Nvidia Nsight Compute

https://youtu.be/8U4zPalF63Y
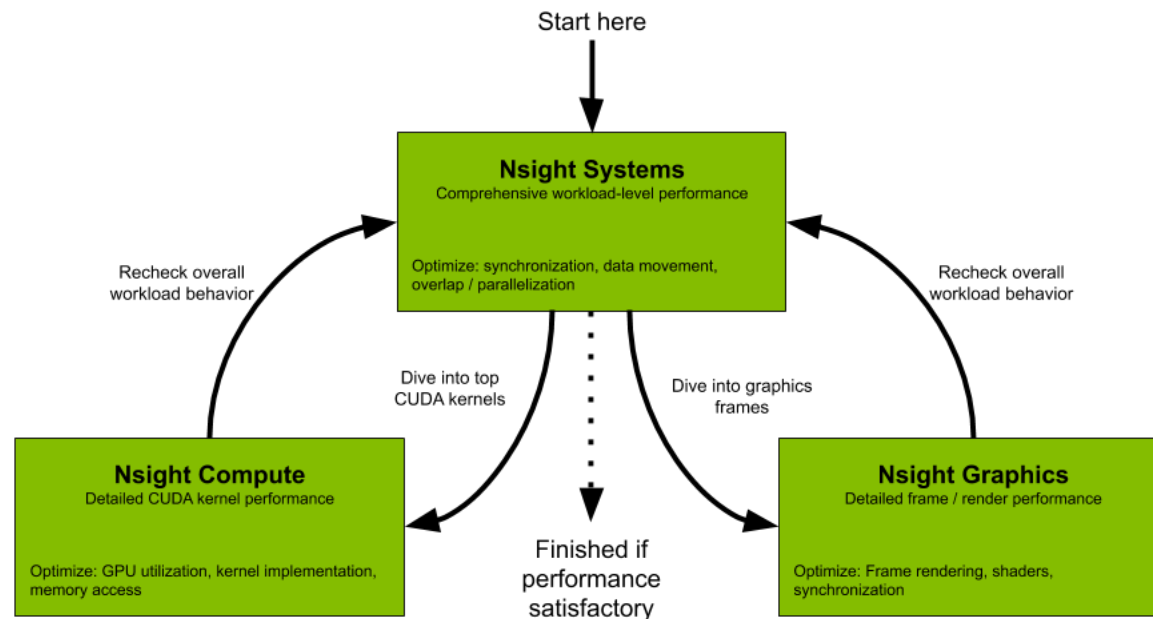
1. Nsight Compute?

- Nsight Compute는 CUDA 애플리케이션을 위한 대화형 커널 프로파일러

- UI 및 command line을 통해서 자세한 성능 메트릭 및 API 디버깅을 제공

1. Nsight Compute?

- Roofline Analysis
  - 시스템의 시각적 표현 메모리와 계산 용량을 제공
- Memory Workload Analysis
  - 프로파일링된 플랫폼의 메모리 처리량을 시각화
- Baseline Comparisons
  - GPU 아키텍처, 커널 시작 매개 변수, 메모리 사용량 등의 변화를 비교할 여러 기준선을 설정
- Run from Nsight Compute GUI or from Console Command Line
- CUDA Task Graph Profiling
- Source Code Correlation

2. Nsight Compute 설치

- https://developer.nvidia.com/tools-overview

NVIDIA® Nsight™ Visual Studio Edition

An application development environment for heterogeneous platforms which brings GPU computing into Microsoft Visual Studio. NVIDIA Nsight™ VSE allows you to build and debug integrated GPU kernels and native CPU code as well as inspect the state of the GPU and memory.

Read More >

**NVIDIA® Nsight™ Tools Integration for Visual Studio**

NVIDIA Nsight Integration is a Visual Studio extension that allows you to access the power of the following NVIDIA Nsight standalone tools from within Visual Studio:

- Nsight Compute : CUDA application interactive kernel profiler
- Nsight Graphics : Graphics application frame debugger and profiler/li>
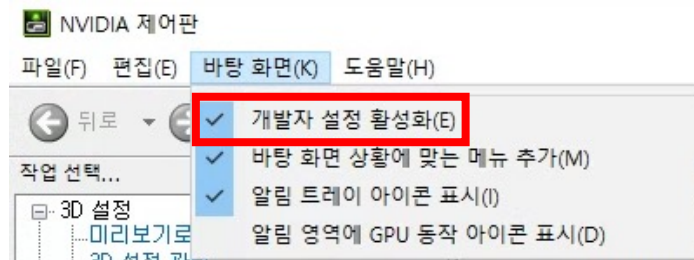- Nsight Systems : System-wide performance analysis tool/li>
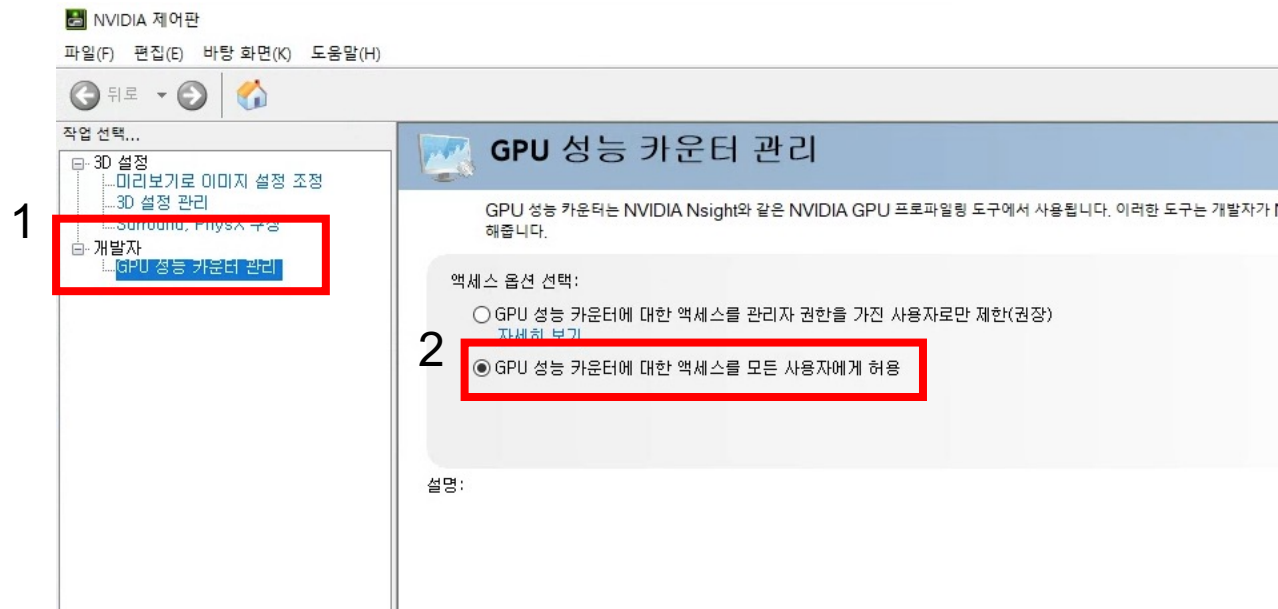
Read More >
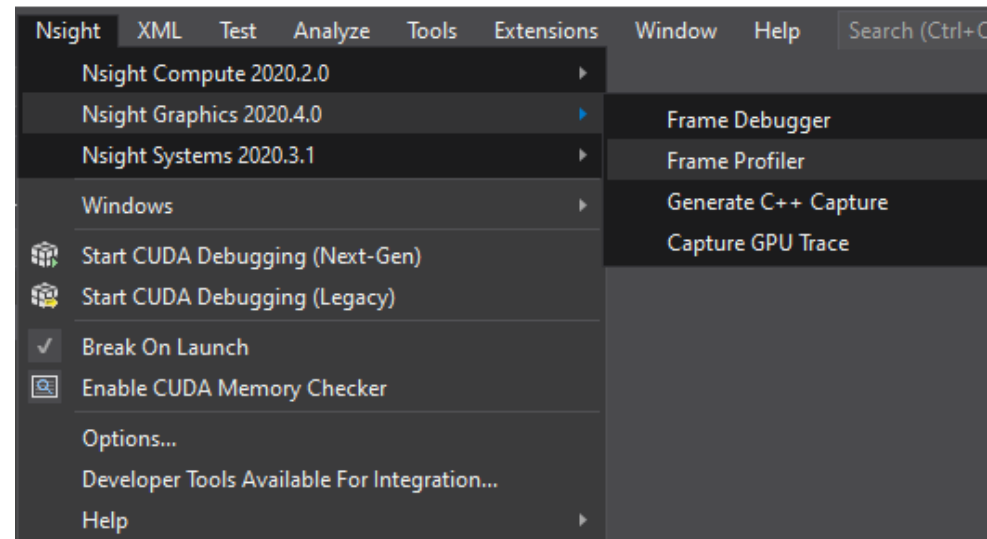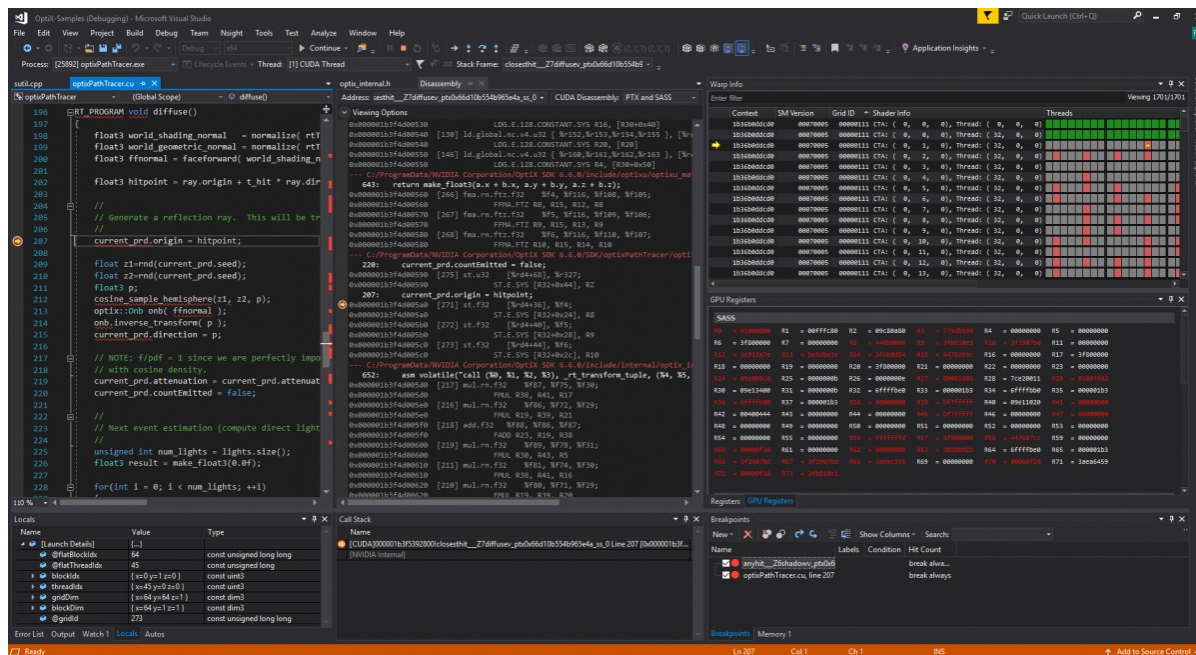
