

Intended Audience: Software Developers

Interested in performance optimizing your application

- Don't need to be a performance expert
- But should be an expert in the application!

Working on a platform with an Intel® Xeon Phi™ code named Knights Landing

Using Intel® VTune™ Amplifier XE performance analyzer

- The performance information here applies to other tools (PTU, etc) but is focused on VTune Amplifier XE
- The last section of this guide also includes information about Intel® Advisor XE



How to Use this Presentation

Read through the slides once, then again while collecting data

Remember performance analysis is a process that may take several iterations

Software Optimization should begin *after you have*:

- Utilized any compiler optimization options (/O2, /QxAVX2, etc)
- Chosen an appropriate workload
- Measured baseline performance



Using Intel® VTune™ Amplifier XE to Tune Software on the Intel® Xeon Phi™ code named Knights Landing (KNL)

Software and Services Group

Ver. 1.1

Optimization Notice



Agenda

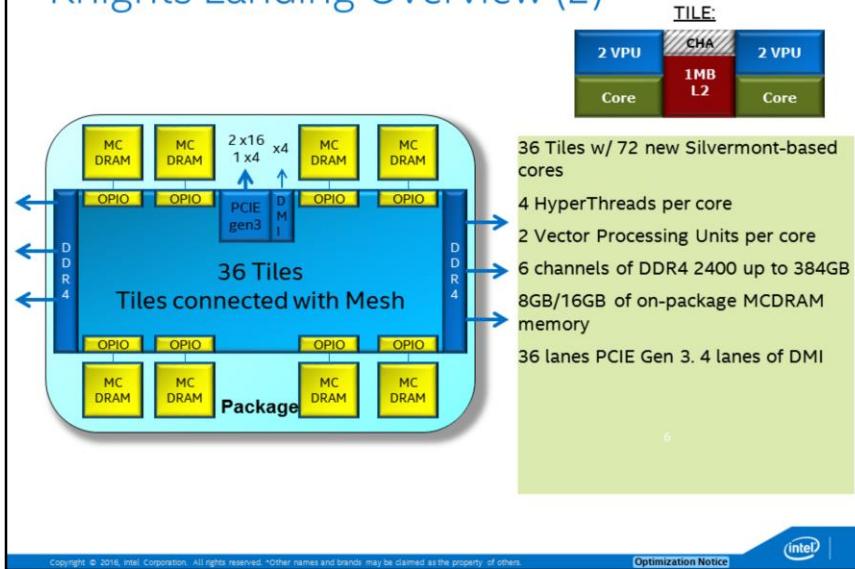
- Intel® Xeon Phi™ Code Named Knights Landing (KNL) Overview
- Intel® VTune™ Amplifier XE
- Software Optimization Steps
 - Profile resource utilization
 - Identify problematic symptoms
 - Locate issues and use recommendations to improve performance
- Additional Tuning Recommendations
- Intel® Advisor



Knights Landing Overview

- Knights Landing is the next Intel Many Core product after Knights Corner
- First self-boot Intel® Xeon Phi™ that is binary compatible with main line IA
- Significant leap in scalar and vector performance improvement over KNC
- Integration of memory on package: Innovative memory architecture for high bandwidth and high capacity
- Integration of fabric on package

Knights Landing Overview (2)



KNL is a highly-parallel architecture with large vector units. To get the most performance out of this platform, it is imperative to take advantage of these strengths.

KNL Tile:



Core: Changed from KNC to KNL. Based on Intel microarchitecture code named Silvermont (SLM) core – with many changes

Selected Important features of the Core

- Out of order 2-wide core: 72 inflight uops. 4 threads/core
- Back to back fetch and issue per thread
- 32KB Icache, 32KB Dcache. 2x 64B Loads ports in Dcache. Larger TLBs than in SLM
- L1 Prefetcher (IPP) and L2 Prefetcher.
- Fast unaligned and cache-line split support. Fast Gather/Scatter support
- 2x BW between Dcache and L2 than in SLM: 1 line Rd and $\frac{1}{2}$ line Wr per cycle

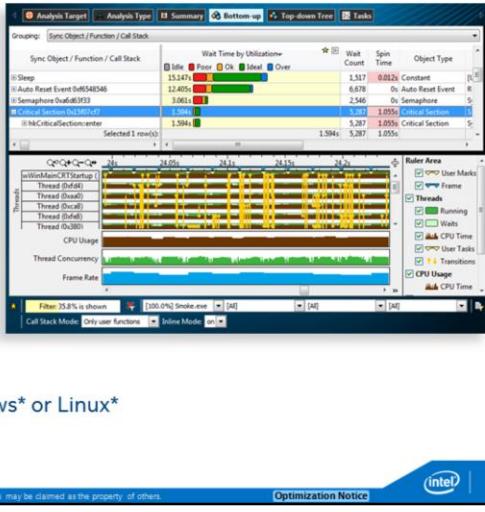
2 VPUs: 2x 512b Vectors. 32SP and 16DP. X87, SSE and EMU support

Intel® VTune™ Amplifier XE



VTune Amplifier XE features:

- Multiple Collection Types
 - Hotspots
 - Bandwidth
 - Event-based Sampling
- Timeline View Integrated into all Analysis Types
- Source/Assembly Viewing
- Compatible with C/C++, Fortran, Java, Assembly, .NET
- Visual Studio Integration, Command-line, or Standalone interface for Windows* or Linux*



Copyright © 2016, Intel Corporation. All rights reserved. *Other names and brands may be claimed as the property of others.

Optimization Notice



Most screenshots in this presentation were taken from Intel® VTune™ Amplifier XE 2016 Update 4. This is the first public version with KNL support.
Screenshots from different versions of the tool may have minor differences.

Running VTune Amplifier from the command-line

On self-boot KNL machines ensure the amplxe-cl command is installed. See the “amplxe-cl –help” command for complete details. To collect:

Hotspots:

```
amplxe-cl -collect advanced-hotspots -- myapp.out
```

General Exploration:

```
amplxe-cl -collect general-exploration -- myapp.out
```

Memory Access:

```
amplxe-cl -collect memory-access -- myapp.out
```

Copyright © 2016, Intel Corporation. All rights reserved. "Other names and brands may be claimed as the property of others.

Optimization Notice



Results will be created in a directory named r##ah, r##ge, or r##macc

Results can be viewed from the command-line or GUI on the KNL machine, but it is generally more efficient to copy results to another machine with the GUI installed for analysis.

It is also recommended to add the –no-auto-finalize flag to collections that will be creating large results. The finalization step is compute intensive and runs serially which may take a long time on the KNL. Finalization can be done on another machine after copying the results off of the KNL.

The data collected may be very large for longer runs with many threads active. If you find that you are reaching the data limit, use the flag -data-limit=<integer>. The default limit is 500MB. The integer specifies the size in MB. Use 0 for no limit.

Advanced Hotspots Analysis

- Supports OpenMP* analysis
- Stack-sampling is enabled. However, call counts and trip counts are not supported.

The screenshot shows the Intel VTune Amplifier XE 2017 interface with the 'Advanced Hotspots' viewpoint selected. The main panel displays performance metrics under the 'Elapsed Time' section:

Metric	Value
CPU Time	566.499s
Effective Time	195.181s
Spin Time	352.865s
Overhead Time	18.453s
Instructions Retired	320,824,000,000
CPI Rate	2.642
CPU Frequency Ratio	1.071
Total Thread Count	130
Paused Time	0s

A red box highlights the 'Total Thread Count' row (130) and contains the annotation: "Total Thread Count much higher on Intel® Xeon Phi™. Threading is vital for performance." An arrow points from the annotation to the 'Total Thread Count' value.

Copyright © 2016, Intel Corporation. All rights reserved. *Other names and brands may be claimed as the property of others.

