Intel[®] oneAPI Performance Analysis Tools

Dmitry Tarakanov Technical Consulting Engineer



Notices & Disclaimers

Performance varies by use, configuration, and other factors. Learn more at <u>www.Intel.com/PerformanceIndex</u>.

Performance results are based on testing as of dates shown in configurations and may not reflect all publicly available updates. See configuration disclosure for details.

Your costs and results may vary.

Intel technologies may require enabled hardware, software or service activation.

© Intel Corporation. Intel, the Intel logo, and other Intel marks are trademarks of Intel Corporation or its subsidiaries. Other names and brands may be claimed as the property of others.

Analysis Tools Overview



Intel[®] VTune[™] Profiler Performance Profiler



Intel[®] Advisor Design and optimize vectorization, threading, accelerator offload and flow graphs.



Intel[®] Inspector Memory & Thread Debugger

Optimize Performance with Intel® VTune™ Profiler



Optimize Performance Intel® VTune™ Profiler

Get the Right Data to Find Bottlenecks

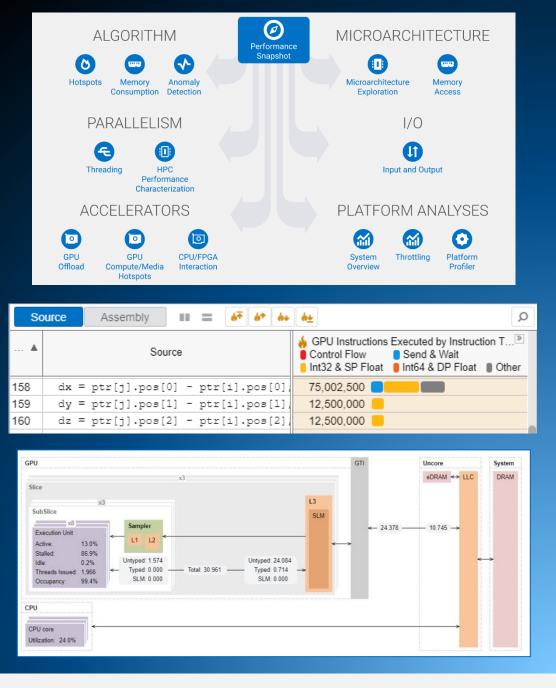
- A suite of profiling for CPU, GPU, FPGA, threading, memory, cache, storage, offload, power...
- DPC++, C, C++, Fortran, Python*, Go*, Java*, or a mix
- Linux, Windows, FreeBSD, Android, Yocto and more

Analyze Data Faster

- See data on your source, in architecture diagrams, as a histogram, on a timeline...
- Filter and organize data to find answers

Work Your Way

- User interface or command line
- Profile locally and remotely
- Install as an application
- Install as a server accessible with a web browser



Rich Set of Profiling Capabilities for Multiple Markets





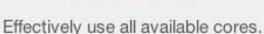
Single Thread

Optimize single-threaded performance.

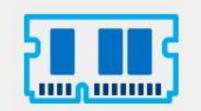


HPC & CLoud

Access specialized, in-depth analyses for HPC and cloud computing.



Multithreaded



Memory & Storage Management

Diagnose memory, storage, and data plane bottlenecks.





System

See a system-level view of application performance.



Analyze & Filter Data Mine data for answers.

Media & OpenCL[™] Applications

Deliver high-performance image and video processing pipelines.



Environment

Fits your environment and workflow.

Two Great Ways to Collect Data Intel® VTune™ Profiler

Software Collector	Hardware Collector
Uses OS interrupts	Uses the on-chip Performance Monitoring Unit (PMU)
Collects from a single process tree	Collect system wide or from a single process tree.
~10ms default resolution	~1ms default resolution (finer granularity - finds small functions)
Either an Intel [®] or a compatible processor	Requires a genuine Intel [®] processor for collection
Call stacks show calling sequence	Optionally collect call stacks
Works in virtual environments	Works in a VM only when supported by the VM
works in virtual environments	(e.g., vSphere*, KVM)
No driver required	Uses Intel driver or perf if driver not installed

No special recompiles - C, C++, DPC++, C#, Fortran, Java, Python, Assembly

Find Answers Fast

Intel[®] VTune[™] Profiler

Adjust Data Grouping

Function / Call Stack Source Function / Function / Call Stack Sync Object / Function / Call Stack Sync Object / Thread / Function / Call Stack ... (Partial list shown)

Double Click Function to View Source Click [▶] for Call Stack Filter by Timeline Selection

(or by Grid Selection)

Zoom In And Filter On Selection Filter In by Selection Remove All Filters

- 5	Analysis Configuration Grouping: Function / Call	Collectio Stack	on Log Summary	Bottom-up	Caller/Callee	Top-dow	n Tree Platf		• 🛠 🔉	
					CPU	Time				
	Function / Call Stack	E/	6	>	>		Ove	rhead Time		
	Tunction / Can Stack		Effective Time by Utilization ▼ Idle ■ Poor ■ Ok ■ Ideal ■ Over			Creation	Scheduling	Reduction	Atomics	
	<pre>grid_intersect</pre>	4.087s			0s	0s	0s	0s	0s	
	sphere_intersect	3.748s			0s	0s	Os	0s	0s	
	grid_intersect	3.748s			0s	0s	0s	0s	0s	
	intersect_objects	3.580s			0s	0s	0s	0s	0s	
	▶ < grid_intersect ← func@0x69e19df0	0.168s			2.021s	0s 0s	0s 0s	0s 0s	0s 0s	
Thread	OMP Worker Thread #	TID: 2 0 (TI	8.25 8.45 8.65 8.85		45 9.65 9.85	10s 10.2s			J Time n and Overho J Sample	e.
I	CPU Utilization							💽 📥 CPU		e.
	FILTER T 100.0%	×	Any Proce V Ar	Thread ▼	Any ModL 🔻	Any U ▼	User functi 🔻	Function •	Show in	li
ter by Process Tuning Opportunities Shown in Pink. Other Controls Hover for Tips										

See Profile Data On Source / Asm

Double Click from Grid or Timeline

View Source / Asm or both CPU Time Right click for instruction reference manual

	So	urce Assembly 🔳 = 👫 😽 🛻	Assembly grouping: Ad	ldress					
	🔺	Source	👍 CPU Time: Total 🛛 🗛 Address 🔺	Sour	Assembly	👍 CPU Time: 1			
	ა 65	curpos = nXp;	0.0028	0x40dcb9 0x40dcbf	570	<pre>pe ux4uaa49 <block 53=""> 3lock 44:</block></pre>	0.2475		
	500	-V +		0x40dcbf	572	est esi, esi	0.007s		
Quick Asm n	avi	aation.		0x40dcc1	572	z 0x40dd07 <block 50=""></block>			
	avi	gation.	0.004s	0x40dcc3		Block 45:			
Salact source	b + c	highlight Asm 🥄		0x40dcc3	573	mov eax, dword ptr [esi+0x4]	0.053s		
Select Source			0.265s 📒	0x40dcc6	573	mov ecx, dword ptr [edi+0x10]	0.750s		
	571	<pre>cur = g->cells[voxindex];</pre>		0x40dcc9	573	mov edx, dword ptr [edi+0xc]	0.020s		
	572	while (cur != NULL) {	0.007s	0x40dccc	573	mov eax, dword ptr [eax]	0.055s		
	573	if (ry->mbox[cur->obj->id] != r	2.058s	0x40dcce	573	<pre>cmp dword ptr [ecx+eax*4], ed</pre>	x 1.177s		
	574	ry->mbox[cur->obj->id] = ry->	0.604s 💼	0x40dcd1	573	<u>jz 0x40dce5 <block 48=""></block></u>	0.003s		
	575	cur->obj->methods->intersect(0.687s	0x40dcd3		Block 46:			
	576	}		0x40dcd3	574	mov dword pt [ecx+eax*4], ed	x 0.604s		
	577	<pre>cur = cur->next;</pre>	0.423s 🛑	0x40dcd6	575	mov eax, dwoil ptr [esi+0x4]	0.175s 📒		
	578	}		0x40dcd9	575	push edi	0.005s		
	579	<pre>curvox.z += step.z;</pre>	0.019s	0x40dcda	575	push eax	0.004s		
	580	if (ry->maxdist < tmax.z curvo	0.011s	0x40dcdb	575	mov ecx, dwor ptr [eax+0x8]	0.027s		
	581	break;		0x40dcde	575	mov eax, dwor ptr [ecx]	0.130s		
	582	<pre>voxindex += step.z*g->xsize*g->ys</pre>		0x40dce0	575	call eax	0.078s		
	- 00		0.000-1	0		51 1 47			

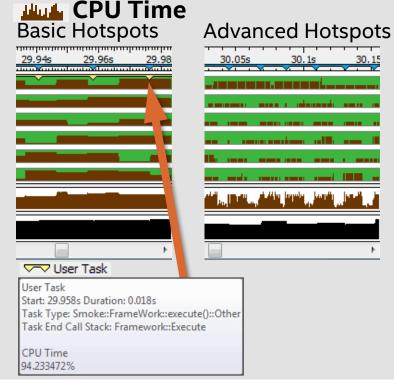
Scroll Bar "Heat Map" is an overview of hot spots

Click jump to scroll Asm

Timeline Visualizes Thread Behavior Intel[®] VTune[™] Profiler

Locks & Waits **** + Ruler Area Q9Q+Q-Q# 29.87s 29.88s 29.89s 29.86s Frame wWinMainCRTStartu... Thread Thread (0x1364) Running 🖸 Thread (0x136c) Vaits Thread (0x1374) User Task ✓ ↑↓ Transition Thread (0x137c) Thread Concurrency Thread (0x1384) Concurrency Thread Concurrency Frames over Time Hule Frame Rate Frames over Time » **P** Transition Frame Transition rame Hovers: Start: 29.858s Duration: 0.017s wWinMainCRTStartup (0x12d4) to Thread (0x138c) (29.899s to 29.899s) Frame: 72 Sync Object: TBB Scheduler Frame Domain: Smoke::Framework::execute() Object Creation File: taskmanagertbb.cpp Frame Type: Good Object Creation Line: 318 Frame Rate: 59.8242179

