



LRZ Workshop – 10.11.2022

Application Profiling with Intel® VTune™ and PTI-GPU

Heinrich Bockhorst



intel.[®]

Agenda

- PTI-GPU tools
- onetrace tool from PTI-GPU
- VTune overview
- Command lines – Playbook
- VTune GPU analysis using GROMACS
- Documentation
- Demo – Hands on

Profiling Tools Interfaces for GPU (PTI for GPU)

- Light weight Open Source tools for GPU profiling (MIT license)
- Github initiated by Intel® engineers but with contributions from community:
<https://github.com/intel/pti-gpu>
- Profiling for Intel® GPUs using OpenCL and Level Zero runtime
- OpenCL*, SYCL*, and OpenMP* offloading supported
- Snapshots taken from Intel® DevCloud using Intel® Iris® Xe GPUs

sysmon

- Linux static and dynamic information about GPU activity
- Can be used like Linux „top“

```
u48655@s011-n001:~  
Every 1.0s: /data/comp/sysmon  
=====  
GPU 0: Intel(R) Iris(R) Xe MAX Graphics [0x4905]    PCI Bus: 0000:69:00.0  
Vendor: Intel(R) Corporation    Driver Version: 1.2.21786    Subdevices: 0  
EU Count: 96    Threads Per EU: 7    EU SIMD Width: 8    Total Memory(MB): 7714.0  
Core Frequency(MHz): 1550.0 of 1550.0    Core Temperature(C): unknown  
=====  
Running Processes: 2  
  PID, Device Memory Used(MB), Shared Memory Used(MB), GPU Engines, Executable  
  221409,           15.9,            0.0,      3D;DMA, /home/u48655/Software/GROMACS/2022/gromacs-master-22.02.03/oneapi/2022.1.2SIMD8USM/bin/gmx  
  221449,           0.2,            0.0,    UNKNOWN, /data/comp/sysmon  
=====  
GPU 1: Intel(R) Iris(R) Xe MAX Graphics [0x4905]    PCI Bus: 0000:1b:00.0  
Vendor: Intel(R) Corporation    Driver Version: 1.2.21786    subdevices: 0  
EU Count: 96    Threads Per EU: 7    EU SIMD Width: 8    Total Memory(MB): 7714.0  
Core Frequency(MHz): 300.0 of 1550.0    Core Temperature(C): unknown  
=====  
Running Processes: 1  
  PID, Device Memory Used(MB), Shared Memory Used(MB), GPU Engines, Executable  
  221449,           0.2,            0.0,    UNKNOWN, /data/comp/sysmon  
=====
```

onetrace

- Profile OpenCL* and Level Zero backend.
- For OpenCL, SYCL* and OpenMP* offloading applications.

usage: onetrace [options] <application> <args>

Options:

--host-timing [-h]	host API calls statistics
--device-timing [-d]	kernel execution timing
--chrome-kernel-timeline	generates trace for chrome://tracing

...

onetrace host statistics (GROMACS SYCL)

- \$ onetrace -h gmx <gmx args>

```
/cygdrive/c/Users/hbockhor/OneDrive - Intel Corporation/Training/2022/22.05.30_ISC_MTA/RESULTS_onetrace

== API Timing Results: ==

    Total Execution Time (ns):          8849359641
    Total API Time for L0 backend (ns): 5083012588

== L0 Backend: ==

      Function,    Calls,    Time (ns),  Time (%),    Average (ns),    Min (ns),    Max (ns)
      zeCommandQueueSynchronize, 2009, 3697535106, 72.74, 1840485, 164, 7145738
      zeModuleCreate, 7, 1030797983, 20.28, 147256854, 83120536, 227200191
      zeCommandQueueExecuteCommandLists, 4226, 146236830, 2.88, 34604, 14481, 104637
      zeCommandListAppendMemoryCopy, 3172, 87847894, 1.73, 27694, 11550, 840112
      zeCommandListReset, 4226, 37004797, 0.73, 8756, 554, 36072
      zeCommandListCreate, 8, 27893013, 0.55, 3486626, 226640, 13670348
      zeCommandListAppendLaunchKernel, 2811, 13505254, 0.27, 4804, 926, 17791
      zeKernelSetArgumentValue, 35215, 9806280, 0.19, 278, 31, 9071
      zeEventCreate, 5983, 6979341, 0.14, 1166, 364, 54293
      zeCommandListClose, 4226, 3455592, 0.07, 817, 140, 15018
      zeMemAllocDevice, 11, 3054949, 0.06, 277722, 56756, 1191462
      zeFenceReset, 4226, 3037219, 0.06, 718, 536, 24105
      zeKernelSetGroupSize, 2811, 2867434, 0.06, 1020, 141, 4410
```

onetrace device statistics (GROMACS SYCL)