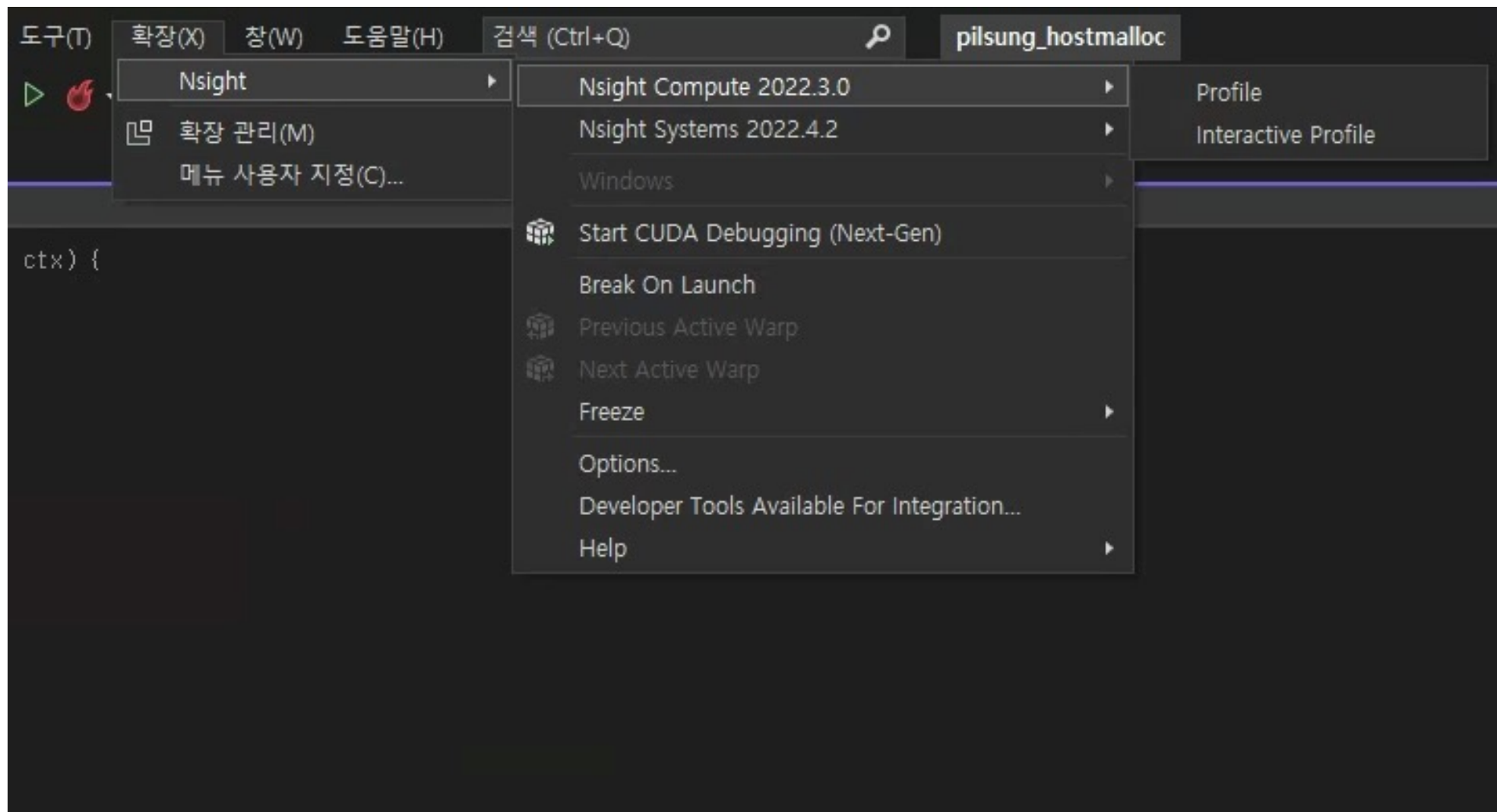# 2. Nsight Compute 설치
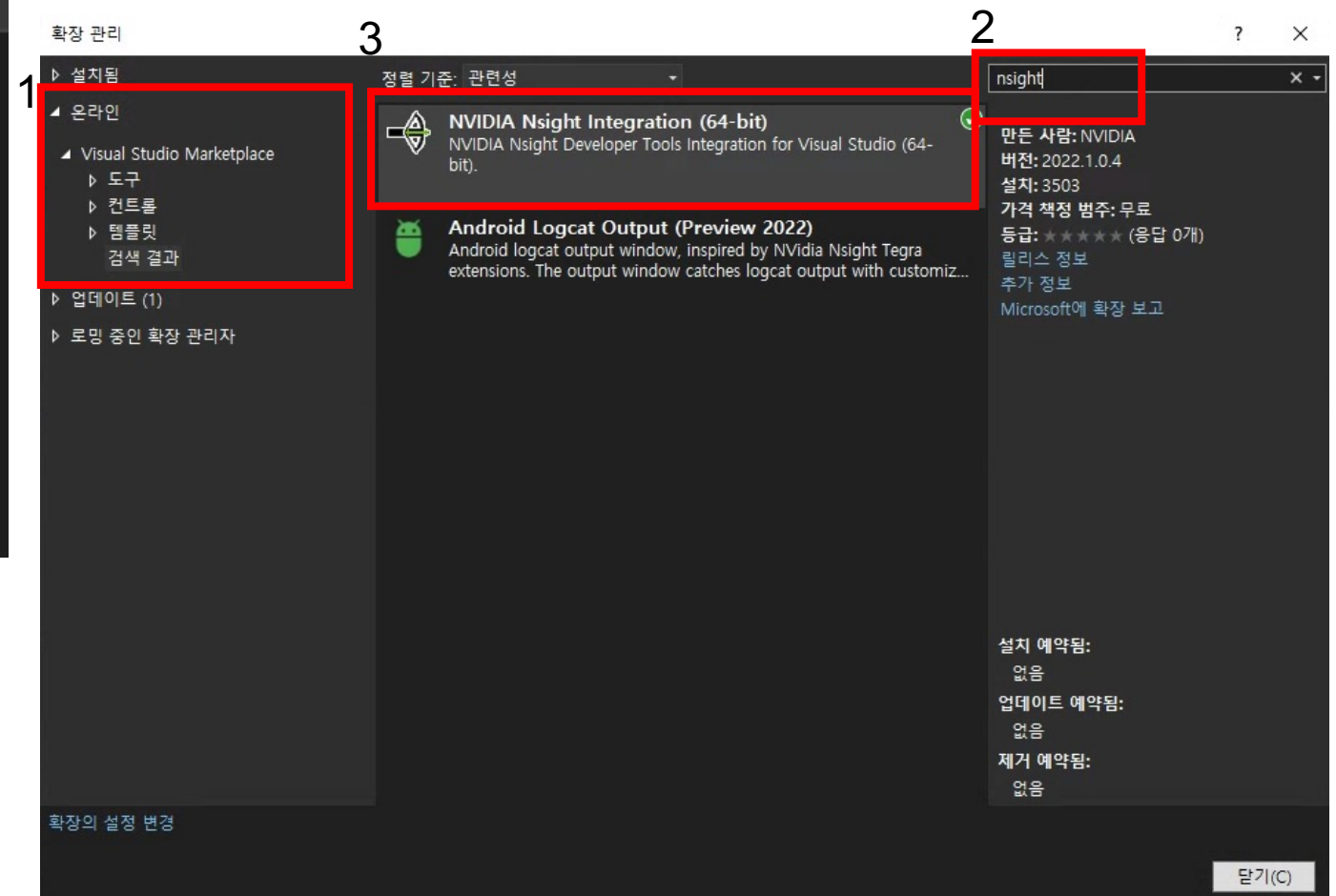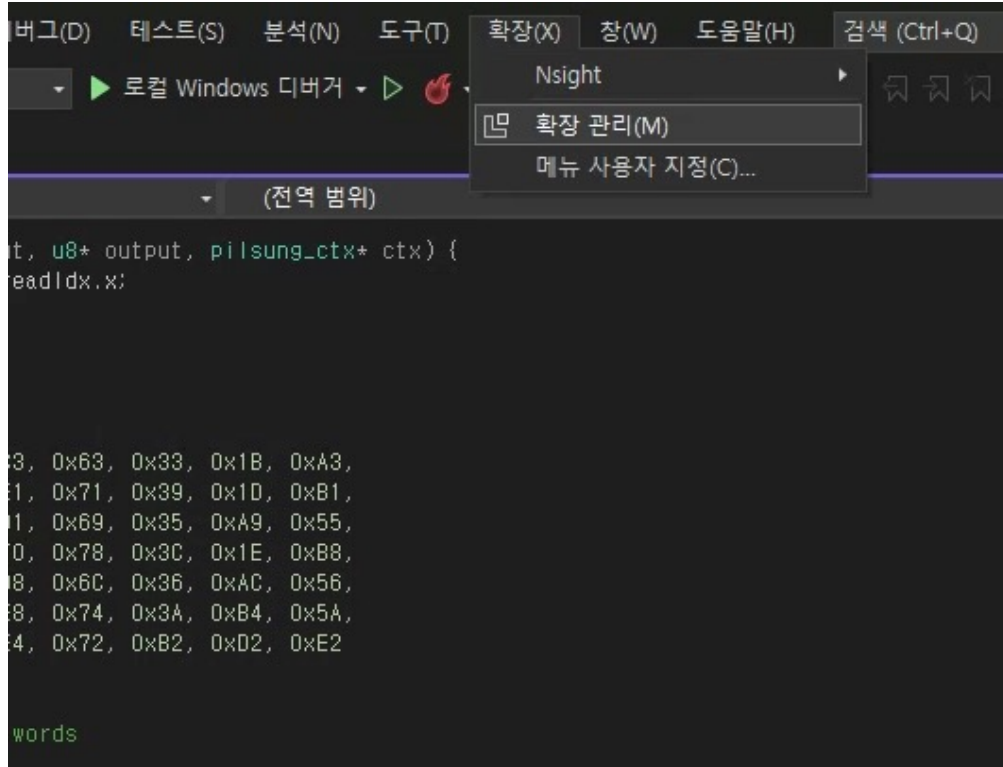
2. Nsight Compute 설치

