Лекция 3

Инструменты профилирования:

- nvprof
- Nsight Compute CLI
- nvvp
- Nsight Compute

nvprof и Nsight Compute CLI

```
__global__ void gInit(float* a, float* b){
  int i=threadIdx.x+blockIdx.x*blockDim.x;
  a[i]=(float)2*i;
  b[i]=(float)(2*i+1);
}

__global__ void gSum(float* a, float *b){
  int i=threadIdx.x+blockIdx.x*blockDim.x;
  a[i]+=b[i];
}
```

Тестовые ядра

```
ip-011@linux-47dw:/home/malkov/WORKSHOP/PGP-2023>
nvprof ./lab3c
```

```
Type  Time(%) Time Calls Avg Min Max Name
GPU activities:
43.59%  2.1760us  1  2.1760us  2.1760us
                                 2.1760us  gSum(int*, int*)
41.67%  2.0800us  1  2.0800us  2.0800us
                                  2.0800us  gInit(int*, int*)
14.74%     736ns  1  736ns   736ns   736ns  [CUDA memcpy DtoH]
API calls:
98.87%  131.54ms  2  65.772ms  6.9650us  131.54ms  cudaMalloc
…………………………………………………………………………………………………..
0.09%  124.46us   2  62.229us  10.561us  113.90us  cudaFree
……………………………………………………………………………………………………
0.01%  14.599us   1  14.599us  14.599us  14.599us

                                     cudaMemcpy
```

```
/Lecture3/Lab3-cuda-gdb # ncu --target-processes all  ./lab3c

 gInit(int *, int *), 2023-Feb-13 15:09:06, Context 1, Stream 7
   Section: GPU Speed Of Light Throughput
--------------------------------------------------------
DRAM Frequency cycle/nsecond                6.40
SM Frequency   cycle/nsecond                1.29
Elapsed Cycles         cycle                3,327
Memory [%]                 %                1.10
DRAM Throughput            %                0.02
Duration             usecond                2.56
 --------------- ----------------------------
   WRN   This kernel grid is too small to fill the available
resources on this device, resulting in only 0.0 full
waves across all SMs. Look at Launch Statistics for more
details.
```

```
/Lecture3/Lab3-cuda-gdb # ncu
--metrics gpu__time_duration.sum  ./lab3c
```

```
gInit(int *, int *), 2023-Feb-13 18:42:52, Context 1, Stream 7
   Section: Command line profiler metrics
   ------ --------------- ------------------------------
gpu__time_duration.sum   usecond                    29.50
-------------------------------------------------------------
gSum(int *, int *), 2023-Feb-13 18:42:52, Context 1, Stream 7
   Section: Command line profiler metrics
-------------------------------------------------------------
gpu__time_duration.sum   usecond                     37.57
----------------------------------------------------------------
```

```
/Lecture3/Lab3-cuda-gdb> nvprof --query-metrics
======== Warning: Skipping profiling on device 0 since
profiling is not supported on devices with compute
capability 7.5 and higher.
      Use NVIDIA Nsight Compute for GPU profiling and NVIDIA
Nsight Systems for GPU tracing and CPU sampling.
        Refer https://developer.nvidia.com/tools-overview
for more details.
```

```
ip-011@linux-47dw:/home/malkov/WORKSHOP/PGP-2023>
nvprof --query-metrics | less
```

```
Available Metrics:   Name                     Description
Device 0 (GeForce GTX 1050):
inst_per_warp:  Average number of instructions executed by each
warp

warp_execution_efficiency:  Ratio of the average active threads
per warp to the maximum number of
threads per warp supported on a multiprocessor
…………………………………………………………………………………………………………………………………………………….
gld_transactions_per_request:  Average number of global memory
load transactions performed for each global memory load.

gst_transactions_per_request:  Average number of global memory
store transactions performed for each global memory store
………………………………………………………………………………………………………………………………………………………
```

```
ip-011@linux-47dw:/home/malkov/WORKSHOP/PGP-2023>
nvprof -m gst_throughput ./lab3c
```

```
Invocations Metric Name  Metric Description  Min   Max   Avg
Device "GeForce GTX 1050 (0)"
   Kernel: gSum(int*, int*)
 1       gst_throughput  Global Store Throughput  40.582MB/s
                                      40.582MB/s  40.582MB/s
   Kernel: gInit(int*, int*)
 1       gst_throughput  Global Store Throughput  71.303MB/s
                                      71.303MB/s  71.302MB/s
```

```
/Лекция4/lab4> ncu --list-sections
Identifier                Display Name              Enabled          Filename


-------------------------------------------
ComputeWorkloadAnalysis   Compute Workload Analysis   yes  ...2024.2.1/Sections/


                                                                 ComputeWorkloadAnalysis.section
InstructionStats      Instruction Statistics        yes      ...2024.2.1/Sections/
                                                                      InstructionStatistics.section
LaunchStats           Launch Statistics             yes   ...2024.2.1/Sections/
                                                                          LaunchStatistics.section
MemoryWorkloadAnalysis   Memory Workload Analysis   yes …
……………………………………………………………………………
```

```
/Лекция4/lab4> ncu --section InstructionStats  ./lab4c
 gInit(float *, float *) (2, 1, 1)x(128, 1, 1), Context 1, Stream 7, Device 0, CC 7.5
   Section: Instruction Statistics
   ---------------------------------------- ----------- ------------
   Metric Name                              Metric Unit Metric Value
   ---------------------------------------- ----------- ------------
   Avg. Executed Instructions Per Scheduler       inst         0,93
   Executed Instructions                          inst          112
   Avg. Issued Instructions Per Scheduler          inst         1,27
   Issued Instructions                             inst          152
   ---------------------------------------- ----------- ------------
gSum(float *, float *) (2, 1, 1)x(128, 1, 1), Context 1,
Stream 7, Device 0, CC 7.5
   Section: Instruction Statistics
_-------------------------------------------------------
```