- \$ onetrace --demangle -d gmx <gmx args>

```
■ /cygdrive/c/Users/hbockhor/OneDrive - Intel Corporation/Training/2022/22.05.30_JSC_MTA/RESULTS_onetrace

== Device Timing Results: ==

    Total Execution Time (ns):          9250282340
    Total Device Time for L0 backend (ns): 5144159294

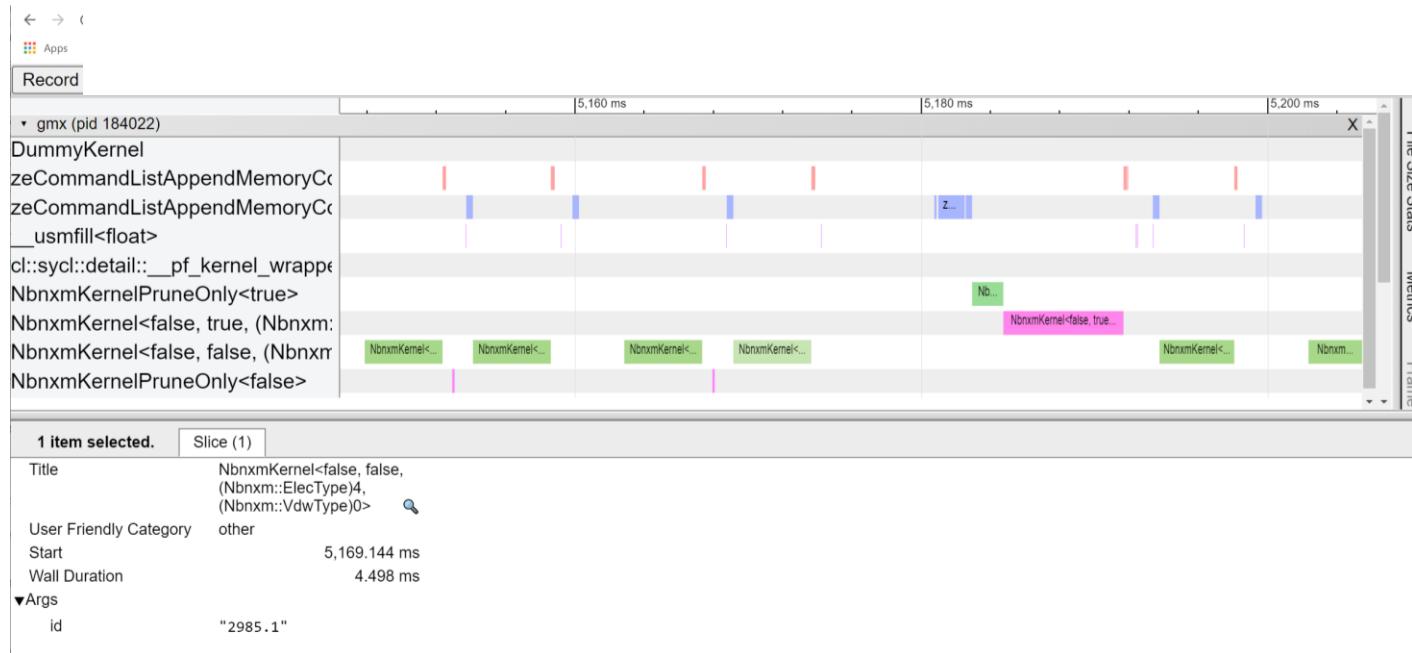
== L0 Backend: ==

      Kernel,      Calls,      Time (ns),      Time (%),      Average (ns),      Min (ns),      Max (ns)
NbnxmKernel<false, false, (Nbnxm::ElecType)4, (Nbnxm::VdwType)0>,           990,      4374811057,      85.04,      4419001,      4290416,      4946458
          zeCommandListAppendMemoryCopy(M2D),           2047,      393601766,       7.65,      192282,      3229,      1283489
          zeCommandListAppendMemoryCopy(D2M),           1125,      183165569,       3.56,      162813,      520,      190104
NbnxmKernel<false, true, (Nbnxm::ElecType)4, (Nbnxm::VdwType)0>,            11,      77472702,       1.51,      7042972,      6985937,      7131562
          NbnxmKernelPruneonly<false>,           490,      67430948,       1.31,      137614,      130416,      151979
          NbnxmKernelPruneonly<true>,           1307,      27828716,       0.54,      21292,      1666,      42500
          cl::sycl::detail::__pf_kernel_wrapper<__usmfill<float>>,           1,      19808433,       0.39,      1800766,      1798125,      1804687
          DummyKernel,           1,      35833,       0.00,      35833,      35833,      35833
                                         ,           1,      4270,       0.00,      4270,      4270,      4270
```

More information available by adding -v (verbose) option

onetrace kernel timeline

- \$ onetrace --demangle --chrome-kernel-timeline <gmx> <args>



PTI-GPU advantages

- No additional drivers necessary
- Low overhead < 3%
- Traces in json format to be loaded by Google Chrome:
`chrome://tracing`

