

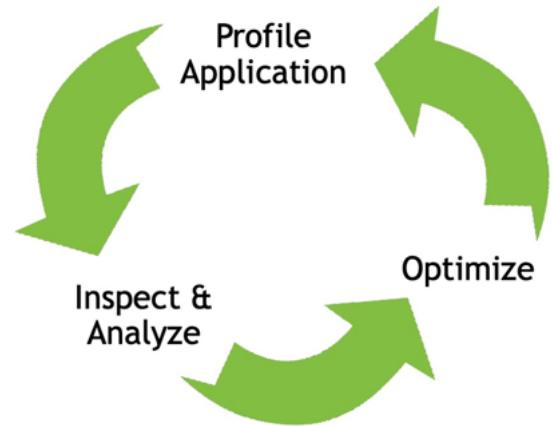
Profilers

Lecture 14

Sunita Chandrasekaran
Associate Professor, University of Delaware
PDC Summer School
Aug 2023

Why profile?

- Identify compute intensive portions in the code
- Generate an overall profile of the runs
- Identify performance bottleneck
- Identify memory usage or memory leak



GNU GPROF profiler

- How to use gprof
 - Using the gprof tool is not at all complex.
 - Have profiling enabled while compiling the code
 - Execute the program code to produce the profiling data
 - Run the gprof tool on the profiling data file (generated in the step above).

Sample test codes in canvas Files->code

Flags to enable (GCC)

- -pg : Generate extra code to write profile information suitable for the analysis program gprof.
 - You must use this option when compiling the source files you want data about, and you must also use it when linking.
- \$ gcc -Wall -pg test.c new_func.c -o test_gprof
- \$ ls
- \$./test_gprof
- \$ gprof test_gprof gmon.out > analysis.txt
- \$ ls

Flat profile:

Each sample counts as 0.01 seconds.

% time	cumulative seconds	self seconds	calls	self s/call	total s/call	name
34.19	7.42	7.42	1	7.42	14.72	func1
33.67	14.72	7.31	1	7.31	7.31	new_func1
33.62	22.02	7.30	1	7.30	7.30	func2
0.14	22.05	0.03				main

% time the percentage of the total running time of the program used by this function.

cumulative seconds a running sum of the number of seconds accounted for by this function and those listed above it.

self seconds the number of seconds accounted for by this function alone. This is the major sort for this listing.

calls the number of times this function was invoked, if this function is profiled, else blank.

self ms/call the average number of milliseconds spent in this function per call, if this function is profiled, else blank.

total ms/call the average number of milliseconds spent in this function and its descendants per call, if this function is profiled, else blank.

name the name of the function. This is the minor sort for this listing. The index shows the location of the function in the gprof listing. If the index is in parenthesis it shows where it would appear in

Profilers

- NVIDIA Nsight Compute and Nsight systems
- Tools Analysis Utilities (TAU)
- Score-P/Vampir
- Arm MAP
- HPCtoolkit
- AMD rocProf
- Intel Vtune

NSIGHT PRODUCT FAMILY

Standalone Performance Tools

Nsight Systems - System-wide application algorithm tuning

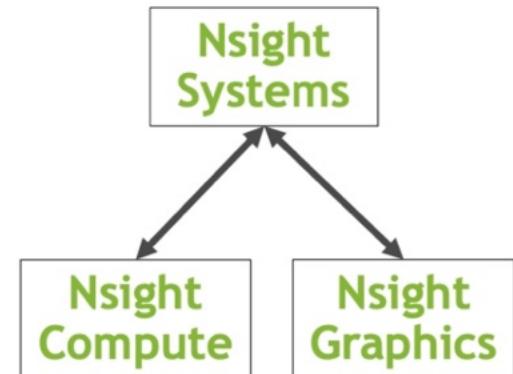
Nsight Compute - Debug/optimize specific CUDA kernel