P Transitions



Optional: Use API to mark frames and user tasks rame volume User Task

Optional: Add a mark during collection





intel 10

Command Line Interface

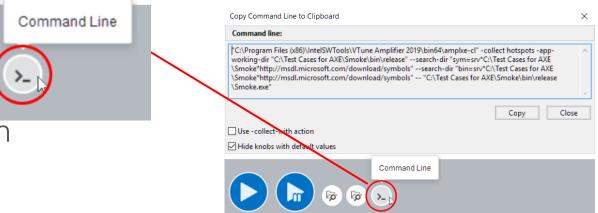
Automate analysis

Set up the environment variables:

- -Windows: <install-dir>\env\vars.bat
- -Linux: <install-dir>/env/vars.sh

Help: vtune -help

Use UI to setup 1) Configure analysis in UI 2) Press "Command Line..." button 3) Copy & paste command



Great for regression analysis – send results file to developer Command line results can also be opened in the UI

Default Intel® VTune™ Profiler Install Directories

In Intel[®] oneAPI Base Toolkit:

- Windows: [Program Files]\Intel\oneAPI\vtune\<version>
- Linux: /opt/intel/oneapi/vtune/<version>

Standalone:

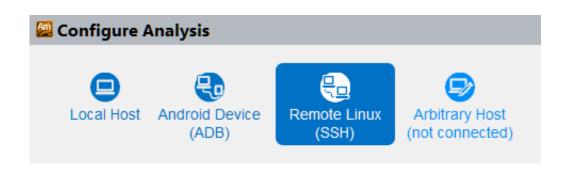
- Windows: [Program Files]\IntelSWTools\VTune Profiler <version>
- Linux: /opt/intel/vtune_profiler_version
- **On Apple* macOS* systems:**
- Applications/Intel VTune Profiler <version>.app

Interactive Remote Data Collection

Performance analysis of remote systems just got a lot easier

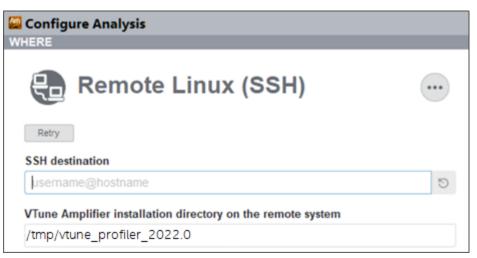
Interactive analysis

- 1) Configure SSH to a remote Linux* target
- 2) Choose and run analysis with the UI



Command line analysis

- 1) Run command line remotely on Windows* or Linux* target
- 2) Copy results back to host and open in UI



Conveniently use your local UI to analyze remote systems

Quickly Find Tuning Opportunities

Intel[®] VTune[™] Profiler

Faster, Scalable Code Faster

Get the Data You Need

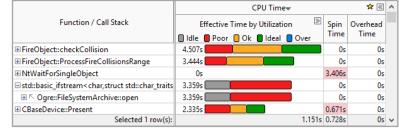
- Hotspot (Statistical call tree), Call counts (Statistical)
- Thread Profiling Concurrency and Lock & Waits Analysis
- Cache miss, Bandwidth analysis...¹
- GPU Offload and OpenCL[™] Kernel Tracing

Find Answers Fast

- View Results on the Source / Assembly
- OpenMP Scalability Analysis, Graphical Frame Analysis
- Filter Out Extraneous Data Organize Data with Viewpoints
- Visualize Thread & Task Activity on the Timeline

Easy to Use

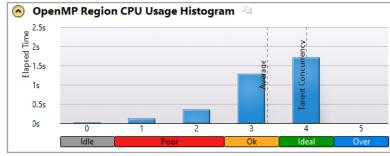
- No Special Compiles C, C++, C#, Fortran, Java, Python, ASM
- Visual Studio* Integration or Stand Alone
- Local & Remote Data Collection, Command Line
- Analyze Windows* & Linux* data on macOS



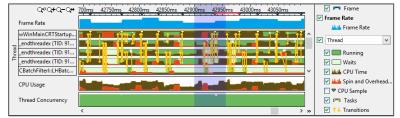
See Results On The Source Code

Source	Assembly 🗐 📰 🐼 🐼 🥙 🖉 🖳	Assembly grouping: Address
Source Line	Source	CPU Time: Total by Utilization
81	<pre>for (int i = 0; i < mem array i max; i++)</pre>	Idle Poor Ok Ideal Over 0.300s
82	{	
83	<pre>for (int j = 0; j < mem_array_j_max; j++)</pre>	4.936s
84	{	
85	<pre>mem_array [j*mem_array_j_max+i] = *fill_val</pre>	7.207s

Tune OpenMP Scalability



Visualize & Filter Data



Intel® VTune[™] Profiler GPU Profiling

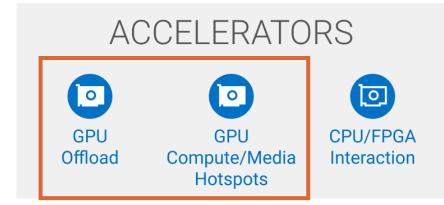
Two GPU Analysis types Intel® VTune[™] Profiler

GPU Offload: Is the offload efficient?

- Find inefficiencies in offload
- Identify if you are CPU or GPU bound
- Find the kernel to optimize first
- Correlate CPU and GPU activity
- Analyze DMA packet execution

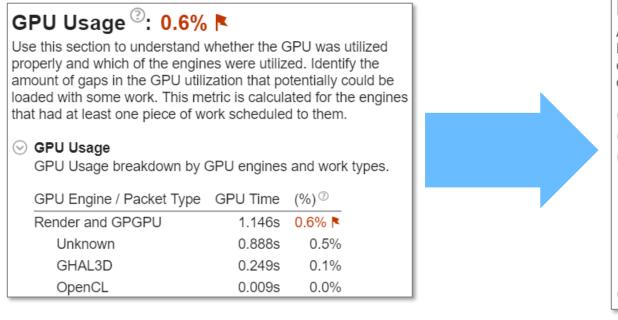
GPU Compute/Media Hotspots: Is the GPU kernel efficient?

- Identify what limits the performance of the kernel
- GPU source/instruction level profiling
- Find memory latency or inefficient kernel algorithms



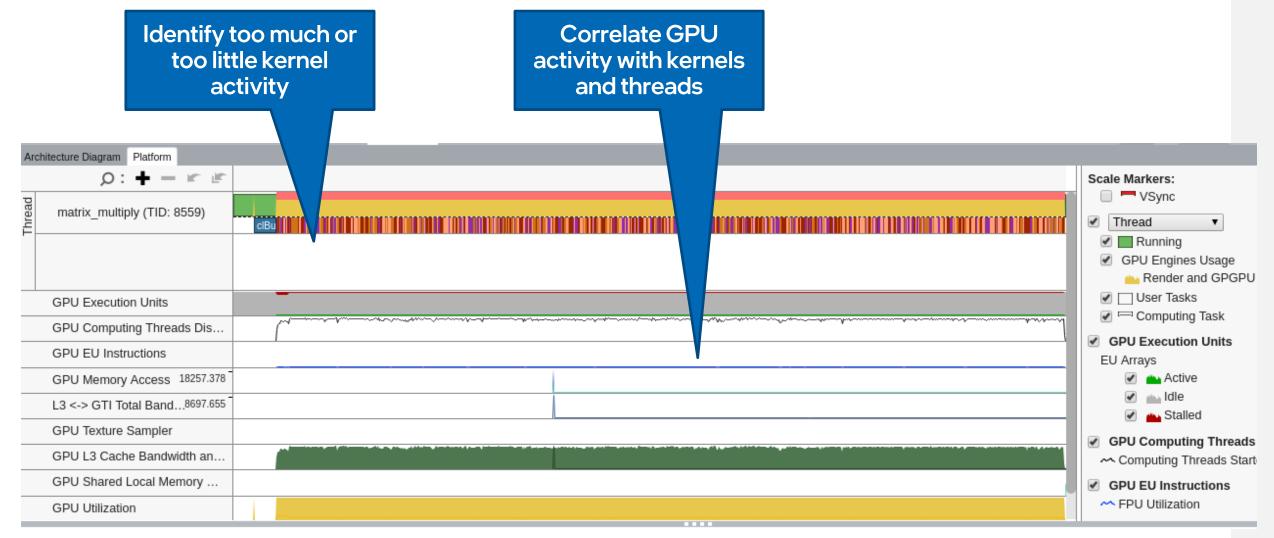
GPU Offload Profiling Intel® VTune™ Profiler

- Simply follow the sections on the Summary page
- Tuning methodology on top of HW metrics



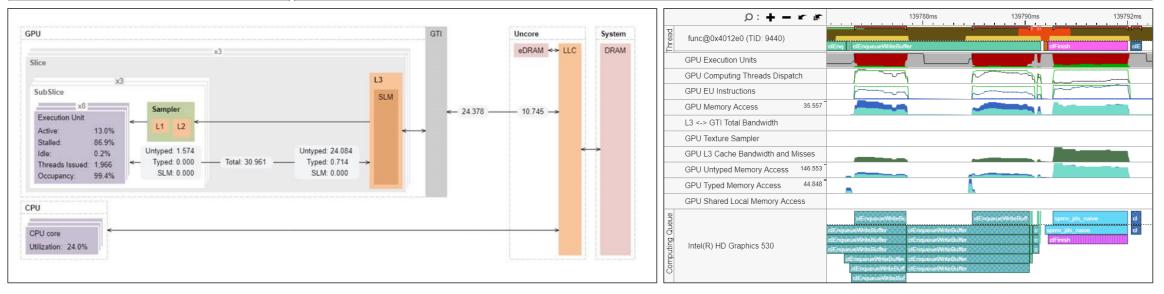
EU Array Stalled/Idle ⁽²⁾: 94.4% ► of Elapsed time Analyze the average value of EU Array Stalled/Idle metric and identify why EUs were waiting for resources instead of doing computations. This metric is critical for compute-bound applications. Explore typical reasons for this kind of inefficiency listed below. GPU L3 Bandwidth Bound ⁽²⁾: 0.5% of peak value DRAM Bandwidth Bound ⁽²⁾: 0.0% of Elapsed time Occupancy ⁽²⁾: 25.8% ► of peak value Identify too large or too small computing tasks with low occupancy that make the EU array idle while waiting for the scheduler. Note that frequent SLM accesses and barriers may affect the maximum possible occupancy. Hottest GPU Computing Tasks with Low Occupancy Sampler Busy ⁽²⁾: 40.6% of peak value