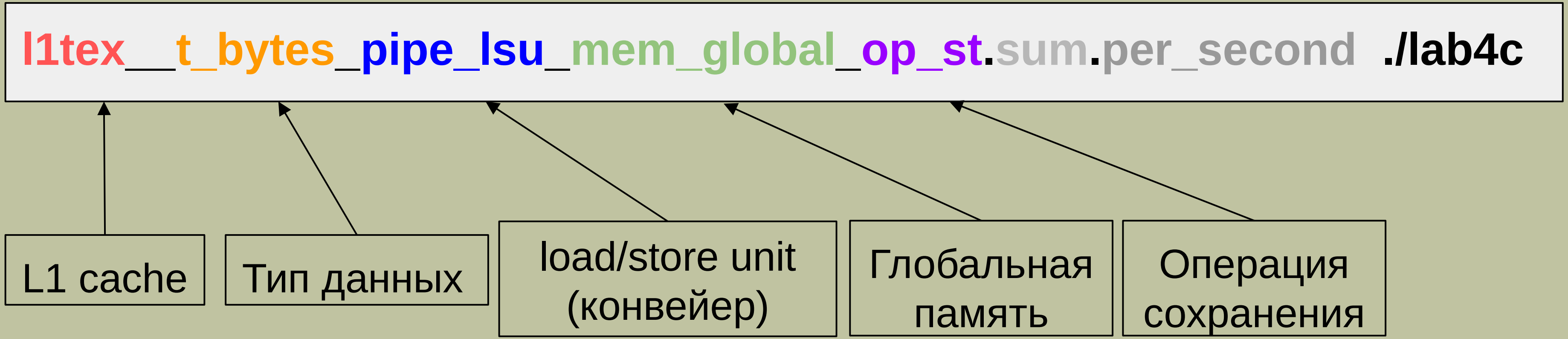
```
/Лекция4/lab4> ncu --section ComputeWorkloadAnalysis ./lab4c
  gSum(float *, float *) (2, 1, 1)x(128, 1, 1), Context 1, Stream 7, Device 0, CC 7.5
    Section: Compute Workload Analysis
    -------------------- ----------- ------------
    Metric Name          Metric Unit Metric Value
    -------------------- ----------- ------------
    Executed Ipc Active   inst/cycle         0,04
    Executed Ipc Elapsed  inst/cycle         0,00
    Issue Slots Busy               %         1,35
    Issued Ipc Active     inst/cycle         0,05
    SM Busy                        %         1,35
    -------------------- ----------- ------------

    OPT   Est. Local Speedup: 99.33%
          All compute pipelines are under-utilized. Either this kernel is very small
or it doesn't issue enough warps  per scheduler. Check the Launch Statistics and
Scheduler Statistics sections for further details.
```

```
/Лекция4/lab4> ncu  --query-metrics > metrics.txt
Device NVIDIA GeForce RTX 2060 (TU104)
------------------------------------------------------------------ ------------- -------------
Metric Name                           Metric Type     Metric Unit       Metric Description


------------------------------------------------------------------ ------------- -------------
dram__bytes                               Counter        byte          # of bytes accessed in DRAM

dram__bytes_read                          Counter        byte          # of bytes read from DRAM

dram__bytes_write                         Counter        byte          # of bytes written to DRAM

………………………………………………………………………..
smsp__average_inst_executed_pipe_lsu_per_warp    Ratio     inst/warp
                                           average # of instructions executed by pipe lsu per warp
……………………………………………………………………………………
```

```
/Лекция4/lab4> ncu --metrics
l1tex__t_bytes_pipe_lsu_mem_global_op_st.sum.per_second  ./lab4c
```

```
gInit(float *, float *) (2, 1, 1)x(128, 1, 1), Context 1, Stream 7, Device 0, CC 7.5
  Section: Command line profiler metrics
  ------------------------------------------------------- ----------- ------------
  Metric Name                                             Metric Unit Metric Value
  ------------------------------------------------------- ----------- ------------
  l1tex__t_bytes_pipe_lsu_mem_global_op_st.sum.per_second     Mbyte/s       688,17
  ------------------------------------------------------- ----------- ------------

gSum(float *, float *) (2, 1, 1)x(128, 1, 1), Context 1, Stream 7, Device 0, CC 7.5
  Section: Command line profiler metrics
  ------------------------------------------------------- ----------- ------------
  Metric Name                                             Metric Unit Metric Value
  ------------------------------------------------------- ----------- ------------
  l1tex__t_bytes_pipe_lsu_mem_global_op_st.sum.per_second     Mbyte/s       347,83
  ------------------------------------------------------- ----------- ------------
```

Кодирование метрики ncu:



```
l1tex__t_bytes_pipe_lsu_mem_global_op_st.sum.per_second  ./lab4c
```

L1 cache

Тип данных

load/store unit (конвейер)

Глобальная память

Операция сохранения

```
ip-011@linux-47dw:/home/malkov/WORKSHOP/PGP-2023>
nvprof -m gld_throughput ./lab3c
```

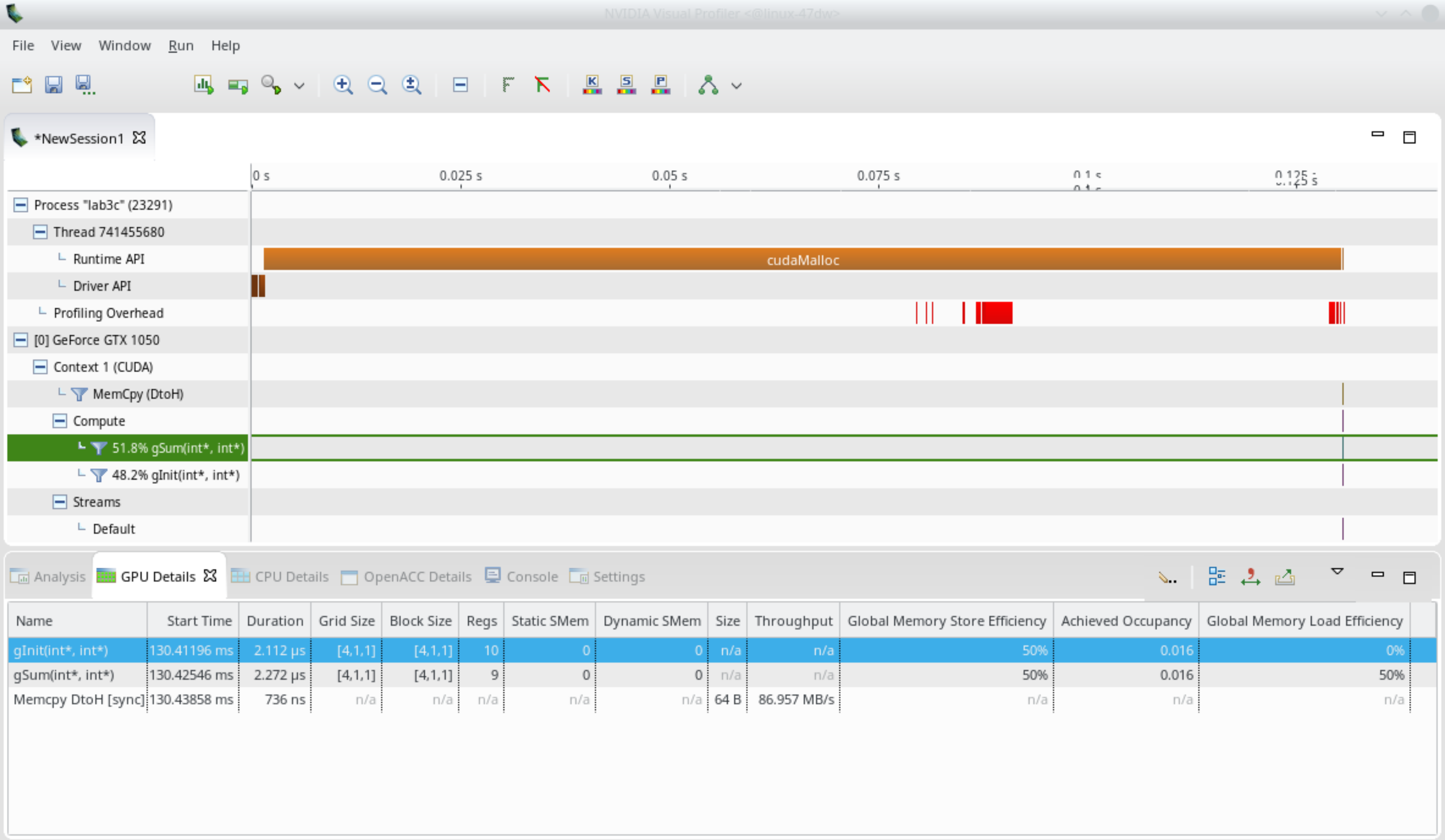
```
Invocations   Metric Name  Metric Description  Min Max   Avg
Device "GeForce GTX 1050 (0)"
   Kernel: gInit(int*, int*)
 1   gld_throughput Global Load Throughput 0.0B/s 0.0B/s  0.0B/s
   Kernel: gSum(int*, int*)
 1   gld_throughput Global Load Throughput  87.694MB/s

                       87.694MB/s  87.694MB/s
```

```
/Lecture3/Lab3-cuda-gdb # ncu --metrics
l1tex__t_bytes_pipe_lsu_mem_global_op_ld.sum.per_second
./lab3c
```

