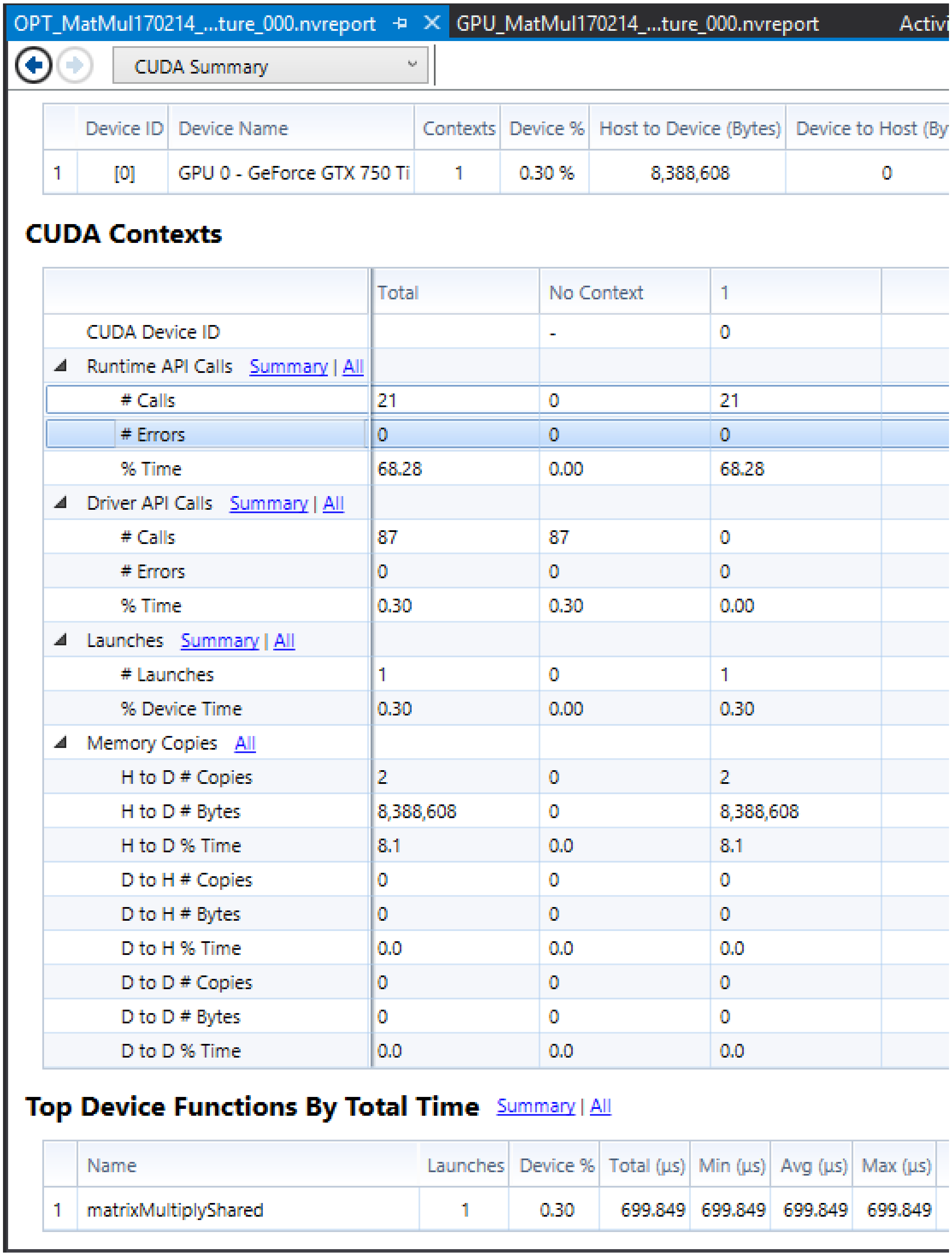
From NSIGHT, the GPU multiplication has a runtime of 41,395us. The OPT multiplication has a runtime of 699us. From this, we can see that the OPT multiplication has improved the speed by a factor of ~60x.

The speedup recorded in grades.html from test.bat is ~42x.

These finding show that the shared memory implementation of matrix multiplication using tiling is significantly faster than using global memory. This is only true when appropriate grid and block dimensions were set. Inappropriate grid and block dimensions have the complication of either slowing the algorithm down, or causing it to return incorrect results for the new matrix. In this lab, I learned that it is important to ensure proper dimensions are set.

I encountered problems as I initially had an inappropriate grid dimension set for the OPT, which was causing me to pass the test cases, but fail due to an insignificant speedup.

I also had another problem while doing this lab, as the antivirus on the lab computers prevented me from profiling the CPU multiplication using NSIGHT.


OPT_MatMul170214_...ture_000.nvreport
GPU_MatMul170214_...ture_000.nvreport
Activi
CUDA Summary