Timeline Correlates GPU and CPU Activity



GPU Hotspots: Aggregated and Overtime Views

Computing Task	Work	Size	Computing Task					
Computing Task	Global	Local	Total Time ▼	Average Time	Instance Count	SIMD Width		
clEnqueueWriteBuffer			0.005s	0.000s	14			
spmv_jds_naive	146944	256	0.003s	0.001s	2	16		
clEnqueueReadBuffer			0.000s	0.000s	2			
[Outside any task]			0s	0s	0			



GPU Compute/Media Hotspots

Tune Inefficient Kernel Algorithms

Analyze GPU Kernel Execution

- Find memory latency or inefficient kernel algorithms
- See the hotspot on the OpenCL[™] or DPC++ source & assembly code
- Analyze DMA packet execution
 - Packet Queue Depth histogram
 - Packet Duration histogram
- GPU-side call stacks

Source			Q
Source 🛦	Source	👌 Estimated GPU Cycles	^
256	#ifdef USE_IMAGE_STORAGE		
257	// Read the node information from the image		
258	<pre>const ushort inx = (nodeData >> 16) * 7;</pre>	0.2%	
259	const ushort iny = (nodeData & 0xffff);		
260	const float4 bboxes_minX = as_float4(read_:	0.8% 📒	
261	const float4 bboxes_maxX = as_float4(read_:	0.7% 📒	
262	const float4 bboxes_minY = as_float4(read_	0.7% 📒	
263	const float4 bboxes_maxY = as_float4(read_	0.7% 📒	
264	const float4 bboxes_minZ = as_float4(read_	0.7% 📒	-
265	const float4 bboxes_maxZ = as_float4(read_	0.7% 📒	
266	const int4 children = as_int4(read_imageui	0.7% 📒	
267			
268	const int4 visit = QBVHNode_BBoxIntersect(13.1%	
269	bboxes_minX, bboxes_maxX,		
270	bboxes minY, bboxes maxY,		
271	bboxes minZ, bboxes maxZ,		

intel

Intel® VTune[™] Profiler Memory Analysis

Intel[®] Vtune[™] Profiler -

What's Using All The Memory?

Memory Consumption Analysis

See What Is Allocating Memory

- Lists top memory consuming functions and objects
- View source to understand cause
- Filter by time using the memory consumption timeline

Standard & Custom Allocators

- Recognizes libc malloc/free, memkind and jemalloc libraries
- Use custom allocators after markup with ITT Notify API

Languages

- Python*
- Linux*: Native C, C++, Fortran

Top Memory-Consuming Objects

This section lists the most memory-consuming objects in your application. Optimizing these objects results in improving an overall application memory consumption.

Memory Object	Memory Consumption
dictobject.c:632 (768 B)	768 B
filedoalloc.c:120 (4 KB)	4 KB
iofopen.c:76 (568 B)	568 B
msort.c:224 (1 KB)	1 KB
dictobject.c:632 (3 KB)	3 KB
[Others]	217 TB

Optimize Memory Access Memory Access Analysis - Intel® VTune™ Profiler

Tune data structures for performance

- Attribute cache misses to data structures (not just the code causing the miss)
- Support for custom memory allocators

Optimize NUMA latency & scalability

- True & false sharing optimization
- Auto detect max system bandwidth
- Easier tuning of inter-socket bandwidth

Easier install, Latest processors

- No special drivers required on Linux*
- Intel[®] Xeon Phi[™] processor MCDRAM (high bandwidth memory) analysis

Top Memory Objects by Latency

This section lists memory objects that introduced the highest latency to the overall application execution.								
Memory Object	Total Latency	Loads	Stores	LLC Miss Count ⁽²⁾				
alloc_test.cpp:157 (30 MB)	65.6%	4,239,327,176	4,475,334,256	0				
alloc_test.cpp:135 (305 MB)	6.8%	411,212,336	441,613,248	0				
alloc_test.cpp:109 (305 MB)	6.3%	439,213,176	449,613,488	0				
alloc_test!I_data_init.436.0.6 (576 B)	5.2%	742,422,272	676,820,304	0				
[vmlinux]	4.6%	173,605,208	116,003,480	0				
[Others]	11.5%	1,533,646,008	1,674,450,232	0				

*N/A is applied to non-summable metrics.

Grouping: Function / Memory Object / Allocation Stack

Function / Memory Object /	Stores	LLC Miss Count 🔻				
Allocation Stack	Stores	Local DRAM Access Count	Remote DRAM Access Count			
doTriad\$omp\$parallel_for@2	40,307,609,1	2,439,273,176	2,430,472,912			
▶ triad!c (152 MB)	19,200,576	1,821,654,648	1,864,855,944			
triad!b (152 MB)	10,400,312	615,218,456	560,816,824			
▶ [Unknown]	7,200,216	2,400,072	3,200,096			
triad!doTriad (2 MB)	15,200,456	0	0			
▶ [Stack]	2,120,063,600	0	1,600,048			
▶ triad!a (152 MB)	38,135,544,0	0	0			
update_blocked_averages	6,400,192	2,400,072	0			

Memory Access Analysis Intel® VTune™ Profiler

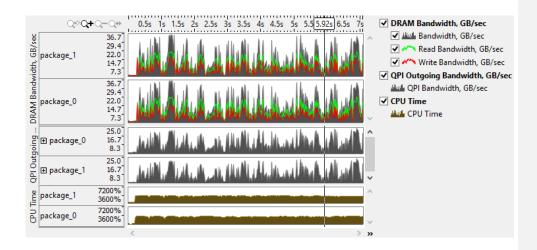
Tune data structures for better performance

Attribute cache misses to data structures

Better Bandwidth Analysis for Non-Uniform Memory

- See Read & Write contributions to Total Bandwidth
- Easier tuning of multi-socket bandwidth

Grouping: Bandwidth Domain / Bandwidth Utilization Type / Memory Object / Allocation Stack 🗸 🗔 🌾									
Bandwidth Domain / Bandwidth Utilization Type / Memory Object / Allocation Stack	Memory Bound	Loads	Stores	LLC Miss Count	Average Latency ▼ (cycles)	^			
🗆 DRAM, GB/sec	0.657	125,874,377,622	16,061,040	130,507,830	40				
🗆 High	0.750	28,236,084,708	5,014,875,	75,304,518	91				
stream.c:180 (76 MB)		900,002,700	654,009,810	18,301,098	495				
		1,050,003,150	667,210,008	33,301,998	487				
stream.c:181 (76 MB)		1,434,004,302	907,213,608	20,101,206	412				
Selected 1 row(s):	1.000	126,000,378	21,600,324	300,018	61	v			



Seeing total bandwidth can suggest data blocking opportunities to change a bandwidth bound app into a compute bound app.

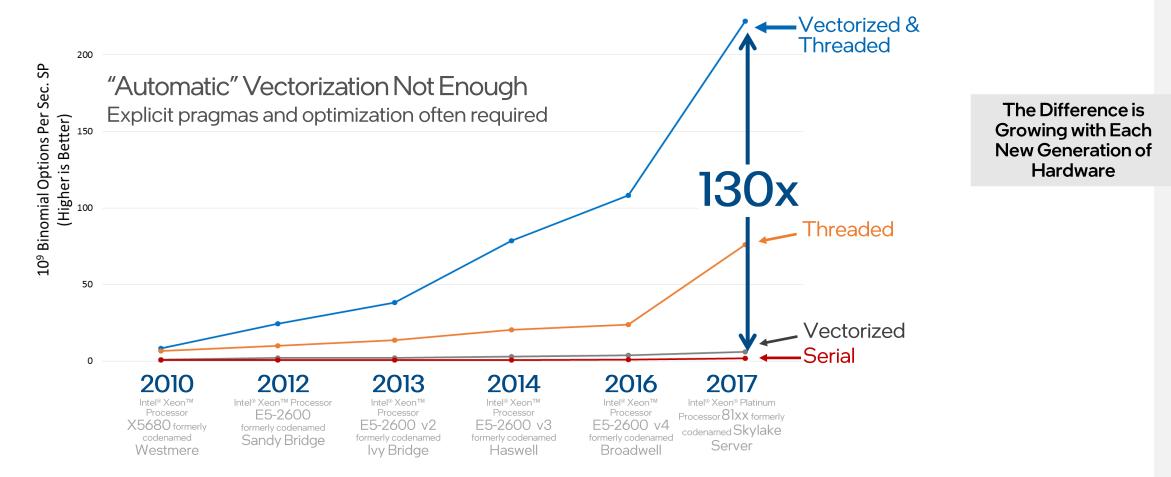
Intel® VTune™ Profiler Demo

Design your code for high-performance with Intel® Advisor



Intel® Advisor: Vectorize & Thread or Performance Dies

Threaded + Vectorized Can Be Much Faster that Either One Alone



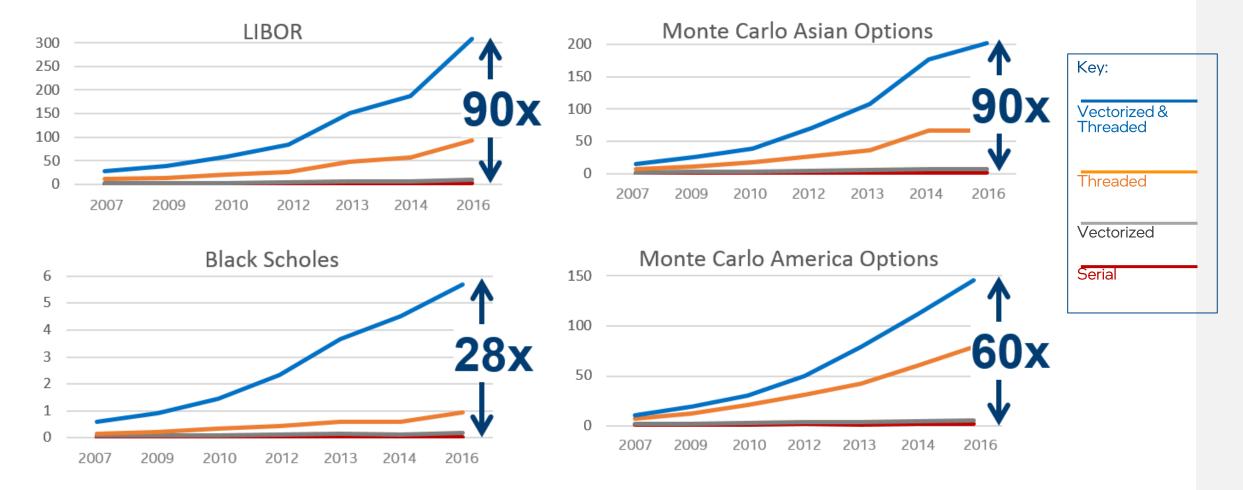
Testing Date: Performance results are based on testing by Intel employees as of 2017 and may not reflect all publicly available security updates.

