Лекция 3

- Инструменты профилирования и отладки
- nvprof
- Nsight Compute CLI
- Nvvp
- Nsight Compute

```
> ~Lecture3/Lab3-gdb> g++ lab3a.cpp  -g3 -o lab3a
> ~Lecture3/Lab3-gdb> gdb lab3a
```

```
(gdb) list main
…….
36        gettimeofday(&t, NULL);
37        Start =(double)t.tv_sec*1000000.0 +
(double)t.tv_usec;
38        hTest(N,a,b);
39        gettimeofday(&t, NULL);
40        Finish =(double)t.tv_sec*1000000.0 +
(double)t.tv_usec;
……
(gdb) b 38
Breakpoint 1 at 0x400865: file lab3a.cpp, line 38.
```

```
(gdb) step
hTest (N=16, a=0x613e70, b=0x613ec0) at lab3a.cpp:7
7          for(int i=0; i<N;i++)
(gdb) list hTest
1        #include <malloc.h>
2        #include <stdio.h>
3        #include <stdlib.h>
4        #include <sys/time.h>
5
6        void hTest(int N, int* a, int* b){
7          for(int i=0; i<N;i++)
8             a[i]+=b[i];
9        }
```

```
(gdb) info args
N = 16
a = 0x613e70
b = 0x613ec0
(gdb) info locals
i = 0
(gdb) next 8
7           for(int i=0; i<N;i++)
(gdb) info locals
i = 3
(gdb) print b[2]
$1 = 5
(gdb) print a[2]
$2 = 9
```

```
gdb) break 8 if i==12
Breakpoint 3 at 0x400705: file lab3a.cpp, line 8.
(gdb) c
Continuing.

Breakpoint 3, hTest (N=16, a=0x613e70, b=0x613ec0) at
lab3a.cpp:8
8             a[i]+=b[i];
(gdb) info locals
i = 12
(gdb) finish
Run till exit from #0  hTest (N=16, a=0x613e70,
b=0x613ec0) at lab3a.cpp:8
main (argc=2, argv=0x7fffffffd9b8) at lab3a.cpp:39
39        gettimeofday(&t, NULL);
```

```
(gdb) x/16d  b
0x613ec0:       1       3       5       7
0x613ed0:       9       11      13      15
0x613ee0:       17      19      21      23
0x613ef0:       25      27      29      31
(gdb) x/16d  a
0x613e70:       1       5       9       13
0x613e80:       17      21      25      29
0x613e90:       33      37      41      45
0x613ea0:       49      53      57      61
(gdb) print a[2]-b[2]
$16 = 4
```

```
(gdb) c
Continuing.
Elapsed time: 9.57138e+06 ms
0       1       1
1       5       3
2       9       5
3       13      7
4       17      9
5       21      11
……..
13      53      27
14      57      29
15      61      31
[Inferior 1 (process 4272) exited normally]

(gdb) quit
```

Отладка многопоточных программ

```
~/Lecture3/Lab3-gdb> gdb lab3b
(gdb) list hTest
16  void* hTest(void* arg){
17      struct targ* s_arg=(struct targ*)arg;
18      int length=s_arg->length;
19      int offset=s_arg->num_thread*length;
20      int i;
21     for(i=0;i<length;i++)
22         a[i+offset]+=/*1000*sin((double)*/b[i+offset];
23     return NULL;
25  }
(gdb) break lab3b.cpp:22
Breakpoint 1 at 0x40083f: file lab3b.cpp, line 22.
```

```
(gdb) run 4 16
Starting program: …/Lecture3/Lab3-gdb/lab3b 4 16
[Thread debugging using libthread_db enabled]
Using host libthread_db library
"/lib64/libthread_db.so.1".