VTune™ -- Short Overview

Optimize Performance

Intel® VTune™ Profiler

Get the Right Data to Find Bottlenecks

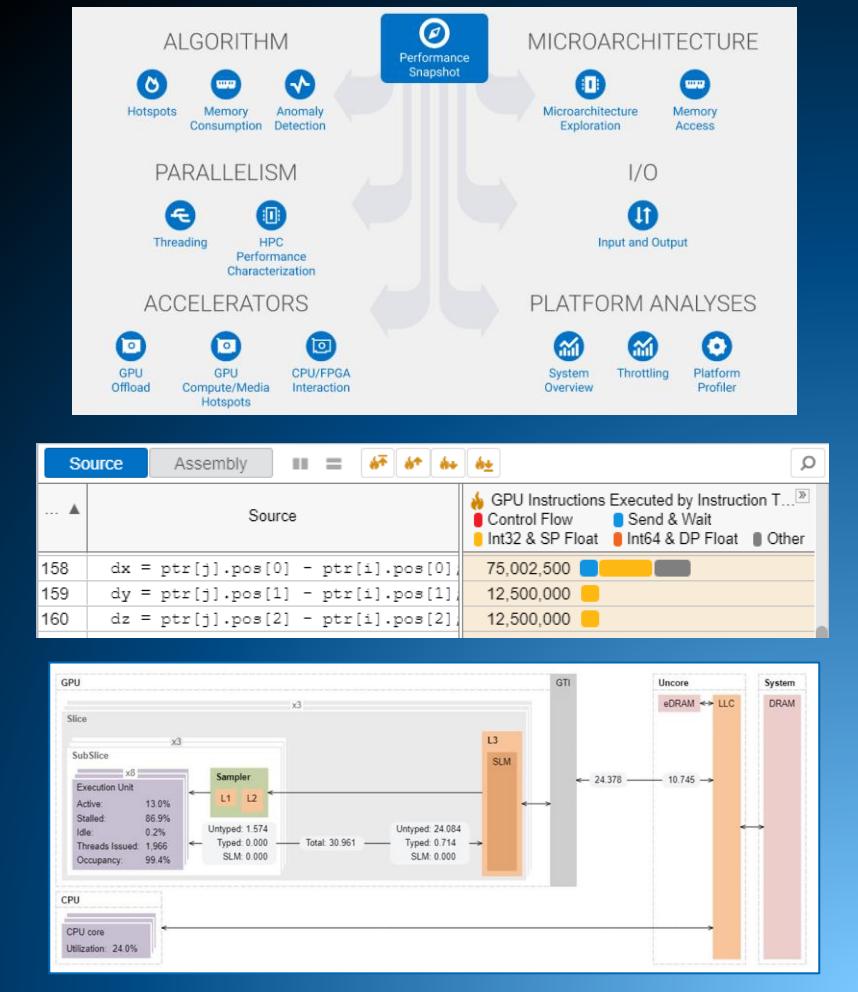
- A suite of profiling for CPU, GPU, FPGA, threading, memory, cache, storage, offload, power...
- DPC++, C, C++, Fortran, Python*, Go*, Java*, or a mix
- Linux, Windows, FreeBSD, Android, Yocto and more

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



Find Answers Fast

Intel® VTune™ Profiler

Adjust Data Grouping

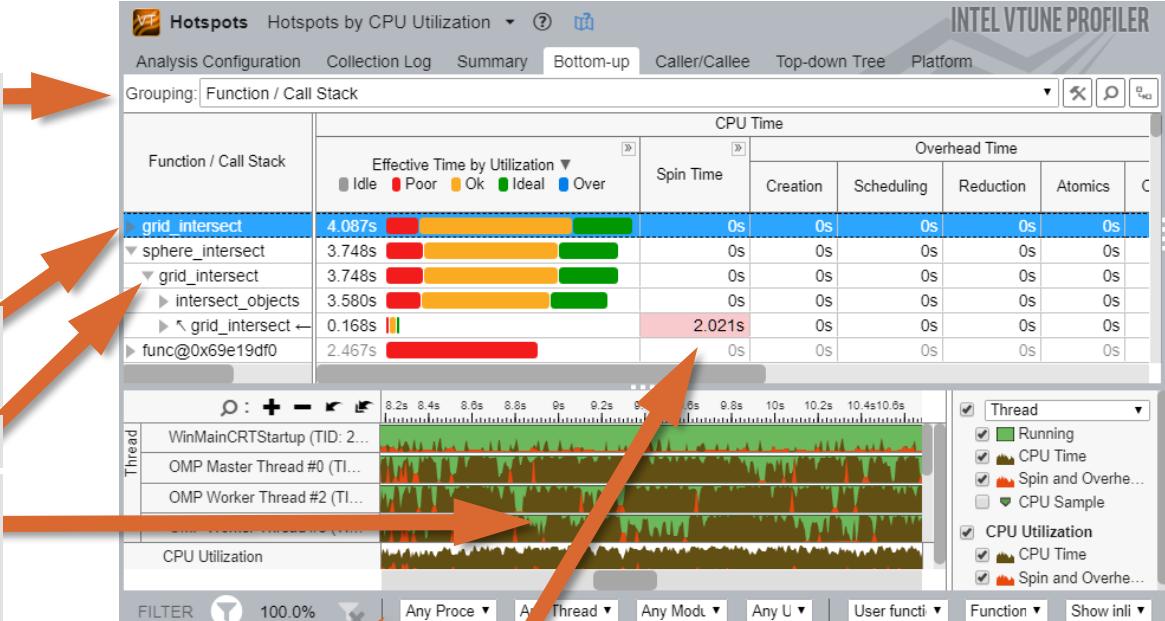
- Function / Call Stack
- Source Function / Function / Call Stack
- Sync Object / Function / Call Stack
- Sync Object / Thread / Function / Call Stack
- ... (Partial list shown)