Configuration Details and Workload Setup: See Vectorize & Thread or Performance Dies Configurations for 2010-2016 Benchmarks in Backup.

Performance results are based on testing as of dates shown in configurations and may not reflect all publicly available updates. See configuration disclosure for details. No product or component can be absolutely secure.

Performance varies by use, configuration, and other factors. Learn more at www.Intel.com/PerformanceIndex. Your costs and results may vary.

Intel[®] Advisor: Vectorization & Threading is Critical on Modern Hardware



Testing Date: Performance results are based on testing by Intel employees as of 2017 and may not reflect all publicly available security updates.

Configuration Details and Workload Setup: See Vectorize & Thread or Performance Dies Configurations for 2010-2016 Benchmarks in Backup.

Performance results are based on testing as of dates shown in configurations and may not reflect all publicly available updates. See configuration disclosure for details. No product or component can be absolutely secure.

Performance varies by use, configuration, and other factors. Learn more at www.Intel.com/PerformanceIndex. Your costs and results may vary.

"Automatic" Vectorization Often Not Enough A good compiler can still benefit greatly from vectorization optimization

- Compiler will not always vectorize
 - Check for Loop Carried Dependencies using <u>Intel[®] Advisor</u>
 - All clear? Force vectorization.
 C++ use: pragma simd, Fortran use: SIMD directive
- Not all vectorization is efficient vectorization
 - Stride of 1 is more cache efficient than stride of 2 and greater. Analyze with <u>Intel® Advisor</u>.
 - Consider data layout changes
 <u>Intel[®] SIMD Data Layout Templates</u> can help

Benchmarks on prior slides did not all "auto vectorize." Compiler directives were used to force vectorization and get more performance.

Arrays of structures are great for intuitively organizing data, but are much less efficient than structures of arrays. Use the Intel® SIMD Data Layout Templates (Intel® SDLT) to map data into a more efficient layout for vectorization.

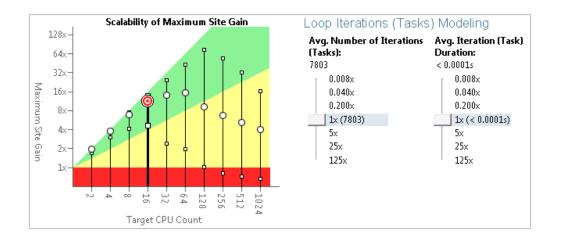
Faster Code Faster with Data Driven Design

Intel® Advisor – Vectorization Optimization and Thread Prototyping

- Faster Vectorization Optimization:
 - Vectorize where it will pay off most
 - Quickly ID what is blocking vectorization
 - Tips for effective vectorization
 - Safely force compiler vectorization
 - Optimize memory stride
- Breakthrough for Threading Design:
 - Quickly prototype multiple options
 - Project scaling on larger systems
 - Find synchronization errors before implementing threading
 - Design without disrupting development

Less Effort, Less Risk and More Impact

🦉 Where should I add						sm? 🚥		ntel Ad	dvisor	
Elapsed time: 54.44s Vector	-	Not Vectorized	-	TER: All I			Sources Y		202011	Q,
		0	Self	Total	Trip 🗵				ed Loops	
Function Call Sites and Loop	۵	Vector Issues	Time+	Time	Counts	Loop Type	Why No Vectorization?	Vecto	Efficiency	
t> O [loop at stl_algo.h:4740 i			0.170s1	0.170;1	1	Scalar	anon-vectorizable I		15 12	1
🖃 💹 [loop at loopstl.cpp:2449		₩ Ineffective peeled	0.170s1	0.170;1	12; 4	Collapse	Collapse	AVX	-100%	
: 😸 (loop at loopstl.cpp:2			0.150s1	0.150;1	12	Vectorized (8	3	AVX		
is 🖑 [loop at loopstl.cpp:2			0.02051	0.02011	4	Remainder				
DO [loop at loopstl.cpp:7900			0.170:1	0.170;1	500	Scalar	vectorization possi			
🗉 📒 [loop at loopstl.cpp:35		💡 <u>1</u> High vector regi	0.160s1	0.160s1	12	Expand	Expand	AVX	-69%	
<										



Part of Intel® Parallel Studio for Windows* and Linux*

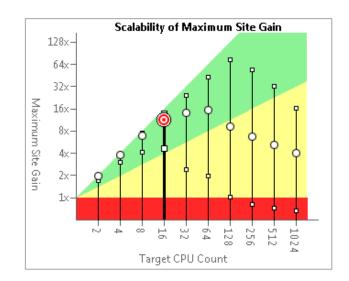
http://intel.ly/advisor-xe

Get Faster Code Faster! Intel[®] Advisor

Have you:

- Threaded an app, but seen little benefit?
- Hit a "scalability barrier"?
- Delayed release due to sync. errors?
- Data Driven Threading Design:
 - Quickly prototype multiple options
 - Project scaling on larger systems
 - Find synchronization errors before implementing threading
 - Design without disrupting development

Add Parallelism with Less Effort, Less Risk and More Impact



"Intel[®] Advisor has allowed us to quickly prototype ideas for parallelism, saving developer time and effort"

Simon Hammond Senior Technical Staff

Sandia National Laboratories

Get Faster Code Faster! Intel® Advisor

Vectorization Optimization

- Have you:
 - Recompiled for AVX2 with little gain
 - Wondered where to vectorize?
 - Recoded intrinsics for new arch.?
 - Struggled with compiler reports?

 Data Driven Ve 	ctorization:
------------------------------------	--------------

- What vectorization will pay off most?
- What's blocking vectorization? Why?
- Are my loops vector friendly?
- Will reorganizing data increase performance?
- Is it safe to just use pragma simd?

Elapsed time: 125.72s O Vectorized			I	OFF	Sn	nart M	lode	nuteon	٩				
TER: All Modules All Sources Summary Survey & Roofline	and the second second		ns 💌 All	Threads •							NIELA	INVISOR	
🕀 🖃 Function Call Sites and Loops		Perfor	Sell Lime *	Total Time	Туре	Why No Vectorization?	Vectorized Loops					Instruction Set	
		Issues					Vect	Efficiency	Gain	VL .	Com	Traits	Da
[loop in main at roofline.cpp:295]			18.538s 🗖	18.538s I	Vectorized (B	s (AVX	-100%	5.34x	4	5.34x		Fic
[Ioop in main at roofline.cpp:310]			18.3945 🚥	18.3945 0	Vectorized (Bo		AVX	-100%	5.34x	4	5.34x		Flo
[loop in main at roofline.cpp:221]			14.741s 📼	14.741s 0	Scalar	a novector dire		500 B					Flo
Iloop in main at roofline.cpp:234]			11.1175	11.117s1	Scalar	a inner loop w							Flo
[loop in main at roofline.cpp:247]			6.967s	6.967s1	Vectorized (Bo		AVX	-31%	1.22x	4	1.22x	Inserts; U	Flo
[00] [loop in main at roofline.cpp:138]			6.949s	6.949s1	Scalar	a novector dire							Flo
[loop in main at roofline.cpp:260]			3.285s	3.28511	Vectorized (Bo		AVX	~ 100%	5.09x	4	5.09x		Fic
🕞 🗇 [loop in main at roofline.cpp:199]			2,454s1	2.454s1	Vectorized (Bo		AVX	- 100%	5.14x	4	5.14x		Flo
🐨 🔿 [loop in main at roofline.cpp:273]			2.258s1	2.258s1	Vectorized (Bo		AVX2	~ 100%	4.73x	4	4.73x	FMA	Flo
Iloop in main at roofline.cpp:151]			1.899s 1	1.899s1	Vectorized (Bo		AVX	-100%	4.80x	4	4.80x		Flo
Iloop in main at roofline.cpp:256]		9 1 Oppo	0.042s1	3.327s1	Scalar	inner loop w		100					
[00] [loop in main at roofline.cpp:304]			0.040s1	18.434s 0	Scalar	inner loop w							
<pre></pre>	=												D.

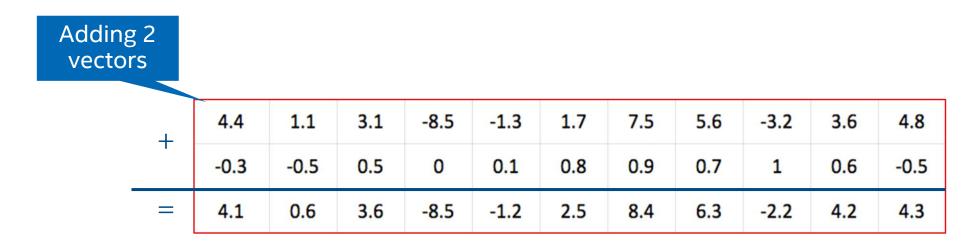
"Intel[®] Advisor's Vectorization Advisor permitted me to focus my work where it really mattered. When you have only a limited amount of time to spend on optimization, it is invaluable."