- NVIDIA 제어판 – 바탕화면 – 개발자 설정 활성화(체크)

- GPU 성능 카운터 관리 – GPU 성능 카운터에 대한 액세스 허용
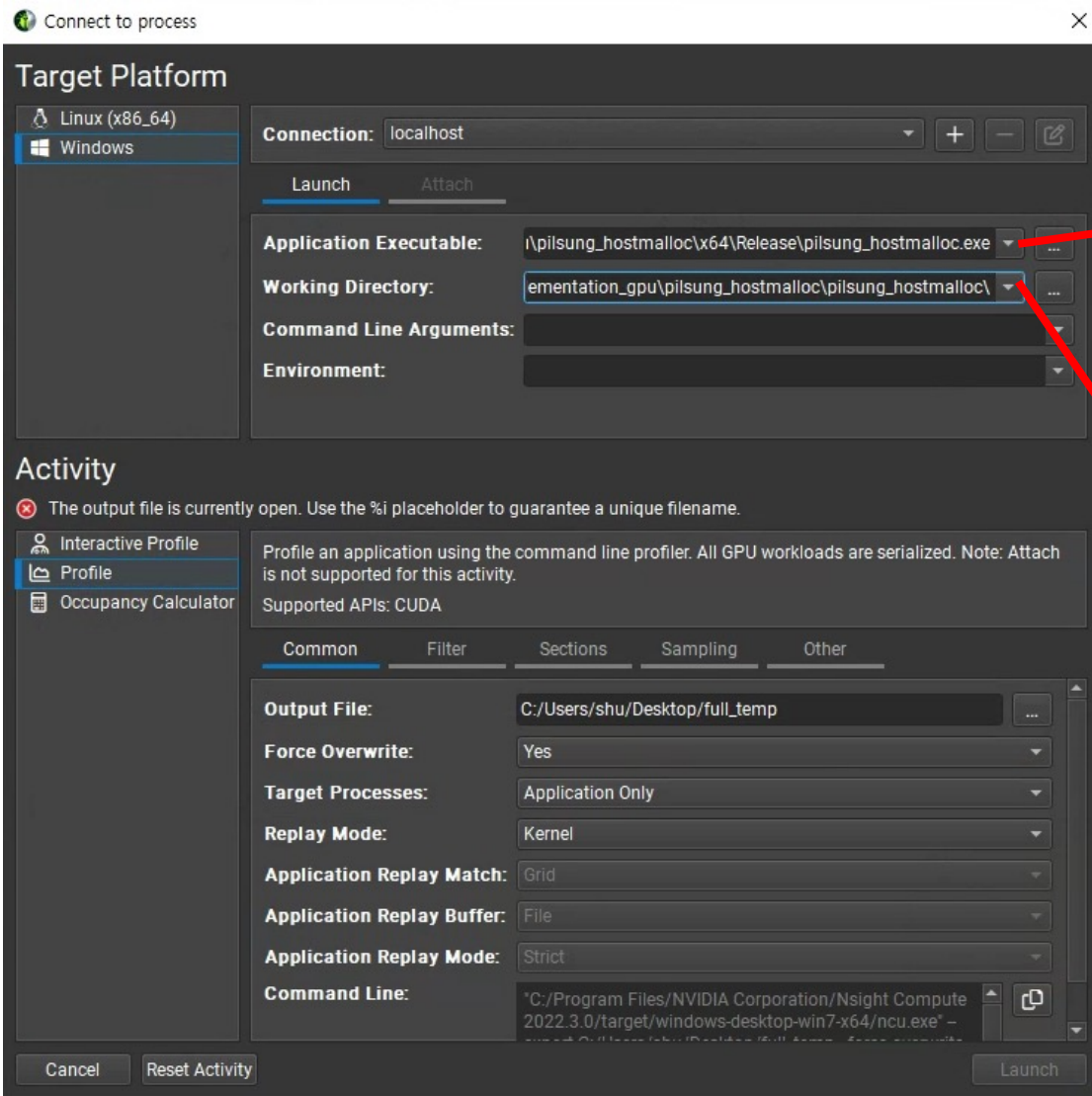
3. Nsight Compute 사용

