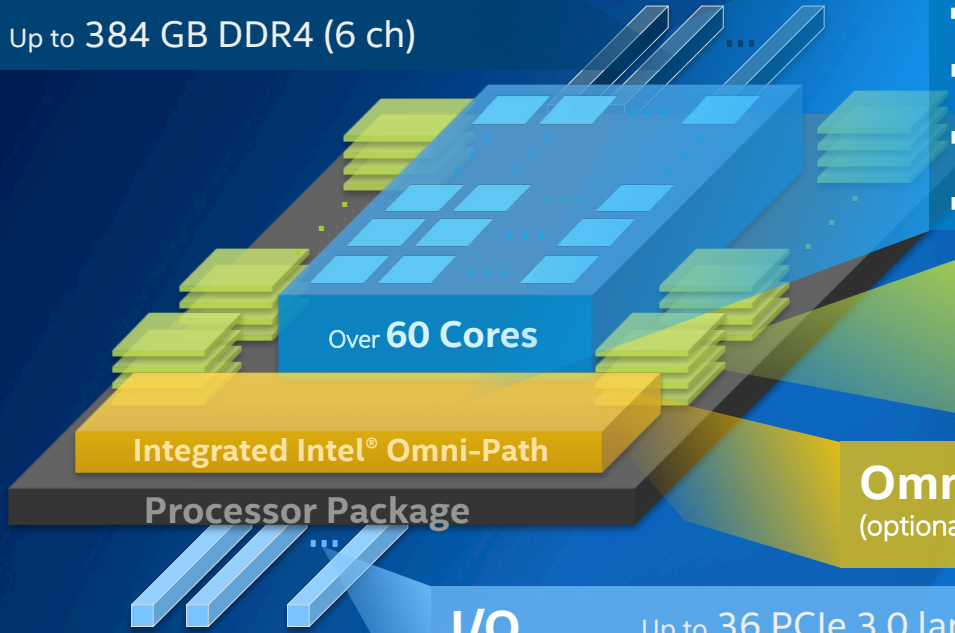
Knights Landing

Holistic Approach to Real Application Breakthroughs

intel inside XEON PHI

Platform Memory

Up to 384 GB DDR4 (6 ch)



Compute

- Intel® Xeon® Processor Binary-Compatible
- **3+ TF**LOPS[1], **3X ST**[2] perf. vs Xeon Phi™ coprocessor
- 2D Mesh Architecture
- Out-of-Order Cores

On-Package Memory

- Over 5x STREAM vs. DDR4[3]
- Up to **16 GB** at launch

Omni-Path (optional)

- **1st** Intel processor to integrate

I/O

Up to 36 PCIe 3.0 lanes

Software and workloads used in performance tests may have been optimized for performance only on Intel microprocessors.
Performance tests 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.