[New Thread 0x7ffff6ed1700 (LWP 10741)]
[New Thread 0x7ffff66d0700 (LWP 10742)]
[New Thread 0x7ffff5ecf700 (LWP 10743)]
[Switching to Thread 0x7ffff6ed1700 (LWP 10741)]

Thread 2 "lab3b" hit Breakpoint 1, hTest (arg=0x614e70)
at lab3b.cpp:22
22          a[i+offset]+=/*1000*sin((double)*/b[i+offset];
```

```
(gdb) info threads
  Id   Target Id                                          Frame
  1    Thread 0x7ffff7fc0740 (LWP 10737) "lab3b" clone ()
at ../sysdeps/unix/sysv/linux/x86_64/clone.S:78
* 2    Thread 0x7ffff6ed1700 (LWP 10741) "lab3b" hTest
(arg=0x614e70) at lab3b.cpp:22
  3    Thread 0x7ffff66d0700 (LWP 10742) "lab3b" hTest
(arg=0x614e7c) at lab3b.cpp:22
  4    Thread 0x7ffff5ecf700 (LWP 10743) "lab3b" clone ()
at ../sysdeps/unix/sysv/linux/x86_64/clone.S:78
```

```
(gdb) print offset
$1 = 0
(gdb) thread 3
[Switching to thread 3 (Thread 0x7ffff66d0700 (LWP
10742))]
#0  hTest (arg=0x614e7c) at lab3b.cpp:22
22
a[i+offset]+=/*1000*sin((double)*/b[i+offset];
(gdb) print offset
$2 = 4
```

```
(gdb) break 22 thread 3
Note: breakpoint 1 (all threads) also set at pc
0x40083f.
Breakpoint 2 at 0x40083f: file lab3b.cpp, line 22.
(gdb) info breakpoints
Num     Type           Disp Enb Address            What
1       breakpoint     keep y   0x000000000040083f in