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



Filter by Process
& Other Controls

Tuning Opportunities Shown in Pink.
Hover for Tips

VTune Playbook – for this session

- Easy access to command lines
- Nbody Example for DevCloud – please try

```
~/INCOMING/22.02.15-EuroCC
-----
Playbook for using VTUNE tool on devcloud or other clusters
-----
Note: command lines start with "$" prompt.

1. Log into DevCloud
-----
$ ssh devcloud
Alternative: open a jupyter notebook and start the terminal

2. Clone samples GitHub
-----
$ git clone https://github.com/oneapi-src/oneAPI-samples.git

3. Start interactive session on a node with DG1 Xe GPU (iris_xe_max)
-----
(People with NDA accounts may use ATS-P gpu)

it is better to compile on compute node because login node has very limited memory etc.

$ qsub -I -l nodes=1:iris_xe_max:ppn=2
4. Check properties
-----
$ sycl-ls --verbose

prints out all backends (GPU device + low level driver level_zero or opencl)
Level Zero shows:

Platform [#4]:
  Version : 1.2
  Name    : Intel(R) Level-Zero
  Vendor  : Intel(R) Corporation
  Devices : 
    Device [#0]:
```

Playbooks on DevCloud:
/data/comp/workshop/vtune_playbook.txt

How to start an Analysis

- VTune offers different analysis types with additional “knobs”
- Application Performance Snapshot (VTune or standalone)
- APS does first analysis and guides to the right VTune analysis or a different tool available in the oneAPI toolkits.
- APS analyses HW counters, OpenMP and MPI
- APS can be used for MPI and/or OpenMP only to avoid too much data being collected

APS usage

- Help menu

```
$ aps -help
```

- Run with APS

```
$ aps <application> <app paramter>
```

- Run with MPI (Intel MPI or compatible)

```
$ mpirun -n <N> aps -- <application> < app parameter>
```

APS HTML output (nbody)

APS report 127.0.0.1:55001/u/NBODY-NEI file:///C:/cygwin64/home/hbockhor/INCOMING/22.02.15-EuroCC/aps_report_20220216_004934.html

Application Performance Snapshot

Application: nbody
Report creation date: 2022-02-16 00:49:08
HW Platform: Intel(R) Xeon(R) Processor code named CascadeLake
Frequency: 3.50 GHz
Logical Core Count per node: 24
Collector type: Event-based sampling driver, Driverless Perf system-wide counting

17.36 s CPU
Elapsed Time 1.77 0
16.93 s GPU IPC Rate SP GFLOPS

0 3.51 GHz DP GFLOPS Average CPU Frequency

Your application might underutilize the available logical CPU cores because of insufficient parallel work, blocking on synchronization, or too much I/O. Perform function or source line-level profiling with tools like Intel® VTune™ Profiler to discover why the CPU is underutilized.

	Current run	Target	Tuning Potential
Physical Core Utilization	11%	>80%	<div style="width: 100%; background-color: red;"></div>
Memory Stalls	3.4%	<20%	<div style="width: 15%; background-color: red;"></div>
Vectorization	0%	>70%	<div style="width: 0%; background-color: red;"></div>
GPU Utilization when Busy	95.3%	>80%	<div style="width: 100%; background-color: red;"></div>

GPU Utilization when Busy 95.3%
EU State % of EUUs
Active 95.3%
Idle 0.8%
Stalled 3.9%
GPU Occupancy 95.9% of Peak Value

Physical Core Utilization 11%
Average Physical Core Utilization 1.32 out of 12 Physical Cores

Memory Footprint
Resident 147 MB
Virtual 68719932 MB

Memory Stalls 3.4% of Pipeline Slots
Cache Stalls 8.4% of Cycles
DRAM Stalls 0.4% of Cycles
DRAM Bandwidth
Average 0.03 GB/s
Peak 6.34 GB/s
Bound 0%
NUMA 0% of Remote Accesses

Vectorization 0%
Instruction Mix
SP FLOPs 0% of uOps
DP FLOPs 0% of uOps
Non-FP 99.8% of uOps
FP Arith/Mem Rd Instr. Ratio 0.01
FP Arith/Mem Wr Instr. Ratio 0.01

intel

Intel® VTune™ Profiler

GROMACS

SYCL version

VTune GPU analysis

- Host API and GPU kernels
- Detailed source view on Kernels
- Bandwidth data with Hierarchical Memory Diagram
- Estimate for timing per source line (basic block timing)
- Estimate for memory latency per source line

SYCL version of GROMACS

- Clone master branch:
\$ git clone <https://gitlab.com/gromacs/gromacs.git>
- Workload: <https://www.mpinat.mpg.de/632209/benchMEM.zip>

```
SRC=$HOME/gromacs  
PRE=$HOME/GROMACS/  
  
cmake $SRC  
-DCMAKE_INSTALL_PREFIX=$PRE  
-DGMX_OPENMP=ON  
-DGMX_OPENMP_MAX_THREADS=128  
-DGMX_GPU=SYCL  
-DGMX_FFT_LIBRARY=mkl  
-DCMAKE_C_COMPILER=icx  
-DCMAKE_CXX_COMPILER=icpx
```

- Run:

```
 ${EXE_DIR}/gmx mdrun -ntmpi $RANKS -ntomp $OMP_NUM_THREADS -nb $MODE -pme cpu -s benchMEM.tpr -nsteps $steps
```