intel

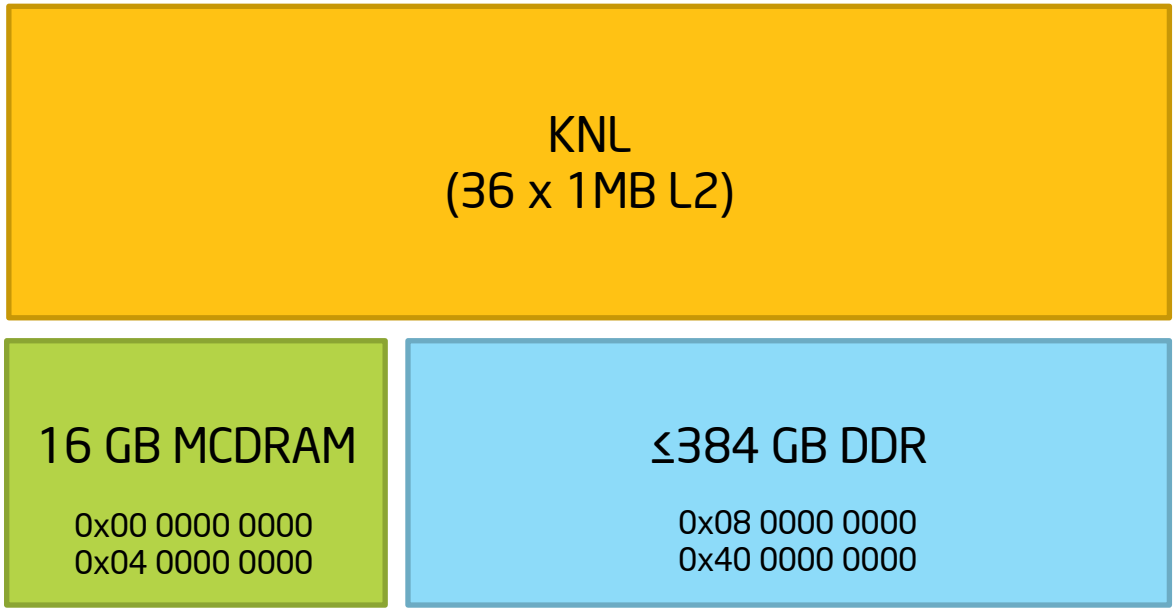
Knights Landing – memory information

- aka ‘high-bandwidth memory’, MCDRAM
  - Partnership with Micron Technology
- Bandwidth STREAM Triad (GB/s)
  - MCDRAM: 400+ GB/s
  - DDR: 90+ GB/s
- Capacity
  - MCDRAM: 16 GB
  - DDR4: 6 channels @ 2400 MHz up to 384GB

https://software.intel.com/en-us/articles/what-disclosures-has-intel-made-about-knights-landing

Memory modes

Flat Mode



Pro

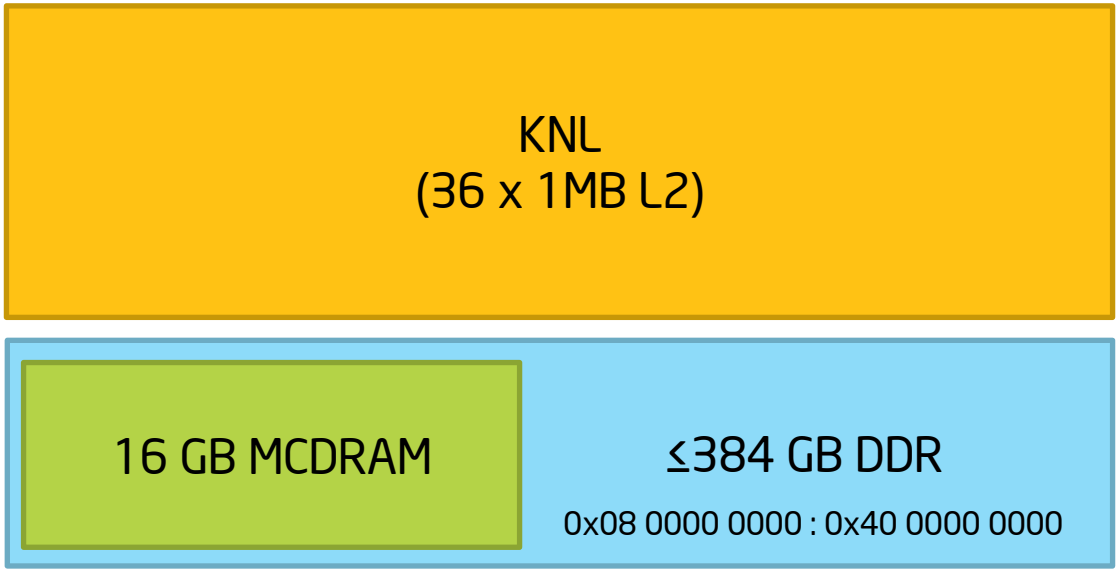
- *Maximum bandwidth and latency performance.*
- Maximum addressable memory.
- No MCDRAM pollution with non-critical data.

Con

- *Software modifications required.*
- Choosing MCDRAM vs DDR hinders application flexibility.

Addresses are given for illustration only. Do not interpret literally.