hTest(void*) at lab3b.cpp:22
       breakpoint already hit 1 time
2       breakpoint     keep y   0x000000000040083f in
hTest(void*) at lab3b.cpp:22 thread 3
       stop only in thread 3
(gdb) delete 1
```

```
(gdb) x/16d a
0x614ee0:       0       2       4       6
0x614ef0:       8       10      12      14
0x614f00:       16      18      20      22
0x614f10:       24      26      28      30
```

```
(gdb) continue
Continuing.

Thread 3 "lab3b" hit Breakpoint 2, hTest (arg=0x614e7c)
at lab3b.cpp:22
22          a[i+offset]+=/*1000*sin((double)*/b[i+offset];
(gdb) c
……
(gdb) x/16d a
0x614ee0:       1       2       4       6
0x614ef0:       17      21      12      14
0x614f00:       33      37      20      22
0x614f10:       49      53      28      30
```

```
(gdb) c
Continuing.
[Thread 0x7ffff5ecf700 (LWP 10743) exited]
[Thread 0x7ffff66d0700 (LWP 10742) exited]
Thread-specific breakpoint 2 deleted - thread 3 no
longer in the thread list.
[Thread 0x7ffff56ce700 (LWP 11167) exited]
[Thread 0x7ffff6ed1700 (LWP 10741) exited]
Elapsed time: 2.22941e+06 ms
0       1       1
1       3       5
2       5       9
……
14      29      57
15      31      61
[Inferior 1 (process 10737) exited normally]
```

```
(gdb)print a[1]
……..
(gdb)info locals
