Лекция 3

Инструменты профилирования:

- nvprof
- nvvp
- Nsight Compute CLI
- Nsight Compute

nvprof и Nsight Compute CLI

```
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
```

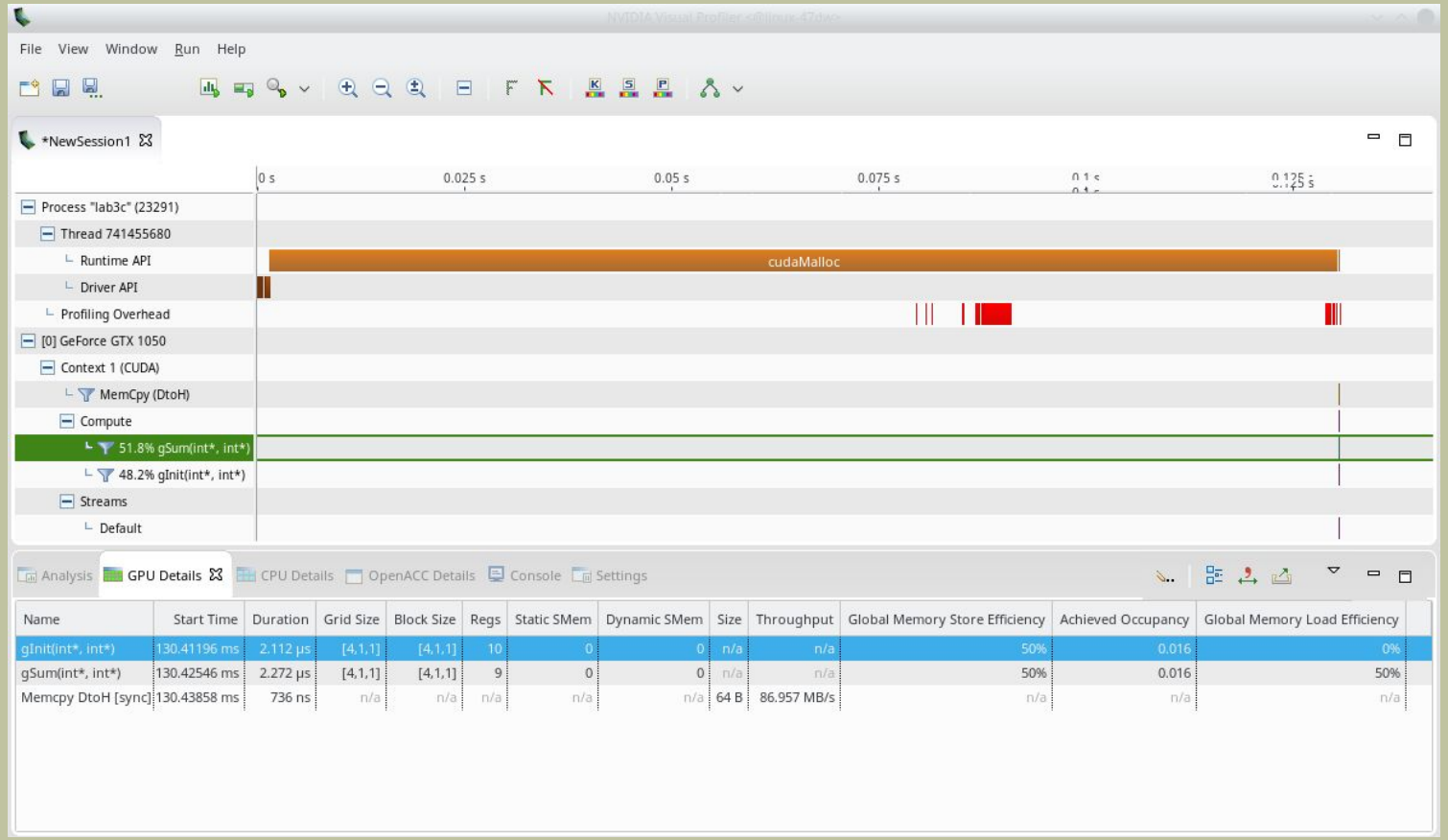
```
Invocations   Metric Name  Metric Description  Min Max    Avg
Device "GeForce GTX 1050 (0)"
   Kernel: gInit(int*, int*)
 1   gld_throughput Global Load Throughput 0.0B/s 0.0B/s  0.0B/s
   Kernel: gSum(int*, int*)
 1   gld_throughput Global Load Throughput  87.694MB/s
                           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

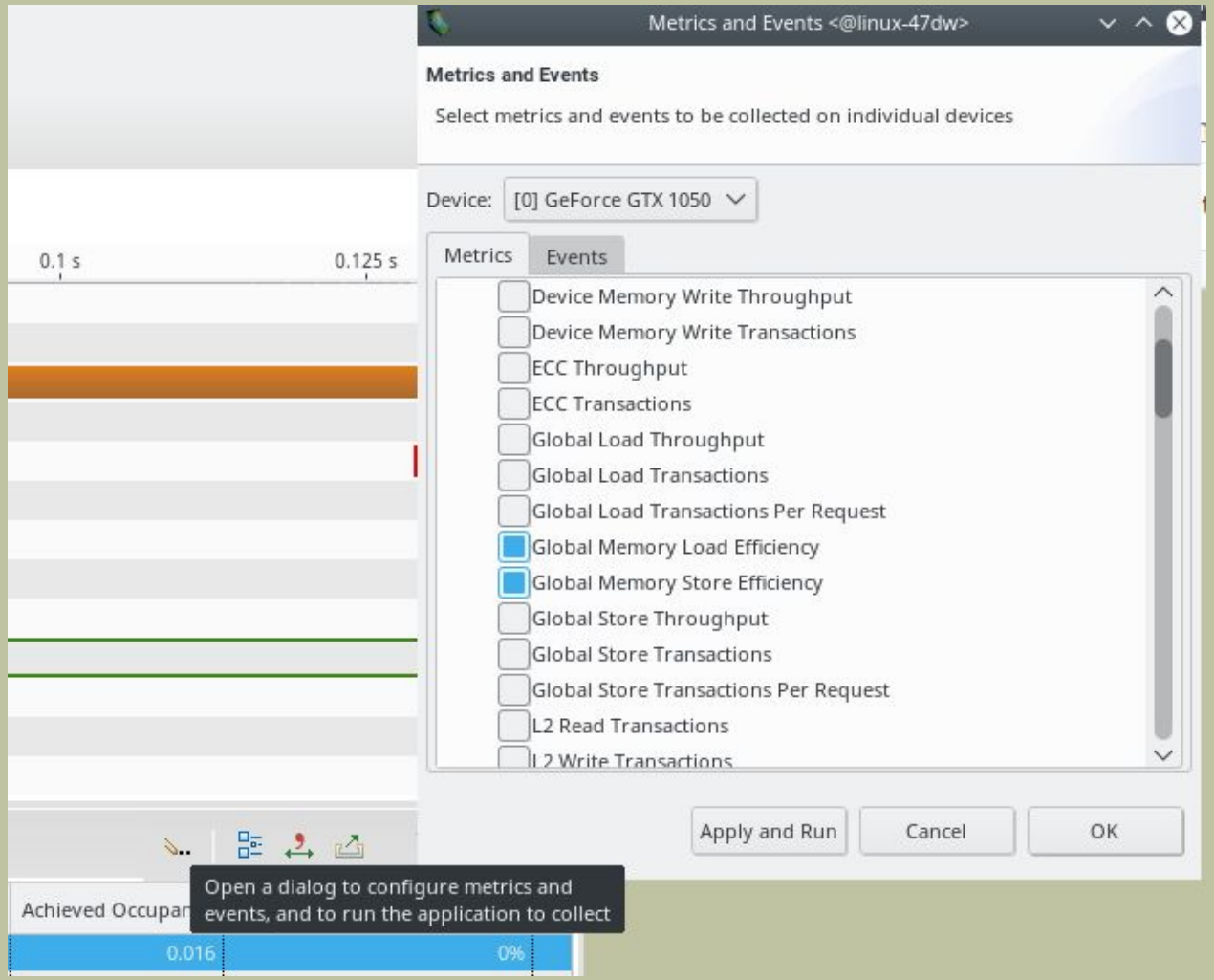
```
/Lecture3/Lab3-cuda-gdb # ncu --metrics
l1tex  t_bytes_pipe_lsu_mem_global_op_ld.sum.per_second
./lab3c
```