Cache Mode



Pro

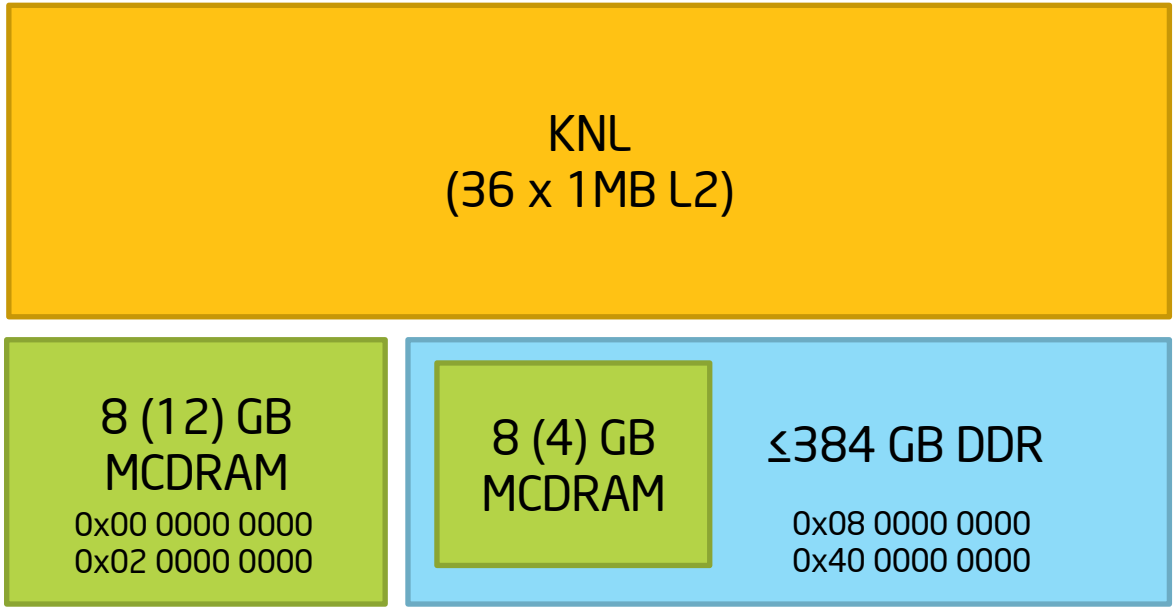
- *No software changes required.*
- Bandwidth benefit (relative to DDR-only).

Con

- *Latency hit to DDR.*
- *Limited sustained bandwidth.*
- All memory is transferred DDR -> MCDRAM -> L2.
- Less addressable memory.

Addresses are given for illustration only. Do not interpret literally.

Hybrid Mode

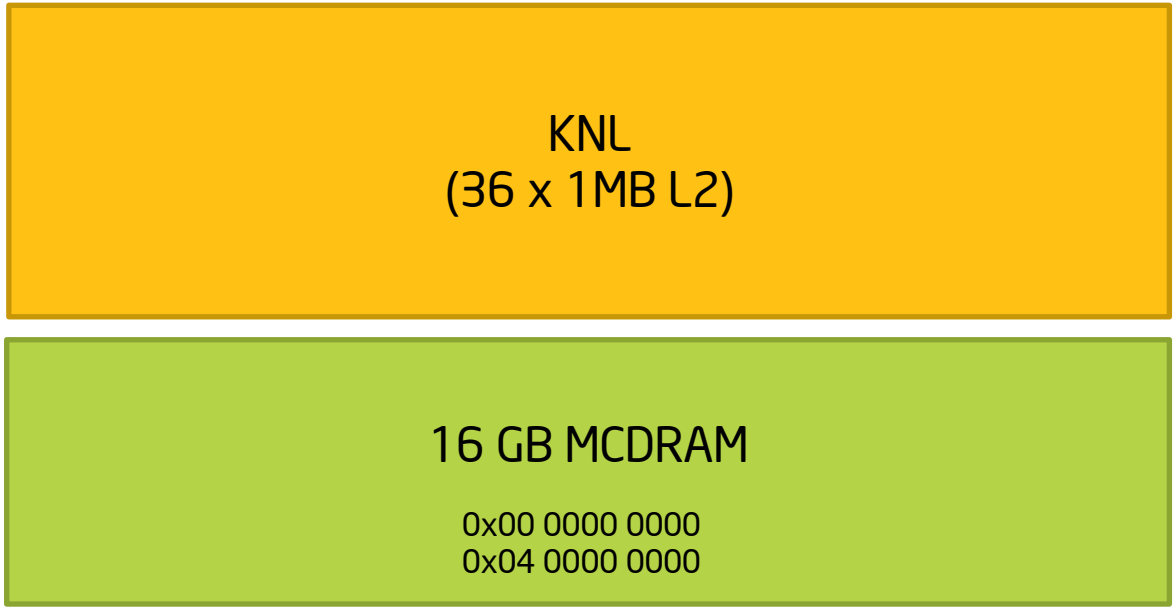


Pro and Con are a linear combination of flat and cache mode.

Addresses are given for illustration only. Do not interpret literally.