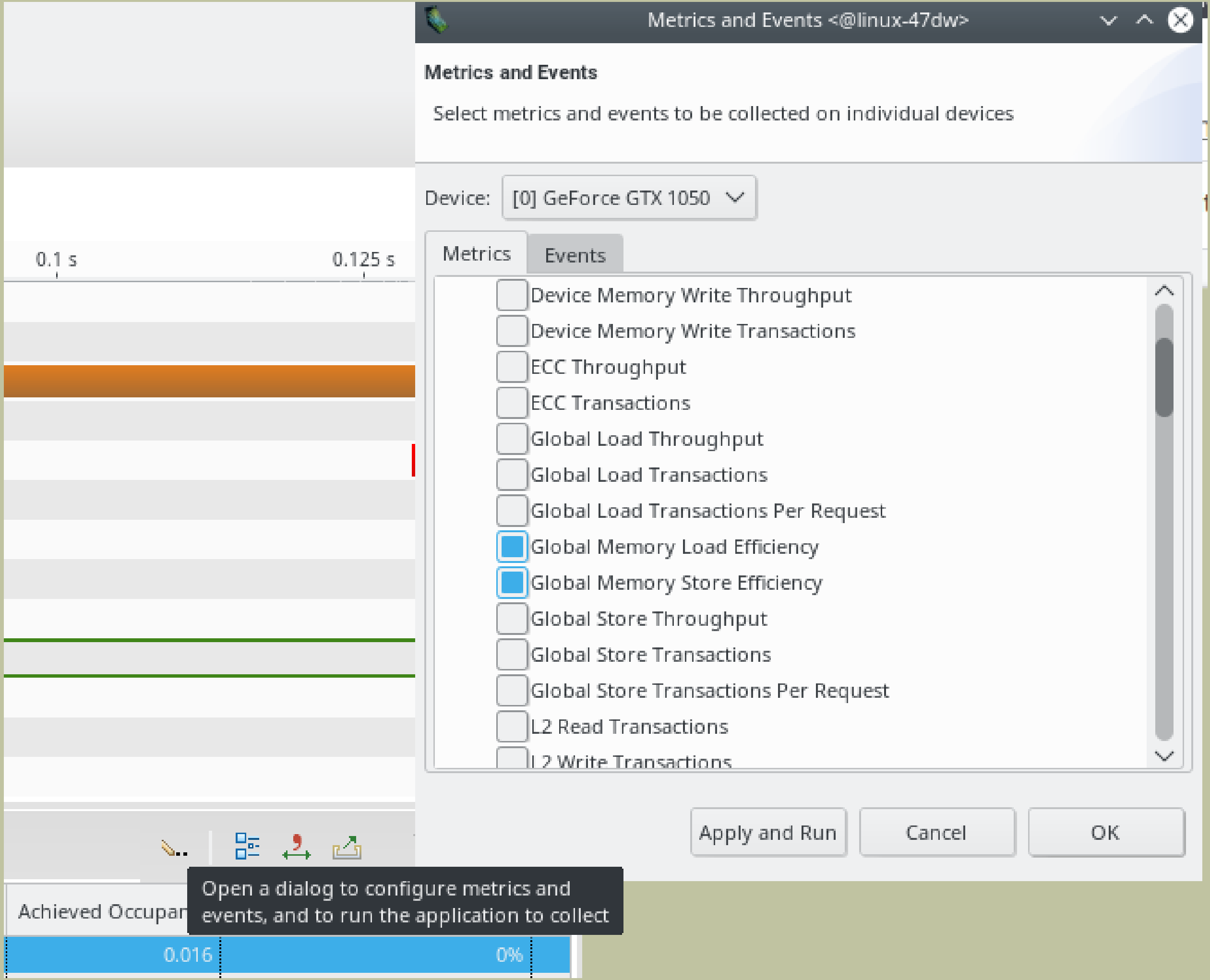
```
gInit(int *, int *), 2023-Feb-13 15:25:41, Context 1, Stream 7
   Section: Command line profiler metrics
 ---------------------------------------------------
l1tex__t_bytes_pipe_lsu_mem_global_op_ld.sum.per_second
         byte/second                              0
   ------------------------------------------------
gSum(int *, int *), 2023-Feb-13 15:25:41, Context 1, Stream 7
   Section: Command line profiler metrics
   ---------------------------------------------------------
l1tex__t_bytes_pipe_lsu_mem_global_op_ld.sum.per_second
        Mbyte/second                          82.47
   -------------------------------------------------------
```

nvvp и Nsight Compute

```
ip-011@linux-47dw:/home/malkov/WORKSHOP/PGP-2023> nvvp ./lab3c
```