Gilles Civario

Senior Software Architect
Irish Centre for High-End Computing

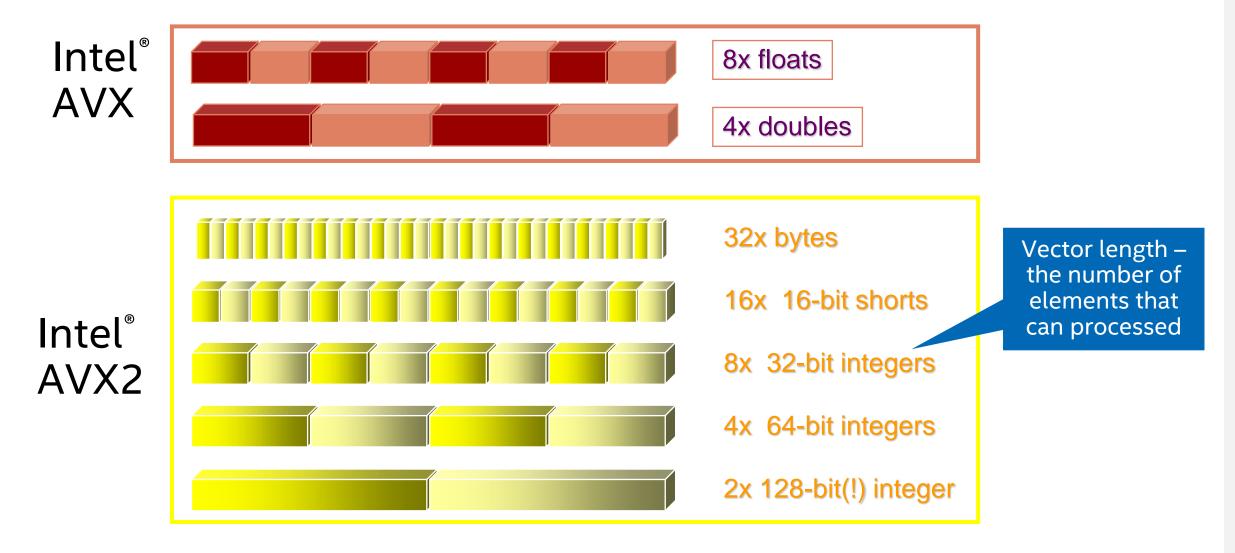
Vector Instructions are Dramatically Faster

Multiple arithmetic operations with a single instruction



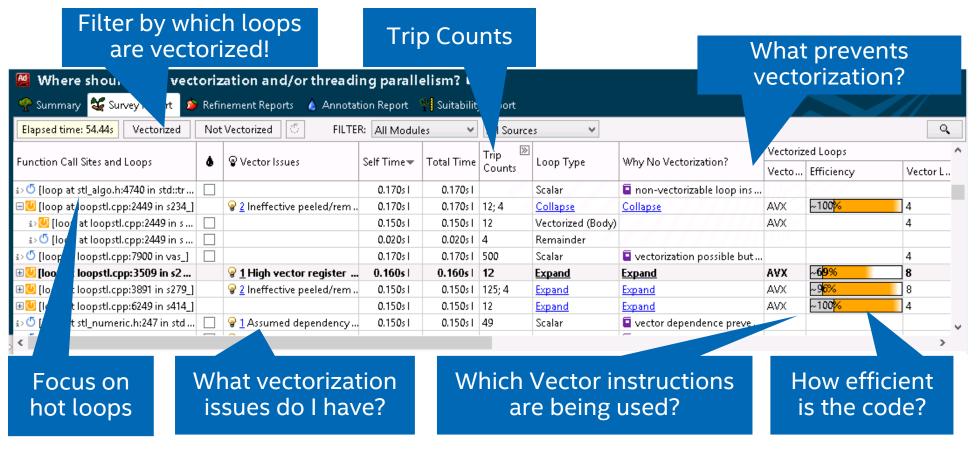
• These instructions are also referred to as Single Instruction Multiple Data (SIMD instructions)

Intel[®] Advanced Vector Extensions (Intel[®] AVX)



The Right Data At Your Fingertips

Get all the data you need for high impact vectorization

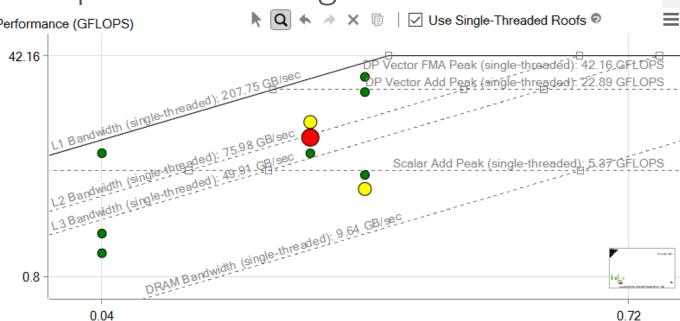


Get Faster Code Faster!

Intel® Advisor Roofline Analysis

What is a Roofline Chart?

- A Roofline Chart plots application performance against hardware limitations.
 Performance (GFLOPS)
 Q A X II Use Single-Threaded Roofs
 - Where are the bottlenecks?
 - How much performance is being left on the table?
 - Which bottlenecks can be addressed, and which *should* be addressed?
 - What's the most likely cause?
 - What are the next steps?



Arithmetic Intensity (FLOP/Byte)

Roofline first proposed by University of California at Berkeley: <u>Roofline: An Insightful</u> <u>Visual Performance Model for Multicore Architectures</u>, 2009

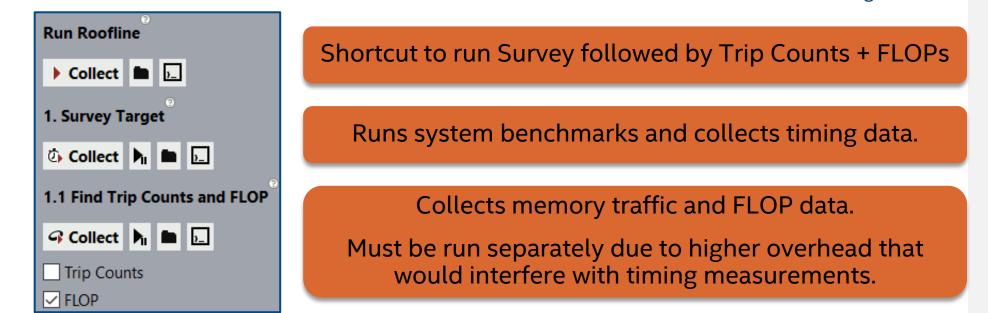
Cache-aware variant proposed by University of Lisbon: <u>Cache-Aware Roofline Model:</u> <u>Upgrading the Loft</u>, 2013

Roofline Metrics

- Roofline is based on FLOPS and Arithmetic Intensity (AI).
 - FLOPS: Floating-Point Operations / Second
 - Arithmetic Intensity: FLOP / Byte Accessed



Collecting this information in Intel[®] Advisor requires two analyses.



Classic vs. Cache-Aware Roofline

 Intel[®] Advisor uses the Cache-Aware Roofline model, which has a different definition of Arithmetic Intensity than the original ("Classic") model.

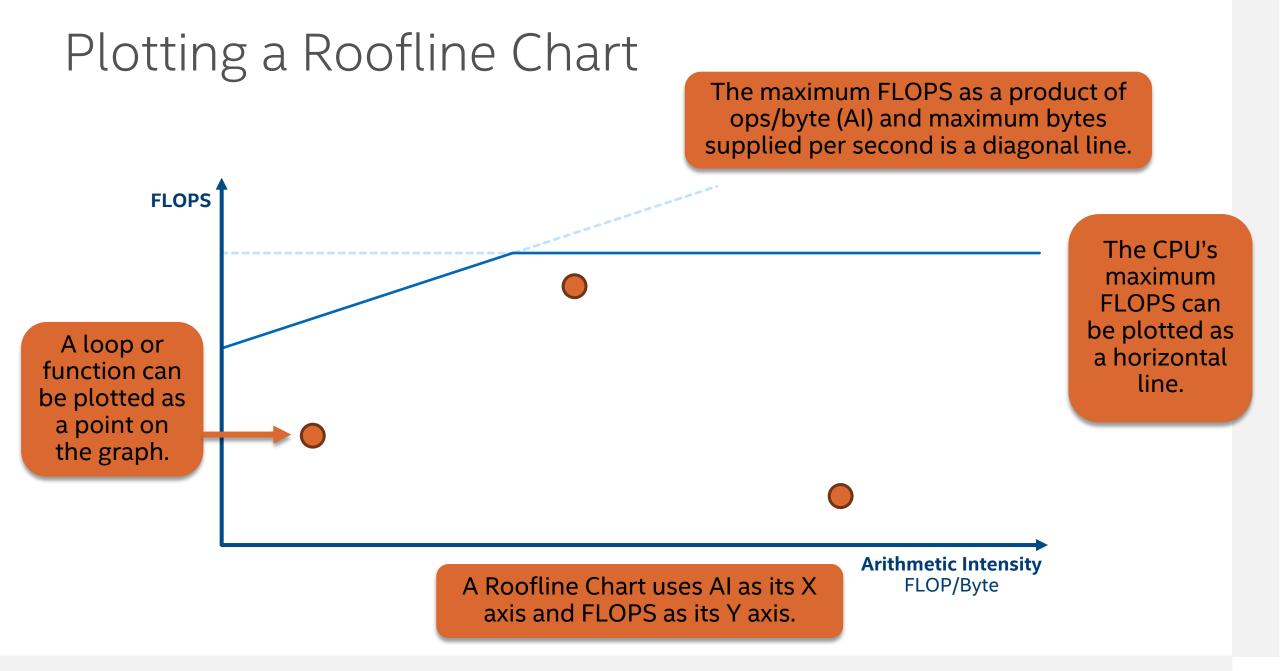
Classical Roofline

- Traffic measured from one level of memory (usually DRAM)
- AI may change with data set size
- AI changes as a result of memory optimizations