```
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
   ---------------------------------------------------
```

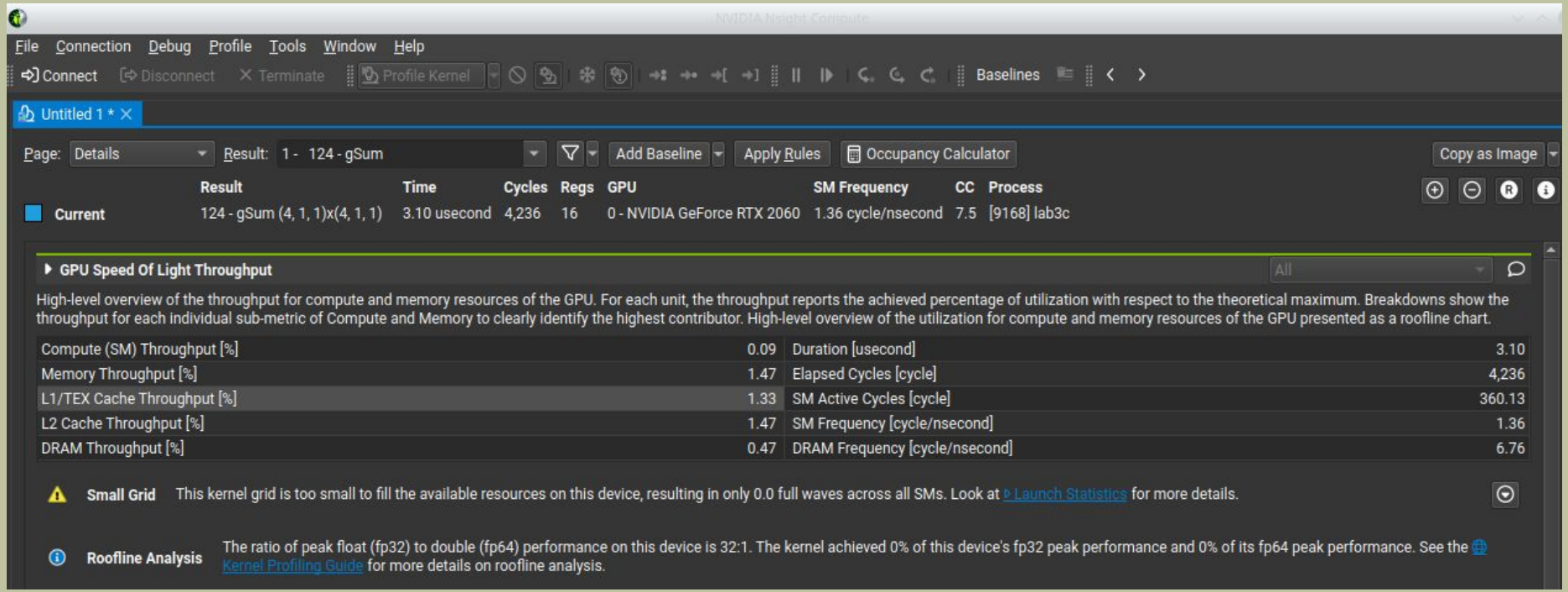
nvvp и Nsight Compute

```
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%

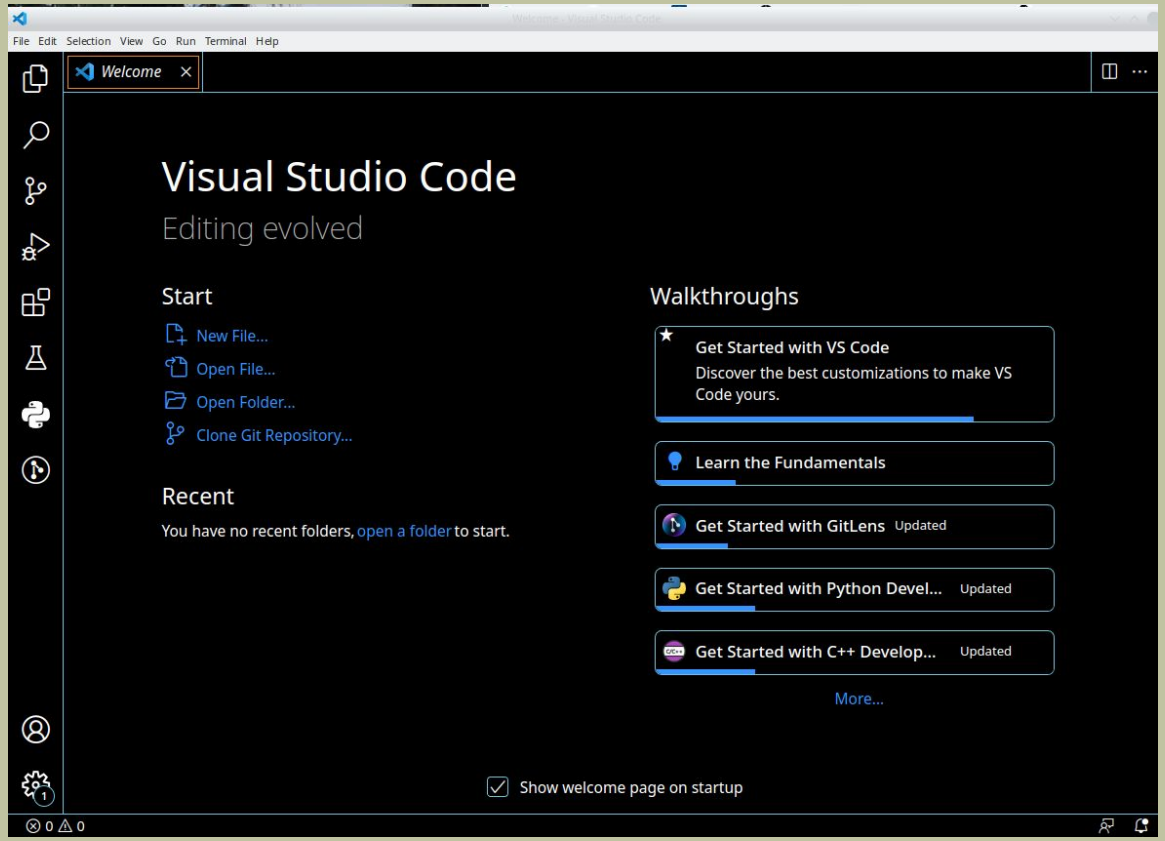
```
/Lecture3/Lab3-cuda-gdb # ncu-ui --target-processes all  ./lab3c
```

