An introduction to Nsight Systems

Mike Giles

1 Overview and client installation

NVIDIA's Nsight Systems profiling tools are usually used in two stages.

First, the user uses a command such as

nsys profile application_code

to run their code and produce an output file with a name of the form filename.nsys-rep.

Then, the user runs an interactive graphics program to display the results of the profiling, using a command such as

nsys-ui filename.nsys-rep

The first command has to be used on a GPU node which is capable of executing your CUDA code. In this case, both nsys and nsys-ui are very likely to have been installed already.

However, the benefit of the two-stage process is that you can copy the report back to your desktop or laptop system to run nsys-ui locally to work with it interactively. In this case you may need to install nsys-ui on your local system even though you don't have an NVIDIA GPU.

The software and installation details are available from NVIDIA here.

On my Ubuntu system at home, for which I have sudo rights, I downloaded the appropriate *.deb file, and installed it using the command:

sudo apt install nsight-systems-2023.4.1_2023.4.1.97-1_amd64.deb

NOTE: this was correct on 6/12/2023 but may change to a newer version in the future.

For desktops within the university it might be better to ask your IT support team to install it for general use, or you could perhaps install it in your own user space just for your use.

2 Using the interactive viewer

Starting the viewer with a command of the form

```
nsys-ui filename.nsys-rep
```

produces a display which looks like

Cooole B BBC B Home NVIDIA Nsight Systems 2023.	n Me Bu Cuford Bu Tech Bu JADF Bu FDI 🔶 Rookmarks Bu Oden Bu AMAS Bu CIMDA Bu Filo 🚳 14th Internati o. Cuford Train S	- 0 8
Elle Yew Icols Help		
report1 nsys-rep ×		
Timeline View -		📾 🔍 🖳 📩 1x 🔥 <u>3 warnings, 11 messages</u>
• (a 1a 2a 3a 4a 5a 6a 7a 8a 9a 10a 11a 12a 13a	14s 15s 16s 17s 🔺
CUDA HW (0000.02.00.0 - NVIDIA		
 Threads (7) 		
+ (38434) laplace3d		
OS runtime libraries		1
CUDA API		
Profiler overhead		ü –
 [38444] cuda-EvtHandir 		
OS runtime libraries		
Profiler overhead		
5 threads hidden+		
		-
		,
Events View *		
		Name *
		Description
	regimentations a stimulate toward peakers show in Events view. To see events neve view	
		Ç.

Clicking on CUDA API, and then right-clicking on "Show in Events View" leads to the bottom Events View pane showing a list of the kernels and CUDA operations carried out on the GPU (including how long they each took):

Google BBC	C In Home In Me In	oxford 🐚 Tech ங JADF ங FDI 🕁 Bookmarks 🐚 Odeo. 📾 AlMS 🛍 CIMDA 🛍 Filo 🚳 14th Internati o Oxford Train S			1	meet
File View Tools	Help					000
report1 nava-rep X						
> Timeline View						s. 11 messages
	- 09	1a 2a 3a 4a 5a 6a 7a 8a 9a 10a 11a 12a		38	14a 15a 16a 17a	
 CUDA HW (000) 	0.02-00.0 - NV/DIA					
· Threads (7)						
+ [38434] laplac	bbe					
OS runtime	Ibraries					
CUDA API		Conference				
Paulin and	CUDA pro					1
Promier over	D.					
+ [38444] cuda-l	EvtHandlr					_
OS runtime l	Ibraries					881
Profiler over	rhead					
5 threads hidd	den +					
	1					*
Evente Mary	-					
Cremy Frem						
					Frame *	
* *	Name	Sart opp	Duration	TID	 Description: 	
-	edwooseeecoolingwooe	ualoora 100005	2.120 µ6	38434	-	
3	cudaDant/Create	0.020/930	517.002 me	38434		
4	cudaEventCreate	13078	3.468 µs	38434		
5	cudaMaloc	1.33764	1.470 ms	38434		
6	cudaMalloc	1 339329	1.334 ms	38434		
7	cudaEventRecord	2.024378	2.808 ms	38434		
8	cudaMemopy	2.02718s	64.689 ms	38434		
9	cudaEventRecord	2.091876	14.964 µs	38434		
10	cudaEventSynchronize	2,091895	68.633 µs	38434	*	
						-
						P .

Clicking on one of those, and then zooming in by ctrl+mouse wheel (which zooms in time relative to the current mouse location, a bit like Google Map zooms) gives the following:

IA Nright	Swrberne 2023 2		FDI T BOOKMARKS	Alles Alles			3						
w Tools	Help												
SUSCED X													
allow by a													
neine View	*											B Q	1x A swarrings, 1
	24 - 2011	s +92.02ms +92.04ms		+92.08ms +92.1ms	+92.12016	+92.14ms +92.16ms	+92.18ms	+92.2ms +92.22ms	+92.24ms	+92.22	ills +92.28ms	+92.3ms	+92.32ms
0000) WH AI	102:00.0 - NV 1												
eads (7)													
434J laplace	e3d			_									
S runtime li	ibraries		2										
UUM API		or comparison	0P0,38960630	0401 05- 05- 05- 0- 05	CI OFULIQUECESO OF COL				cosseventsynchron	20			
rofiler overt	head												
144] curla F	NtHandle												
nuntime li	ibratian						and .						
files quest	head												
eads hidd	en +												
	4												
0ew	*												
												Name	•
* 1	Name							Start	Duration	TID	 Description: 		
	cudaEventSynch	ronize						2.09189	68.633 µs	38434	Call to GPU_laplac	e3d	
	cudaEventRecom	8						2.09199	2.504 µs	38434	Kernel launcher Begins: 2.09205s		
	GPU_laplace3d							2.0928	46.829 µs	38434	Ends: 2.09208s (+)	29.120 µs)	
	GPU_laplace3d							2.09205	29.120 µs	38434	GPU: 0000.02.00.0	- NVIDIA GeForce RTX 2080	n
	GPU_laplace3d							2.09208	5.057 µs	38434	Stream: 7 Latency: 4,548 ms	-	
	uPU_isplace3d							2.09208	4.481 µs	38434	Correlation ID: 476		
	GPU_laplace3d							2.09209	3.961 µs	35434	-		
	GPU Japlace3d							2.09209	a.791 µs	20434	- 1		
	or operation							2.09210	0.007 pb	28/24	- 1		

Clicking on the keyboard symbol in the top-right gives other ways of navigating the primary Timeline View pane. There is also a slider bar which controls the vertical spacing.

For more examples of the use of nsys and nsys-ui see these slides from ISC 2023.