Optimization Notice

Intel

Advanced Hotspots Analysis

Advanced OpenMP performance analysis

OpenMP Analysis. Collection Time: 4.506

OpenMP Region	OpenMP Potential Gain (%)	OpenMP Region Time
mainomp\$parallel@unknown:19:25	2.182s (48.4%)	3.399s

Top OpenMP Regions by Potential Gain

This section lists OpenMP regions with the highest potential for performance improvement. The Potential Gain metric shows the elapsed time that could be saved if the region was optimized to have no load imbalance assuming no runtime overhead.

Function	Module	CPU Time
_kmp_wai_template<kmp_flag_64>	libiomp5.so	282.763s
mainomp\$parallel_for@19	testout	186.521s
kmp_basic_flag<unsigned long long>_notdone_check	libiomp5.so	13.521s
_kmp_hierarchical_barrier_release	libiomp5.so	8.720s
_kmp_yield	libiomp5.so	7.838s
[Others]	N/A*	66.935s

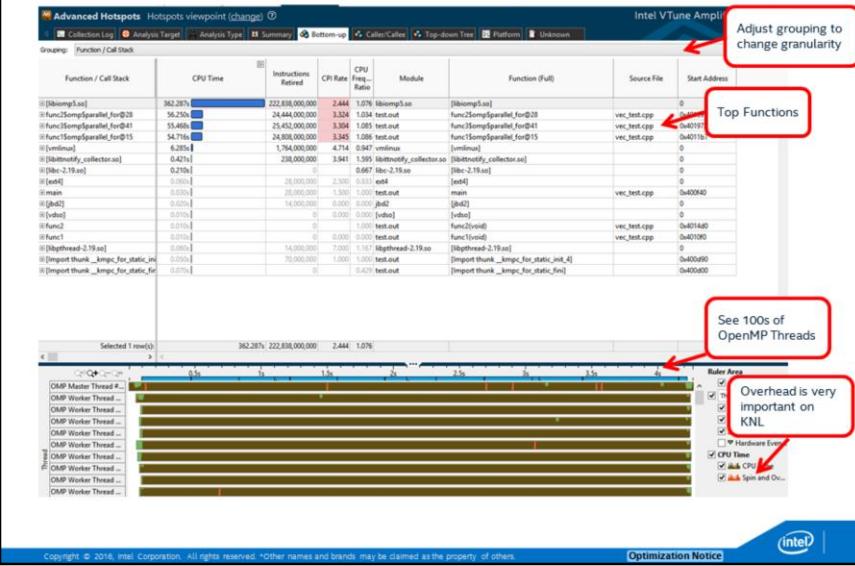
*NA is applied to non-summable metrics.

Optimization Notice

List of the hottest functions

The Advanced Hotspots Analysis will show where your application is spending its time, including information related to OpenMP parallelism. Ensure that the OpenMP runtime library used in the application (e.g. libiomp5.so) is available on the system doing the analysis. This is required to accurately analyze OpenMP overhead.

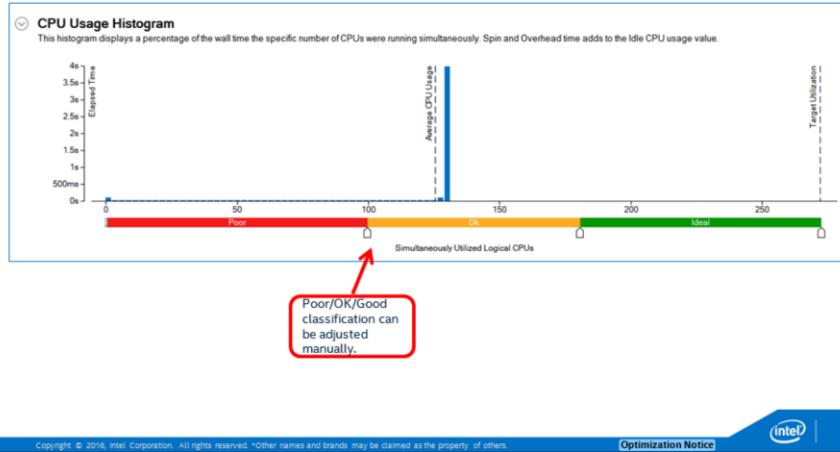
Advanced Hotspots Analysis



Use the Bottom-up view to see time spent at various granularities; for example Function or Module granularities. This can be changed in the Grouping drop-down menu. Focus tuning efforts on the hot portions of your application.

Profile Resource Utilization

Advanced Hotspots > Summary Tab



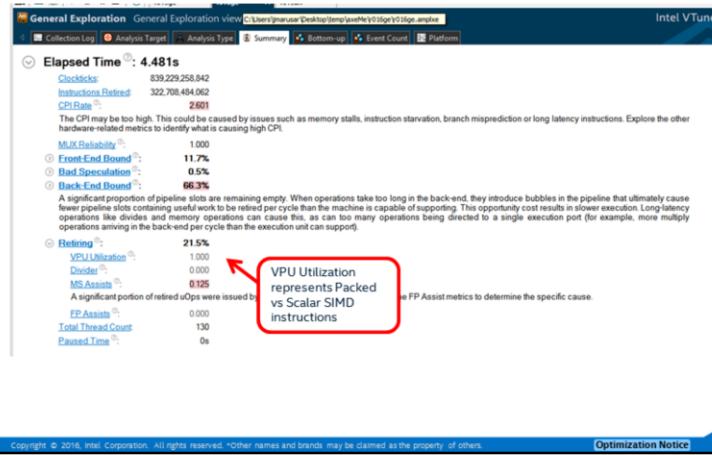
To get the best performance from KNL, it is important to have highly threaded parallel applications. The CPU Usage Histogram in the Summary shows how much time was spent with various numbers of logical cores active. As a general guideline, the vast majority of time should be spent with more than 50% of all available logical cores active. Because each KNL core has 4 HyperThreads, it isn't always beneficial to have all logical cores active if the bottleneck is the execution core, which is shared between HyperThreads. If memory accesses are the bottlenecks, more threads may alleviate the problem.

Memory Bandwidth may not be helped by more threads, but Memory Latency can.

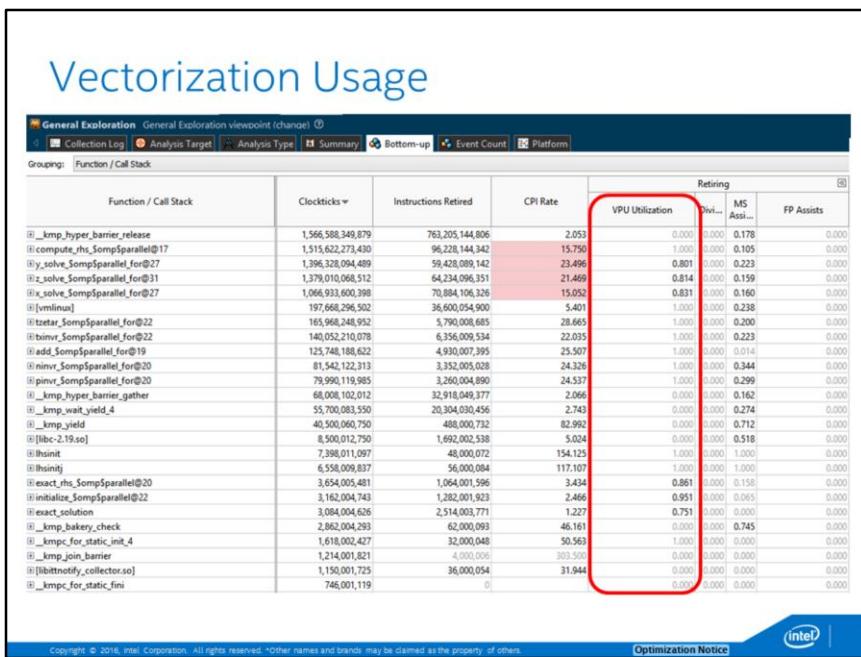
To identify memory latency as the issue look at L2 misses. If L2 misses are high and bandwidth is high, bandwidth may be the bottleneck. If L2 misses are high, but bandwidth is low, latency may be the issue, and more threads may help.