MCDRAM-only (“lean mode”*)

* This is my name for this talk. Don’t expect anyone else at Intel to understand this term.



Pro

- *Maximum bandwidth and latency performance.*
- *No software changes required.*

Con

- *Memory capacity is limited.*

Addresses are given for illustration only. Do not interpret literally.

Choosing the right mode

	DDR-only	Lean	Cache	Flat	Hybrid
Software changes	None	None	None	More data structures	Fewer data structures
Performance	Latency	Bandwidth, Latency	Bandwidth, ~Latency	Bandwidth, Latency	Bandwidth, ~Latency
Capacity	DDR	MCDRAM	DDR	DDR+ MCDRAM	DDR+ %MCDRAM

~latency: DDR latency include cost of cache miss, unlike flat mode. Cache hit sees MCDRAM latency.

Software enabling

How to manage two memories

- Easy model: two memories, two mallocs.
- Advanced model: memory management system that can manage any number of memories.

The advanced model is not just about KNL!

JEMALLOC+MEMKIND address a long-standing software gap. There is no good solution for managing multiple kinds of memory – DRAM, MMIO, RDMA slabs, symmetric heaps – in an OS/RT-agnostic way.

A Heterogeneous Memory Management Framework

MEMKIND

- Defines a plug-in architecture.
- Each plug-in is called a “kind” of memory.
- Built on top of jemalloc: the FreeBSD OS default heap manager.
- Partition is defined by functions that provide inputs for operating system calls.
- High level memory management functions can be over-ridden as well.
- https://github.com/memkind

HBWMALLOC

- Implements easy model for KNL.
- Implemented using memkind; simplifies plug-in (kind) selection.
- Provides support for 2MB and 1GB pages.
- Select fallback behavior when on package memory does not exist or is exhausted.
- Check for existence of on package memory.

MEMKIND details

http://memkind.github.io/memkind/ links to:

- memkind architecture document
- OpenSuSE RPMs
- Fedora packages
- Mailing list
- Git repos

memkind: An Extensible Heap Manager for Heterogeneous Memory Platforms and Mixed Memory Policies

Christopher Cantalupo
Intel Corporation
christopher.m.cantalupo@intel.com

Vishwanath Venkatesan
Intel Corporation
vishwanath.venkatesan@intel.com

Jeff R. Hammond
Intel Corporation
jeff.r.hammond@intel.com

Krzysztof Czuryło
Intel Corporation
krzysztof.czurylo@intel.com

Simon Hammond
Sandia National Laboratories
sdhammo@sandia.gov

No Hammonds were harmed in the writing of this code!
Chris, Vish, Krzysztof and others at Intel deserve the credit.

NAME

hbwmalloc - The high bandwidth memory interface

SYNOPSIS

```
#include <hbwmalloc.h>
...
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);
void hbw_set_policy(int mode);
```

> man https://github.com/memkind/memkind/blob/dev/man/hbwmalloc.3

NAME

memkind - Heap manager that enables allocations to memory with different properties.

SYNOPSIS

```
#include <memkind.h>
...
HEAP MANAGEMENT:
void *memkind_malloc(memkind_t kind, size_t size);
void *memkind_calloc(memkind_t kind, size_t num, size_t size);
void *memkind_realloc(memkind_t kind, void *ptr, size_t size);
int memkind_posix_memalign(memkind_t kind, void **memptr, size_t alignment, size_t size);
void memkind_free(memkind_t kind, void *ptr);

ALLOCATOR CALLBACK FUNCTION:
void *memkind_partition_mmap(int partition, void *addr, size_t size);
...
```

> man https://github.com/memkind/memkind/blob/dev/man/memkind.3

End Goal Usage: Code Snippets

Heap allocation in C

```
float * fv1 = malloc(sizeof(float) * 1000);
float * fv2 = hbw_malloc(sizeof(float) * 1000);
```

Allocatable arrays in Fortran

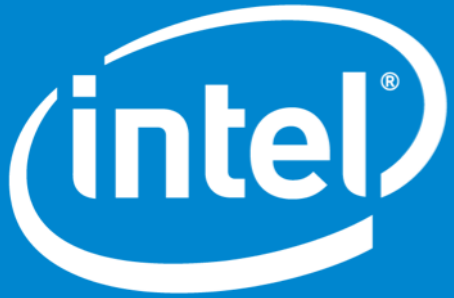
```
    REAL, ALLOCATABLE :: A(:), B(:), C(:)
!DIR$ ATTRIBUTES FASTMEM :: A
    NSIZE=1000
!   allocate array 'A' from MCDRAM
    ALLOCATE (A(1:NSIZE))
!   Allocate arrays that will come from DDR
    ALLOCATE  (B(1:NSIZE), C(1:NSIZE))
```

Automatic variables will be allocated in DDR in flat mode.