APS report 127.0.0.1:55001/u/NBODY-NE

file:///C:/cygwin64/home/hbockhor/INCOMING/22.02.15-EuroCC/aps_gromacs.html

67%

Application Performance Snapshot

Application: gromacs
Report creation date: 2022-02-16 05:09:15
OpenMP threads per Process: 12
HW Platform: Intel® Xeon® Processor code named CascadeLake
Frequency: 3.50 GHz
Logical Core Count per node: 24
Collector type: Event-based sampling driver/Driverless Perf system-wide counting

Performance Metrics:

- CPU: Elapsed Time **24.12 s**, IPC Rate **0.92**, SP GFLOPS **7.81**, DP GFLOPS **0.13**, GPU Utilization when Busy **7.85 s**
- GPU: Average CPU Frequency **3.59 GHz**

Serial Time Analysis: Your application has significant serial time that can prevent application scalability. Use OpenMP profiling tools like [Intel® VTune™ Profiler](#) to explore serial hotspots and opportunities on parallelization.

	Current run	Target	Tuning Potential
Serial Time	83.42% >15%	<10%	
OpenMP Imbalance	1.08% <10%		
Memory Stalls	20.65% >20%		
Vectorization	56.1% >70%		
GPU Utilization when Busy	58.2% >80%		

GPU Utilization when Busy: 58.2%
EU State % of EU
Active 58.2%
Idle 19.1%
Stalled 22.7%
GPU Occupancy 80.6% of Peak Value

Serial Time: 20.12 s, 83.42% of Elapsed Time

OpenMP Imbalance: 0.26 s, 1.08% of Elapsed Time

Memory Footprint:
Resident 882 MB
Virtual 34362707 MB

Memory Stalls: 20.6% of Pipeline Slots
Cache Stalls 14.8% of Cycles
DRAM Stalls 5.8% of Cycles
DRAM Bandwidth
Average 3.18 GB/s
Peak 14.62 GB/s
Bound 0%
NUMA 0% of Remote Accesses

Vectorization: 56.1%
Instruction Mix
SP FLOPs 7% of uOps
Packed: 55.8% from SP FP
128-bit: 11.6%
256-bit: 31.8%
512-bit: 12.3%
Scalar: 44.2% from SP FP
DP FLOPs 0.2% of uOps
Packed: 66% from DP FP
128-bit: 3.5%
256-bit: 62.6%
512-bit: 0%
Scalar: 34% from DP FP
Non-FP 92.7% of uOps
FP Arith/Mem Rd Instr. Ratio 0.26
FP Arith/Mem Wr Instr. Ratio 0.94

intel

Command lines used

- Basic gpu analysis:

```
vtune -collect gpu-hotspots -r <your-result-dir> -- <executable><args>
```

- Full instrumentation:

```
vtune -c gpu-hotspots -knob characterization-mode=instruction-count \
       -r <your-result-dir> -- <executable><args>
```

- Source Instrumentation with timing of basic blocks:

```
vtune -c gpu-hotspots -knob profiling-mode=source-analysis \
       -r <your-result-dir> -- <executable><args>
```

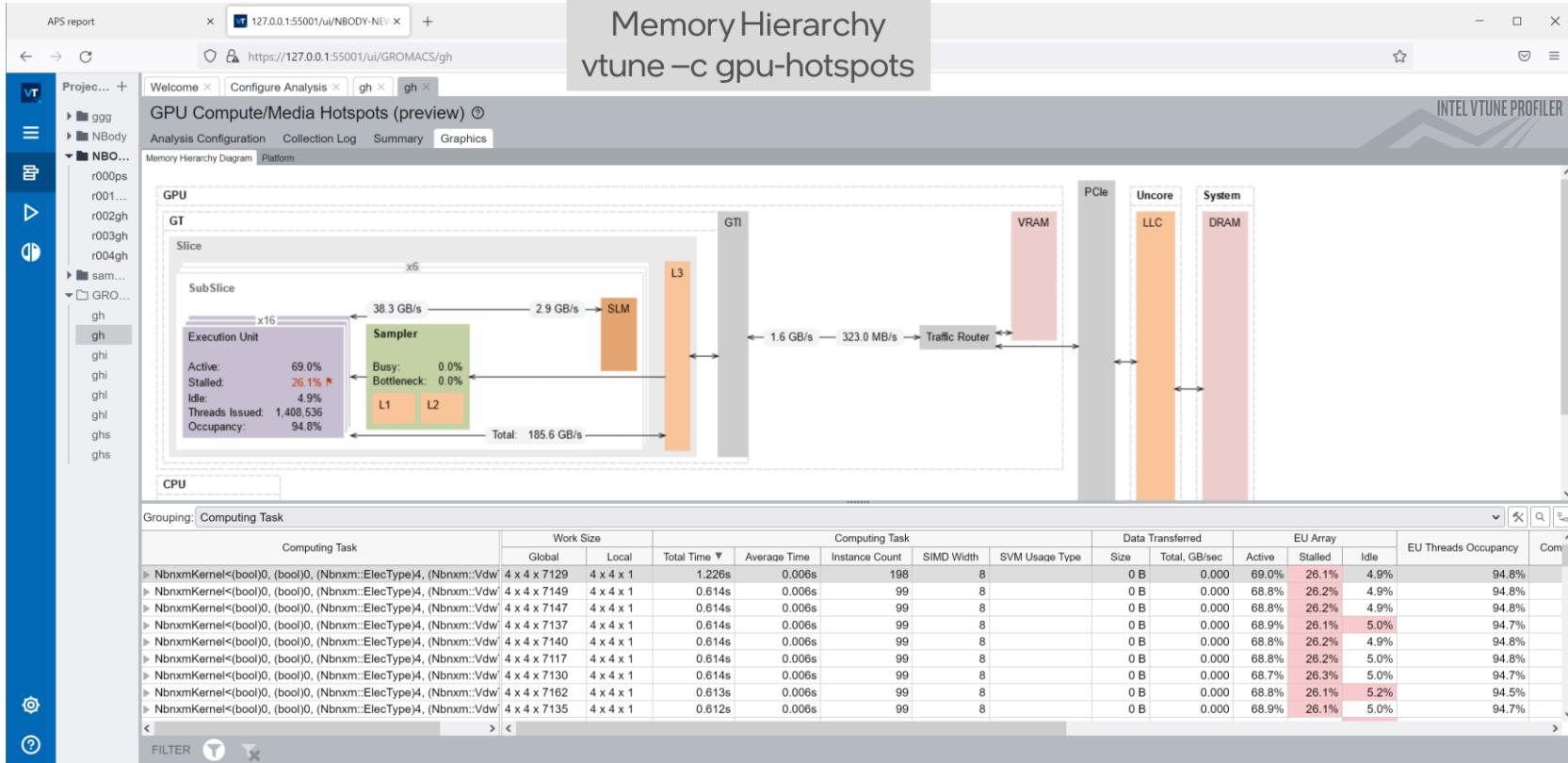
- Source Instrumentation with only memory inst. timed:

```
vtune -c gpu-hotspots -knob profiling-mode=source-analysis \
       -knob source-analysis=mem-latency -r ...
```