3. Nsight Compute 사용

3. Nsight Compute 사용

3. Nsight Compute 사용

3. Nsight Compute 사용

3. Nsight Compute 사용

ID	Issues Detecte	Function Name	Demangled Name	Process	Device Name	Grid Size	Block Size	Cycles [cycle]	Duration [msecond]	Compute Thrc	Memory Throughput	# Registers [register/thread]
0	10	pilsungCuda_Kernel	pilsungCuda_Kernel(_	[10888] pilsung_hostmalloc.exe	NVIDIA GeForce RT_	32768, 1, 1	32, 1, 1	8,107,402	9.01	14.47	14.47	40
1	12	pilsungCuda_Kernel	pilsungCuda_Kernel(uns_	[10888] pilsung_hostmalloc.exe	NVIDIA GeForce RTX 3_	32768, 1, 1	64, 1, 1	8,634,812	9.59	27.15	27.15	40
2	12	pilsungCuda_Kernel	pilsungCuda_Kernel(uns_	[10888] pilsung_hostmalloc.exe	NVIDIA GeForce RTX 3_	32768, 1, 1	128, 1, 1	9,703,325	10.79	48.29	48.29	40
3	12	pilsungCuda_Kernel	pilsungCuda_Kernel(uns_	[10888] pilsung_hostmalloc.exe	NVIDIA GeForce RTX 3_	32768, 1, 1	256, 1, 1	11,888,890	13.21	78.81	78.81	40
4	13	pilsungCuda_Kernel	pilsungCuda_Kernel(uns_	[10888] pilsung_hostmalloc.exe	NVIDIA GeForce RTX 3_	32768, 1, 1	512, 1, 1	20,592,438	22.89	90.99	90.99	40
5	15	pilsungCuda_Kernel	pilsungCuda_Kernel(uns_	[10888] pilsung_hostmalloc.exe	NVIDIA GeForce RTX 3_	32768, 1, 1	1024, 1, 1	49,292,499	54.77	76.02	76.02	40

3. Nsight Compute 사용

Page: Details | Result: 0 - 543 - pilsungCuda_Kernel | Add Baseline | Apply Rules | Occupancy Calculator | Copy as Image

	Result	Time	Cycles	Regs	GPU	SM Frequency	CC	Process
Current	543 - pilsungCuda_Kernel (32768, 1, 1)x(32, 1, 1)	9.01 msecond	8,107,402	40	0 - NVIDIA GeForce RTX 3060 Laptop GPU	899.92 cycle/usecond	8.6	[10888] pilsung_hostmalloc.exe

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.