Nsight Graphics - Debug/optimize specific graphics shader

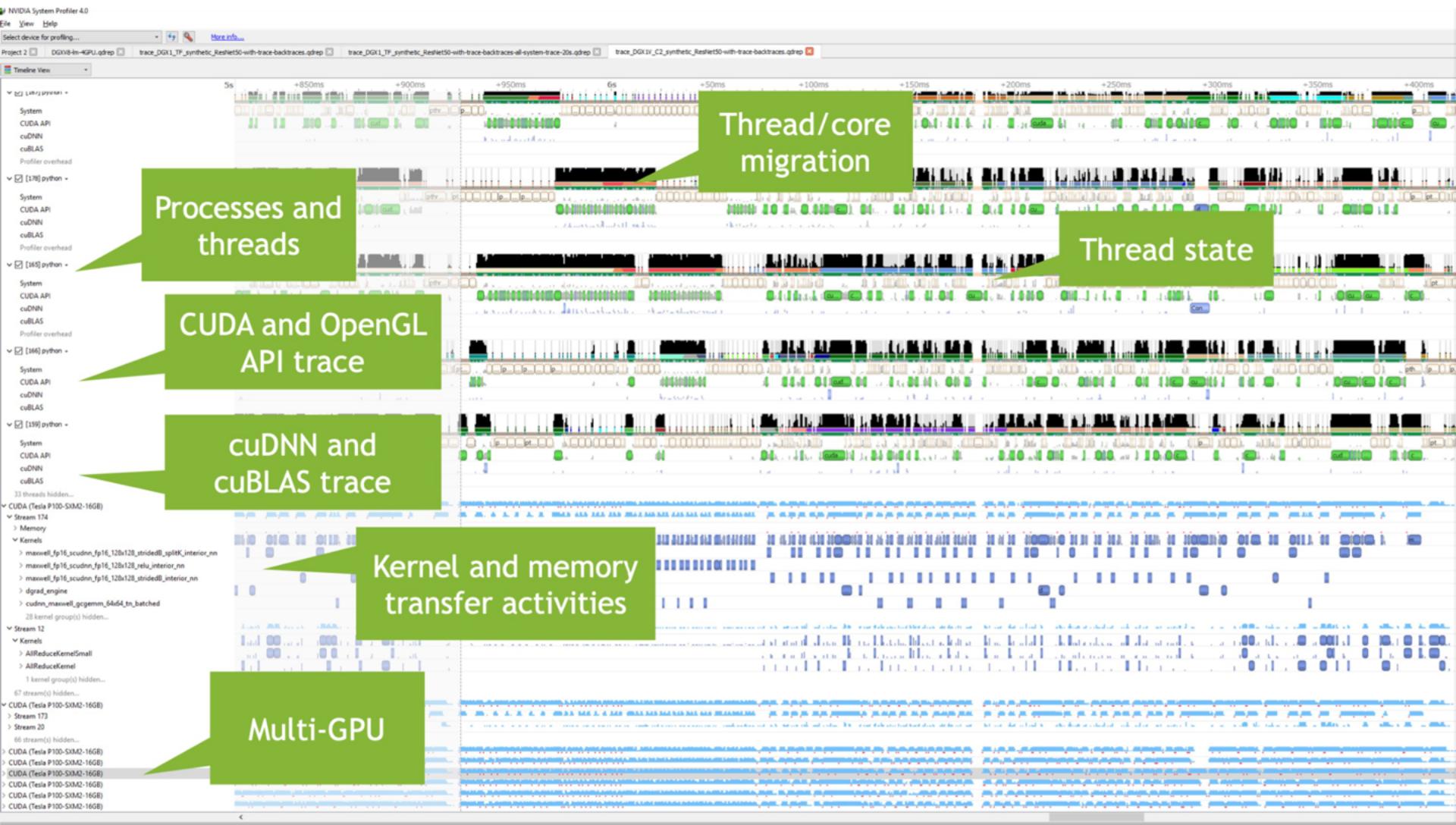
IDE Plugins

Nsight Eclipse Edition/Visual Studio - editor, debugger, some perf analysis

Workflow



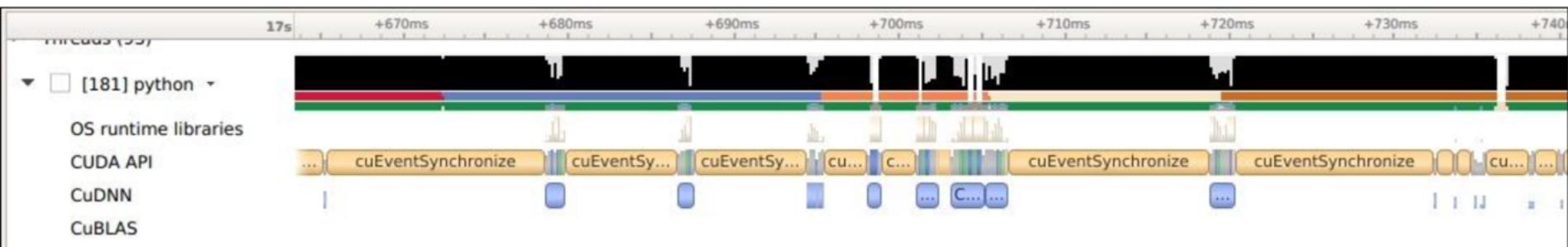
Nsight Systems



CPU functionalities

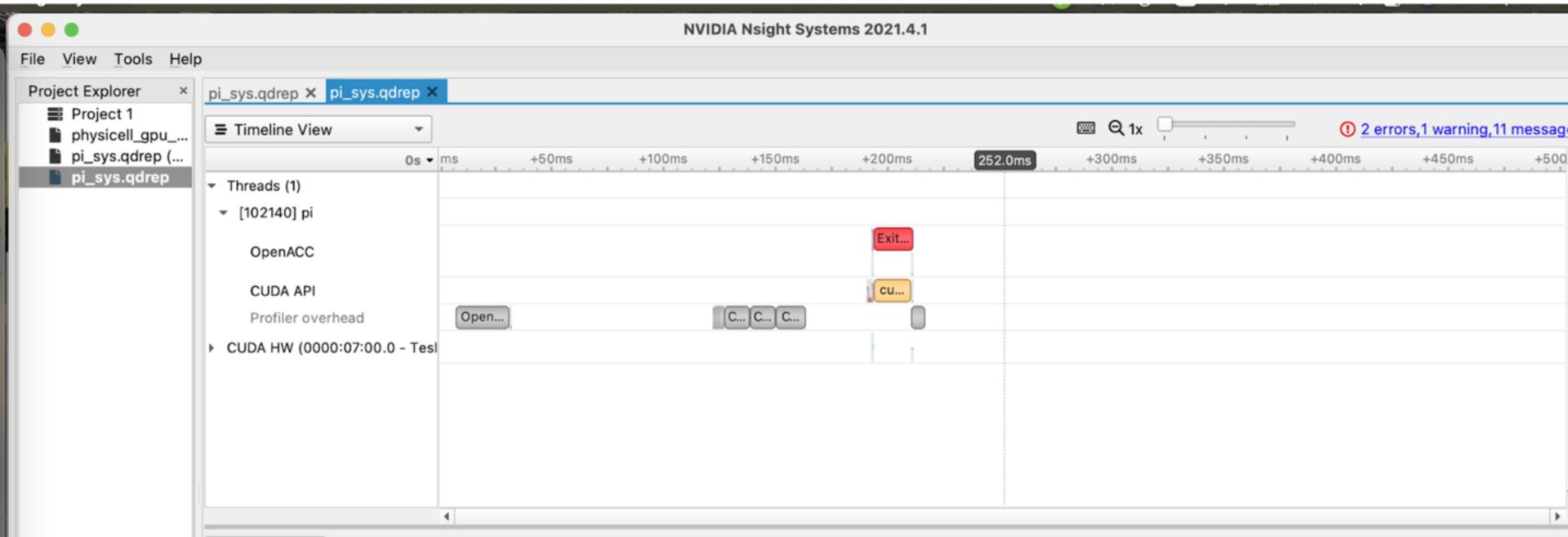
Get an overview of each thread's activities

- Which core the thread is running and the utilization
- CPU state and transition
- OS runtime libraries usage: pthread, file I/O, etc.
- API usage: CUDA, cuDNN, cuBLAS, TensorRT, ...



GPU functionality

- Trace OpenACC/CUDA API calls
- See when kernels are dispatched





NVIDIA NSIGHT COMPUTE

Next-Gen Kernel Profiling Tool

Key Features:

- Interactive CUDA API debugging and kernel profiling
- Graphical profile report
- Comparison of multiple kernel reports
- Fully Customizable (Reports and Analysis Rules)
- Command Line, Standalone, IDE Integration

OS: Linux, Windows, ARM, MacOSX (host only)

GPUs: Pascal (GP10x), Volta, Turing