Vectorization Usage

General Exploration > Summary Tab



KNL supports 512 bit vector instructions. To optimize for KNL, an application should take advantage of these large vector units with heavily vectorized code. Look at the metric VPU Utilization to determine the areas of high and low vectorization in your application.



The VPU Utilization metric is also available in the Bottom-up view of the General Exploration viewpoint. Locate hotspots with low VPU Utilization and try to improve their usage of the AVX512 capabilities.

Identify the Hotspots

What: Hotspots are where your application spends the most time

Why: You should aim your optimization efforts there!

- Why improve a function that only takes 2% of your application's runtime?

How: VTune Amplifier XE Advanced Hotspots analysis type

- Usually hotspots are defined in terms of the CPU_CLK_UNHALTED.THREAD event (aka "clockticks")

Copyright © 2016, Intel Corporation. All rights reserved. *Other names and brands may be claimed as the property of others.

Optimization Notice



For this processor, the CPU_CLK_UNHALTED.THREAD counter measures unhalted clockticks on a per hardware thread basis. The CPU_CLK_UNHALTED.THREAD counter allows you to see where cycles are being spent on each individual hardware thread.

There is also a CPU_CLK_UNHALTED.REF counter, which counts unhalted clockticks per thread, at the reference frequency for the CPU. In other words, the CPU_CLK_UNHALTED.REF counter should not increase or decrease as a result of frequency changes due to throttling. This counter can be useful for removing the variance introduced due to throttling when comparing multiple analyses.

The “Software on Hardware” Tuning Process

For each Hotspot

- Determine efficiency
- If inefficient:
 - Determine primary bottleneck
 - Identify architectural reason for inefficiency
 - Optimize the issue

Repeat



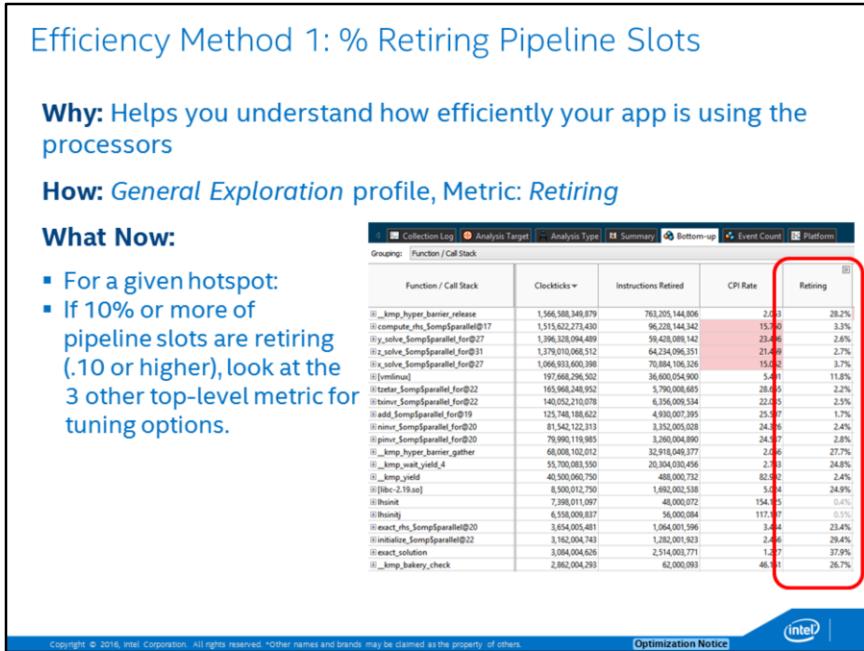
Efficiency Method 1: % Retiring Pipeline Slots

Why: Helps you understand how efficiently your app is using the processors

How: General Exploration profile, Metric: *Retiring*

What Now:

- For a given hotspot:
- If 10% or more of pipeline slots are retiring (.10 or higher), look at the 3 other top-level metric for tuning options.



Formula:

$$(\text{UOPS_RETIRED.ALL} / (2 \times \text{CPU_CLK_UNHALTED.THREAD}))$$

Thresholds: Investigate if -
% Retiring < .10

This metric is based on the fact that when operating at peak performance, the pipeline on this CPU should be able to retire 2 micro-operations per clock cycle (or "clocktick"). The formula looks at "slots" in the pipeline for each core, and sees if the slots are filled, and if so, whether they contained a micro-op that retired.

Efficiency Method 2: Changes in Cycles per Instruction (CPI)

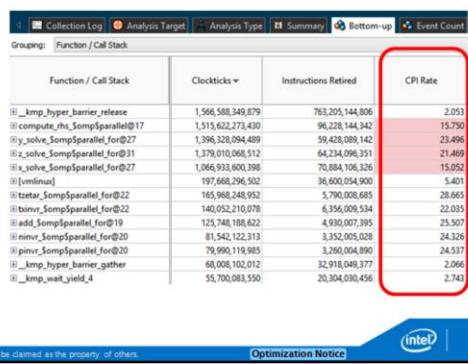
Why: Another measure of efficiency that can be useful when comparing 2 sets of data

- Shows average time it takes one of your workload's instructions to execute

How: General Exploration profile, Metric: CPI Rate

What Now:

- CPI can vary widely depending on the application and platform!
- If code size stays constant, optimizations should focus on reducing CPI



Formula:

CPU_CLK_UNHALTED.THREAD/INST_RETIRIED.ANY

Threshold:

In the interface, CPI will be highlighted if it is greater than 6. This is a very general rule based on the fact that many tuned applications should be able to get below this threshold. However, many applications will naturally have a CPI of over 6 – it is very dependent on workload and platform. It is best used as a comparison factor – know your app's CPI and see if over time it is moving upward (that is bad) or reducing (good!).

Note that CPI is a ratio! Cycles per instruction. So if the code size changes for a binary, CPI will change. In general, if CPI reduces as a result of optimizations, that is good, and if it increases, that is bad. However there are exceptions. Some code can have a very low CPI but still be inefficient because more instructions are executed than are needed.

Additionally, CPI can be affected if using Intel® Hyper-threading. In a serial workload, or a workload with Intel® Hyper-threading disabled the theoretical best CPI on a hardware thread is 0.5 because the core can allocate and retire 2 instructions per cycle. In a workload with Intel® Hyper-threading enabled which utilizes all 4 hardware threads effectively, the ideal CPI per-thread would be 2 instead of 0.5. This is because the hardware threads share allocation and

retirement resources on the core.

Note: Optimized code (e.g. with AVX512 instructions) may actually increase the CPI, and increase stall % – but improve the performance. This is because a single vector instruction will generally take more cycles than a single scalar instruction, but it also often performs more work. For example, a vector instruction may take twice as many cycles, but perform the work of four scalar instructions. In that case, the average CPI will increase, but the application will still be running faster.

CPI is just a general efficiency metric – the real measure of efficiency is work taking less time.

The “Software on Hardware” Tuning Process

For each Hotspot

- Determine efficiency
 - If inefficient:
 - Determine primary bottleneck
 - Identify architectural reason for inefficiency
 - Optimize the issue

Repeat



Determine the Primary Bottleneck

If Methods 1 or 2 are used to determine code is inefficient, first determine the primary bottleneck.

The Top-Down hierarchy implemented in General Exploration classifies your application's utilization of the CPU cores into 4 categories:

- Front-End Bound
- Back-End Bound
- Bad Speculation
- Retiring

The primary bottleneck has the highest fraction of pipeline slots, and should be investigated first!

Copyright © 2016, Intel Corporation. All rights reserved. *Other names and brands may be claimed as the property of others.