……..
(gdb) info args
arg = 0x614e7c
(gdb) print ((struct targ*)arg)->length
$1 = 4
```

Отладка программ, выполняемых на GPU

https://docs.nvidia.com/cuda/archive/11.2.0/cuda-gdb/index.html

```
(cuda-gdb) info cuda threads
BlockIdx ThreadIdx To BlockIdx ThreadIdx Count  VirtualPC
                                          Filename  Line
Kernel 0
*  (0,0,0)   (0,0,0)      (0,0,0)   (3,0,0)      4
0x00007fffe525c3e0
…/Lecture3/Lab3-cuda-gdb/lab3c.cu      9
   (1,0,0)   (0,0,0)      (3,0,0)   (3,0,0)     12
0x00007fffe525c2b0
…/Lecture3/Lab3-cuda-gdb/lab3c.cu      8
```

```
cuda-gdb) cuda block 2 thread 3
[Switching focus to CUDA kernel 0, grid 1, block
(2,0,0), thread (3,0,0), device 0, sm 2, warp 0, lane 3]
8          b[i]=2*i+1;
(cuda-gdb) print i
$2 = 11
```

```
(cuda-gdb) n
9        }
(cuda-gdb) x/16d b
0x7fffecc00200: 1        3        5        7
0x7fffecc00210: 0        0        0        0
0x7fffecc00220: 0        0        0        0
0x7fffecc00230: 0        0        0        0
```

```
malkov@192:~> ssh cyber.sibsutis.ru
```

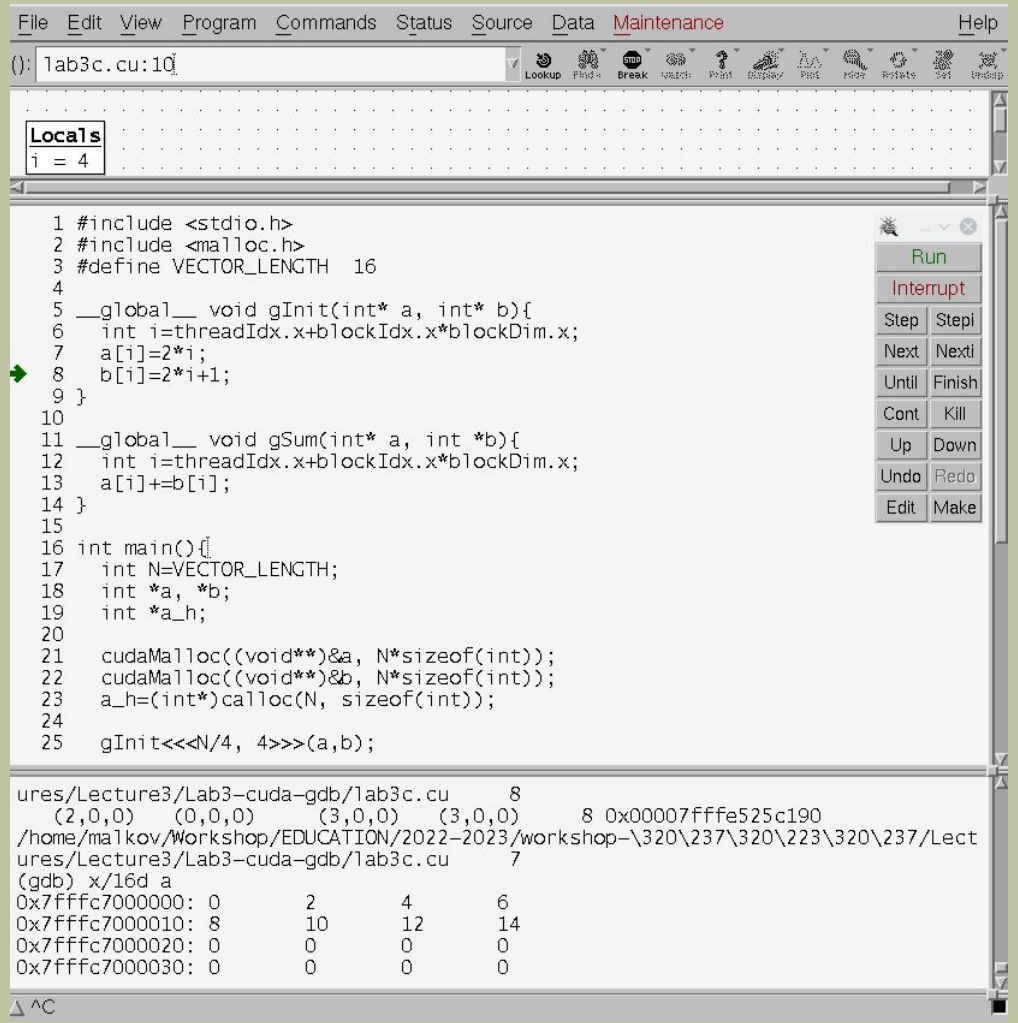
```
malkov@linux-47dw: ~/WORKSHOP/PGP-2023> cuda-gdb  lab3c
```

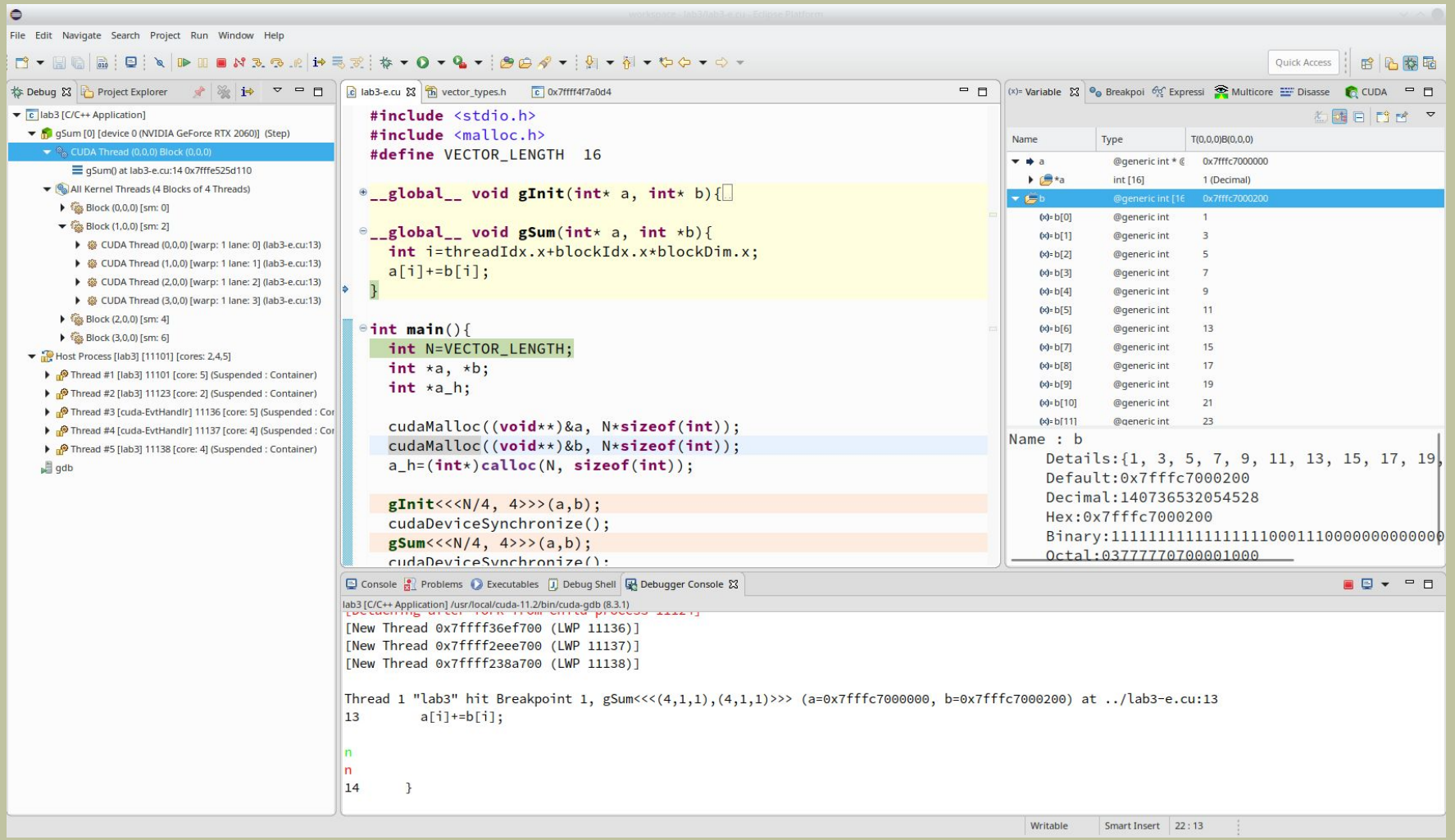
```
(cuda-gdb) break 8
Breakpoint 1 at 0x403851: file lab3c.cu, line 8.