This means you may need to convert from automatic to heap arrays or use hybrid mode if such data is used in a bandwidth-intensive way.

Standard containers in C++ (not documented upstream yet)

```
std::vector<float, hbwmalloc::hbwmalloc_allocator<float> > vec;
```


intel

Legal Disclaimers

INFORMATION IN THIS DOCUMENT IS PROVIDED IN CONNECTION WITH INTEL PRODUCTS. NO LICENSE, EXPRESS OR IMPLIED, BY ESTOPPEL OR OTHERWISE, TO ANY INTELLECTUAL PROPERTY RIGHTS IS GRANTED BY THIS DOCUMENT. EXCEPT AS PROVIDED IN INTEL'S TERMS AND CONDITIONS OF SALE FOR SUCH PRODUCTS, INTEL ASSUMES NO LIABILITY WHATSOEVER AND INTEL DISCLAIMS ANY EXPRESS OR IMPLIED WARRANTY, RELATING TO SALE AND/OR USE OF INTEL PRODUCTS INCLUDING LIABILITY OR WARRANTIES RELATING TO FITNESS FOR A PARTICULAR PURPOSE, MERCHANTABILITY, OR INFRINGEMENT OF ANY PATENT, COPYRIGHT OR OTHER INTELLECTUAL PROPERTY RIGHT.

A "Mission Critical Application" is any application in which failure of the Intel Product could result, directly or indirectly, in personal injury or death. SHOULD YOU PURCHASE OR USE INTEL'S PRODUCTS FOR ANY SUCH MISSION CRITICAL APPLICATION, YOU SHALL INDEMNIFY AND HOLD INTEL AND ITS SUBSIDIARIES, SUBCONTRACTORS AND AFFILIATES, AND THE DIRECTORS, OFFICERS, AND EMPLOYEES OF EACH, HARMLESS AGAINST ALL CLAIMS COSTS, DAMAGES, AND EXPENSES AND REASONABLE ATTORNEYS' FEES ARISING OUT OF, DIRECTLY OR INDIRECTLY, ANY CLAIM OF PRODUCT LIABILITY, PERSONAL INJURY, OR DEATH ARISING IN ANY WAY OUT OF SUCH MISSION CRITICAL APPLICATION, WHETHER OR NOT INTEL OR ITS SUBCONTRACTOR WAS NEGLIGENT IN THE DESIGN, MANUFACTURE, OR WARNING OF THE INTEL PRODUCT OR ANY OF ITS PARTS.

Intel may make changes to specifications and product descriptions at any time, without notice. Designers must not rely on the absence or characteristics of any features or instructions marked "reserved" or "undefined". Intel reserves these for future definition and shall have no responsibility whatsoever for conflicts or incompatibilities arising from future changes to them. The information here is subject to change without notice. Do not finalize a design with this information.

The products described in this document 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.

Contact your local Intel sales office or your distributor to obtain the latest specifications and before placing your product order.
Copies of documents which have an order number and are referenced in this document, or other Intel literature, may be obtained by calling 1-800-548-4725, or go to:
http://www.intel.com/design/literature.htm

Knights Landing and other code names featured are used internally within Intel to identify products that are in development and not yet publicly announced for release. Customers, licensees and other third parties are not authorized by Intel to use code names in advertising, promotion or marketing of any product or services and any such use of Intel's internal code names is at the sole risk of the user

Intel, Look Inside, Xeon, Intel Xeon Phi, Pentium, Cilk, VTune and the Intel logo are trademarks of Intel Corporation in the U.S. and other countries.

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

Legal Disclaimers

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 SSE3 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

Legal Disclaimers

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 information go to http://www.intel.com/performance.

Intel® Advanced Vector Extensions (Intel® AVX)* provides higher throughput to certain processor operations. Due to varying processor power characteristics, utilizing AVX instructions may cause a) some parts to operate at less than the rated frequency and b) some parts with Intel® Turbo Boost Technology 2.0 to not achieve any or maximum turbo frequencies. Performance varies depending on hardware, software, and system configuration and you can learn more at http://www.intel.com/go/turbo.