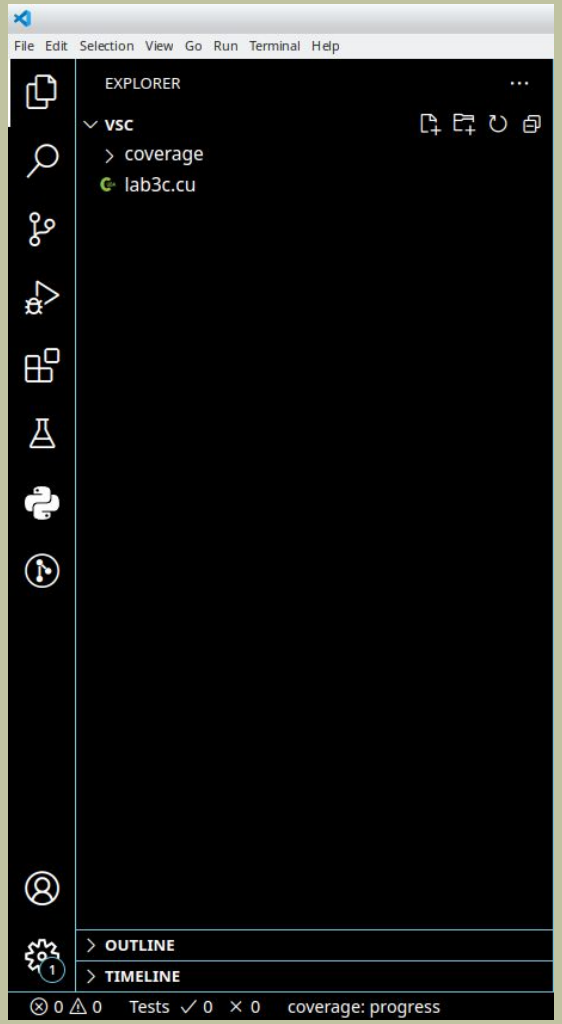


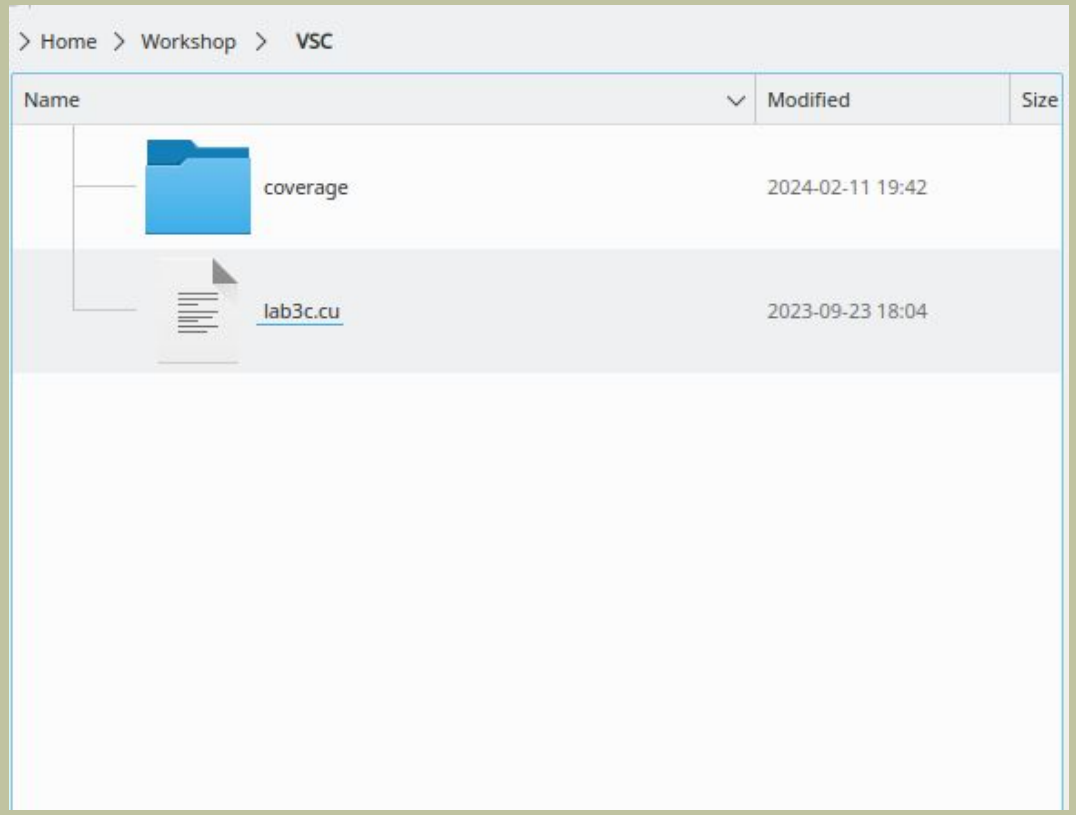
Nsight Visual Studio Code Edition

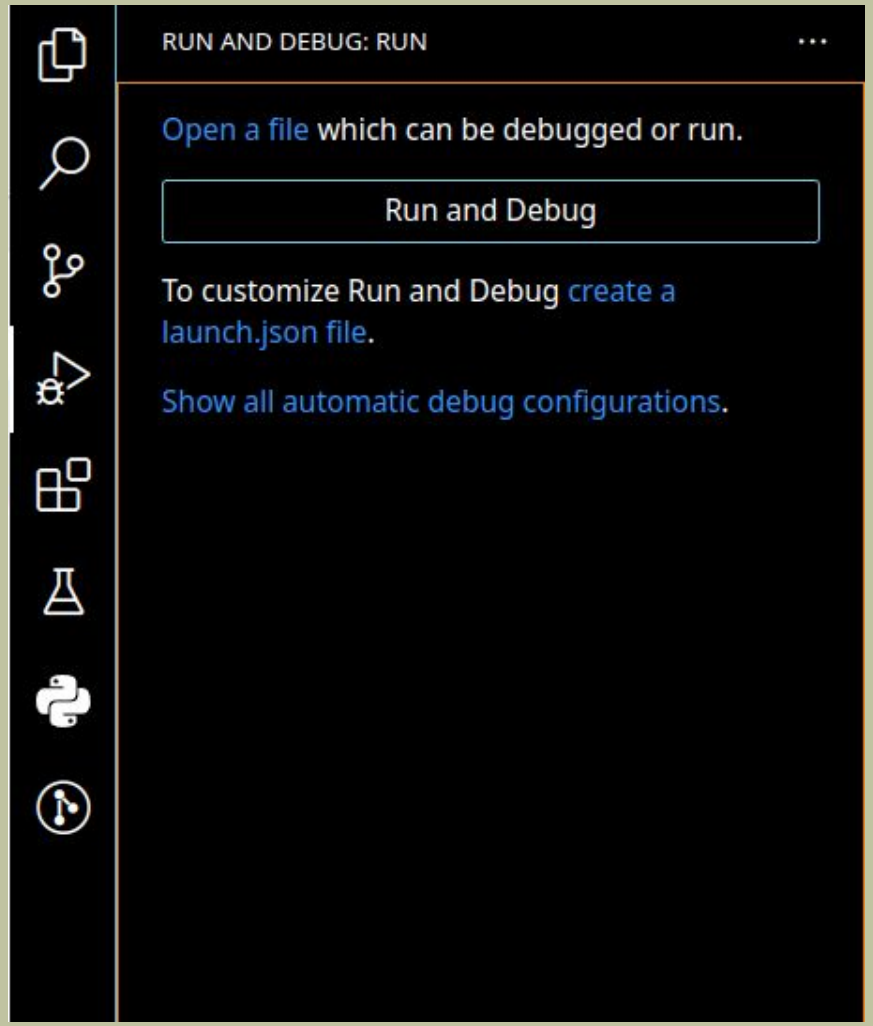
https://developer.nvidia.com/**nsight-visual-studio-code-edition**