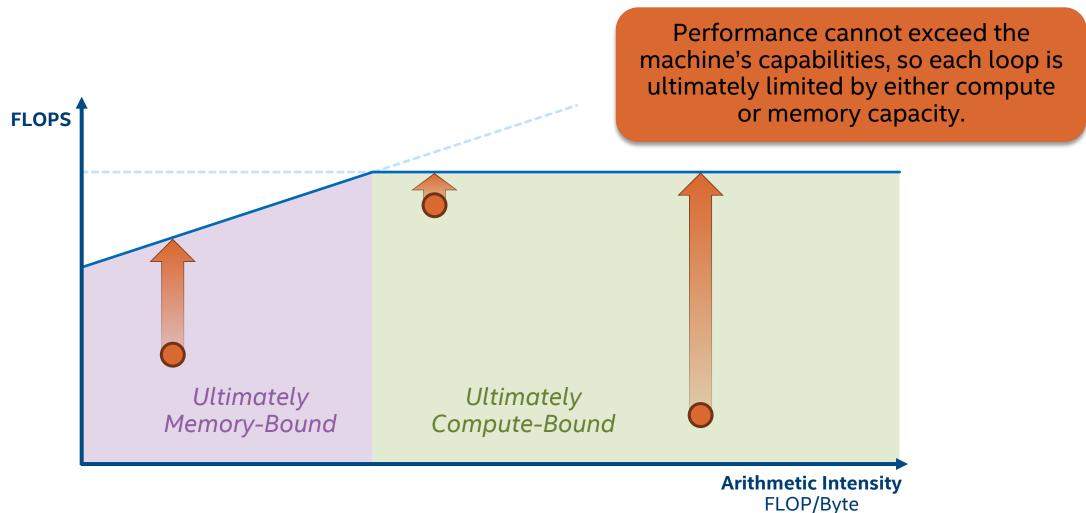
Cache-Aware Roofline

- Traffic measured from all levels of memory
- AI is tied to the algorithm and will not change with data set size
- Optimization does not change AI*, only the performance

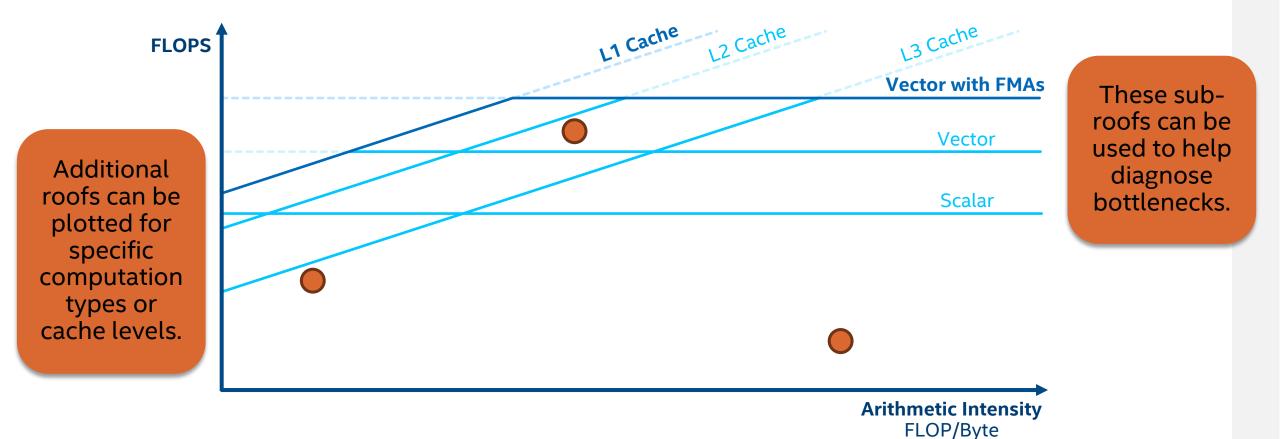
*Compiler optimizations may modify the algorithm, which may change the AI.



Ultimate Performance Limits

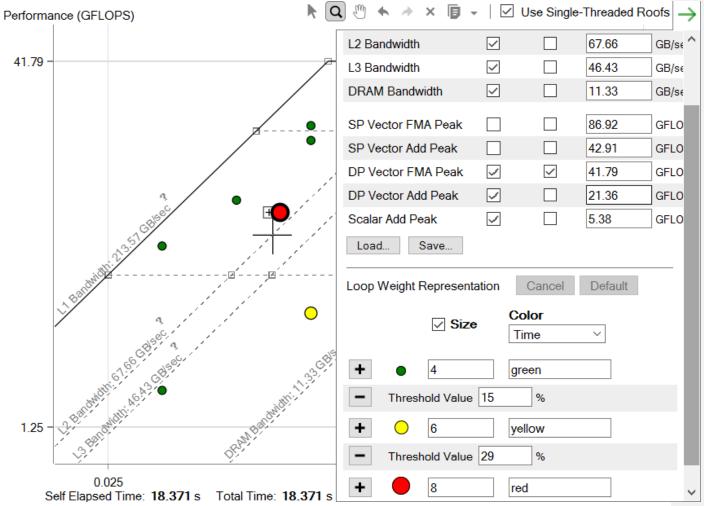


Sub-Roofs and Current Limits



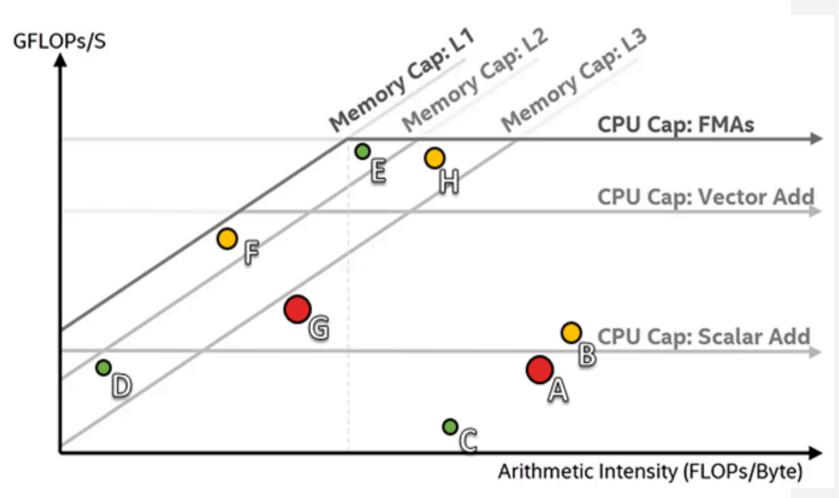
The Intel® Advisor Roofline Interface

- Roofs are based on benchmarks run before the application.
 - Roofs can be hidden, highlighted, or adjusted.
- Intel[®] Advisor has size- and colorcoding for dots.
 - Color code by duration or vectorization status
 - Categories, cutoffs, and visual style can be modified.



Identifying Good Optimization Candidates

- Focus optimization effort where it makes the most difference.
 - Large, red loops have the most impact.
 - Loops far from the upper roofs have more room to improve.



Intel® Advisor Offload Advisor

Intel[®] Advisor - Offload Advisor

(intel)

Intel® Advisor Beta OFFLOAD ADVISOR

Starting from an optimized binary (running on CPU):

- Helps define which sections of the code should run on a given accelerator
- Provides performance projection on accelerators

30	ninary Onioaded Reg	gions Non Offloaded Regions	Can free Conligu	Tation Logs
Program me	etrics ⑦			
Original ⑦ Accelerated ⑦	0.780s			
Target Platform	Gen9 GT2	Time on Host ⑦	0.100s	
Number of Offloa Speed Up for Acc	ds ⑦ 1 elerated Code ⑦ 1.8x	Time on Target ③	0.377s	21%
Amdahl's Law Sp	eed Up ⑦ 1.9x	Data Transfer Tax (?)	0s	79%
Fraction of Accele	erated Code ⑦ 95%	Kernel Launch Tax 🕐	0.00000520s	

Speed Up for Accelerated Code (?)	1.8x Number of Office	ads 1 Fraction of Accelerated 95%
Offloads bounde	ed by ⑦	Gen9 GT2 configuration … 🛓 🗐 💍
Compute ③ L3 Cache BW ④ LLC BW ④ Memory BW ③ Data Transfer ③ Invoke Tax ③ Transfer Tax ③ Dependency ③ Trip Count ③ Unknown ③ Non Offloaded ③	0% 95% 0% 0% 0% 0% 95% 0% 0% 0% 5%	1.15 GHz frequency (2) 24 EU (2) 512 0 KB L3 (2) 220 8 GB/s L3 bandwidth (2) 24 GB/s DRAM bandwidth (2) Integrated GPU (2)

Top offloaded ?

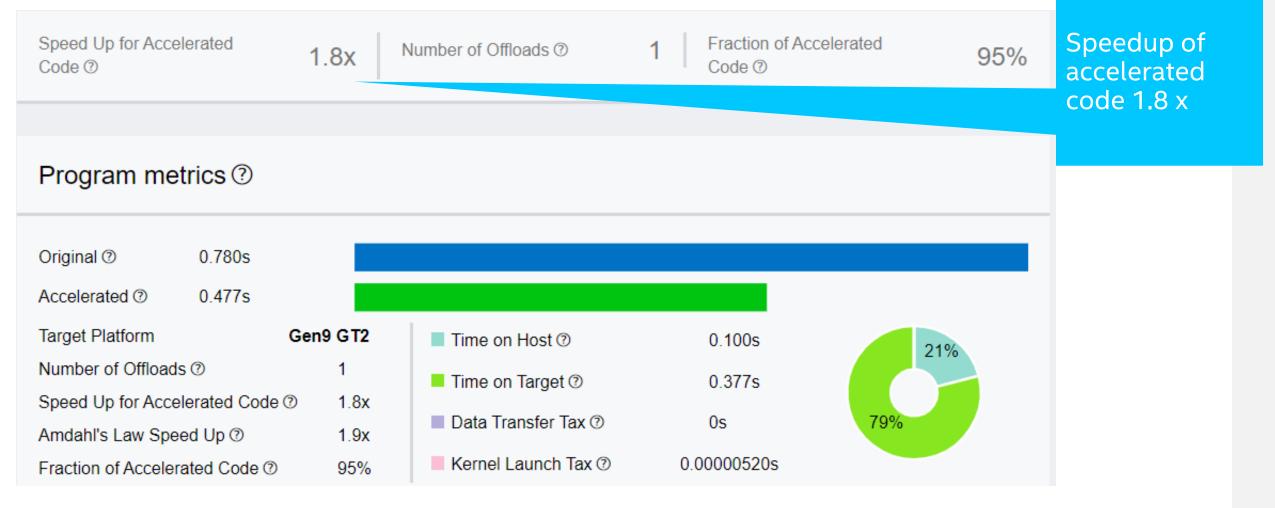
Location ⑦	Speed Up ⑦	Bounded By ⑦	Data Transfer ⑦
[loop in multiply1\$omp\$parallel@201 at multiply.c:202]	1.80x CPU 0.680s GPU 0.377s	L3_BW	25.17MB

Top non offloaded ?