Docs/product: <https://developer.nvidia.com/nsight-compute>

File Connection Debug Profile Tools Window Help

Connect Disconnect Terminate Profile Kernel

API Stream

11236 > vectorAdd

API Stream

Next Trigger: vector

ID	API Name	Details	Fu
: 186	cuDeviceGetAttribute		CU
: 187	cuDeviceGetAttribute		CU
: 188	cuDeviceGetAttribute		CU
: 189	cuDeviceGetAttribute		CU
: 190	cuDeviceGetAttribute		CU
: 191	cuDeviceGetAttribute		CU
: 192	cudaMalloc		CU
: 193	cuCtxGetCurrent		CU
: 194	cuCtxSetCurrent		CU
: 195	cuDevicePrimaryCtxRe_		CU
: 196	cuCtxGetCurrent		CU
: 197	cuCtxGetDevice		CU
: 198	cuModuleGetFunction		CU
: 199	cuMemAlloc_v2		CU
: 200	cudaMalloc		CU
: 201	cuMemAlloc_v2		CU
: 202	cudaMalloc		CU
: 203	cuMemAlloc_v2		CU
: 204	cudaMemcpy		CU
: 205	cuMemcpyHtoD_v2		CU
: 206	cudaMemcpy		CU
: 207	cuMemcpyHtoD_v2		CU
: 208	cudaConfigureCall		CU
: 209	cudaSetupArgument		CU
: 210	cudaSetupArgument		CU
: 211	cudaSetupArgument		CU
: 212	cudaSetupArgument		CU
213	cudaLaunch		CU
214	cuLaunchKernel		CU
215	vectorAdd	vectorAdd	CU