Name	Start Time	Duration	Grid Size	Block Size	Regs	Static SMem	Dynamic SMem	Size	Throughput	Global Memory Store Efficiency	Achieved Occupancy	Global Memory Load Efficiency
gInit(int*, int*)	130.41196 ms	2.112 μs	[4,1,1]	[4,1,1]	10	0	0	n/a	n/a	50%	0.016	0%
gSum(int*, int*)	130.42546 ms	2.272 μs	[4,1,1]	[4,1,1]	9	0	0	n/a	n/a	50%	0.016	50%
Memcpy DtoH [sync]	130.43858 ms	736 ns	n/a	n/a	n/a	n/a	n/a	64 B	86.957 MB/s	n/a	n/a	n/a



Metrics and Events

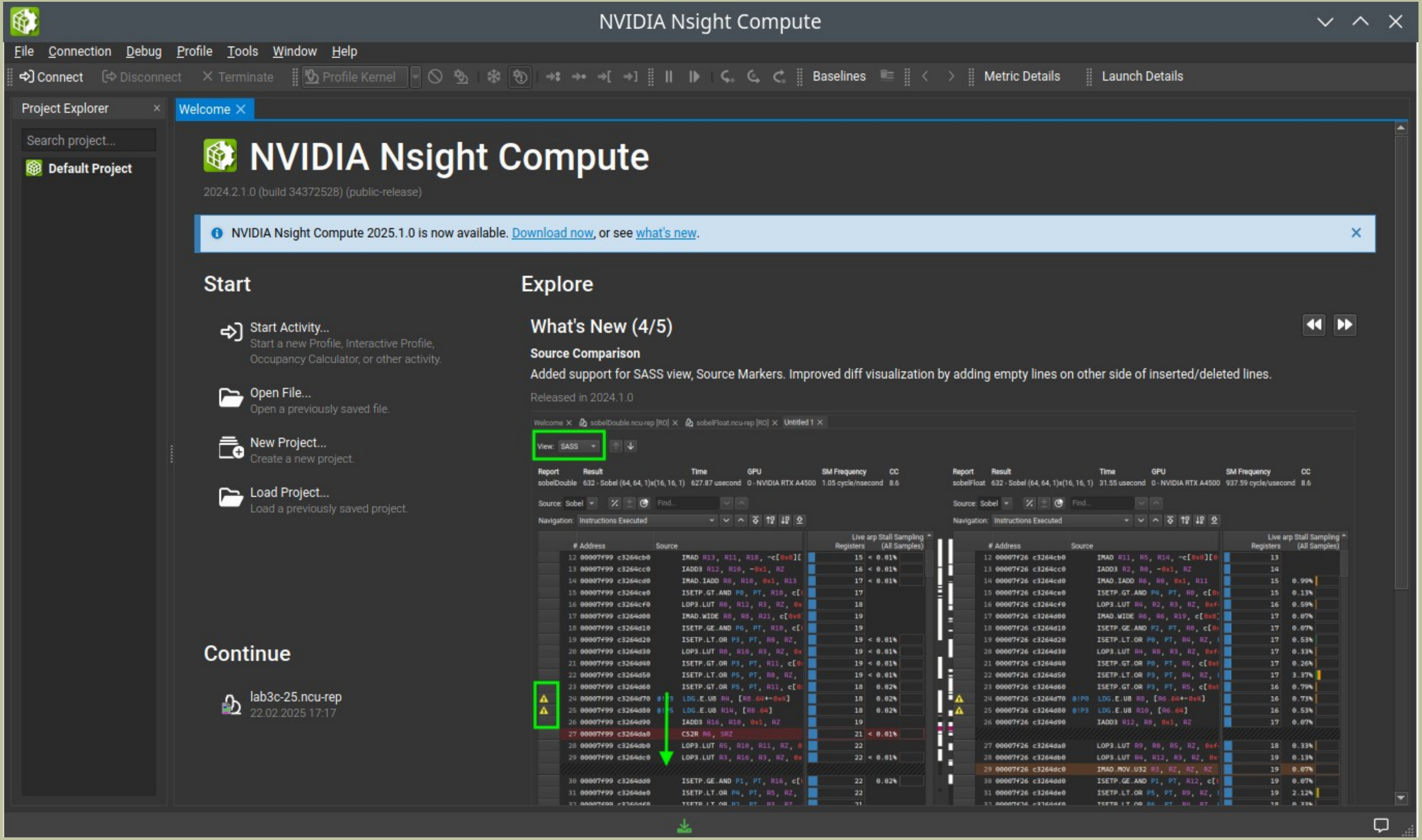
Select metrics and events to be collected on individual devices

Device: [0] GeForce GTX 1050

- □ Device Memory Write Throughput
- □ Device Memory Write Transactions
- □ ECC Throughput
- □ ECC Transactions
- □ Global Load Throughput
- □ Global Load Transactions
- □ Global Load Transactions Per Request
- ☑ Global Memory Load Efficiency
- ☑ Global Memory Store Efficiency
- □ Global Store Throughput
- □ Global Store Transactions
- □ Global Store Transactions Per Request
- □ L2 Read Transactions
- □ L2 Write Transactions

Open a dialog to configure metrics and events, and to run the application to collect

```
/Лекция4/lab4> ncu-ui  &
```



NVIDIA Nsight Compute
File Connection Debug Profile Tools Window Help
Connect Disconnect Terminate Profile Kernel Baselines Metric Details Launch Details
Project Explorer
Search project...
Default Project
Welcome

NVIDIA Nsight Compute

2024.2.1.0 (build 34372528) (public-release)

NVIDIA Nsight Compute 2025.1.0 is now available. Download now, or see what's new.

Start

Start Activity...
Start a new Profile, Interactive Profile, Occupancy Calculator, or other activity.

Open File...
Open a previously saved file.

New Project...
Create a new project.

Load Project...
Load a previously saved project.