Estimated Results Benchmark Disclaimer:
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.

Software Source Code Disclaimer:
Any software source code reprinted in this document is furnished under a software license and may only be used or copied in accordance with the terms of that license.

Permission is hereby granted, free of charge, to any person obtaining a copy of this software and associated documentation files (the "Software"), to deal in the Software without restriction, including without limitation the rights to use, copy, modify, merge, publish, distribute, sublicense, and/or sell copies of the Software, and to permit persons to whom the Software is furnished to do so, subject to the following conditions:

THE SOFTWARE IS PROVIDED "AS IS", WITHOUT WARRANTY OF ANY KIND, EXPRESS OR IMPLIED, INCLUDING BUT NOT LIMITED TO THE WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE AND NONINFRINGEMENT. IN NO EVENT SHALL THE AUTHORS OR COPYRIGHT HOLDERS BE LIABLE FOR ANY CLAIM, DAMAGES OR OTHER LIABILITY, WHETHER IN AN ACTION OF CONTRACT, TORT OR OTHERWISE, ARISING FROM, OUT OF OR IN CONNECTION WITH THE SOFTWARE OR THE USE OR OTHER DEALINGS IN THE SOFTWARE.

Legal Disclaimers

The above statements and any others in this document that refer to plans and expectations for the third quarter, the year and the future are forward-looking statements that involve a number of risks and uncertainties. Words such as “anticipates,” “expects,” “intends,” “plans,” “believes,” “seeks,” “estimates,” “may,” “will,” “should” and their variations identify forward-looking statements. Statements that refer to or are based on projections, uncertain events or assumptions also identify forward-looking statements. Many factors could affect Intel’s actual results, and variances from Intel’s current expectations regarding such factors could cause actual results to differ materially from those expressed in these forward-looking statements. Intel presently considers the following to be the important factors that could cause actual results to differ materially from the company’s expectations. Demand could be different from Intel's expectations due to factors including changes in business and economic conditions; customer acceptance of Intel’s and competitors’ products; supply constraints and other disruptions affecting customers; changes in customer order patterns including order cancellations; and changes in the level of inventory at customers. Uncertainty in global economic and financial conditions poses a risk that consumers and businesses may defer purchases in response to negative financial events, which could negatively affect product demand and other related matters. Intel operates in intensely competitive industries that are characterized by a high percentage of costs that are fixed or difficult to reduce in the short term and product demand that is highly variable and difficult to forecast. Revenue and the gross margin percentage are affected by the timing of Intel product introductions and the demand for and market acceptance of Intel's products; actions taken by Intel's competitors, including product offerings and introductions, marketing programs and pricing pressures and Intel’s response to such actions; and Intel’s ability to respond quickly to technological developments and to incorporate new features into its products. The gross margin percentage could vary significantly from expectations based on capacity utilization; variations in inventory valuation, including variations related to the timing of qualifying products for sale; changes in revenue levels; segment product mix; the timing and execution of the manufacturing ramp and associated costs; start-up costs; excess or obsolete inventory; changes in unit costs; defects or disruptions in the supply of materials or resources; product manufacturing quality/yields; and impairments of long-lived assets, including manufacturing, assembly/test and intangible assets. Intel's results could be affected by adverse economic, social, political and physical/infrastructure conditions in countries where Intel, its customers or its suppliers operate, including military conflict and other security risks, natural disasters, infrastructure disruptions, health concerns and fluctuations in currency exchange rates. Expenses, particularly certain marketing and compensation expenses, as well as restructuring and asset impairment charges, vary depending on the level of demand for Intel's products and the level of revenue and profits. Intel’s results could be affected by the timing of closing of acquisitions and divestitures. Intel's results could be affected by adverse effects associated with product defects and errata (deviations from published specifications), and by litigation or regulatory matters involving intellectual property, stockholder, consumer, antitrust, disclosure and other issues, such as the litigation and regulatory matters described in Intel's SEC reports. An unfavorable ruling could include monetary damages or an injunction prohibiting Intel from manufacturing or selling one or more products, precluding particular business practices, impacting Intel’s ability to design its products, or requiring other remedies such as compulsory licensing of intellectual property. A detailed discussion of these and other factors that could affect Intel’s results is included in Intel’s SEC filings, including the company’s most recent reports on Form 10-Q, Form 10-K and earnings release.

Rev. 7/17/13