https://docs.nvidia.com/**nsight-visual-studio-code-edition**/cuda-debugger/index.html



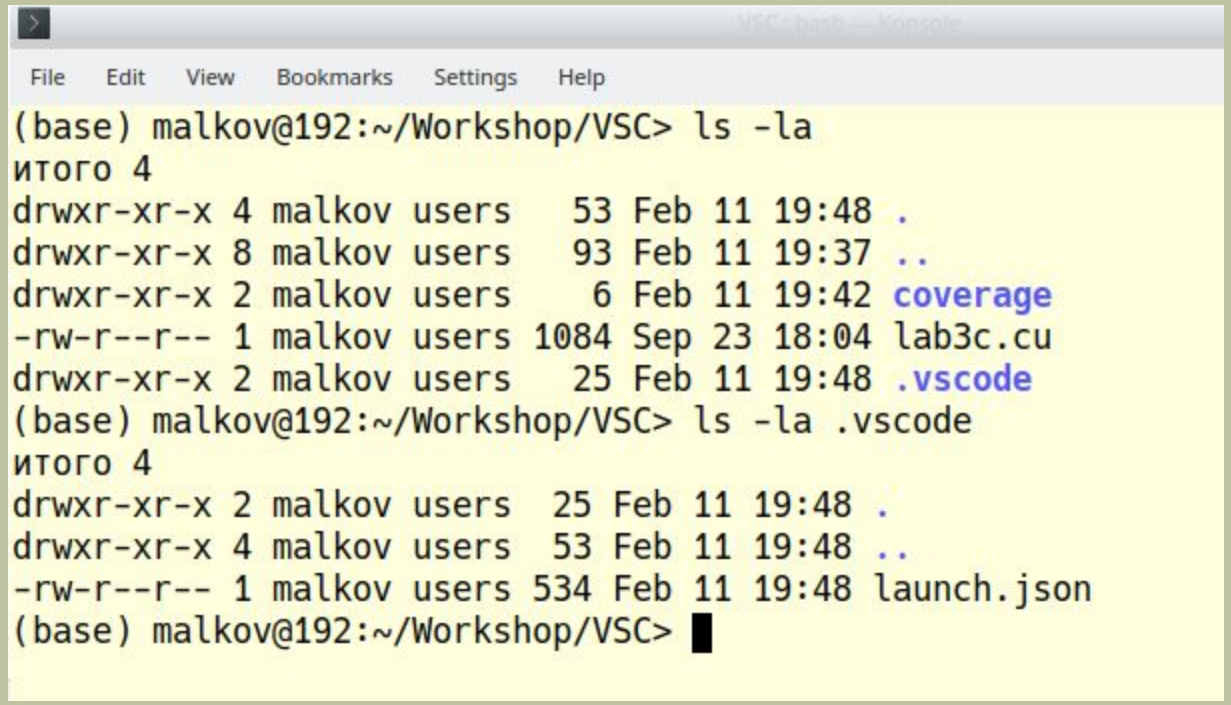

File Edit Selection View Go Run Terminal Help
EXPLORER
vsc
coverage
lab3c.cu
OUTLINE
TIMELINE
0 0 Tests 0 0 coverage: progress

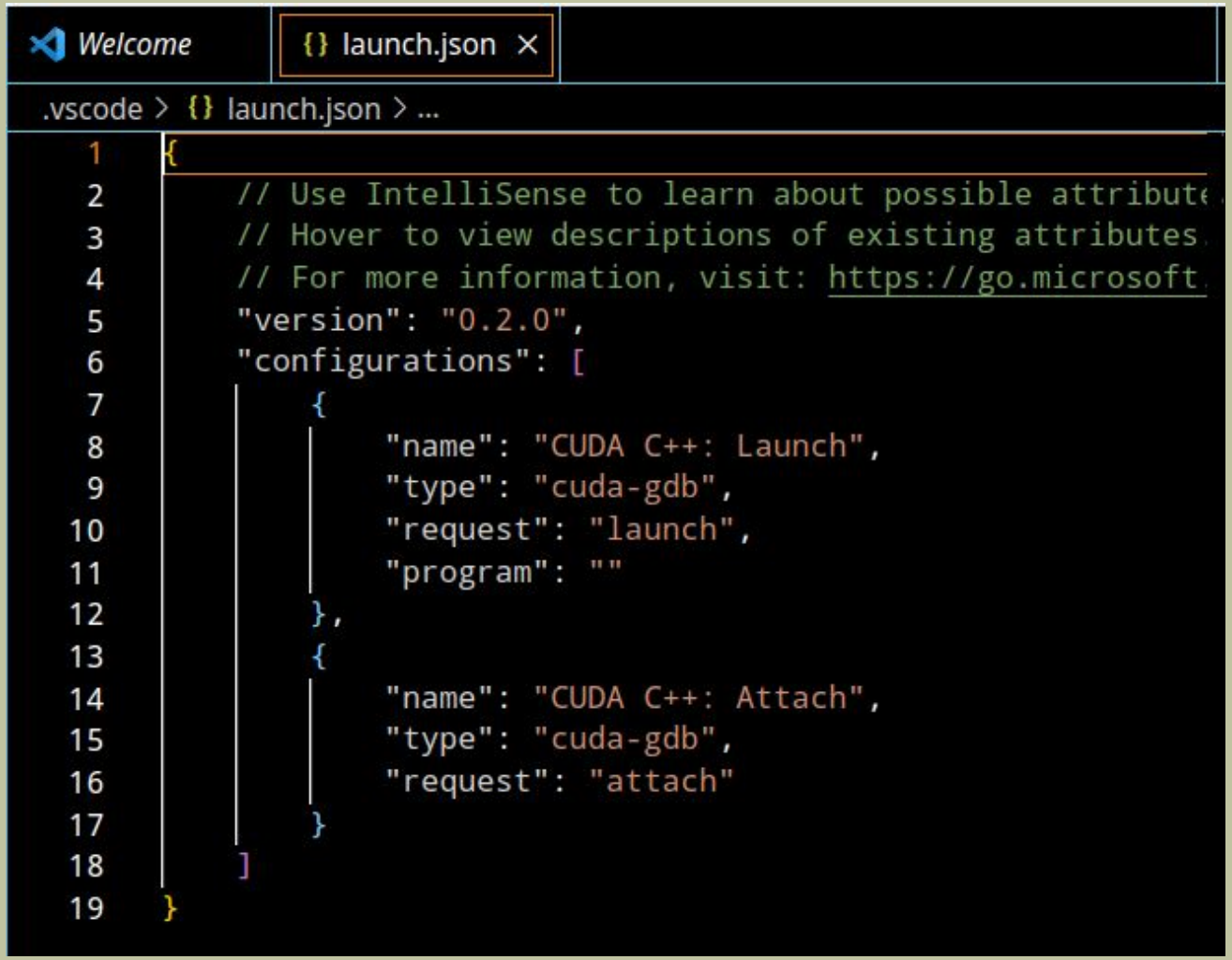

Home > Workshop > VSC
Name
Modified
Size
coverage
2024-02-11 19:42
lab3c.cu
2023-09-23 18:04