Sections/Rules Info Reload Enable All Disable All

Enter filter

Name	Pri
Memory Workload Analysis	7
Memory Workload Analysis Chart	7
Memory Workload Analysis Tables	7
Scheduler Statistics (1)	8
Warp State Statistics (17)	9
Instruction Statistics	10

Sections/Rules Info API Statistics NVTX

GPU SOL section

Launch: 0 - 215 - vectorAdd (50176, 1, 1)

Frequency: 883.21 cycle/usecond CC: 7.0 Process: [11236] vectorAdd

Copy as Image

GPU Utilization

Speed Of Light [%]	SM [%]	Memory [%]
0.0 - 10.0	3.72	38.46
10.0 - 20.0		
20.0 - 30.0		
30.0 - 40.0		
40.0 - 50.0		
50.0 - 60.0		
60.0 - 70.0		
70.0 - 80.0		
80.0 - 90.0		
90.0 - 100.0		

Recommendations

Bottleneck [Warning] This kernel grid is too small to fill the available resources on this device. Look at 'Launch Statistics' for more details.

Compute Workload Analysis

Detailed analysis of the compute resources of the streaming multiprocessors (SM), including the achieved instructions per clock (IPC) and the utilization of each available pipeline. Pipelines with very high utilization might limit the overall performance.

Executed Ipc Elapsed [inst/cycle]	6.61
Executed Ipc Active [inst/cycle]	3.67
Received Ipc Active [inst/cycle]	4.48

Memory workload analysis section

Memory Workload Analysis

Detailed analysis of the memory resources of the GPU. Memory can become a limiting factor for the overall kernel performance when fully utilizing the involved hardware units (Mem Busy), exhausting the available communication bandwidth between those units (Max Bandwidth), or by reaching the maximum throughput of issuing memory instructions (Mem Pipes Busy). Detailed chart of the memory units. Detailed tables with data for each memory unit.

Memory Throughput [Gbyte/second]	185.04	Mem Busy [%]	9.77
L1 Hit Rate [%]	0	Max Bandwidth [%]	38.46
L2 Hit Rate [%]	34.75	Mem Pipes Busy [%]	2.32

Scheduler Statistics

Summary of the activity of the schedulers issuing instructions. Each scheduler maintains a pool of warps that it can issue instructions for. The upper bound of warps in the pool (Theoretical Warps) is limited by the launch configuration. On every cycle each scheduler checks the state of the allocated warps in the pool (Active Warps). Active warps that are not stalled (Eligible Warps) are ready to issue their next instruction. From the set of eligible warps the scheduler selects a single warp from which to issue one or more instructions (Issued Warp). On cycles with no eligible warps, the issue slot is skipped and no instruction is issued. Having many skipped issue slots indicates poor latency hiding.

Active Warps Per Scheduler [warp]	4.26	Instructions Per Active Issue Slot [inst/cycle]	96.17
Eligible Warps Per Scheduler [warp]	0.05	No Eligible [%]	3.83
Issued Warp Per Scheduler	0.04	One or More Eligible [%]	

Warp State Statistics

Analysis of the states in which all warps spent cycles during the kernel execution. The warp states describe a warp's readiness or inability to issue its next instruction. The warp cycles per instruction define the latency between two consecutive instructions. The higher the value, the more warp parallelism is required to hide this latency. For each warp state, the chart shows the average number of cycles spent in that state per issued instruction. Stalls are not always impacting the overall performance nor are they completely avoidable. Only focus on stall reasons if the schedulers fail to issue every cycle.

Warp Cycles Per Issued Instruction	111.19	Avg. Active Threads Per Warp	31.99
------------------------------------	--------	------------------------------	-------

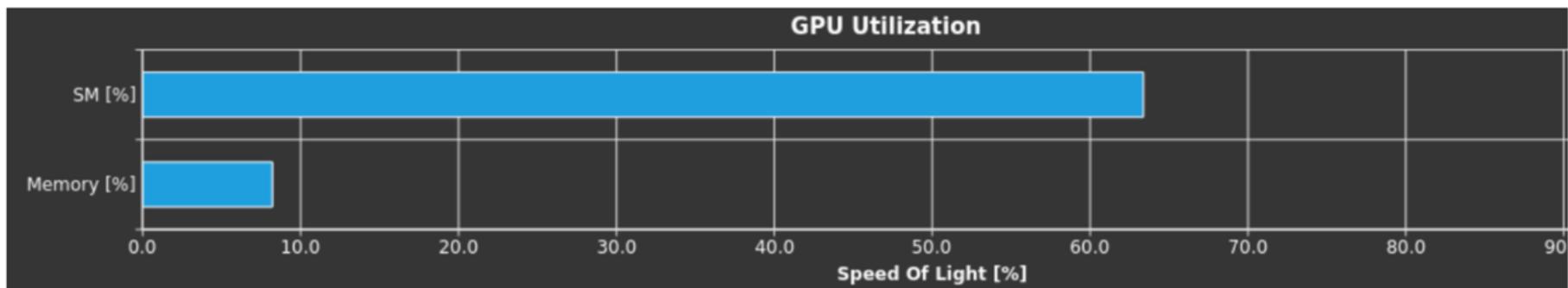
vectorAdd [11236]