Optimization Notice



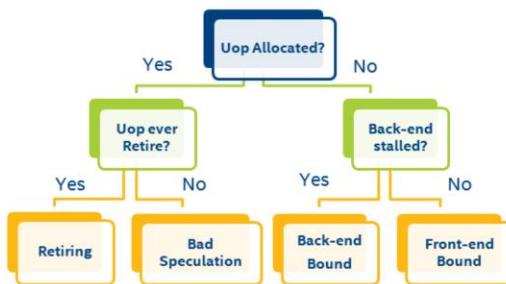
For a hotspot that is inefficient, determining the primary bottleneck is the first step. Optimizing code to fix issues outside the primary bottleneck category may not boost performance – the biggest boost will come from resolving the biggest bottleneck. Generally, if Retiring is the primary bottleneck, that is good. See next slides.

Issue Classification

A Pipeline Slot is an abstract concept – it represents the hardware resources needed to process one micro-operation

On this CPU, there are 2 pipeline slots available on each core, each cycle

Performance is classified according to what happened for each slot available to the application or hotspot:



Copyright © 2016, Intel Corporation. All rights reserved. *Other names and brands may be claimed as the property of others.

Optimization Notice



Note the way that this methodology allows us to classify what percentage of all pipeline slots end up in each category, for each cycle and for each core. It is possible that for a given dataset, there may be a significant percentage of pipeline slots in multiple categories that merit investigation. Start with the category with the highest percentage of pipeline slots. Ideally a large percentage of slots will fall into the "Retiring" category, but even then, it may be possible to make your code more efficient.

The “Software on Hardware” Tuning Process

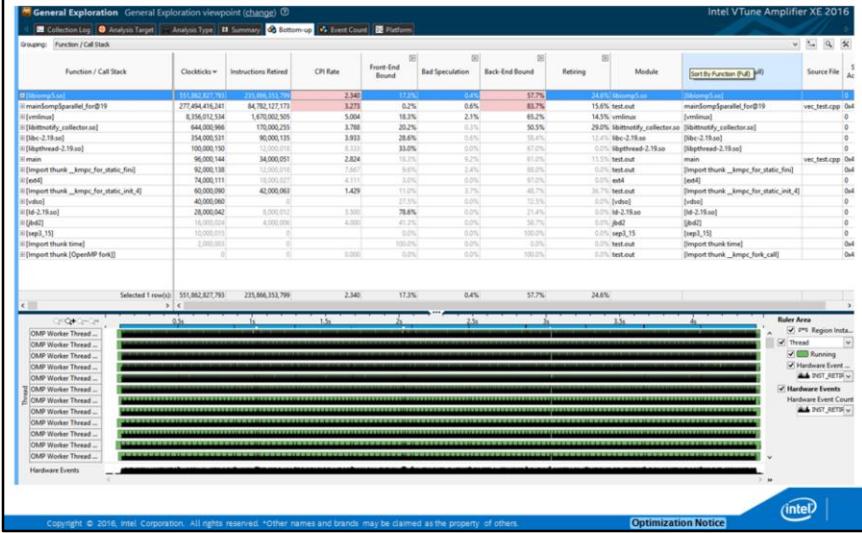
For each Hotspot

- Determine efficiency
 - If inefficient:
 - Determine primary bottleneck
 - Identify architectural reason for inefficiency
 - Optimize the issue

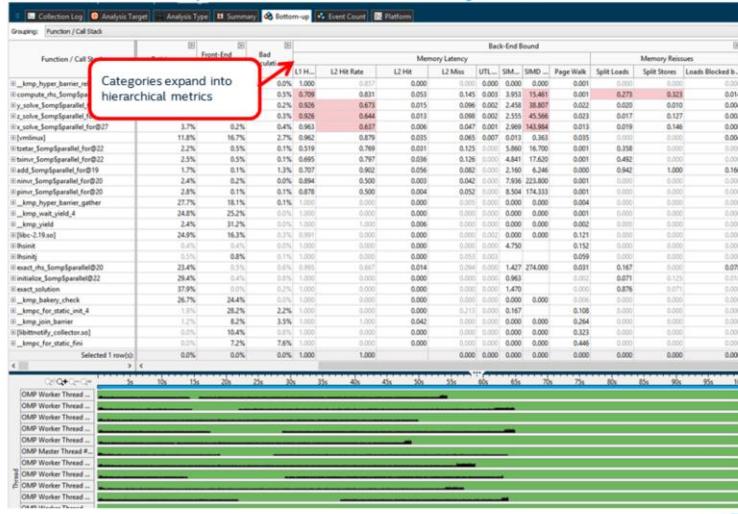
Repeat



General Exploration Analysis



General Exploration Analysis



Copyright © 2016, Intel Corporation. All rights reserved. *Other names and brands may be claimed as the property of others.



General Exploration Analysis

Front-End Bound

ICache Misses

Description:

Missing instruction fetches from the Instruction Cache (ICache) causes stalls in the pipeline. This may be the result of branch-heavy code or poor code layout by the compiler.

Formula:

$$\text{ICache Misses} = \frac{\text{FETCH_STALL.ICYCLE_FILL_PENDING_CYCLES}}{\text{INST_RETIRED.ANY}}$$

Threshold:

Investigate if > 0.1



General Exploration Analysis

Bad Speculation

Branch Mispredict

Description:

Mispredicting branch targets causes the processor to execute instructions that will never retire, because they are on the incorrect code path. This represents wasted work and should be minimized.

Formula:

Branch Mispredict =
$$(2 * \text{NO_ALLOC_CYCLES.MISPREDICTS}) / (2 * \text{CPU_CLK_UNHALTED.THREAD})$$

Threshold:

Investigate if > 0.05



General Exploration Analysis

Back-End Bound

L2 Hit Rate

Description:

The L2 is the last, and longest-latency, level in the memory hierarchy before DRAM or MCDRAM. This metric provides the ratio of demand load requests that hit in L2 to the total number of demand load requests serviced by L2. This metric does not include instruction fetches.

Formula:

$$\text{L2 Hit Rate} = \frac{\text{MEM_UOPS_RETIRED.L2_HIT_LOADS_PS}}{\text{MEM_UOPS_RETIRED.L2_HIT_LOADS_PS} + \text{MEM_UOPS_RETIRED.L2_MISS_LOADS_PS}}$$

Threshold:

Investigate if < 0.80



General Exploration Analysis

Back-End Bound

L2 Hit

Description:

The L2 is the last, and longest-latency, level in the memory hierarchy before DRAM or MCDRAM. While L2 hits are serviced much more quickly than hits in DRAM, they can still incur a significant performance penalty. This metric provides the ratio of cycles spent in servicing demand load requests that hit in L2 to the total number of cycles.

Formula:

$$\text{L2 Hit Penalty} = (17 * \text{MEM_UOPS_RETIRED.L2_HIT_LOADS_PS} / \text{CPU_CLK_UNHALTED.THREAD})$$

Threshold:

Investigate if > 0.10



General Exploration Analysis

Back-End Bound

L2 Miss

Description:

The L2 is the last and longest-latency level in the memory hierarchy before the main memory (DRAM) and MCDRAM. Any memory requests missing here must be serviced by either DRAM or MCDRAM, with significant latency. The L2 Miss metric shows ratio of cycles spent in servicing demand load requests that miss in L2 to the total number of cycles.

Formula:

L2 Miss Penalty =
$$(230 * \text{MEM_UOPS_RETIRED.L2_MISS_LOADS_PS} / \text{CPU_CLK_UNHALTED.THREAD})$$

Threshold:

Investigate if > 0.15



General Exploration Analysis

Retiring

VPU Utilization

Description:

This metric measures the fraction of micro-ops (uops) that performed packed vector operations of any vector length and any mask. VPU utilization metric can be in conjunction with the compiler's vectorization report to assess VPU utilization and to understand the compiler's judgement about the code. Note that this metric includes integer packed SIMD uops but does not account for loads and stores. Also, this metric does not take into consideration the uop masking behavior or vector length of the uops.