Location ⑦	Data Transfer ⑦	Execution Time ⑦	Why Not Offloaded ⑦
[loop in start_thread]		CPU 0.100s GPU n/a	Cannot be modelled: Outside of Marked Region
[loop in _INTERNAL27dd4e00:: [OpenMP worker at z_Linux_util.cpp:589]		CPU 0.100s GPU n/a	Cannot be modelled: Outside of Marked Region

Intel® Advisor Beta, build 605774

Intel[®] Advisor - Offload Advisor Find code that can be profitably offloaded

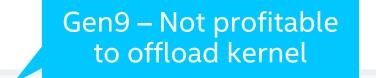


How to Run Intel[®] Advisor – Offload Advisor

- source <advisor_install_dir>/advixe-vars.sh
- advixe-python \$APM/collect.py advisor_project --config gen9 --/home/test/matrix
- advixe-python \$APM/analyze.py advisor_project --config gen9 --out-dir /home/test/analyze
- View the report.html generated (or generate a command-line report)

Analyze for a specific GPU config

Compare Acceleration on Different GPUs



Speed Up for Accelerated Code ⑦

1.0x

Number of Offloads ?

Fraction of Accelerated Code @

0

Speed Up for 0% Accelerated Code ⑦

Number of Offloads 1.6x \bigcirc

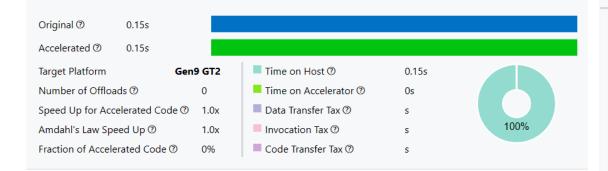
Fraction of Accelerated

Gen11 – 1.6x speedup

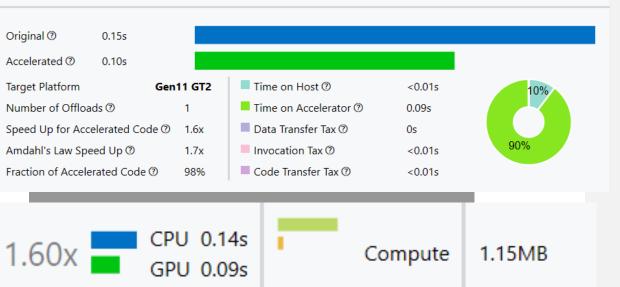
Code ⑦

98%

Program metrics ⑦



Program metrics ⑦



CPU 0.14s GPU 0.167

Not profitable: Computation Time is high despite the full use of Target Device capabilities

Intel® Advisor

Deliver reliable applications with Intel® Inspector

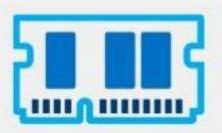


Threading, Memory and Persistence Debugger Intel® Inspector





Debug hard-to-find data races and deadlocks.





Memory Debugger

Detect memory leaks, invalid accesses, and more.

Persistent Memory Debugger

Find persistence errors that include redundant cache flushes.

Race Conditions Are Difficult to Diagnose They only occur occasionally and are difficult to reproduce

Correct Answer

Incorrect Answer

Thread 1	Thread 2		Shared Counter	Thread 1	Thread 2		Shared Counter
			0				0
Read count		←	0	Read count		←	0
Increment			0		Read count	←	0
Write count		→	1	Increment			0
	Read count	←	1		Increment		0
	Increment		1	Write count		→	1
	Write count	→	2		Write count	→	1

Intel[®] Inspector

Find & Debug Memory and Threading Errors

- Correctness Tools Increase ROI by 12%-21%¹
 - Errors found earlier are less expensive to fix
 - Races & deadlocks not easily reproduced
 - Memory errors are hard to find without a tool
- Faster Diagnosis with Debugger Breakpoints
 - Breakpoint set just before the problem occurs
 - Examine variables and threads with the debugger
- Debug Persistent Memory Errors
 - Missing cache flushes / store fences and more
- New in 2021 release:
 - Preview: Memory and threading errors analysis for DPC++ and OpenMP offloaded codes, executed on CPU target.

<u>1</u>	.ocat	e Deadlo	cks and Dat	a Races		INTEL	INSPECTOR
4 💮	Target	A Analysis	s Type 🔂 Coll	ection Log	Summary		
Proble	ms						1
ID 🔺	۹	Туре	Sources		Modules		State
± P1	0	Data race	find_and_fix_thr	eading_errors.c	pp find_and	_fix_threading_errors.ex	e R New
±P2	0	Data race	winvideo.h		find_and	_fix_threading_errors.e	🛯 🏲 Confirmed
₫ 1			1	of 10 D All	Code Loca	tions: Data race	4
Descript	tion	Source	Function	Module		Variable	
Read		winvideo.h:	201 loop_once	find_and_fix_	threading_er	rors.exe g_updates	
19 20 20 20 20	0 1 2	if(int up g_upo	n update not: pdates = g_up dates = 0; _video->updat	dates) {	rips += up	find_and_fix_thr find_and_fix_thr	
Write		winvideo.h:	270 next_frame	find_and_fix_	threading_er	rors.exe g_updates	
26 26 27	9 0	g_updates	ing) return 1 3++; // Fast aded) while(1	but inaccur		<pre>find_and_fix_thr find_and_fix_thr find_and_fix_thr find_and_fix_thr find_and_fix_thr</pre>	eading_errors.es
27	2		_handles[1])			find_and_fix_thr	

54

Debug Memory & Threading Errors Intel® Inspector

- Find and eliminate errors
 - Memory leaks, invalid access...
 - Races & deadlocks
 - C, C++ and Fortran (or a mix)
- Simple, Reliable, Accurate
 - No special recompiles
 Use any build, any compiler¹
 - Analyzes dynamically generated or linked code
 - Inspects 3rd party libraries without source
 - Productive user interface + debugger integration
 - Command line for automated regression analysis

Рго	blems									1
	٩	Туре		Sources		Object Size	: St	ate	Modules	1
± P5	i 🔕	Mismatched allocation	/deallocat	. gdivideo.c	pp		R	New	find_a	1
⊞ P6	3	Memory leak		find_and_f	fix_me	28672	P	Confirmed	find_a	
± P7	7 🔕	Memory leak		gdiplusgra	phics.h	507904	P	New	find_a	
± P8	8 🔕 🛛	Memory leak		mlock.c		32	P	New	tbb_de	
± P9) 🙆 🗌	Invalid memory access		dynamic_l	ink.c		~	Fixed	find_a	
±Ρ.	Δ	Memory not deallocate	ed	api.cpp; ut	til.cpp	10376	R	New	find_a	
4	1		1 of 1	7 🕨 📶	Code Lo	cations: M	emory	leak		G
-	cripti	Source	Funct N	/lodule		Object	Off	Variable		
Des				1 10		224		block alloca	ated at find	d.
	locat	. find_and_fix_memory	opera fi	nd_and_fix_	mem	224				
	locat 161	.find_and_fix_memory unsigned in			mem		d_fix	_memory_e	errors.ex	ĸe
		/ _	nt serial	=1;		find_an	_	_memory_e		
	161 162	unsigned in unsigned in	nt serial: nt mboxsi:	=l; ze = size	of (uns	find_an find_an	d_fix	memory_e	errors.e	
	161	unsigned in	nt serial: nt mboxsi: nt * loca	=1; ze = size 1_mbox =	of (uns <mark>(unsig</mark>	find_an find_an find_an find_an	d_fix d_fix d_fix		errors.ex errors.ex errors.ex	ĸ

Clicking an error instantly displays source code snippets and the call stack

Fits your existing process

Productive User Interface Saves Time Intel® Inspector

Select a problem set

🧾 Dete	ect Memory Problems					INTE	L INSPECTOR
🛛 🕀 Targ	et Å Analysis Type 🛃 Collect	ion Log 🥥 \varTheta Summary					
Problems				8	Filters		Sort 🔻 🦮
🔺 🗞	Туре	Sources	State	M	Severity		
IP1 🔕	Mismatched allocation/deallocati	on find_and_fix_memory	🖻 Confir	med	Error		3 item(s)
P2 🔕	Memory leak	find_and_fix_memory	P Deferr	ed fi.	ming		1 item(s)
P3 🔕	Invalid memory access	find_and_fix_memory	New New	fi.	Туре		
P4 🛆	Memory not deallocated	api.cpp; mlock.c; util.c	New	fi.	Invalid mem	ory acc.	1 item(s)
	Memory not deallocated	video.cpp:82	New New	fi.	Memory lea		1 item(s)
	Memory not deallocated	util.cpp:163	New New	fi.	· · · · ·	t deallocated	1. m(s)
	Memory not deallocated	api.cpp:218	New New	fi.	Mismatched	l allocation/deal	loc 1 item(s)
	Memory not deallocated	mlock.c:347	New New	tb.	Source		
					api.cpp		1 item(s)
			—		find and fin		2 :+(-)
4 1 📄	1 of •	4 D All Code Locations:	Mismatch	ed allocation/	deallocation		
escription	Source	Funct	tion M	odule		Object Size	Offset
Mismatche	ed deallocation site find_and_fix_m	emory_errors.cpp:175 opera	tor() fin	d_and_fix_me	mory_errors.exe		
173	drawing	->put_pixel(c);				find_and_fix	x_memory_error
174	}						<pre>x_memory_error</pre>
175		g); //Memory Error: u	se delet	e instead o	of free		<pre>k_memory_error</pre>
176	//delete dra	awing;					<pre>k_memory_error</pre>
177	}					tbb_debug.d	ll!local_wait_
Allocation	site find_and_fix_m	emory_errors.cpp:170 opera	tor() fin	d_and_fix_mer	mory_errors.exe		
168	for (int y = r.beging	n(); y != r.end(); ++y) {				<pre>x_memory_error</pre>
169	{						<pre>x_memory_error</pre>
170		a * drawing = new draw		(startx, to	otaly-y, st		<pre>x_memory_error</pre>
171		<pre>= startx ; x < stopx;</pre>					<pre>k_memory_error</pre>
172	color t	c = render one pixel	(X, Y, ⊥	ocal mbox,	serial, st	too debug.d.	ll!local wait

Filters let you focus on a module, or error type, or just the new errors or...