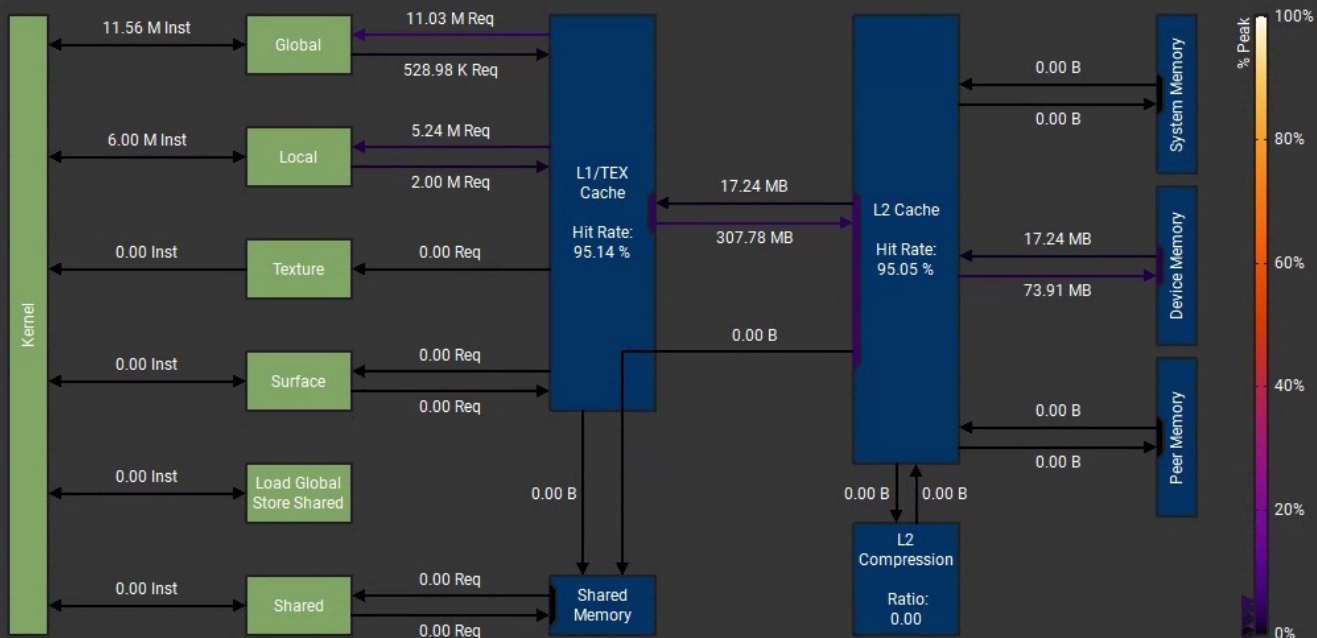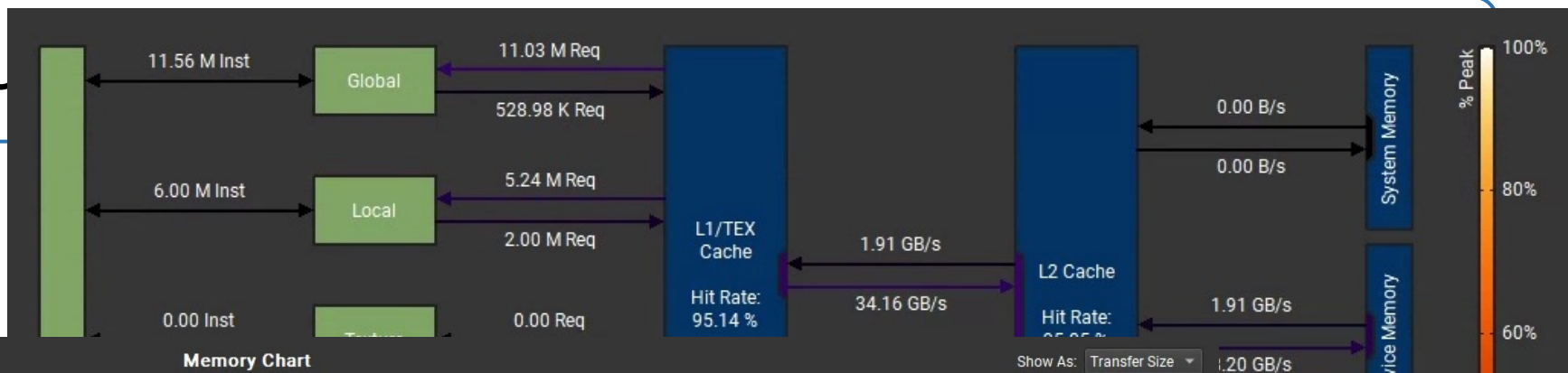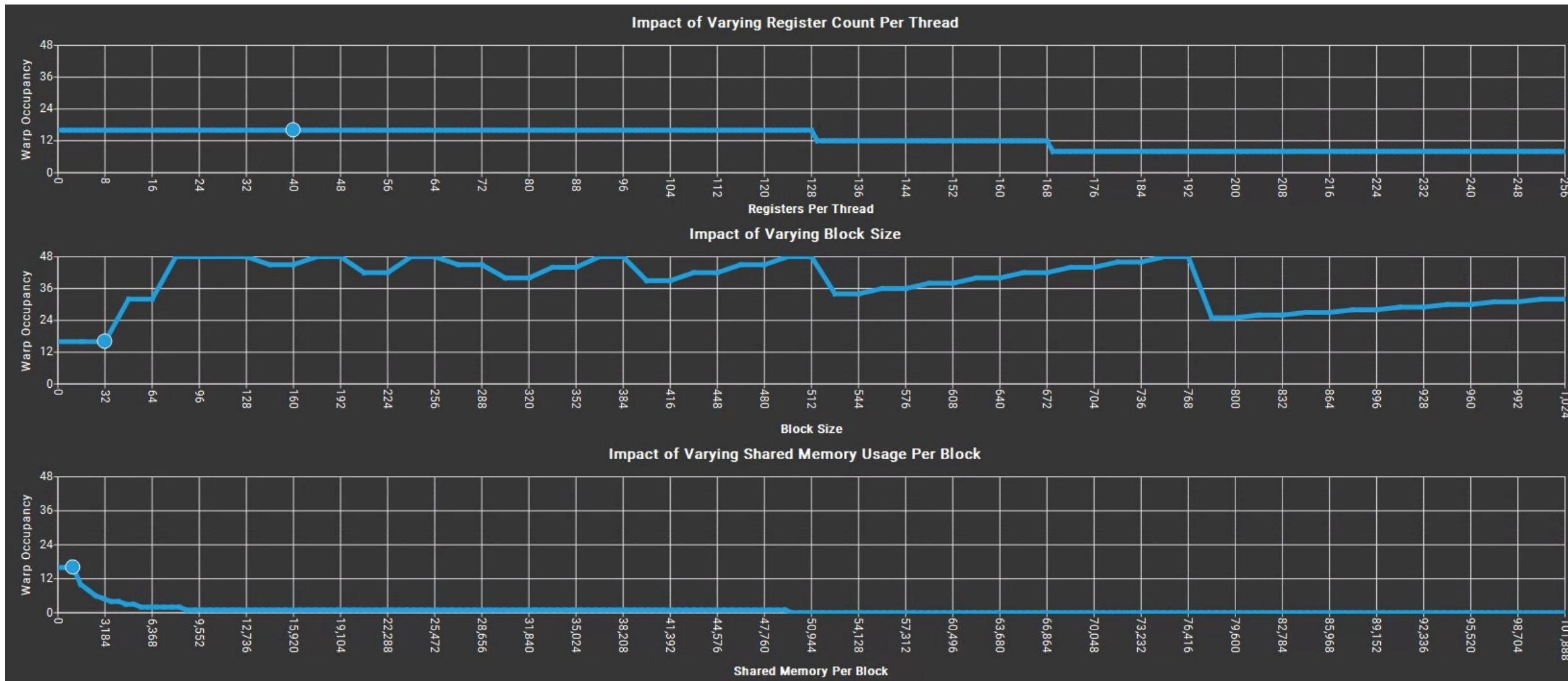|			
Compute (SM) Throughput [%]	14.47	Duration [msecond]	9.01
Memory Throughput [%]	14.47	Elapsed Cycles [cycle]	8,107,402
L1/TEX Cache Throughput [%]	77.36	SM Active Cycles [cycle]	1,516,891.20
L2 Cache Throughput [%]	5.50	SM Frequency [cycle/usecond]	899.92
DRAM Throughput [%]	3.05	DRAM Frequency [cycle/nsecond]	6.99