(cuda-gdb) run
Starting program: /home/malkov/WORKSHOP/PGP-2023/lab3c
……
[Switching focus to CUDA kernel 0, grid 1, block
(0,0,0), thread (0,0,0), device 0, sm 0, warp 0, lane 0]
Thread 1 "lab3c" hit Breakpoint 1,
gInit<<<(4,1,1),(4,1,1)>>> (a=0x7fffe6800000,
b=0x7fffe6800200) at lab3c.cu:8
8           b[i]=2*i+1;
```

```
…/Lecture3/
Lab3-cuda-gdb>
ddd cuda-gdb lab3c
```

```
malkov@192:~> ssh
cyber.sibsutis.ru -X
```

```
malkov@linux-47dw:
~/WORKSHOP/PGP-2023>
ddd cuda-gdb lab3c
```




File Edit Navigate Search Project Run Window Help
Quick Access
Debug
Project Explorer
lab3 [C/C++ Application]
gSum [0] [device 0 (NVIDIA GeForce RTX 2060)] (Step)
CUDA Thread (0,0,0) Block (0,0,0)
gSum() at lab3-e.cu:14 0x7fffe525d110
All Kernel Threads (4 Blocks of 4 Threads)
Block (0,0,0) [sm: 0]
Block (1,0,0) [sm: 2]
CUDA Thread (0,0,0) [warp: 1 lane: 0] (lab3-e.cu:13)
CUDA Thread (1,0,0) [warp: 1 lane: 1] (lab3-e.cu:13)
CUDA Thread (2,0,0) [warp: 1 lane: 2] (lab3-e.cu:13)
CUDA Thread (3,0,0) [warp: 1 lane: 3] (lab3-e.cu:13)
Block (2,0,0) [sm: 4]
Block (3,0,0) [sm: 6]
Host Process [lab3] [11101] [cores: 2,4,5]
Thread #1 [lab3] 11101 [core: 5] (Suspended : Container)
Thread #2 [lab3] 11123 [core: 2] (Suspended : Container)
Thread #3 [cuda-EvtHandlr] 11136 [core: 5] (Suspended : Cor
Thread #4 [cuda-EvtHandlr] 11137 [core: 4] (Suspended : Cor
Thread #5 [lab3] 11138 [core: 4] (Suspended : Container)
gdb
lab3-e.cu
vector_types.h
0x7ffff4f7a0d4
#include <stdio.h>
#include <malloc.h>
#define VECTOR_LENGTH 16
__global__ void gInit(int* a, int* b){
__global__ void gSum(int* a, int *b){
int i=threadIdx.x+blockIdx.x*blockDim.x;
a[i]+=b[i];
}
int main(){