SOL SECTION

Sections

SOL Section (case 1: Compute Bound)

- High-level overview of the utilization for compute and memory resources of the GPU. For each unit, the Speed Of Light (SOL) reports the achieved percentage of utilization with respect to the theoretical maximum

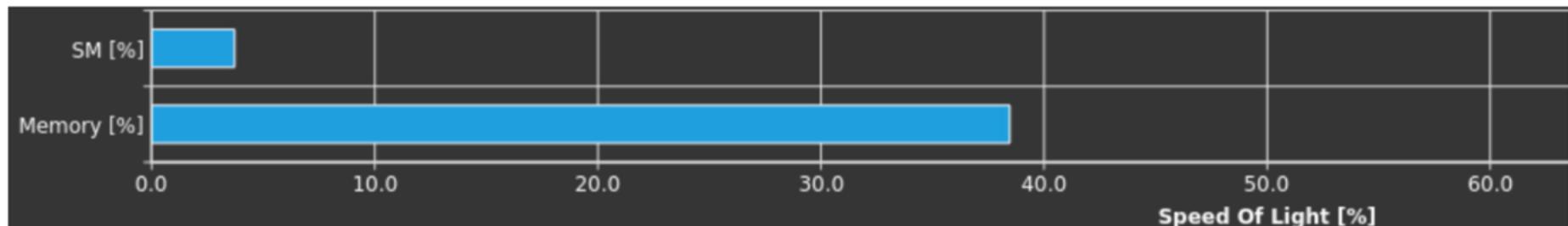


SOL SECTION

Sections

SOL Section (case 2: Latency Bound)

- High-level overview of the utilization for compute and memory resources of the GPU. For each unit, the Speed Of Light (SOL) reports the achieved percentage of utilization with respect to the theoretical maximum