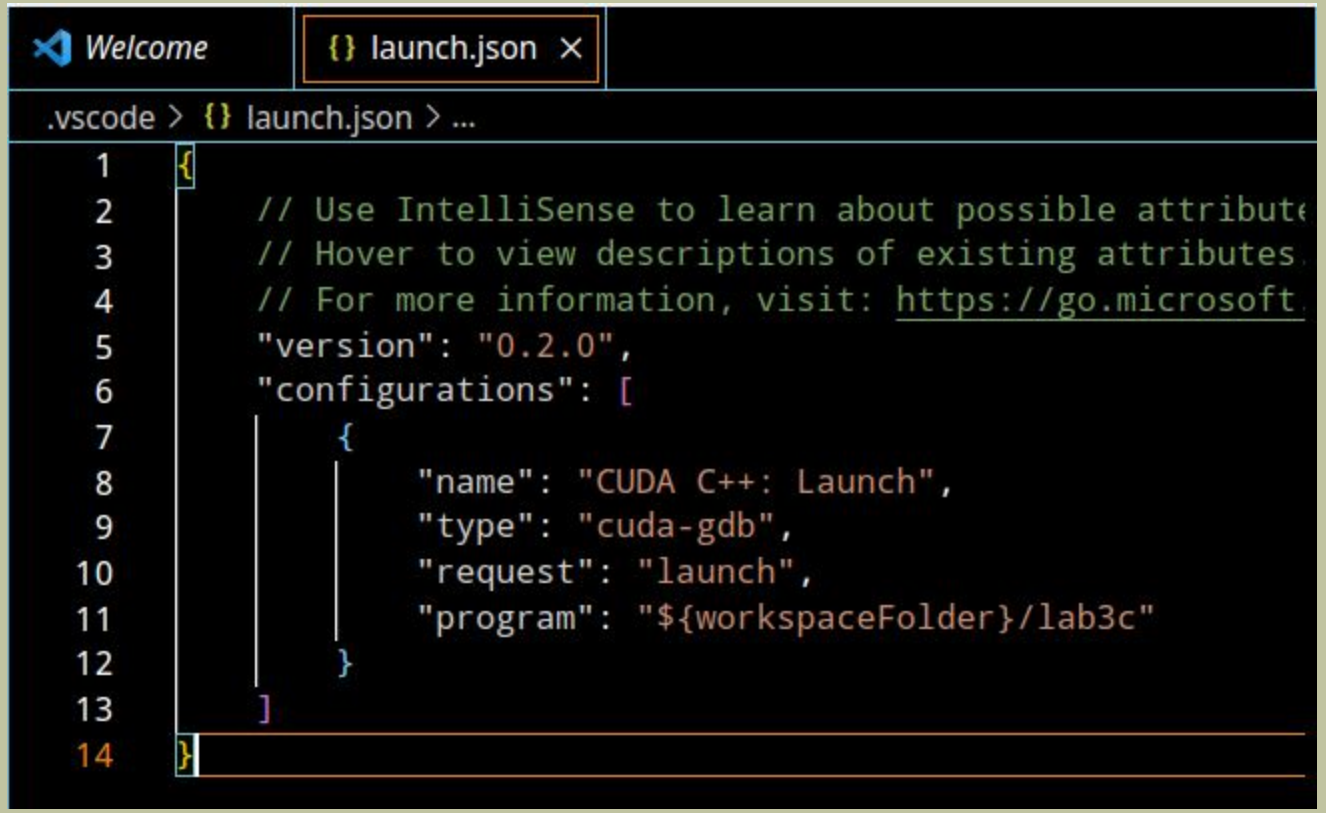

RUN AND DEBUG: RUN
Open a file which can be debugged or run.
Run and Debug
To customize Run and Debug create a launch.json file.
Show all automatic debug configurations.