Continue

lab3c-25.ncu-rep
22.02.2025 17:17

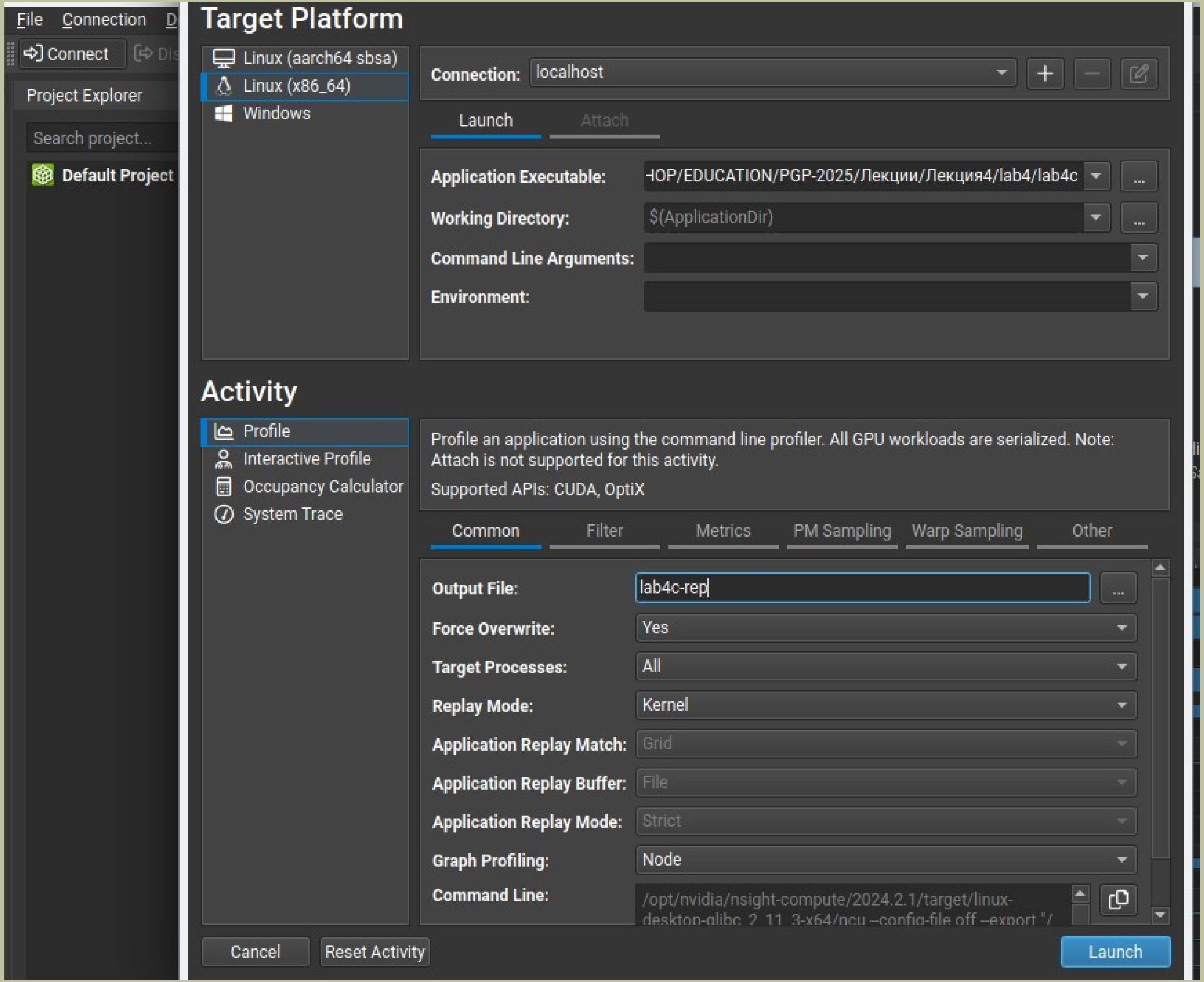
Explore

What's New (4/5)

Source Comparison

Added support for SASS view, Source Markers. Improved diff visualization by adding empty lines on other side of inserted/deleted lines.

Released in 2024.1.0



Target Platform

Connection: localhost

Application Executable: ·ЮP/EDUCATION/PGP-2025/Лекции/Лекция4/lab4/lab4c

Working Directory: $(ApplicationDir)

Command Line Arguments:

Environment:

Activity

Profile an application using the command line profiler. All GPU workloads are serialized. Note: Attach is not supported for this activity.