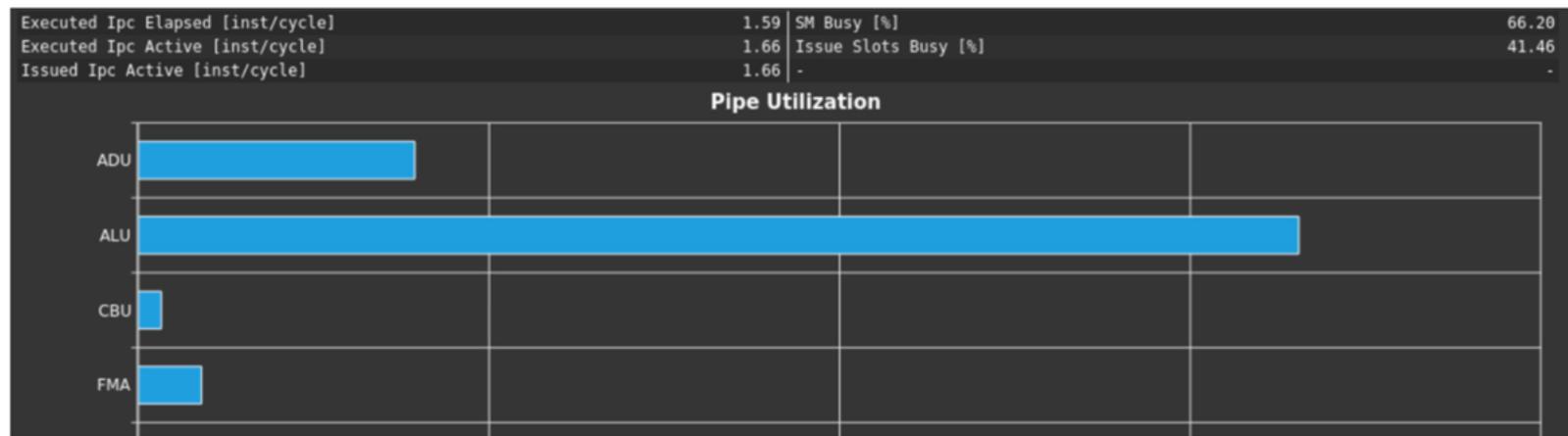


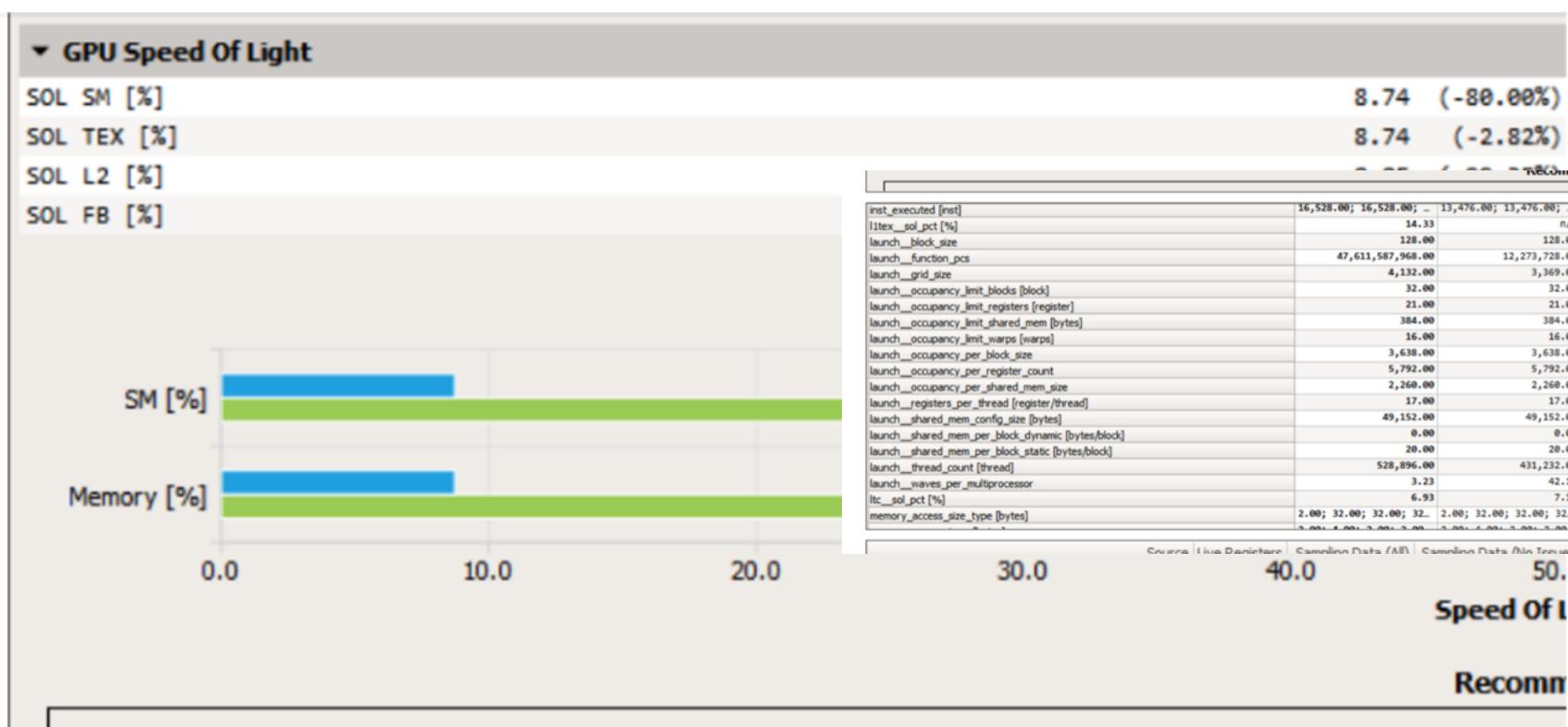
COMPUTE WORKLOAD ANALYSIS

Sections

Compute Workload Analysis (case 1)