int N=VECTOR_LENGTH;
int *a, *b;
int *a_h;
cudaMalloc((void**)&a, N*sizeof(int));
cudaMalloc((void**)&b, N*sizeof(int));
a_h=(int*)calloc(N, sizeof(int));
gInit<<<N/4, 4>>>(a,b);
cudaDeviceSynchronize();
gSum<<<N/4, 4>>>(a,b);
cudaDeviceSynchronize();
Variable
Breakpoi
Expressi
Multicore
Disasse
CUDA
Name
Type
T(0,0,0)B(0,0,0)
a
@generic int * @
0x7fffc7000000
*a
int [16]
1 (Decimal)
b
@generic int [16
0x7fffc7000200
b[0]
@generic int
1
b[1]
@generic int
3
b[2]
@generic int
5
b[3]
@generic int
7
b[4]
@generic int
9
b[5]
@generic int
11
b[6]
@generic int
13
b[7]
@generic int
15
b[8]
@generic int
17
b[9]
@generic int
19
b[10]
@generic int
21
b[11]
@generic int
23
Name : b
Details:{1, 3, 5, 7, 9, 11, 13, 15, 17, 19,
Default:0x7fffc7000200
Decimal:140736532054528
Hex:0x7fffc7000200
Binary:11111111111111111000111000000000000
Octal:03777770700001000
Console
Problems
Executables
Debug Shell
Debugger Console
lab3 [C/C++ Application] /usr/local/cuda-11.2/bin/cuda-gdb (8.3.1)
[New Thread 0x7ffff36ef700 (LWP 11136)]
[New Thread 0x7ffff2eee700 (LWP 11137)]
[New Thread 0x7ffff238a700 (LWP 11138)]
Thread 1 "lab3" hit Breakpoint 1, gSum<<<(4,1,1),(4,1,1)>>> (a=0x7fffc7000000, b=0x7fffc7000200) at ../lab3-e.cu:13
13 a[i]+=b[i];
n
n
14 }
Writable
Smart Insert
22 : 13

```
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
```

```
/Lecture3/Lab3-cuda-gdb # ncu --list-sections

/Lecture3/Lab3-cuda-gdb # ncu --query-metrics
```

https://docs.nvidia.com/nsight-compute/NsightComputeCli/index.html#nvprof-metric-collection

```
/Lecture3/Lab3-cuda-gdb # ncu --metrics
l1tex  t_bytes_pipe_lsu_mem_global_op_st.sum.per_second
./lab3c
```

```