Formula:

$$\text{Vector VPU Compute Percentage} = \frac{(\text{UOPS_RETIRED.PACKED_SIMD})}{(\text{UOPS_RETIRED.PACKED_SIMD} + \text{UOPS_RETIRED.SCALAR_SIMD})}$$

Threshold:

Investigate if < 0.5



General Exploration Analysis

Retiring

Divider

Description:

Not all arithmetic operations take the same amount of time. Divides and square roots, both performed by the DIV unit, take considerably longer than integer or floating point addition, subtraction, or multiplication. This metric measures the fraction of total cycles when DIV unit was active. Note that this metric accounts only for the following division operations: integer div, x87 div, divss, divsd, sqrtss, sqrttsd.

Formula:

$$\text{Divider} = (\text{CYCLES_DIV_BUSY.ALL}) / (\text{CPU_CLK_UNHALTED.THREAD})$$

Threshold:

Investigate if > 0.05



General Exploration Analysis

Retiring

FP Assists

Description:

Certain floating point operations cannot be handled natively by the execution pipeline and must be performed by microcode (small programs injected into the execution stream). For example, when working with very small floating point values (so-called denormals), the floating-point units are not set up to perform these operations natively. Instead, a sequence of instructions to perform the computation on the denormal is injected into the pipeline. Since these microcode sequences might be hundreds of instructions long, these microcode assists are extremely detrimental to performance. This metric also accounts for other FP assists such as Flush-To-Zero (FTZ).

Formula:

$$\text{FP Assists} = (\text{MACHINE_CLEAR.S.FP_ASSIST}) / (\text{INST_RETIRED.ANY})$$

Threshold:

Investigate if > 0.05

Copyright © 2016, Intel Corporation. All rights reserved. *Other names and brands may be claimed as the property of others.

Optimization Notice



Additional Topic: Metric Reliability

Function / Call Stack	Clocktic...▼	Instructions Retired	CPI Rate	Filled Pipeline Slots		Unfilled Pipeline Slots (Stalls)	
				Retiring	Bad Speculation	Back-End Bound	Front-End Bound
grid_intersect	14,076,021,114	12,468,018,702	1.129	0.210	0.076	0.650	0.063
sphere_intersect	9,306,013,999	9,206,013,809	1.011	0.282	0.038	0.615	0.065
grid_bounds_intersect	1,098,001,647	690,001,035	1.591	0.123	0.020	0.781	0.075
func@0x1002e3d5	922,001,383	700,001,050	1.317	0.000	0.000	1.000	0.000
_kmp_x86_pause	354,000,531	212,000,318	1.670	0.000	0.000	1.000	0.000
tri_intersect	222,000,333	152,000,228	1.461	0.405	0.000	0.561	0.101
pos2grid	212,000,318	186,000,279	1.140	0.248	0.000	0.717	0.035
__main__	303,000,202	269,000,402	0.781	0.196	0.260	0.490	0.074
Selected 1 row(s):	14,076,021,114	12,468,018,702	1.129	0.210	0.076	0.650	0.063

Grayed out metric values represent low reliability of the metrics for each value in the grid.

Copyright © 2016, Intel Corporation. All rights reserved. *Other names and brands may be claimed as the property of others.

Optimization Notice

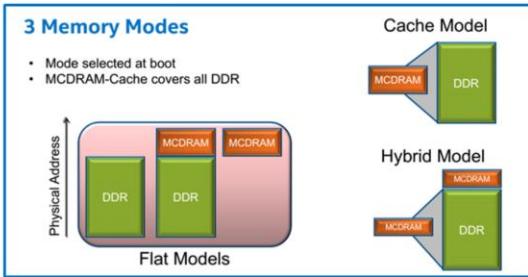


The General Exploration analysis type multiplexes hardware events during collection, which can result in imprecise results if too few samples are collected. The GUI will gray out metrics if the reliability is low based on the number of samples collected. If a metric is grayed out for your area of interest, consider increasing the runtime of the analysis or allowing multiple runs via the project properties.

Previous versions of the tool used a MUX Reliability metric for each row, however this was unable to distinguish between different metrics on the same row.

Memory Access Analysis

- Provides individual bandwidth information for both, MCDRAM and DDR.
- VTune cannot yet identify the system configuration: cluster mode and memory modes. Hence, shows the bandwidth information for both cache and flat mode. Users need to choose the correct data based on system configuration.



Copyright © 2016, Intel Corporation. All rights reserved. *Other names and brands may be claimed as the property of others.

Optimization Notice



Memory Access Analysis

- Flat Mode

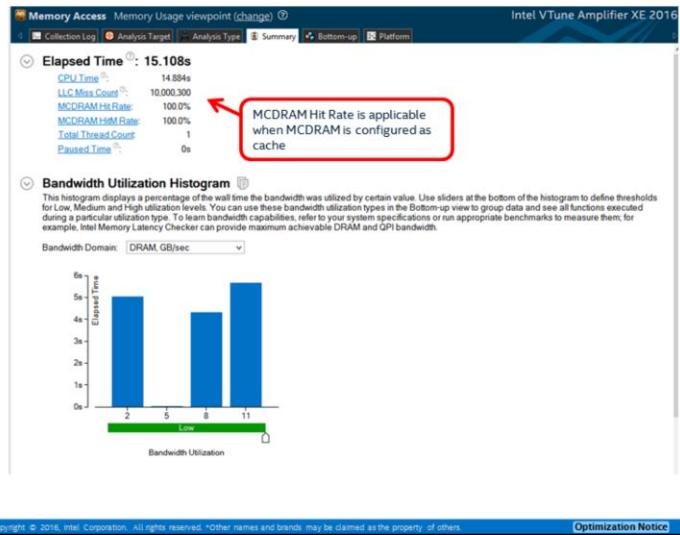
- Allows the developers to explicitly control which data structures are in MCDRAM vs. DDR.
- Requires code modification, otherwise DDR will be used by default.

- Cache Mode

- No code modification. Hardware will use L1, then L2, then MCDRAM cache.
- Data allocated into MCDRAM cache needs to be highly reusable to see performance benefits.
- Average L2 miss latency is higher because misses in MCDRAM cache then go to DDR
- May have aliasing issues if multiple pages are mapped into same cache lines in MCDRAM Cache. Non deterministic (not repeatable)
- Streaming stores are negatively impacted. Expect lower bandwidth



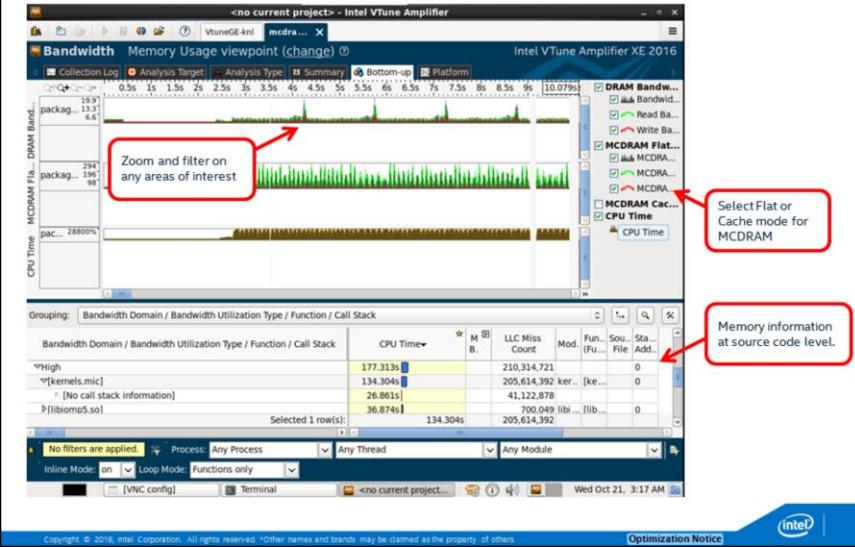
Memory Access Analysis (2)