Supported APIs: CUDA, OptiX

Output File: lab4c-rep

Force Overwrite: Yes

Target Processes: All

Replay Mode: Kernel

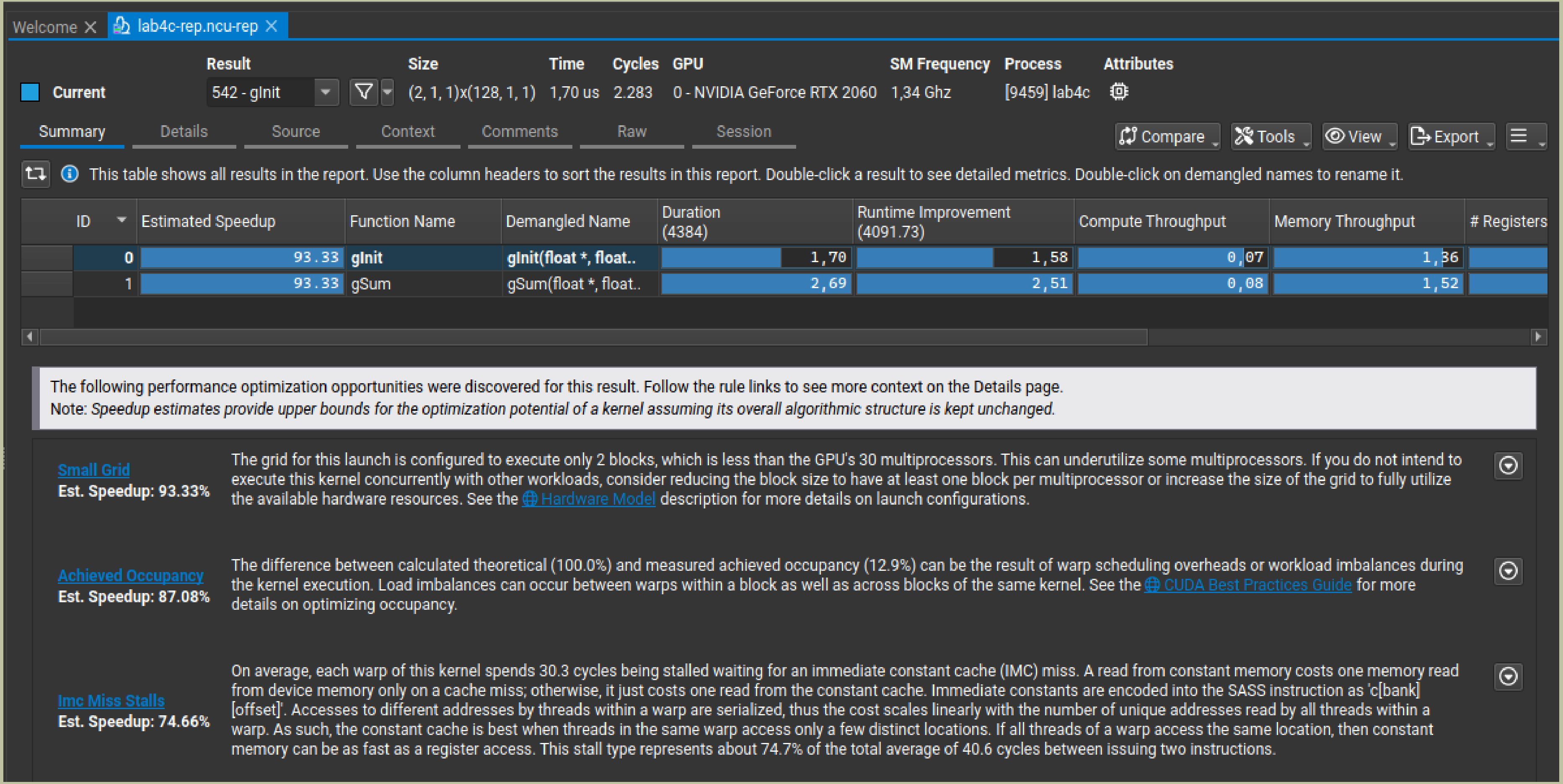
Application Replay Match: Grid

Application Replay Buffer: File

Application Replay Mode: Strict

Graph Profiling: Node

Command Line: /opt/nvidia/nsight-compute/2024.2.1/target/linux-desktop-glibc_2_11_3-x64/ncu --config-file off --export "/



	Result	Size	Time	Cycles	GPU	SM Frequency	Process	Attributes
Current	542 - gInit	(2, 1, 1)x(128, 1, 1)	1,70 us	2.283	0 - NVIDIA GeForce RTX 2060	1,34 Ghz	[9459] lab4c	

This table shows all results in the report. Use the column headers to sort the results in this report. Double-click a result to see detailed metrics. Double-click on demangled names to rename it.

ID	Estimated Speedup	Function Name	Demangled Name	Duration (4384)	Runtime Improvement (4091.73)	Compute Throughput	Memory Throughput	# Registers
0	93.33	gInit	gInit(float *, float..	1,70	1,58	0,07	1,36	
1	93.33	gSum	gSum(float *, float..	2,69	2,51	0,08	1,52	

The following performance optimization opportunities were discovered for this result. Follow the rule links to see more context on the Details page.
Note: *Speedup estimates provide upper bounds for the optimization potential of a kernel assuming its overall algorithmic structure is kept unchanged.*

Small Grid
Est. Speedup: 93.33%

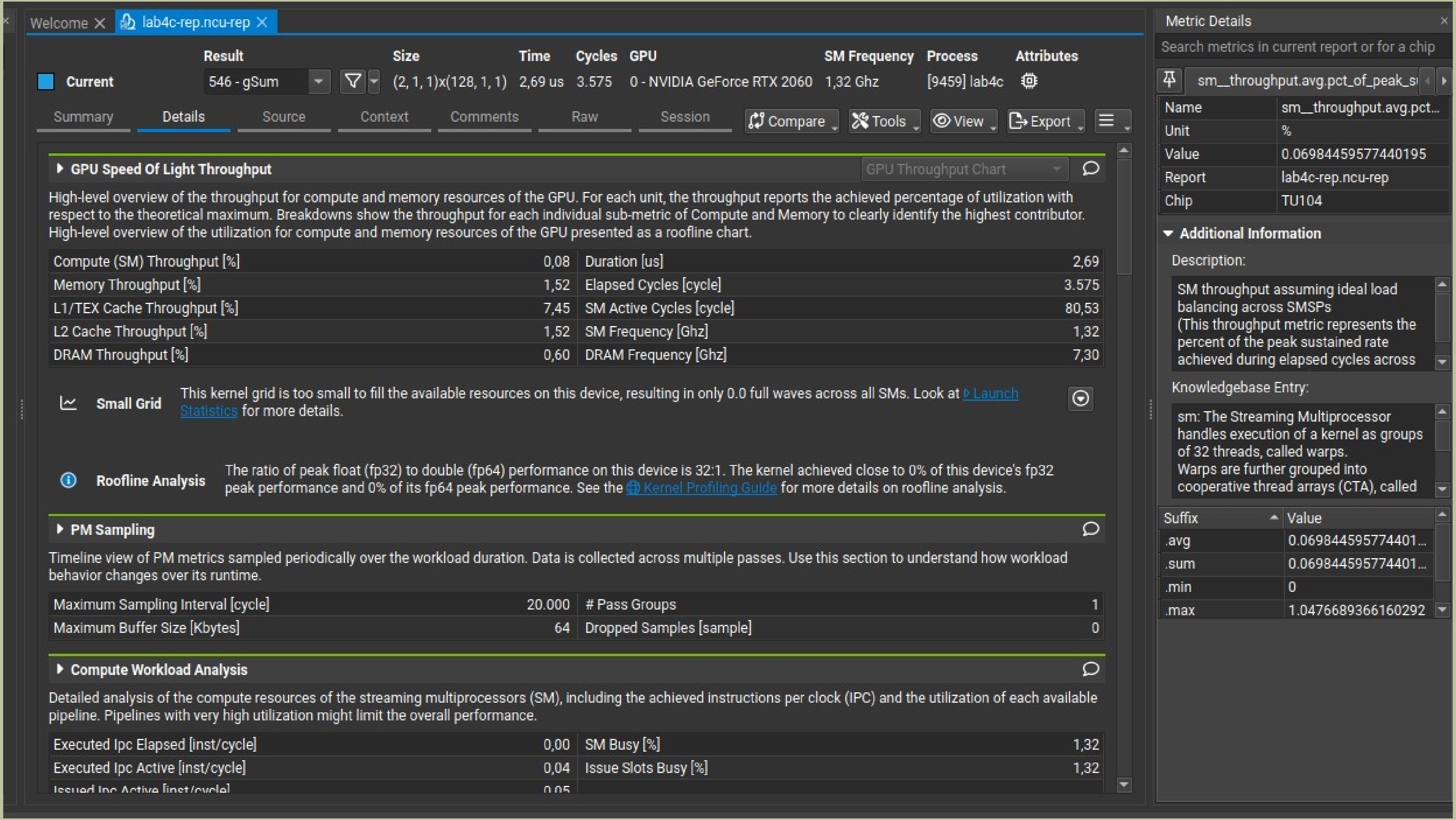
The grid for this launch is configured to execute only 2 blocks, which is less than the GPU's 30 multiprocessors. This can underutilize some multiprocessors. If you do not intend to execute this kernel concurrently with other workloads, consider reducing the block size to have at least one block per multiprocessor or increase the size of the grid to fully utilize the available hardware resources. See the Hardware Model description for more details on launch configurations.