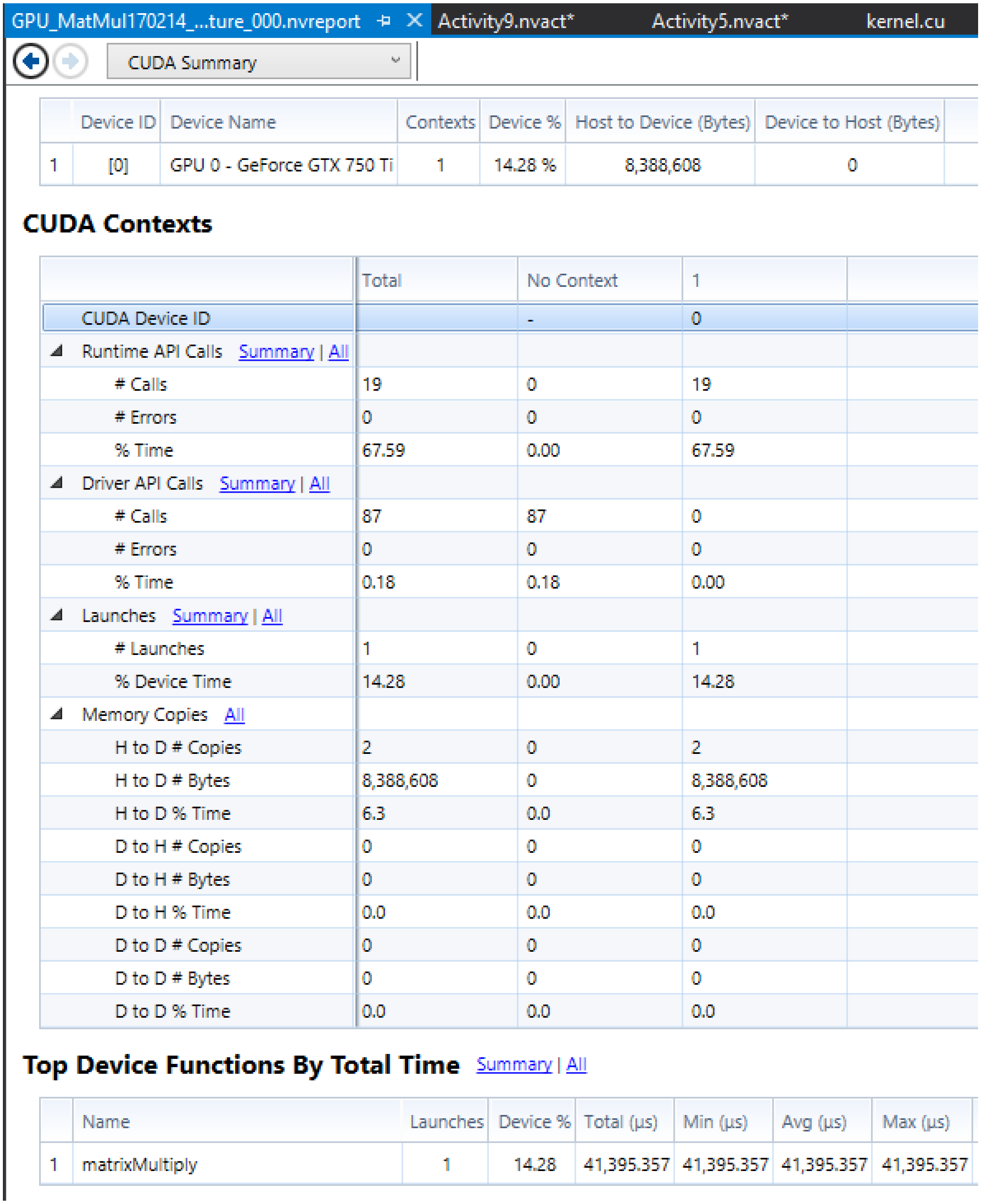
	Device ID	Device Name	Contexts	Device %	Host to Device (Bytes)	Device to Host (By
1	[0]	GPU 0 - GeForce GTX 750 Ti	1	0.30 %	8,388,608	0

CUDA Contexts

	Total	No Context	1
CUDA Device ID		-	0
Runtime API Calls Summary | All			
# Calls	21	0	21
# Errors	0	0	0
% Time	68.28	0.00	68.28
Driver API Calls Summary | All			
# Calls	87	87	0
# Errors	0	0	0
% Time	0.30	0.30	0.00
Launches Summary | All			
# Launches	1	0	1
% Device Time	0.30	0.00	0.30
Memory Copies All			
H to D # Copies	2	0	2
H to D # Bytes	8,388,608	0	8,388,608
H to D % Time	8.1	0.0	8.1
D to H # Copies	0	0	0
D to H # Bytes	0	0	0
D to H % Time	0.0	0.0	0.0
D to D # Copies	0	0	0
D to D # Bytes	0	0	0
D to D % Time	0.0	0.0	0.0

Top Device Functions By Total Time Summary | All

	Name	Launches	Device %	Total (µs)	Min (µs)	Avg (µs)	Max (µs)
1	matrixMultiplyShared	1	0.30	699.849	699.849	699.849	699.849



	Device ID	Device Name	Contexts	Device %	Host to Device (Bytes)	Device to Host (Bytes)
1	[0]	GPU 0 - GeForce GTX 750 Ti	1	14.28 %	8,388,608	0

CUDA Contexts

	Total	No Context	1
CUDA Device ID		-	0
◢ Runtime API Calls Summary | All			
# Calls	19	0	19
# Errors	0	0	0
% Time	67.59	0.00	67.59
◢ Driver API Calls Summary | All			
# Calls	87	87	0
# Errors	0	0	0
% Time	0.18	0.18	0.00
◢ Launches Summary | All			
# Launches	1	0	1
% Device Time	14.28	0.00	14.28
◢ Memory Copies All			
H to D # Copies	2	0	2
H to D # Bytes	8,388,608	0	8,388,608
H to D % Time	6.3	0.0	6.3
D to H # Copies	0	0	0
D to H # Bytes	0	0	0
D to H % Time	0.0	0.0	0.0
D to D # Copies	0	0	0
D to D # Bytes	0	0	0
D to D % Time	0.0	0.0	0.0

Top Device Functions By Total Time Summary | All

	Name	Launches	Device %	Total (µs)	Min (µs)	Avg (µs)	Max (µs)
1	matrixMultiply	1	14.28	41,395.357	41,395.357	41,395.357	41,395.357