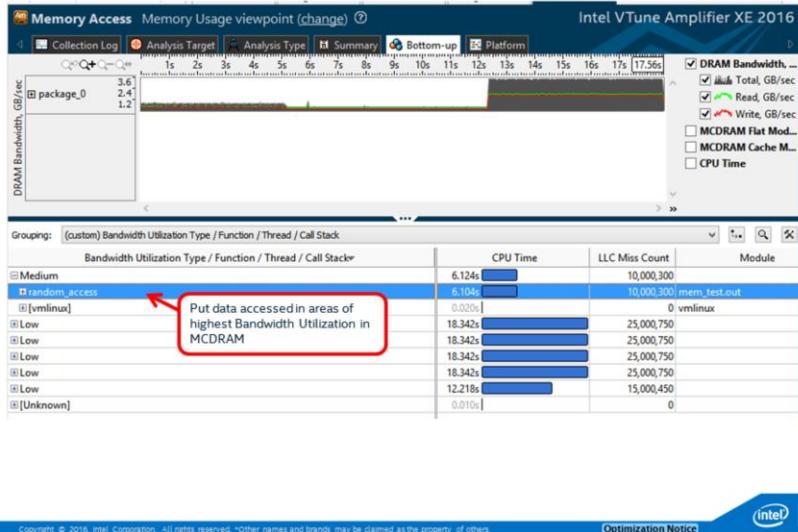
1. Notes about MCDRAM Hit Rate

1. This rate counts loads and streaming stores, e.g. `vmovnt` (non-temporal), but not stores/writebacks
2. If you have streaming stores, your MCDRAM Hit Rate may be lower than expected because streaming stores are expected to miss MCDRAM Cache

Memory Access Analysis (3)



Identifying Objects to put in MCDRAM



Run a Memory Access analysis with MCDRAM configures in flat mode and all allocations occurring in DDR (not using MCDRAM).

Create a custom grouping in the Memory Access analysis to see functions causing Medium or High bandwidth utilization. Objects accessed within these functions may be candidates to move into MCDRAM.

KNL Cluster Mode Performance Tuning

- Quadrant Cluster Mode
 - This configuration allows for an increase in usable bandwidth on the mesh because there is less traffic crossing quadrant boundaries. It is generally expected to offer better performance than all-to-all mode but does require that all DDR channels be populated identically.
- Sub-NUMA Cluster Mode (SNC4)
 - This mode is expected to be preferable when threads running on the chip can be grouped and affinitized to specific quadrants of tiles and they mostly access their own data. A single data structure or array will normally be mapped to only a pair of MCDRAM channels, or half of the DDR channels. Therefore, the accessible bandwidth to that structure will be less than what it would be in quadrant mode because it is not spread evenly across all channels. While this may seem undesirable, it is important to remember that if the chip is being used to run multiple MPI ranks, or multiple processes, then the total available bandwidth of the system is likely to be highest in this mode. Note that if you try and allocate more memory than is available in your local cluster, the additional memory will be allocated on another cluster. This is expected and does not cause an exception or some other error to occur.
- All-to-All Cluster Mode
 - This mode is rarely used for performance. This is the fallback mode in the event of system asymmetries or irregularities.



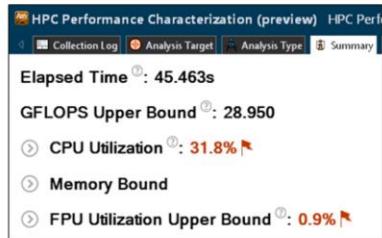
HPC Performance Characterization

Two characterization metrics

- Elapsed Time
- GFLOPs Upper Bound*

Three performance aspects

- CPU Utilization
- Memory Bound
- FPU Utilization Upper Bound*



*Calculated based on FLOP HW counters assuming full vector utilization

Copyright © 2016, Intel Corporation. All rights reserved. *Other names and brands may be claimed as the property of others.

Optimization Notice



HPC Performance Characterization provides performance information that is especially important for High Performance Computing (HPC) applications. This analysis can be run from the GUI or using the command line flag “–collect hpc-performance”

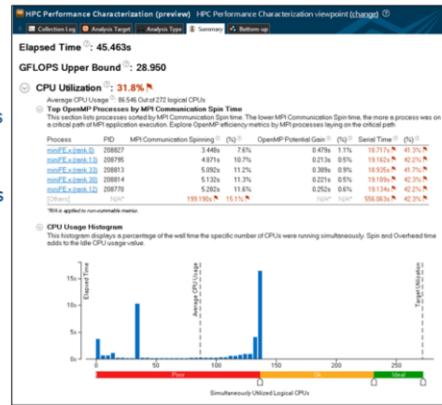
HPC Performance Characterization

CPU Utilization

- % of "Effective" CPU usage by the application under profiling (threshold 90%)
 - Under assumption that the app should use all available logical cores on a node
 - Subtracting spin/overhead time spent in MPI and threading runtimes

Metrics in CPU utilization section

- Average CPU usage
- Additional MPI and OpenMP scalability metrics impacting effective CPU utilization
- CPU usage histogram



Copyright © 2016, Intel Corporation. All rights reserved. *Other names and brands may be claimed as the property of others.

Optimization Notice

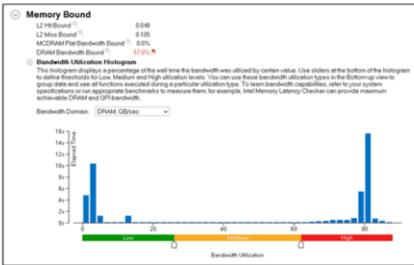


The CPU Utilization metrics provide another way to determine how busy all of the cores are during the performance analysis.

HPC Performance Characterization

Metrics in Memory Bound section

- L2 Hit Bound
 - Cost of L1 misses served in L2
- L2 Miss Bound
 - Cost of L2 misses
- MCDRAM Bandwidth Bound
 - % of app elapsed time consuming high MCDRAM Bandwidth
- MCDRAM Bandwidth Bound
 - % of app elapsed time consuming high MCDRAM Bandwidth
- Bandwidth utilization histogram



Copyright © 2016, Intel Corporation. All rights reserved. *Other names and brands may be claimed as the property of others.

Optimization Notice



The Memory Bound metrics provide information about how the application is utilizing, and possibly bottlenecked by, the memory subsystem. If issues are exposed here, the Memory Access analysis may provide even more detailed information.

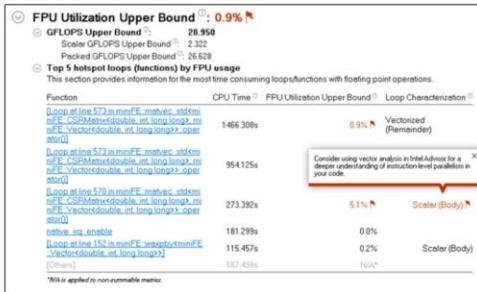
HPC Performance Characterization

FPU Utilization Upper Bound

- % of FPU load (100% when FPU is fully loaded, threshold 50%)

Metrics in FPU utilization section

- GFLOPs broken down by scalar and packed
- Top 5 loops/functions by FPU usage
 - Dynamically generated issue descriptions on low FPU usage help to define the cause and next steps



Copyright © 2016, Intel Corporation. All rights reserved. *Other names and brands may be claimed as the property of others.

Optimization Notice



The Floating Point Units (FPU) on KNL are important for getting the best performance. The HPC Performance Characterization provides an upper bound estimate of the utilization. This is an upper bound because the events used are not able to account for masking, and the metric assumes all vector lanes are used in each instruction.

VTune Amplifier Tips

- **VTune Finalization:**
 - Finalization is very slow on KNL. Finalize on Xeon.
 - Disable auto finalization with: -no-auto-finalize
- **Large amount of raw data collected:**
 - Appropriately select the app run duration using: -target-duration-type=<veryshort/short/medium/long>
 - Change the default data limit as required.
- **Power throttling:**
 - Keep an eye on the CPU frequency ratio. If this ratio changes significantly during the run then you might be seeing throttling or turbo effects.



VTune Amplifier Tips (cont.)

- Event multiplexing:
 - Similar to KNC, KNL has only 2 general purpose counters. Hence, when collecting a large number of events the data might be statistically invalid.
 - Try changing the target duration type or allow multiple runs.