Achieved Occupancy
Est. Speedup: 87.08%

The difference between calculated theoretical (100.0%) and measured achieved occupancy (12.9%) can be the result of warp scheduling overheads or workload imbalances during the kernel execution. Load imbalances can occur between warps within a block as well as across blocks of the same kernel. See the CUDA Best Practices Guide for more details on optimizing occupancy.

Imc Miss Stalls
Est. Speedup: 74.66%

On average, each warp of this kernel spends 30.3 cycles being stalled waiting for an immediate constant cache (IMC) miss. A read from constant memory costs one memory read from device memory only on a cache miss; otherwise, it just costs one read from the constant cache. Immediate constants are encoded into the SASS instruction as 'c[bank][offset]'. Accesses to different addresses by threads within a warp are serialized, thus the cost scales linearly with the number of unique addresses read by all threads within a warp. As such, the constant cache is best when threads in the same warp access only a few distinct locations. If all threads of a warp access the same location, then constant memory can be as fast as a register access. This stall type represents about 74.7% of the total average of 40.6 cycles between issuing two instructions.



	Result	Size	Time	Cycles	GPU	SM Frequency	Process	Attributes
Current	546 - gSum	(2, 1, 1)x(128, 1, 1)	2,69 us	3.575	0 - NVIDIA GeForce RTX 2060	1,32 Ghz	[9459] lab4c	

GPU Speed Of Light Throughput

High-level overview of the throughput for compute and memory resources of the GPU. For each unit, the throughput reports the achieved percentage of utilization with respect to the theoretical maximum. Breakdowns show the throughput for each individual sub-metric of Compute and Memory to clearly identify the highest contributor. High-level overview of the utilization for compute and memory resources of the GPU presented as a roofline chart.

Compute (SM) Throughput [%]	0,08	Duration [us]	2,69
Memory Throughput [%]	1,52	Elapsed Cycles [cycle]	3.575
L1/TEX Cache Throughput [%]	7,45	SM Active Cycles [cycle]	80,53
L2 Cache Throughput [%]	1,52	SM Frequency [Ghz]	1,32
DRAM Throughput [%]	0,60	DRAM Frequency [Ghz]	7,30

Small Grid This kernel grid is too small to fill the available resources on this device, resulting in only 0.0 full waves across all SMs. Look at Launch Statistics for more details.

Roofline Analysis The ratio of peak float (fp32) to double (fp64) performance on this device is 32:1. The kernel achieved close to 0% of this device's fp32 peak performance and 0% of its fp64 peak performance. See the Kernel Profiling Guide for more details on roofline analysis.

PM Sampling

Timeline view of PM metrics sampled periodically over the workload duration. Data is collected across multiple passes. Use this section to understand how workload behavior changes over its runtime.

Maximum Sampling Interval [cycle]	20.000	# Pass Groups	1
Maximum Buffer Size [Kbytes]	64	Dropped Samples [sample]	0

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]	0,00	SM Busy [%]	1,32
Executed Ipc Active [inst/cycle]	0,04	Issue Slots Busy [%]	1,32
Issued Ipc Active [inst/cycle]	0,05		

Name	sm__throughput.avg.pct...
Unit	%
Value	0.06984459577440195
Report	lab4c-rep.ncu-rep
Chip	TU104

Additional Information

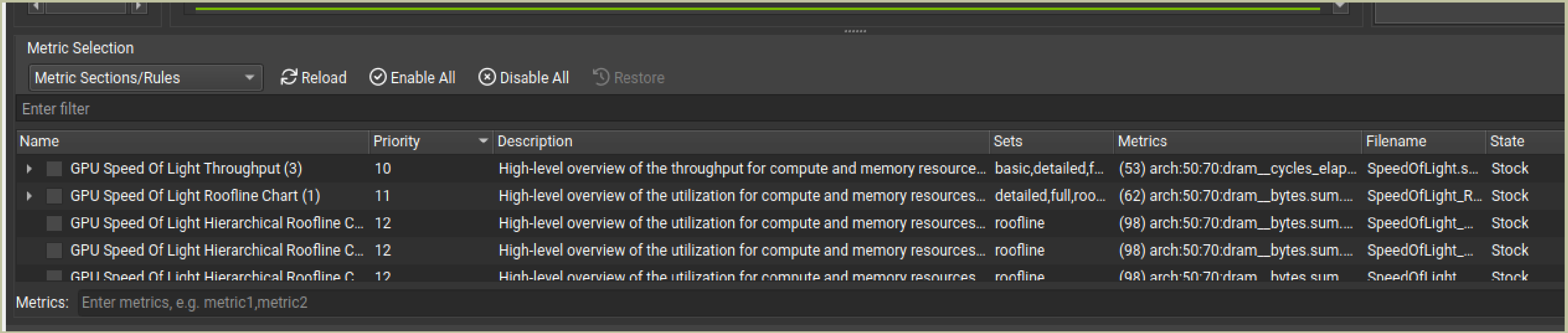
Description:

SM throughput assuming ideal load balancing across SMSPs
(This throughput metric represents the percent of the peak sustained rate achieved during elapsed cycles across

Knowledgebase Entry:

sm: The Streaming Multiprocessor handles execution of a kernel as groups of 32 threads, called warps.
Warps are further grouped into cooperative thread arrays (CTA), called