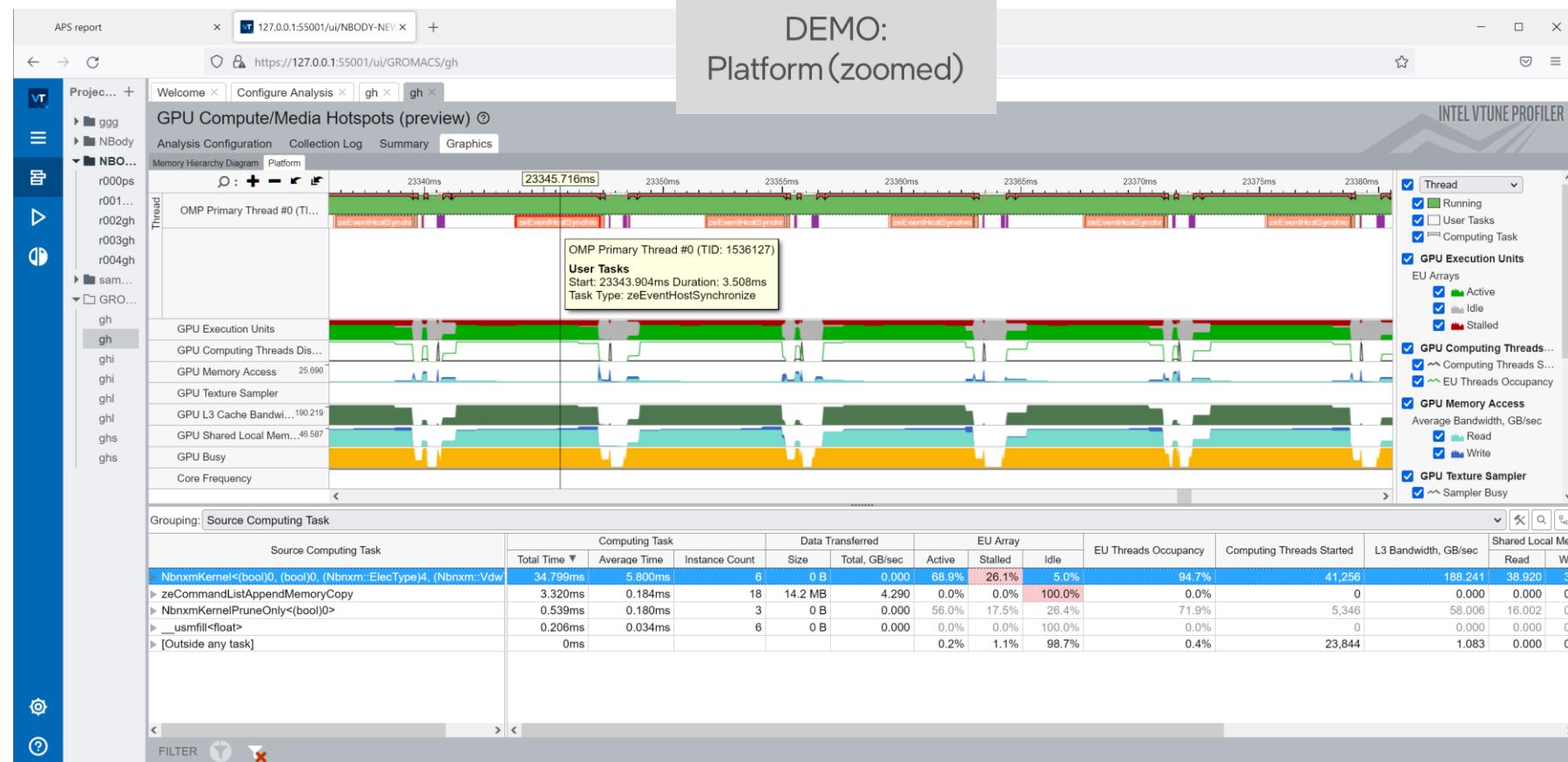
DEMO:

Memory Hierarchy

vtune -c gpu-hotspots



DEMO: Platform (zoomed)



DEMO:
characterization-mode
=instruction-count

APS report x 127.0.0.1:55001/ui/NBODY-NEV +
https://127.0.0.1:55001/ui/GROMACS/ghi

Project... + Welcome x Configure Analysis x gh x gh x ghi x GPU Compute/Media Hotspots (preview) ② ⓘ Analysis Configuration Collection Log Summary Graphics nbnxm_sycl_kernel.cpp

Source Assembly

Source Line ▲ Source

GPU Instructions Executed by Instruction Type

Control Fl... Se... Synchronizat... Int16 & HP Fl... Int32 & SP F

Line Number	Source Code	Instruction Type	Count
847	// cutoff & exclusion check	Control Flow	15,247,802,964
848		Se...	15,247,802,964
849	const bool notExcluded = doExclusionForces ? (nonSelfInteraction (ci != cj)) : (wexcl & maskJI);	Synchronization	15,247,802,964
850		Int16 & HP Flow	15,247,802,964
851		Int32 & SP Flow	15,247,802,964
852	// SYCL-TODO: Check optimal way of branching here.	Control Flow	3,158,901,217
853	if ((r2 < rCoulombSq) && notExcluded)	Se...	3,158,901,217
854	{	Synchronization	3,158,901,217
855	const float q1 = xqi[3];	Int16 & HP Flow	3,158,901,217
856	int atomTypeI; // Only needed if (!props.vdwComb)	Int32 & SP Flow	3,158,901,217
857	float sigma, epsilon;	Control Flow	89,057,962,872
858	Float2 c6c12;	Se...	89,057,962,872
859		Synchronization	89,057,962,872
860	if constexpr (!props.vdwComb)	Int16 & HP Flow	89,057,962,872
861	{	Int32 & SP Flow	89,057,962,872
862	/* LJ 6*C6 and 12*C12 */	Control Flow	89,057,962,872
863	atomTypeI = sm_atomTypeI[i * c_clsSize + tidxi];	Se...	89,057,962,872
864	c6c12 = a_nbfp[numTypes * atomTypeI + atomTypeJ];	Synchronization	89,057,962,872
865	}	Int16 & HP Flow	89,057,962,872
866	else	Int32 & SP Flow	89,057,962,872
867	{	Control Flow	89,057,962,872
868	const Float2 ljCombI = sm_ljCombI[i * c_c1Size + tidxi];	Se...	89,057,962,872
869	if constexpr (props.vdwCombGeom)	Synchronization	89,057,962,872
870	{	Int16 & HP Flow	89,057,962,872
871	c6c12 = Float2(ljCombI[0] * ljCombJ[0], ljCombI[1] * ljCombJ[1]);	Int32 & SP Flow	89,057,962,872
872	}	Control Flow	89,057,962,872
873	else	Se...	89,057,962,872
874	{	Synchronization	89,057,962,872
875	static_assert(props.vdwCombLB);	Int16 & HP Flow	89,057,962,872
876	// ... other code here ...	Int32 & SP Flow	89,057,962,872