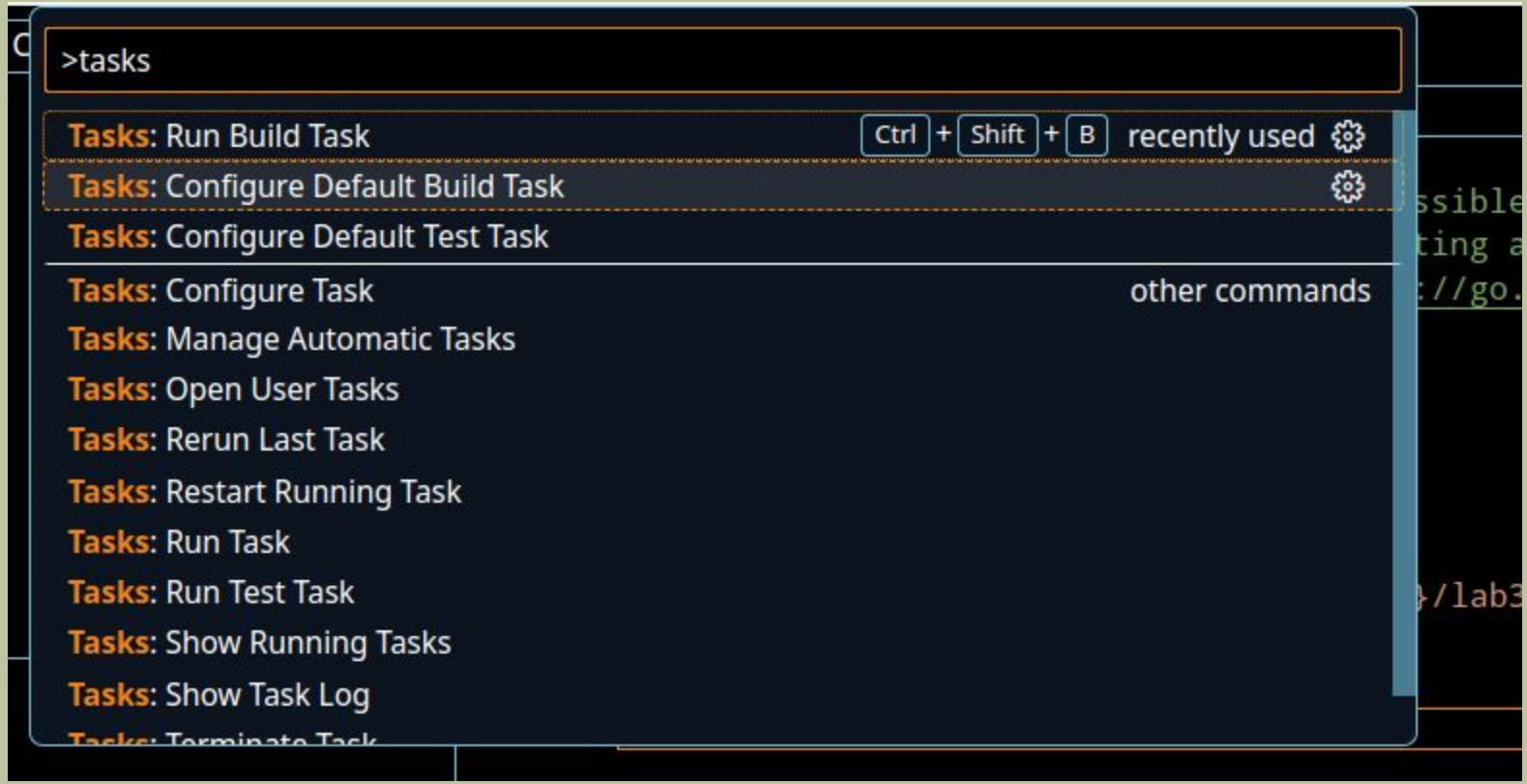

RUN AND DEBUG: RUN
Open a file which can
Run a
To customize Run and
launch.json file.
Show all automatic de
Extensions (Ctrl+Shift+X)
Select debugger
CUDA C++ (CUDA-GDB)
CUDA C++ (CUDA-GDBSERVER)
CUDA C++ QNX (CUDA-GDBSERVER)
CMake Debugger
Node.js
Python
Web App (Chrome)
Web App (Edge)
Install extension...

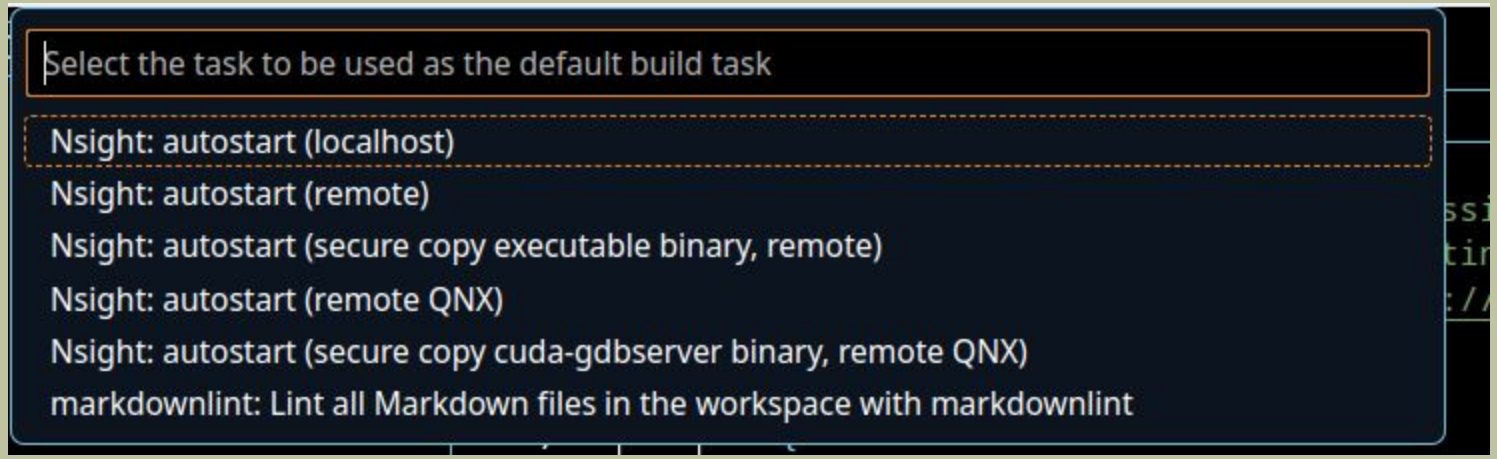

RUN AND DEBUG: RUN
Open a file which can
Run a
To customize Run and
launch.json file.
Show all automatic de
Select debugger
CUDA C++ (CUDA-GDB) Suggested
CUDA C++ (CUDA-GDBSERVER)
CUDA C++ QNX (CUDA-GDBSERVER)
CMake Debugger
Node.js
Python
Web App (Chrome)
Web App (Edge)
Install extension...