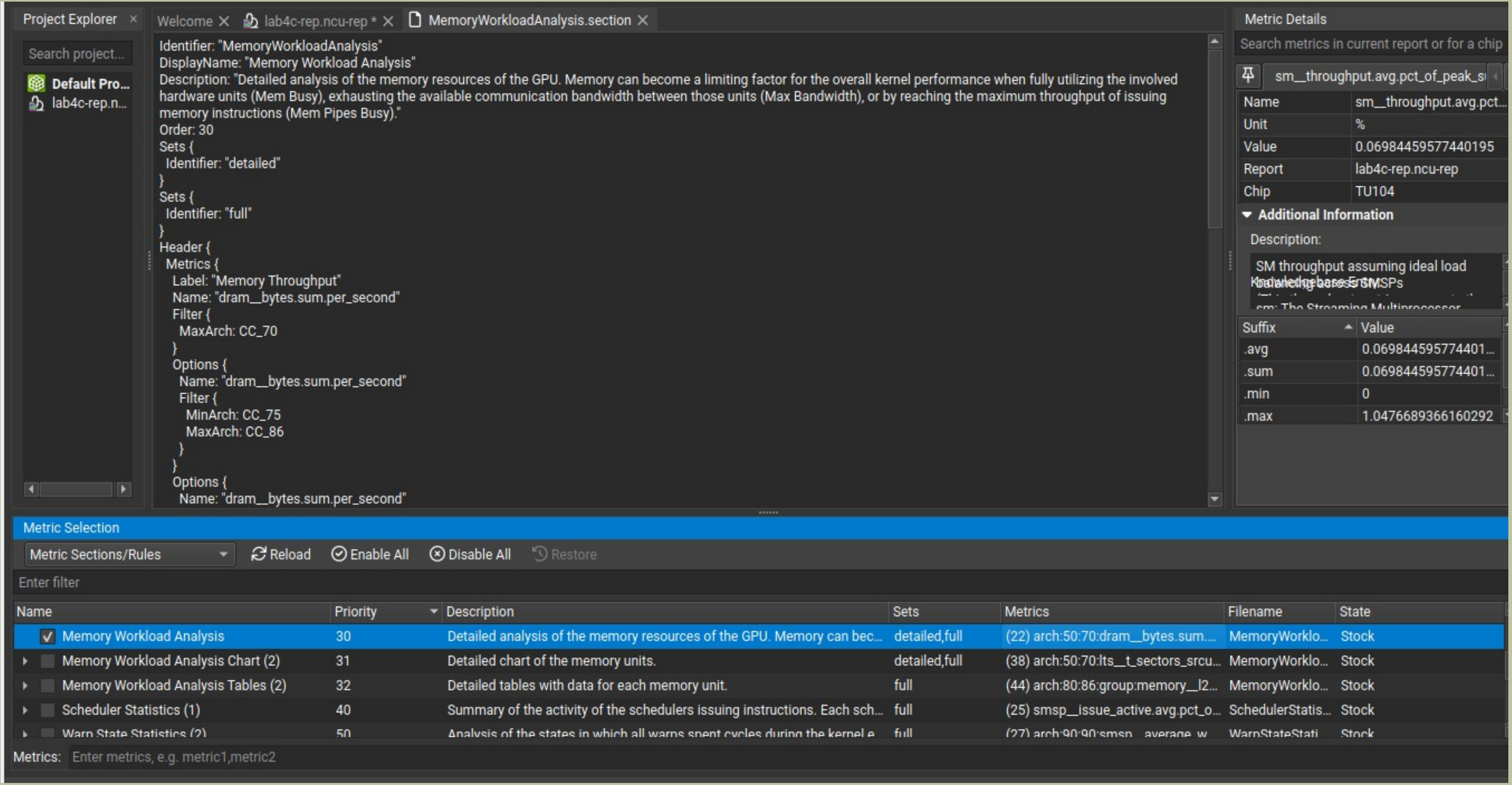
Suffix	Value
.avg	0.069844595774401...
.sum	0.069844595774401...
.min	0
.max	1.0476689366160292



Metric Sections/Rules

Name	Priority	Description	Sets	Metrics	Filename	State
▸ □ GPU Speed Of Light Throughput (3)	10	High-level overview of the throughput for compute and memory resource...	basic,detailed,f...	(53) arch:50:70:dram__cycles_elap...	SpeedOfLight.s...	Stock
▸ □ GPU Speed Of Light Roofline Chart (1)	11	High-level overview of the utilization for compute and memory resources...	detailed,full,roo...	(62) arch:50:70:dram__bytes.sum....	SpeedOfLight_R...	Stock
□ GPU Speed Of Light Hierarchical Roofline C...	12	High-level overview of the utilization for compute and memory resources...	roofline	(98) arch:50:70:dram__bytes.sum....	SpeedOfLight_...	Stock
□ GPU Speed Of Light Hierarchical Roofline C...	12	High-level overview of the utilization for compute and memory resources...	roofline	(98) arch:50:70:dram__bytes.sum....	SpeedOfLight_...	Stock
□ GPU Speed Of Light Hierarchical Roofline C...	12	High-level overview of the utilization for compute and memory resources...	roofline	(98) arch:50:70:dram__bytes.sum...	SpeedOfLight...	Stock

Metrics: Enter metrics, e.g. metric1,metric2



```
Identifier: "MemoryWorkloadAnalysis"
DisplayName: "Memory Workload Analysis"
Description: "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)."
Order: 30
Sets {
 Identifier: "detailed"
}
Sets {
 Identifier: "full"
}
Header {
 Metrics {
  Label: "Memory Throughput"
  Name: "dram__bytes.sum.per_second"
  Filter {
   MaxArch: CC_70
  }
  Options {
   Name: "dram__bytes.sum.per_second"
   Filter {
    MinArch: CC_75
    MaxArch: CC_86
   }
  }
  Options {
   Name: "dram__bytes.sum.per_second"
```

Name	sm__throughput.avg.pct...
Unit	%
Value	0.06984459577440195
Report	lab4c-rep.ncu-rep
Chip	TU104

Additional Information

Description:

SM throughput assuming ideal load balancing across SMSPs

Knowledgebase Entry

sm: The Streaming Multiprocessor

Suffix	Value
.avg	0.069844595774401...
.sum	0.069844595774401...
.min	0
.max	1.0476689366160292

Name	Priority	Description	Sets	Metrics	Filename	State
☑ Memory Workload Analysis	30	Detailed analysis of the memory resources of the GPU. Memory can bec...	detailed,full	(22) arch:50:70:dram__bytes.sum....	MemoryWorklo...	Stock
▸ □ Memory Workload Analysis Chart (2)	31	Detailed chart of the memory units.	detailed,full	(38) arch:50:70:lts__t_sectors_srcu...	MemoryWorklo...	Stock
▸ □ Memory Workload Analysis Tables (2)	32	Detailed tables with data for each memory unit.	full	(44) arch:80:86:group:memory__l2...	MemoryWorklo...	Stock
▸ □ Scheduler Statistics (1)	40	Summary of the activity of the schedulers issuing instructions. Each sch...	full	(25) smsp__issue_active.avg.pct_o...	SchedulerStatis...	Stock
▸ □ Warp State Statistics (2)	50	Analysis of the states in which all warps spent cycles during the kernel e...	full	(27) arch:90:90:smsp__average_w...	WarpStateStati...	Stock

Metrics: Enter metrics, e.g. metric1,metric2