Boost Vectorization with Intel® Advisor

Intel Advisor XE has a new feature to help analyze existing vectorization and guide you through improving vectorization use.

The screenshot displays the Intel Advisor XE interface with a blue header bar containing the Intel logo and the word "Advisor". Below the header, there are five numbered sections:

- 1. Compiler diagnostics + Performance Data + SIMD efficiency information**: Shows a table of compiler diagnostics with columns for Site, Total, and Time, along with SIMD efficiency information.
- 2. Guidance: detect problem and recommend how to fix it**: Displays a tooltip for a warning icon regarding "Nested Remainder loops present" and provides a recommendation to align memory access.
- 3. "Accurate" Trip Counts + FLOPs: understand utilization, parallelism granularity & overheads**: Shows trip counts for various loops and a table of problems and messages.
- 4. Loop-Carried Dependency Analysis**: A yellow box highlighting a section of the interface.
- 5. Memory Access Patterns Analysis**: Shows memory access patterns for sites, including site names, site functions, loop-carried dependencies, steady distribution, and access patterns.

At the bottom left, the text "Copyright © 2016, Intel Corporation. All rights reserved. *Other names and brands may be claimed as the property of others." is visible. At the bottom right, the text "Optimization Notice" is displayed next to the Intel logo.

Use this 5 step process to determine how well you are vectorizing and where you can improve.

Intel® Advisor is available at: <https://software.intel.com/en-us/intel-advisor-xe>

Survey Analysis

Function Call Sites and Loops

	Vector Issues	Vect... Efficiency	Gain...	VL...	Trails
[loop in s151, at lo...]		AVX 2.7%	7.76	8	Floating32
[loop in s152, at lo...]		AVX2 2.5%	7.71x	8	Floating32
[loop in s452, at lo...]	⚠️ Data type conversions present	AVX2 2.0%	7.71x	8	FMA; Type Con...
[loop in s413, at lo...]	⚠️ 1 Ineffective peeled/remainder	AVX2 1.6%	7.66x	4; 8	FMA
[loop in s273, at lo...]	⚠️ 3 Possible inefficient memory a...	AVX 2.0%	7.69x	8	FMA; Masked St...
[loop in s253, at lo...]	⚠️ 3 Possible inefficient memory a...	AVX 2.0%	7.69x	8	Blends; FMA
[loop in s255, at lo...]	⚠️ 2 Possible inefficient memory a...	AVX2 2.0%	7.30x	8	Blends; FMA
[loop in s271, at lo...]	⚠️ 2 Possible inefficient memory a...	AVX2 2.0%	7.23x	8	FMA
[loop in v7, at loop...]	⚠️ 3 Possible inefficient memory a...	AVX 2.0%	7.16x	4; 8	FMA; Masked St...
[loop in s274, at lo...]	⚠️ 3 Possible inefficient memory a...	AVX 2.0%	6.29x	8	Blends; FMA; M...
[loop in std::vector<...>]		AVX 2.1%	5.81x	8	Floating32
[loop in SET1D at m...]	⚠️ Data type conversions present	AVX2 2.0%	5.37x	8	Divisions; Type ...

Instruction Set Analysis

Vectorized **Not Vectorized**

Sort – Look at your hottest vectorized loops

Efficiency – use as a performance thermometer

Recommendations – get tips on how to improve performance

Issue: Assumed dependency present

All or some source loop iterations are not executing in the loop body. Improve performance by moving source code around.

Recommendation: Add data padding

The trip count is not a multiple of vector length. To fix: Do one of the following:

- Increase the size of objects and add iterations so the trip count is a multiple of vector length.
- Increase the size of static and automatic objects, and use a compiler option to add data padding

Windows* OS **Linux* OS**

/Ogpt-assume-safe-padding -Ogpt-assume-safe-padding

Note: These compiler options apply only to Intel® Many Integrated Core Architecture (Intel® MIC Arch).

When you use one of these compiler options, the compiler does not add any padding for static and aut application. To satisfy this assumption, you must increase the size of static and automatic objects in your code.

Optional: Specify the trip count, if it is not constant, using a directive: #pragma loop_count

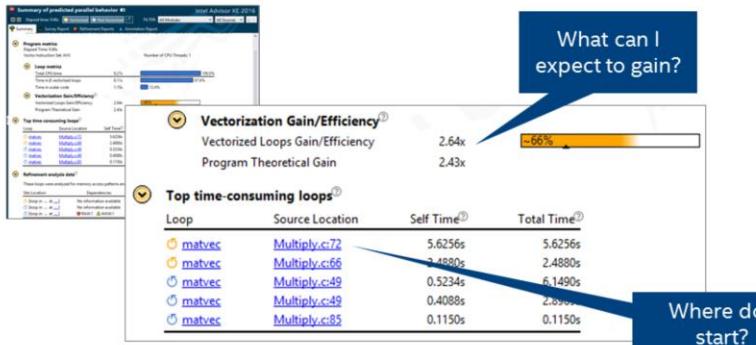
Read More: [#gpt-assume-safe-padding, #gpt-assume-safe-padding: loop_count](#)

Copyright © 2016, Intel Corporation. All rights reserved. *Other names and brands may be claimed as the property of others.

Optimization Notice

Intel

Summary View: Plan Your Next Steps



Copyright © 2016, Intel Corporation. All rights reserved. *Other names and brands may be claimed as the property of others.

Optimization Notice



49

Factors That Can Affect Efficiency

2.19x Vectorization Gain

1.A. Indirect memory access

```
for (int i=0; i<N; i++)
    A[B[i]] = C[i]*D[i]
```

1.B Memory sub-system Latency / Throughput

```
void scale(int *x, int *y)
{
    for (int i = 0; i < VEC0_SIZE; i++)
        x[i] = y * A[i][i];
    b[i] = y * A[i][i];
}
```

2.2 Check Memory Access Patterns

Command Line

3. Small trip counts that are not a multiple of the Vector Length

```
void doit(int *a, int *b, int unknown_size)
{
    for (int i = 0; i < unknown_size;
        i++)
        a[i] = i*b[i];
}
```

4. Branchy codes, outer vs. inner loops

```
for (i = 0; i < MAX; i++)
    if (D[i] < N)
        do_thing();
    else if (D[i] > N)
        do_that();
//...
```

5. MANY others: spill/fill, FP accuracy trade-offs, FMA, DIV/SQRT, Unrolling

Copyright © 2016, Intel Corporation. All rights reserved. *Other names and brands may be claimed as the property of others.

Optimization Notice

Analyze the hot loops for the common issues that can impact vectorization. Use the Memory Access Patterns Analysis and Recommendations to identify problematic behaviors and ways to correct them.

Check if it is Safe to Vectorize

Loop-Carried Dependencies Analysis Verifies Correctness

The screenshot shows the Intel Advisor XE 2016 interface. The main window displays a table of loop analysis results. A blue arrow points from the 'Select loop for Dependency Analysis and press play!' button to a specific row in the table, which is highlighted with a yellow background. Another blue arrow points from the 'Vector Dependence prevents Vectorization!' message to the same row.

Function Call Sites and Loops	Self Time	Total Time	Trip Counts	Compiler Vectorization	Loop Type	Why No Vectorization?
i: [loop at Multiply.c:53 in matvec]	0.047s	0.047s	1	0	Vectorized (Body)	Scalar
i: [loop at Multiply.c:53 in matvec]	0.413s	0.413s	101	0	Vectorized (Body)	Scalar
i: [loop at Multiply.c:53 in matvec]	0.030s	12.37s	1	0	Vectorized (Body)	Cellwise
i: [loop at Multiply.c:53 in matvec]	0.078s	11.93s	12	0	Vectorized (Body)	Scalar
i: [loop at Multiply.c:53 in matvec]	0.031s	0.444s	2	0	Vectorized (Body)	Remainder
[loop in Driver.c:148 in main]	0.016s	12.483s	1000000	0	Scalar	vector dependence prevents vectoriz...

Copyright © 2016, Intel Corporation. All rights reserved. *Other names and brands may be claimed as the property of others.