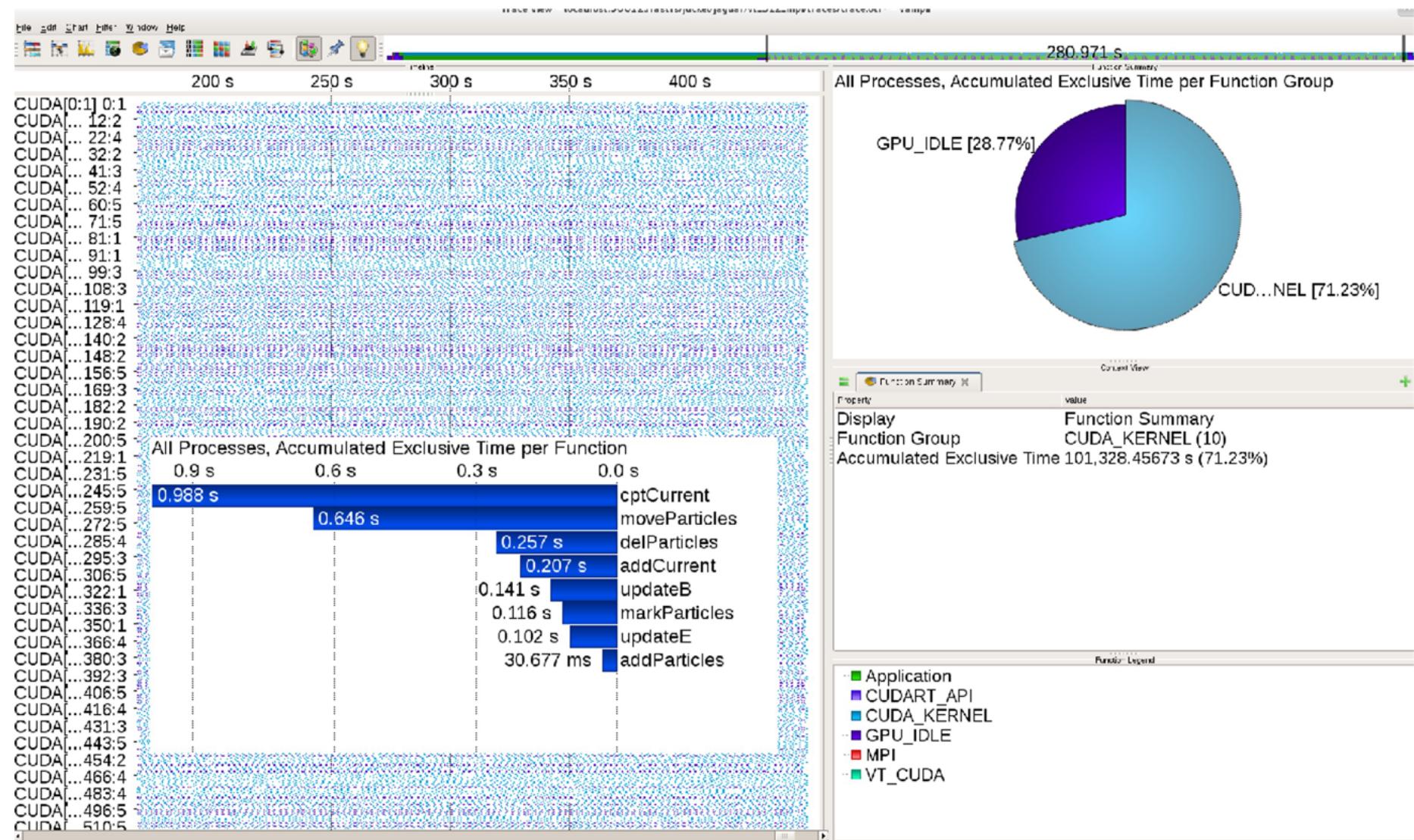
- Detailed analysis of the compute resources of the streaming multiprocessors (SM), including the achieved instructions per clock (IPC) and the utilization of each available pipeline. Pipelines with very high utilization might limit the overall performance