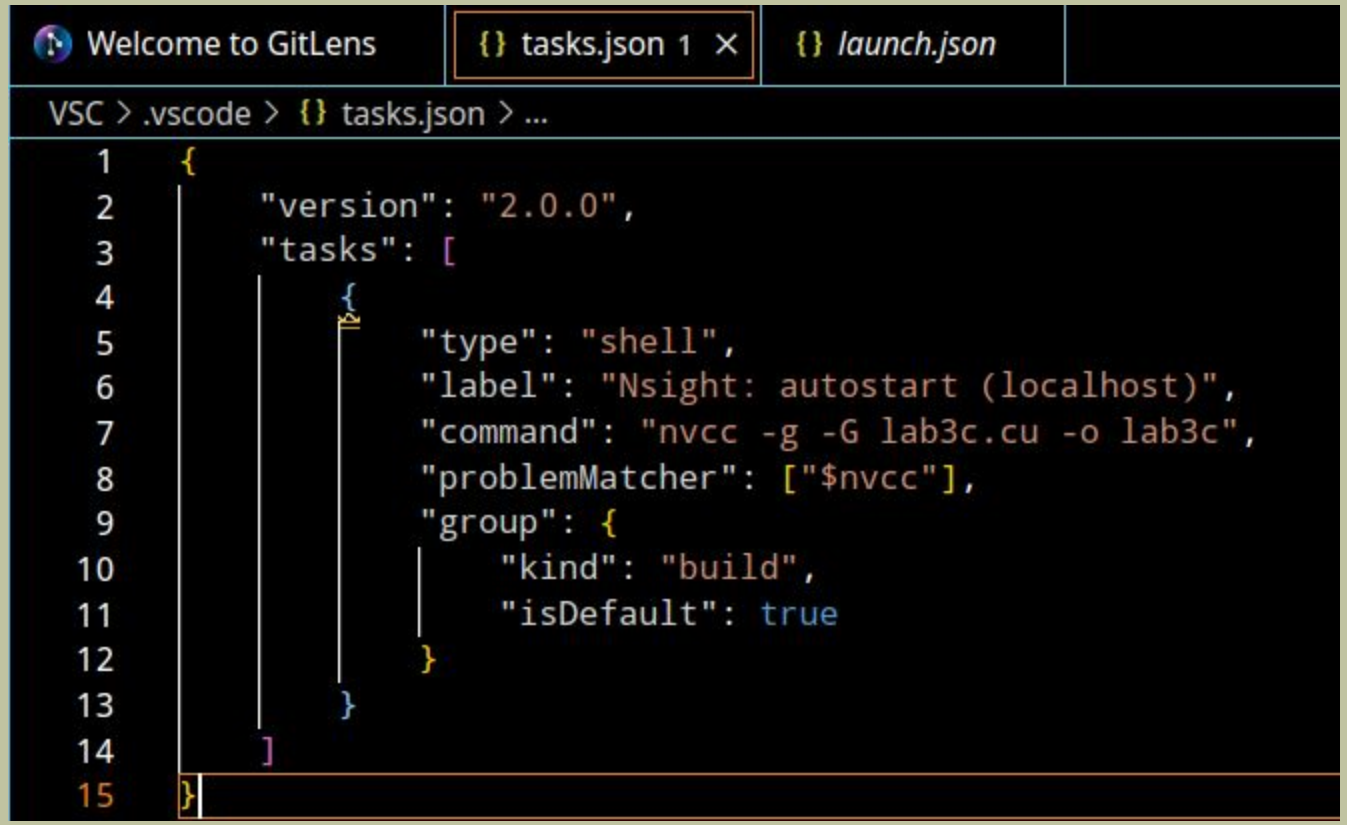

File Edit View Bookmarks Settings Help
(base) malkov@192:~/Workshop/VSC> ls -la
итого 4
drwxr-xr-x 4 malkov users   53 Feb 11 19:48 .
drwxr-xr-x 8 malkov users   93 Feb 11 19:37 ..
drwxr-xr-x 2 malkov users   6 Feb 11 19:42 coverage
-rw-r--r-- 1 malkov users 1084 Sep 23 18:04 lab3c.cu
drwxr-xr-x 2 malkov users   25 Feb 11 19:48 .vscode
(base) malkov@192:~/Workshop/VSC> ls -la .vscode
итого 4
drwxr-xr-x 2 malkov users 25 Feb 11 19:48 .
drwxr-xr-x 4 malkov users 53 Feb 11 19:48 ..
-rw-r--r-- 1 malkov users 534 Feb 11 19:48 launch.json
(base) malkov@192:~/Workshop/VSC>

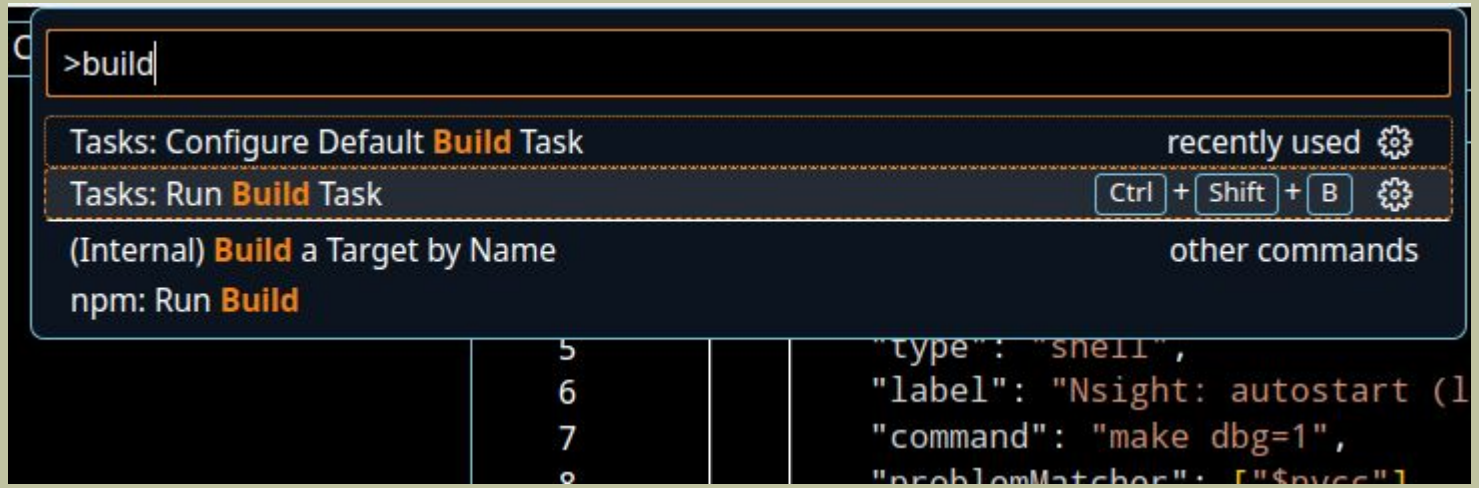

Welcome
{} launch.json
.vscode > {} launch.json > ...
{
// Use IntelliSense to learn about possible attribute
// Hover to view descriptions of existing attributes
// For more information, visit: https://go.microsoft
"version": "0.2.0",
"configurations": [
{
"name": "CUDA C++: Launch",
"type": "cuda-gdb",
"request": "launch",
"program": ""
},
{
"name": "CUDA C++: Attach",
"type": "cuda-gdb",
"request": "attach"
}
]
}