Problem States: New, Not Fixed, Fixed, Confirmed, Not a problem, Deferred, Regression

Double Click for Source & Call Stack

Intel[®] Inspector

Source code

displayed for selected

locations

problem

🧾 Mis	matched allocation/deallocation	INTEL INSPECTOR
🛛 🕀 Targ	get 🖄 Analysis Type 🖪 Collection Log 🛛 🥥 Summary 🛛 🖓 Sources	
Mismatche	d deallocation site - Thread thread_video (4596) (find_and_fix_memory_errors.exe!ope	erator() - find_and_fix_memory_errors.cp 💡 💻
find_and_fi	ix_memory_errors.cpp Disassembly (find_and_fix_memory_errors.exe!0x46d6)	Call Stack
164 165 166 167 168 169 170 171 172 173 174	<pre>for (unsigned int i=0;i<=(mboxsize/(sizeof(unsigned int)));i++) local_mbox[i]=0; //Memory Error: C declared arrays go from for (int y = r.begin(); y != r.end(); ++y) { { drawing_area * drawing = new drawing_area(startx, totaly for (int x = startx ; x < stopx; x++) { color_t c = render_one_pixel (x, y, local_mbox, seri drawing->put_pixel(c); } </pre>	find_and_fix_memory_errors.exelexecute nara tbb_debug.dll!local_wait_for_all - custom_s tbb_debug.dll!local_spawn_root_and_wait - sc tbb_debug.dll!spawn_root_and_wait - schedul find and fix memory errors.exelspawn root a
An cation s	ite - Thread thread_video (4596) (find_and_fix_memory_errors.exe!operator() - find_a	nd_fix_memory_errors.cpp:170) 🛛 💡 🕞
find_and_f	ix_memory_errors.cpp Disassembly (find_and_fix_memory_errors.exe!0x4613)	Call Stack
170	<pre>drawing_area * drawing = new drawing_area(startx, totaly</pre>	- find_and_fix_memory_errors.exeloperator() - f
171	for (int $x = startx ; x < stopx; x++$) {	find_and_fix_memory_errors.exe!run_body - p
172	<pre>color_t c = render_one_pixel (x, y, local_mbox, seri</pre>	find_and_fix_memory_errors.exelexecute <class< td=""></class<>
173	drawing->put_pixel(c);	find_and_fix_memory_errors.exelexecute - par
174	}	tbb_debug.dll!local_wait_for_all - custom_sch
175	<pre>free(drawing); //Memory Error: use delete instead of fi</pre>	
176	<pre>//delete drawing;</pre>	tbb_debug.dll!spawn_root_and_wait - schedul

Call Stack

Easy Problem Management

Quickly see new problems and regressions

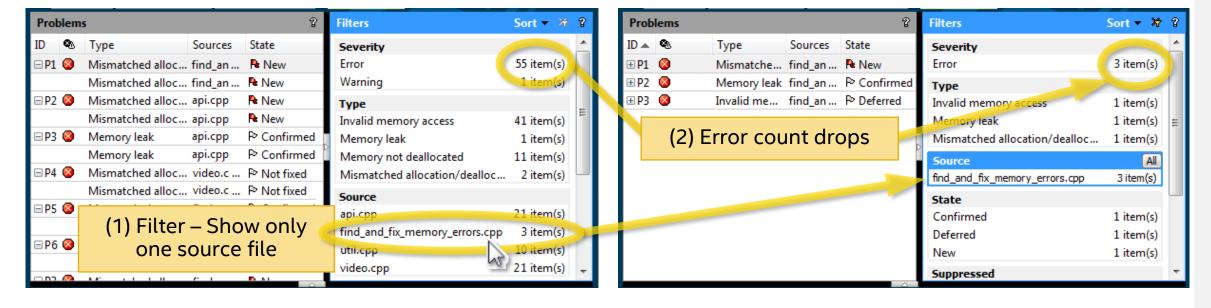
State	Description
New	Detected by this run
Not Fixed	Previously seen error detected by this run
Not a Problem	Set by user (tool will <u>not</u> change)
Confirmed	Set by user (tool will <u>not</u> change)
Fixed	Set by user (tool <u>will</u> change)
Regression	Error detected with previous state of "Fixed"

√ € Prob	Targe	ect Memory Problems et 🗛 Analysis Type 🖪 Collection	Log 📄 🤊 Summary	INTEL INS	PECTOR 2017 > *	View Source Edit Source Copy to Clipboard Explain Problem	
ID 🔺	٩	Туре	Sources	State	Modules	Create Problem Report	
⊞ P1	8	Mismatched allocation/deallocation	find_and_fix_memoryrs	P Confirmed	find_and	 Debug This Problem	Not fixed
⊞ P2	8	Memory leak	find_and_fix_memory_errons	P Deferred	find_and	Change State 🔹 🕨	Confirmed
⊞ P3	8	Invalid memory access	find_and_fix_memory_errors	ବ New	find_and	Merge States	Fixed
⊞ P4	Δ	Memory not deallocated	api.cpp; mlock.c; util.cpp; vi	Rew New	find_and		Not a problem
							Deferred

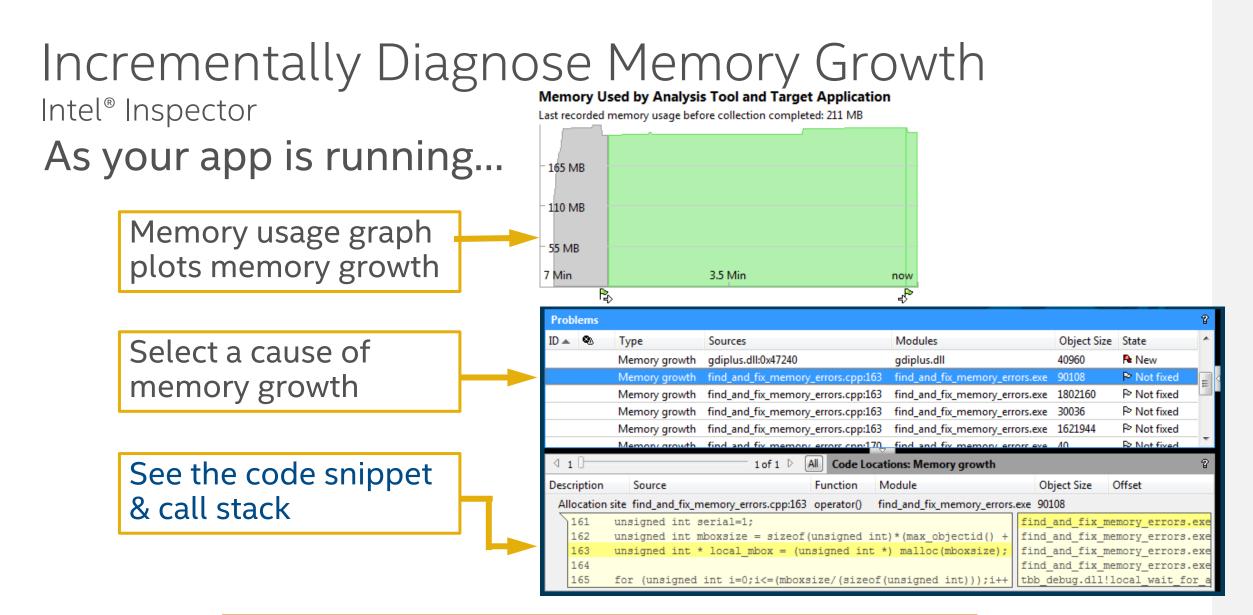
Filtering – Focus on What's Important Example: See only the errors in one source file

Before – All Errors

After – Only errors from one source file



Tip: Set the "Investigated" filter to "Not investigated" while investigating problems. This removes from view the problems you are done with, leaving only the ones left to investigate.



Speed diagnosis of difficult to find heap errors

Automate Regression Analysis

Command Line Interface

inspxe-cl is the command line:

-Windows: C:\Program Files\Intel\Inspector XE \bin[32|64]\inspxe-cl.exe

-Linux: /opt/intel/inspector_xe/bin[32|64]/inspxe-cl

Help:

inspxe-cl -help

- Set up command line with GUI
- Command examples:
 - 1.inspxe-cl -collect-list
 - 2. inspxe-cl -collect ti2 -- MyApp.exe
 - 3. inspxe-cl -report problems

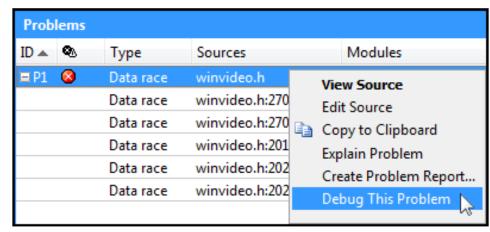
Send results file to developer to analyze with the UI

Break At Just The Right Time Intel[®] Inspector - Memory & Thread Debugger

Memory Errors

Problems								
ID 🔺 🕙	Туре	Sources						
🗄 P1 🔕	Mismatched allocation/deallo	llo View Source						
🗄 P2 🥝	Memory leak	Edit Source						
🖃 P3 🥝	Invalid memory access	Copy to Clipboard						
	Invalid memory access	Explain Problem						
🗄 P4 🧘	Memory growth	Create Problem Report						
🗄 P5 🧘	Memory growth	Debug This Problem						
🗄 P6 🧘	Memory growth	V						

Threading Errors



- Break into the debugger just before the error occurs.
- Examine the variables and threads.
- Diagnose the problem.

Save time. Find and diagnose errors with less effort.

Productive Memory & Threading Debugger		
Intel [®] Inspector	Memory Analysis	Threading Analysis
View Context of Problem Stack Multiple Contributing Source Locations	\checkmark	\checkmark
Collapse multiple "sightings" to one error (e.g., memory allocated in a loop, then leaked is 1 error)	\checkmark	\checkmark
Suppression, Filtering, and Workflow Management	\checkmark	\checkmark
Visual Studio* Integration (Windows*)	\checkmark	\checkmark
Command line for automated tests	\checkmark	\checkmark
Timeline visualization	\checkmark	\checkmark
Memory Growth during a transaction	\checkmark	
Trigger Debugger Breakpoint	\checkmark	\checkmark

Easier & Faster Debugging of Memory & Threading Errors

#