Optimization Notice

Intel | 51

Data dependencies between loop iterations make it difficult for the compiler to vectorize a loop. For example:

```
for (i = 1; i < N; i++) {  
    A[i] = A[i-1] + C[i];  
}
```

Each iteration is dependent on the value calculated in the previous iteration. Use Advisor to detect these dependencies.

Improve Vectorization

Memory Access Pattern Analysis

The screenshot shows the Intel Advisor interface with the following details:

Where should I add vectorization and/or threading parallelism?

Elapsed time: 8.52s

Function Call Sites and Loops

Loop	Type	Time	Reason
Loop at fractal.cpp:179 in <lambda>::op...	Vector	0.013s	12.020s
Loop at fractal.cpp:179 in <lambda>::op...	Pealed	0.000s	0.163s
Loop at fractal.cpp:179 in <lambda>::op...	Remainder	0.000s	0.576s
Loop at fractal.cpp:177 in <lambda>::oper...	Scalar	0.010s	12.030s

2.2 Check Memory Access Patterns
Identify and explore complex memory accesses for marked loops. Fix the reported problems.

Run Memory Access Patterns analysis to check how memory is used in the loop and the called function

Copyright © 2016, Intel Corporation. All rights reserved. *Other names and brands may be claimed as the property of others.

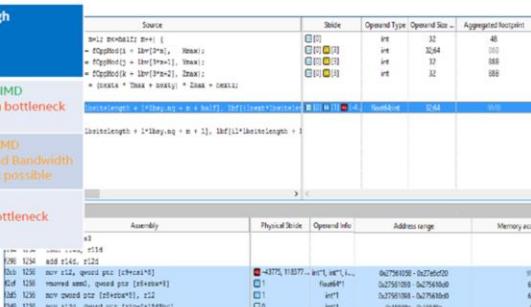
Optimization Notice | 52

The strides of memory accesses can affect vectorization. Determine the patterns to learn which loops may be difficult to vectorize.

Memory Analysis Is Critical

Determine Possible Bandwidth or Latency Issues

Footprint	Small enough	Big enough
Access Pattern		
Unit Stride	Effective SIMD No Latency and BW bottlenecks	Effective SIMD Bandwidth bottleneck
Const stride	Medium SIMD Latency bottleneck possible	Medium SIMD Latency and Bandwidth bottleneck possible
Irregular Access, Gather/Scatter	Bad SIMD Latency bottleneck possible	Bad SIMD Latency bottleneck



The screenshot shows the Intel Memory Advisor interface. It displays assembly code at the top, followed by memory access analysis. The analysis table includes columns for Source, Slice, Operand Type, Operand Size, and Aggregated footprint. Below this is a timeline showing memory operations over time. At the bottom, there are tabs for Assembly, Physical Drives, General Info, Address Range, and Memory Access.

Copyright © 2016, Intel Corporation. All rights reserved. *Other names and brands may be claimed as the property of others.

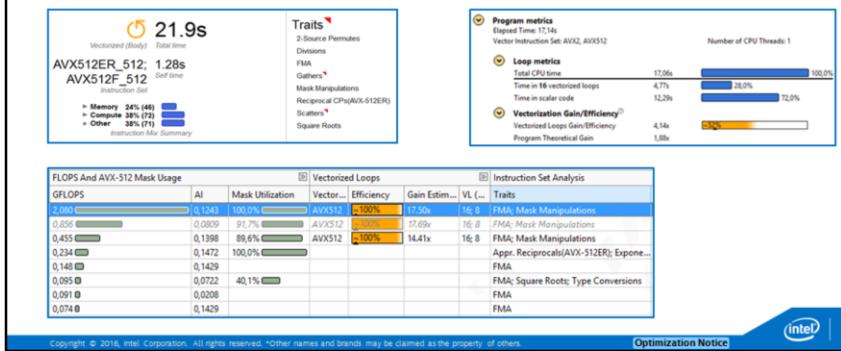
Optimization Notice



53

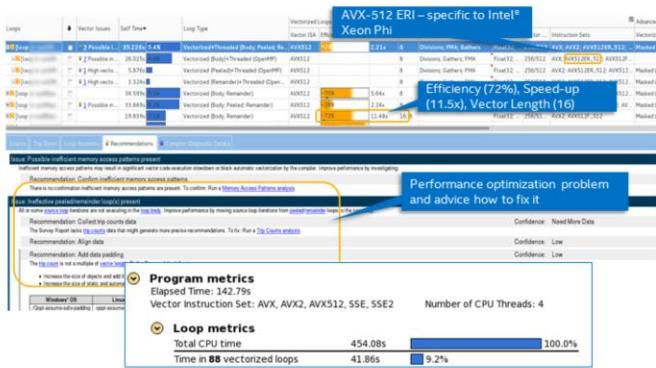
AVX-512 Specifics for KNL

1. Native AVX-512 profiling on KNL
2. Precise FLOPs and Mask Utilization profiler
3. AVX-512 Advice and Traits
4. AVX-512 Gather/Scatter Profiler



Vectorization Advisor on KNL AVX-512

See the Intel Advisor tutorials and documentation to learn how to analyze your KNL application.



Copyright © 2016, Intel Corporation. All rights reserved. *Other names and brands may be claimed as the property of others.

Optimization Notice



Good Luck! For more information:

VTune Amplifier XE Videos, Forums, and Resources:

<http://software.intel.com/en-us/intel-vtune-amplifier-xe/#pid-3659-760/>

Intel® 64 and IA-32 Architecture Software Developer's Manuals:

<http://www.intel.com/products/processor/manuals/index.htm>

VTune Amplifier XE Tuning Guides for Other microarchitectures:

<http://software.intel.com/en-us/articles/processor-specific-performance-analysis-papers>

Copyright © 2016, Intel Corporation. All rights reserved. *Other names and brands may be claimed as the property of others.

Optimization Notice



Legal Disclaimer

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 THE 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.

UNLESS OTHERWISE AGREED IN WRITING BY INTEL, THE INTEL PRODUCTS ARE NOT DESIGNED NOR INTENDED FOR ANY APPLICATION IN WHICH THE FAILURE OF THE INTEL PRODUCT COULD CREATE A SITUATION WHERE PERSONAL INJURY OR DEATH MAY OCCUR.

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 by visiting Intel's website.

Intel® Hyper-Threading Technology requires a computer system with a processor supporting HT Technology and an HT Technology-enabled chipset, BIOS and operating system. Performance will vary depending on the specific hardware and software you use. For more information including details on which processors support HT Technology, see [here](#).

Intel® Virtualization Technology requires a computer system with an enabled Intel® processor, BIOS, virtual machine monitor (VMM) and, for some uses, certain computer system software enabled for it. Functionality, performance or other benefits will vary depending on hardware and software configurations and may require a BIOS update. Software applications may not be compatible with all operating systems. Please check with your application vendor.

64-bit computing on Intel architecture requires a computer system with a processor, chipset, BIOS, operating system, device drivers and applications enabled for Intel® 64 architecture. Performance will vary depending on your hardware and software configurations. Consult with your system vendor for more information.

"Intel® Turbo Boost Technology requires a PC with a processor with Intel Turbo Boost Technology capability. Intel Turbo Boost Technology performance varies depending on hardware, software and overall system configuration. Check with your PC manufacturer on whether your system delivers Intel Turbo Boost Technology. For more information, see <http://www.intel.com/technology/turboboost>".

Intel, the Intel logo, Xeon, Xeon Inside, VTune, inTru, and Core are trademarks or registered trademarks of Intel Corporation or its subsidiaries in the United States and other countries.

*Other names and brands are the property of their respective owners.

Copyright © 2014, Intel Corporation

Copyright © 2014, Intel Corporation. All rights reserved. *Other names and brands may be claimed as the property of others.

Optimization Notice



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

Copyright © 2016, Intel Corporation. All rights reserved. *Other names and brands may be claimed as the property of others.

Optimization Notice