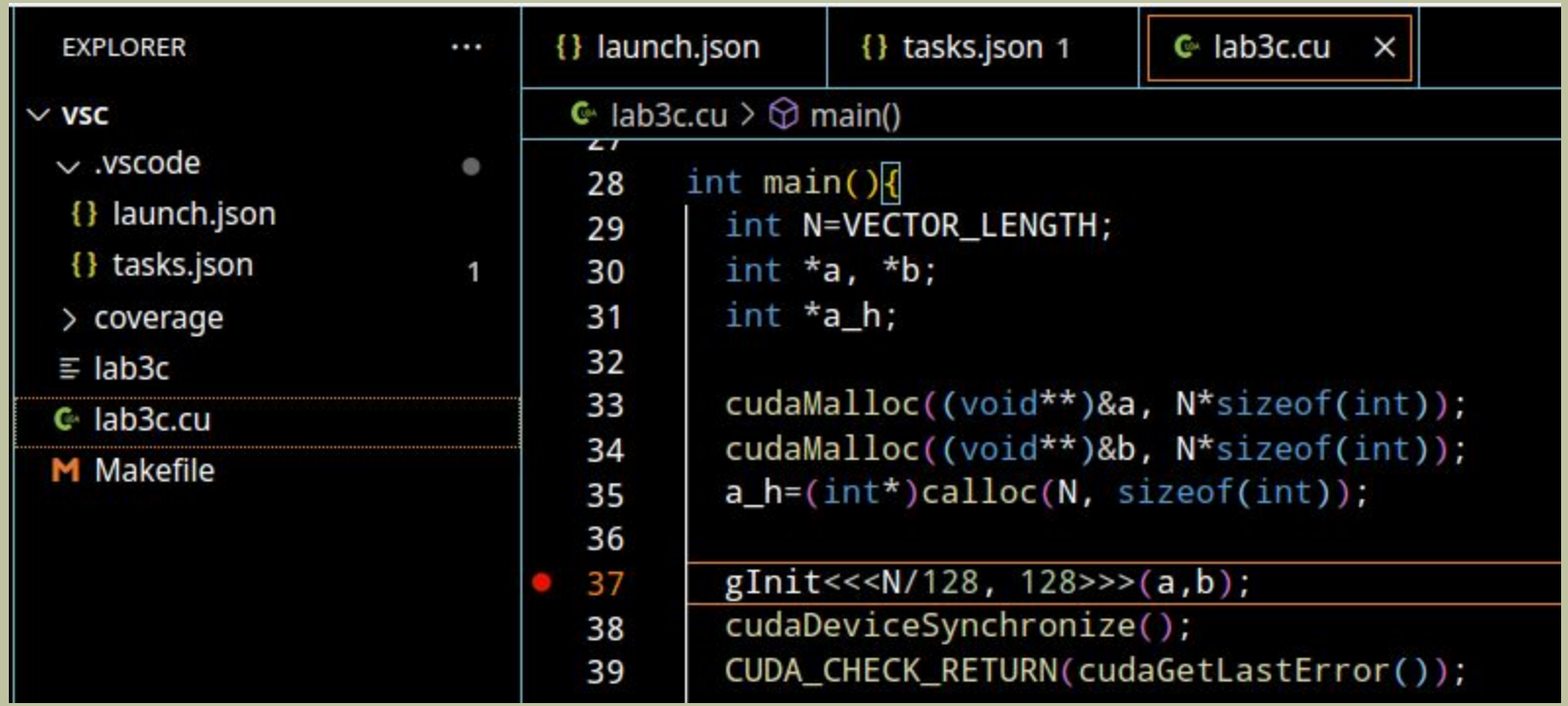

Welcome
launch.json
.vscode > launch.json > ...
{
// Use IntelliSense to learn about possible attribute
// Hover to view descriptions of existing attributes.
// For more information, visit: https://go.microsoft.
"version": "0.2.0",
"configurations": [
{
"name": "CUDA C++: Launch",
"type": "cuda-gdb",
"request": "launch",
"program": "${workspaceFolder}/lab3c"
}
]
}

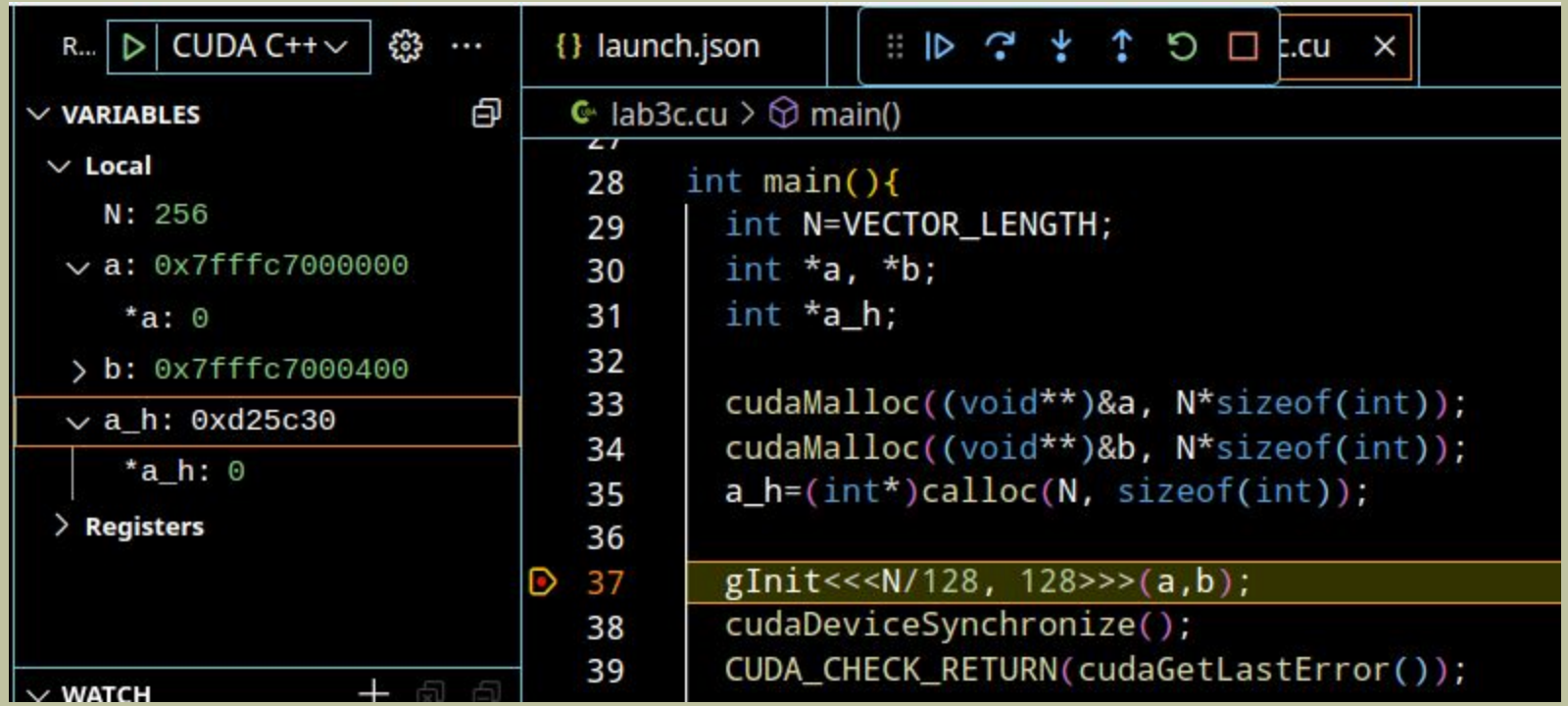

>tasks
Tasks: Run Build Task
Ctrl + Shift + B
recently used
Tasks: Configure Default Build Task
Tasks: Configure Default Test Task
Tasks: Configure Task
other commands
Tasks: Manage Automatic Tasks
Tasks: Open User Tasks
Tasks: Rerun Last Task
Tasks: Restart Running Task
Tasks: Run Task
Tasks: Run Test Task
Tasks: Show Running Tasks
Tasks: Show Task Log


Select the task to be used as the default build task
Nsight: autostart (localhost)
Nsight: autostart (remote)
Nsight: autostart (secure copy executable binary, remote)
Nsight: autostart (remote QNX)
Nsight: autostart (secure copy cuda-gdbserver binary, remote QNX)
markdownlint: Lint all Markdown files in the workspace with markdownlint


Welcome to GitLens
{} tasks.json 1
{} launch.json
VSC > .vscode > {} tasks.json > ...
{
"version": "2.0.0",
"tasks": [
{
"type": "shell",
"label": "Nsight: autostart (localhost)",
"command": "nvcc -g -G lab3c.cu -o lab3c",
"problemMatcher": ["$nvcc"],
"group": {
"kind": "build",
"isDefault": true
}
}
]
}


>build
Tasks: Configure Default Build Task
recently used
Tasks: Run Build Task
Ctrl + Shift + B
(Internal) Build a Target by Name
other commands
npm: Run Build
5
6
7
"type": "shell",
"label": "Nsight: autostart (l
"command": "make dbg=1",


EXPLORER
launch.json
tasks.json 1
lab3c.cu
vsc
.vscode
launch.json
tasks.json 1
coverage
lab3c
lab3c.cu
Makefile
lab3c.cu > main()
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/128, 128>>>(a,b);
cudaDeviceSynchronize();
CUDA_CHECK_RETURN(cudaGetLastError());


R...
CUDA C++
launch.json
.cu
VARIABLES
Local
N: 256
a: 0x7fffc7000000
*a: 0
b: 0x7fffc7000400
a_h: 0xd25c30
*a_h: 0
Registers
WATCH
lab3c.cu > main()
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/128, 128>>>(a,b);
cudaDeviceSynchronize();
CUDA_CHECK_RETURN(cudaGetLastError());