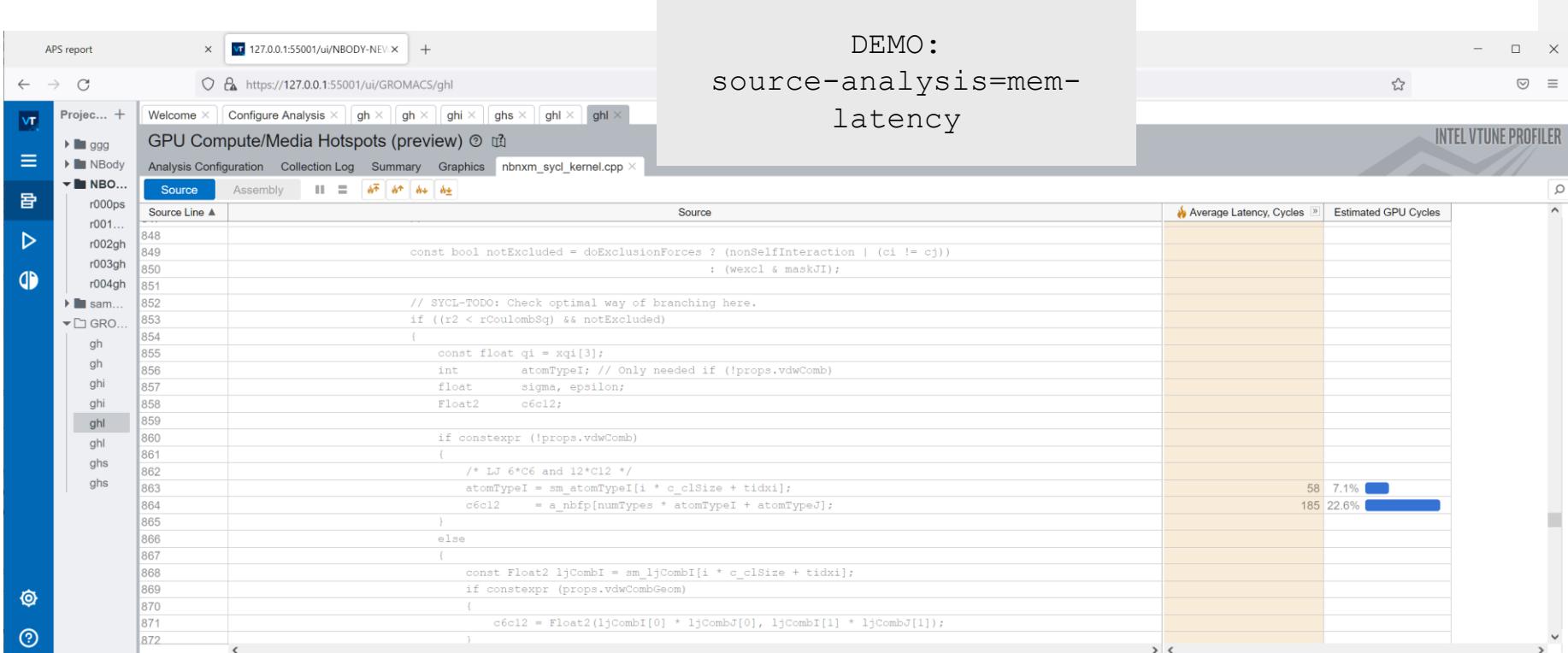
intel

DEMO: profiling-mode=source- analysis

The screenshot shows the Intel VTune Profiler interface. On the left, there's a sidebar with project navigation and a toolbar with various icons. The main area has tabs for 'Welcome', 'Configure Analysis', 'gh', 'ghx', 'ghi', and 'ghs'. Below these tabs, the title is 'GPU Compute/Media Hotspots (preview)'. Underneath, there are tabs for 'Analysis Configuration', 'Collection Log', 'Summary', and 'Graphics'. The 'Source' tab is selected, and the file 'nbnxm_sycl_kernel.cpp' is open. The code editor displays the following C++ code:

```
Source Line ▲ Source Estimated GPU Cycles
852 // SYCL-TODO: Check optimal way of branching here.
853 if ((r2 < rCoulombSq) && notExcluded)
854 {
855     const float qi = xqi[3];
856     int atomTypeI; // Only needed if (!props.vdwComb)
857     float sigma, epsilon;
858     Float2 c6c12;
859
860     if constexpr (!props.vdwComb)
861     {
862         /* LJ 6*C6 and 12*C12 */
863         atomTypeI = sm_atomTypeI[i * c_clSize + tidxi];
864         c6c12 = a_nbfp[numTypes * atomTypeI + atomTypeJ];
865     }
866     else
867     {
868         const Float2 ljCombI = sm_ljCombI[i * c_clSize + tidxi];
869         if constexpr (props.vdwCombGeom)
870         {
871             c6c12 = Float2(ljCombI[0] * ljCombJ[0], ljCombI[1] * ljCombJ[1]);
872         }
873         else
874         {
875             static_assert(props.vdwCombLB);
876             // LJ 2*(1/6)*sigma and 12*epsilon
877             sigma = ljCombI[0] + ljCombJ[0];
878             epsilon = ljCombI[1] * ljCombJ[1];
879             if constexpr (doCalcEnergies)
880             {
881                 c6c12 = convertSigmaEpsilonToC6C12(sigma, epsilon);
882             }
883         }
884     }
885 }
```

The line 864, which contains the assignment to 'c6c12', is highlighted with a thick blue bar, indicating it is the current source being analyzed. To the right of the code, there's a column for 'Estimated GPU Cycles' with values: 1.1% (for line 853), 1.4% (for line 862), and 17.3% (for line 864). The rest of the lines have no visible bars.



Documentation

VTune cookbooks:

<https://www.intel.com/content/www/us/en/develop/documentation/vtune-cookbook/top.html>

Vtune on DevCloud:

<https://www.intel.com/content/www/us/en/develop/documentation/vtune-cookbook/top/configuration-recipes/using-vtune-server-with-vs-code-intel-devcloud.html>

GPU profiling:

<https://www.intel.com/content/www/us/en/develop/documentation/vtune-help/top/analyze-performance/accelerators-group/gpu-compute-media-hotspots-analysis.html>

Notices & Disclaimers

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

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

Your costs and results may vary.

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

Intel does not control or audit third-party data. You should consult other sources to evaluate accuracy.

© 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.

The Intel logo is displayed in white against a solid blue background. The word "intel" is written in a lowercase, sans-serif font. A small, solid cyan square is positioned above the top of the letter "i". The letter "t" has a vertical stroke on its left side. The letter "e" has a vertical stroke on its right side. The letter "l" has a vertical stroke on its left side. A registered trademark symbol (®) is located at the bottom right of the letter "l".

intel®