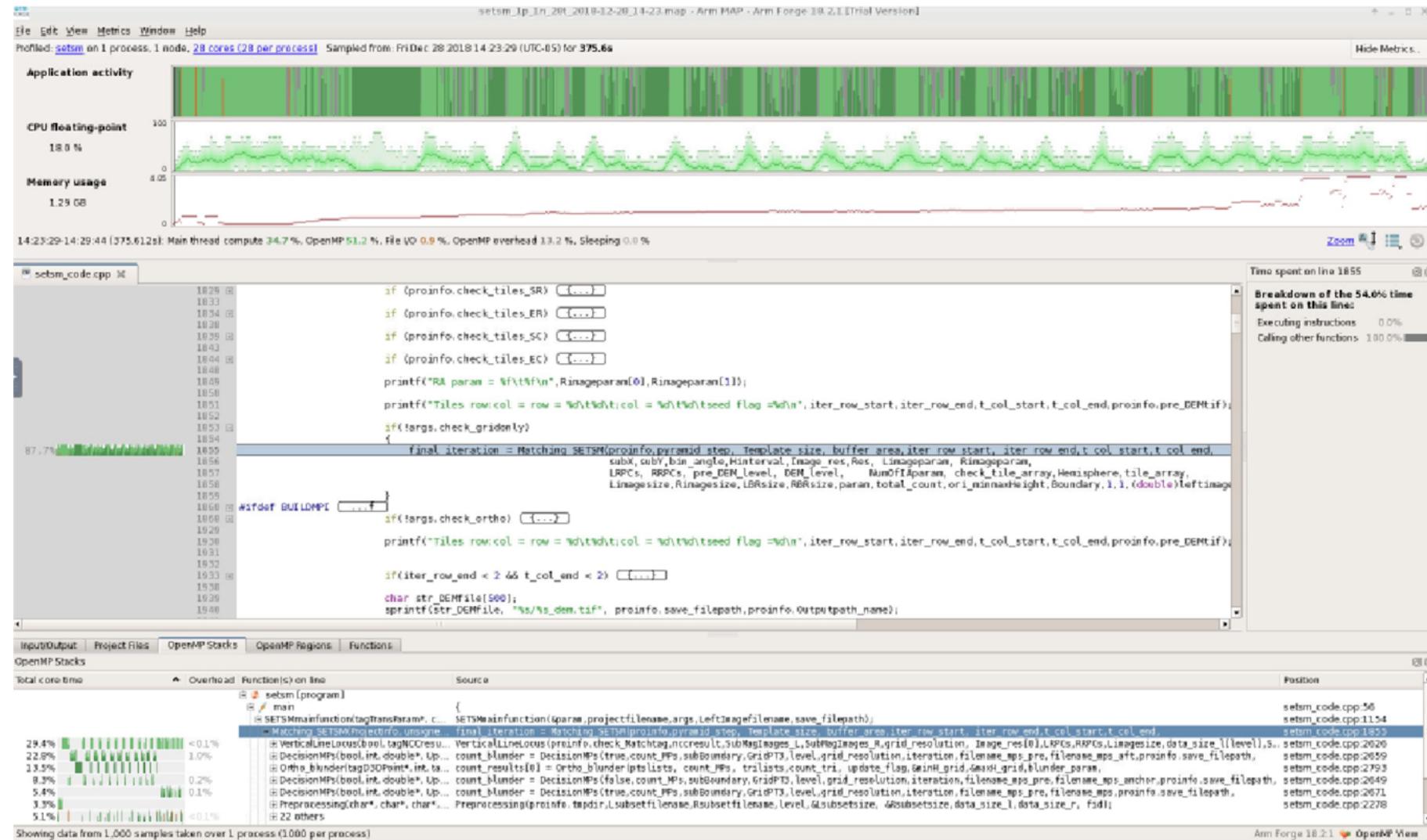


	Source	Live Registers	Sampling Data (All)	Sampling Data (No Issue)
@!PT	SHFL.IDX PT, RZ, RZ, RZ, RZ;	0	223	0
	MOV R1, c[0x0][0x28];	1	13	44
	S2R R0, SR_CTAID.X;	2	143	75
	S2R R2, SR_TID.X;	3	0	38
	IMAD R0, R0, c[0x0][0x0], R2;	3	599	94
	ISETP.GE.AND P0, PT, R0, c[0x0][0x170]	2	125	26
@P0	EXIT;	2	259	86
	MOV R2, R0;	3	386	29
@!PT	SHFL.IDX PT, RZ, RZ, RZ, RZ;	2	0	0
	MOV R4, 0x4;	3	0	0
	IMAD.WIDE R4, R2, R4, c[0x0][0x160];	4	0	0
	LDG.E.SYS R3, [R4];	3	0	0
	BSSY B0, 0xb00976780;	3	0	0
	SHF.R.S32.HI R0, RZ, 0x1f, R2;	4	0	0

Score-P Vampir Trace



Arm MAP



Arm MAP

Summary: wave_c is **Compute-bound** in this configuration

Compute 99.9%



Time spent running application code. High values are usually good. This is **very high**; check the CPU performance section for advice.

MPI 0.1%

Time spent in MPI calls. High values are usually bad. This is **very low**; this code may benefit from a higher process count.

I/O 0.0%

Time spent in filesystem I/O. High values are usually bad. This is **negligible**; there's no need to investigate I/O performance.

This application run was **Compute-bound**. A breakdown of this time and advice for investigating further is in the **CPU** section below.

As very little time is spent in **MPI** calls, this code may also benefit from running at larger scales.

CPU

A breakdown of the **99.9%** CPU time:

Scalar numeric ops 16.0%

Vector numeric ops 0.0%

Memory accesses 25.2%

The per-core performance is memory-bound. Use a profiler to identify time-consuming loops and check their cache performance.

No time is spent in **vectorized instructions**. Check the compiler's vectorization advice to see why key loops could not be vectorized.

MPI

A breakdown of the **0.1%** MPI time:

Time in collective calls 100.0%

Time in point-to-point calls 0.0%

Effective process collective rate 8.92 kB/s

Effective process point-to-point rate 0.00 bytes/s

Most of the time is spent in **collective calls** with a very low transfer rate. This suggests load imbalance is causing synchronization overhead; use an MPI profiler to investigate.

I/O

A breakdown of the **0.0%** I/O time:

Time in reads 0.0%

Time in writes 0.0%

Effective process read rate 0.00 bytes/s

Effective process write rate 0.00 bytes/s

No time is spent in **I/O** operations. There's nothing to optimize here!

Threads

A breakdown of how multiple threads were used:

Computation 0.0%

Synchronization 0.0%

Physical core utilization 5.0%

System load 5.1%

No measurable time is spent in multithreaded code.

Physical core utilization is low. Try increasing the number of processes to improve performance.

Memory

Per-process memory usage may also affect scaling:

Mean process memory usage 66.8 MB

Peak process memory usage 74.9 MB

Peak node memory usage 3.0%

The peak node memory usage is very low. Running with fewer MPI processes and more data on each process may be more efficient.

TAU (Tools Analysis and Utilities)



HPCToolkit