gInit(int *, int *), 2023-Feb-13 17:13:20, Context 1, Stream 7
   Section: Command line profiler metrics
---------------------------------------------------------
l1tex  t bytes pipe lsu mem global op st.sum.per_second
Mbyte/second                                      89.89
---------------------------------------------------------
 gSum(int *, int *), 2023-Feb-13 17:13:20, Context 1, Stream 7
   Section: Command line profiler metrics
---------------------------------------------------------
   l1tex  t bytes pipe lsu mem global op st.sum.per_second
Mbyte/second                                      41.24
---------------------------------------------------------
```

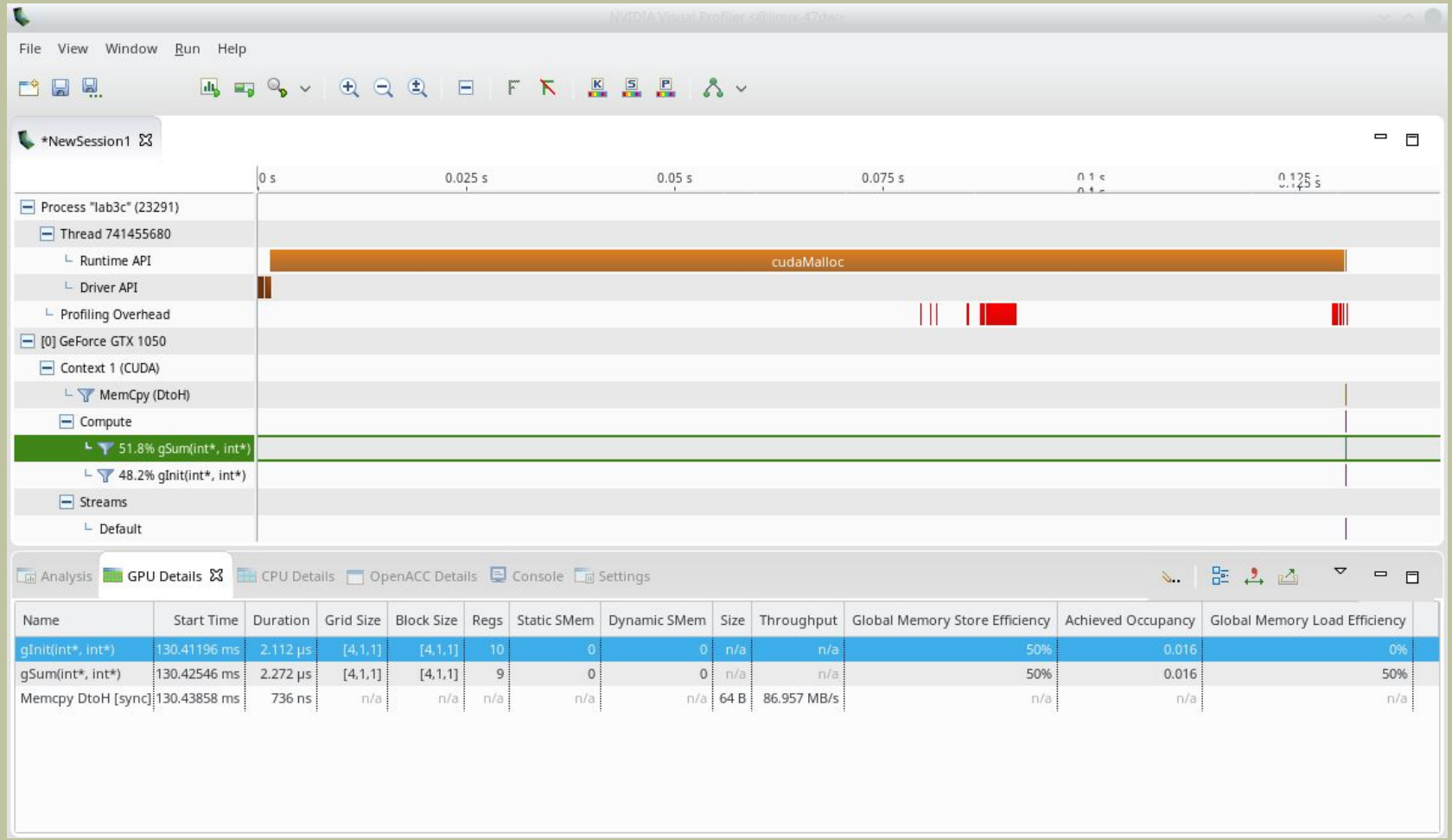
```
ip-011@linux-47dw:/home/malkov/WORKSHOP/PGP-2023>
nvprof -m gld_throughput ./lab3c
```

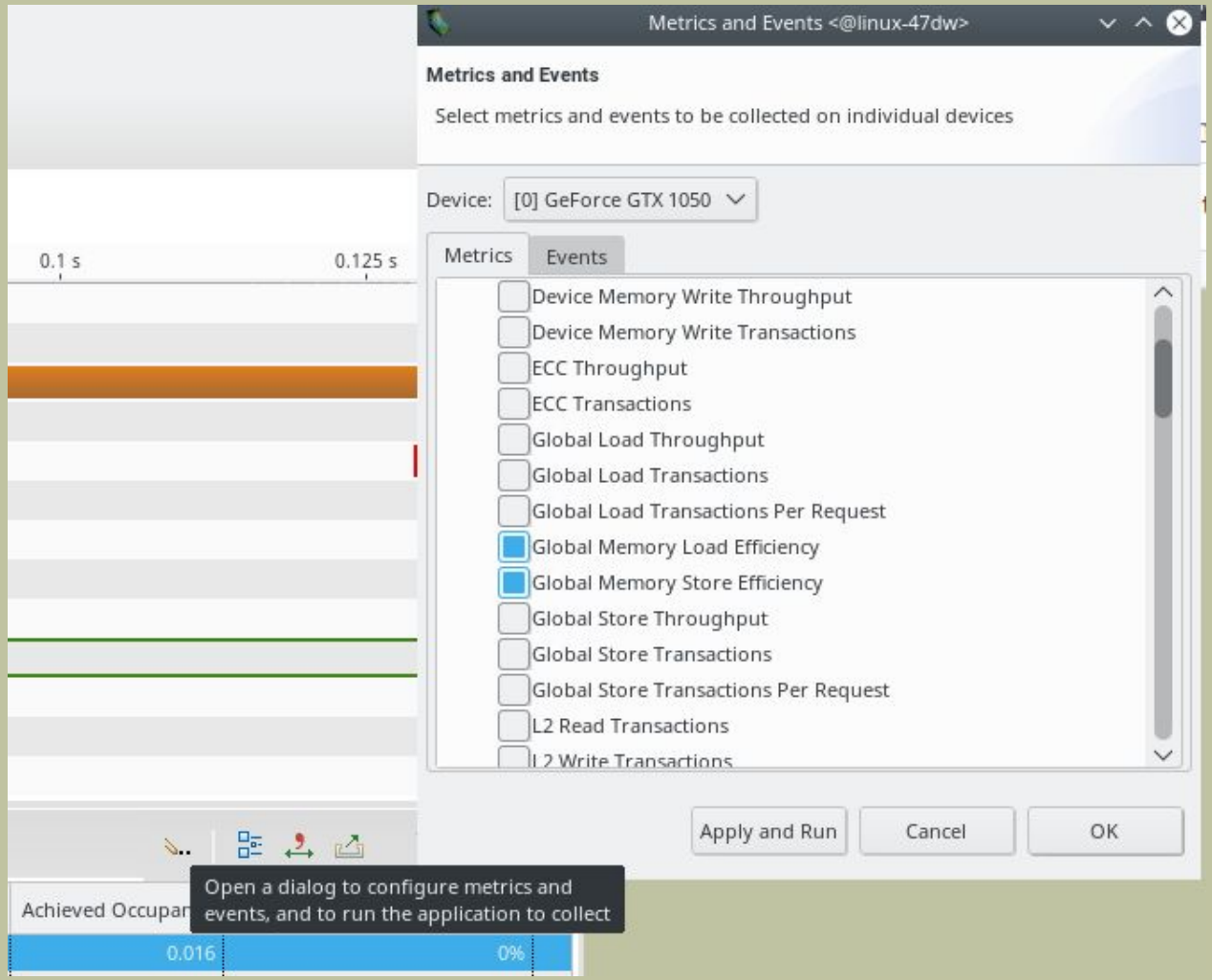
```
Invocations   Metric Name  Metric Description  Min Max    Avg
Device "GeForce GTX 1050 (0)"
   Kernel: gInit(int*, int*)
 1   gld_throughput Global Load Throughput 0.0B/s 0.0B/s  0.0B/s
   Kernel: gSum(int*, int*)
 1   gld_throughput Global Load Throughput  87.694MB/s
                           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

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




Metrics and Events <@linux-47dw>
Metrics and Events
Select metrics and events to be collected on individual devices
Device: [0] GeForce GTX 1050
Metrics
Events
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
Apply and Run
Cancel
OK
0.1 s
0.125 s
Open a dialog to configure metrics and events, and to run the application to collect
Achieved Occupan
0.016
0%

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