⚠ **Latency Issue** This kernel exhibits low compute throughput and memory bandwidth utilization relative to the peak performance of this device. Achieved compute throughput and/or memory bandwidth below 60.0% of peak typically indicate latency issues. Look at ▸ Scheduler Statistics and ▸ Warp State Statistics for potential reasons.

ⓘ **Roofline Analysis** The ratio of peak float (fp32) to double (fp64) performance on this device is 64:1. The kernel achieved 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.

GPU Throughput

Compute (SM) [%]
Memory [%]

Speed Of Light (SOL) [%] 0.0 10.0 20.0 30.0 40.0 50.0 60.0 70.0 80.0 90.0 100.0

Compute Throughput Breakdown

SM: Inst Executed Pipe Lsu [%]	14.47
SM: Pipe Alu Cycles Active [%]	11.92
SM: Issue Active [%]	10.63
SM: Inst Executed [%]	10.62
SM: Mio Inst Issued [%]	4.84
SM: Pipe Fmaheavy Cycles Active [%]	3.89
SM: Mio2rf Writeback Active [%]	3.36
SM: Pipe Fma Cycles Active [%]	1.95
SM: Inst Executed Pipe Uniform [%]	1.66
SM: Mio Pq Read Cycles Active [%]	1.04
SM: Mio Pq Write Cycles Active [%]	1.04
SM: Inst Executed Pipe Cbu Pred On Any [%]	0.06
SM: Inst Executed Pipe Adu [%]	0.05
IDC: Request Cycles Active [%]	0.00
SM: Inst Executed Pipe Ipa [%]	0
SM: Inst Executed Pipe Tex [%]	0
SM: Inst Executed Pipe Xu [%]	0
SM: Pipe Fp64 Cycles Active [%]	0

Memory Throughput Breakdown

L1: Lsuin Requests [%]	14.47
L1: Data Pipe Lsu Wavefronts [%]	9.88
L1: Lsu Writeback Active [%]	6.70
L2: Xbar2lts Cycles Active [%]	5.50
L2: T Sectors [%]	5.45
L1: M L1tex2xbar Req Cycles Active [%]	4.20
DRAM: Dram Sectors [%]	3.05
DRAM: Cycles Active [%]	3.01
L1: Data Bank Reads [%]	1.84
L2: D Sectors [%]	1.75
L2: T Tag Requests [%]	1.41
L1: Data Bank Writes [%]	0.55
L2: Lts2xbar Cycles Active [%]	0.29
L2: D Sectors Fill Device [%]	0.29
L1: M Xbar2l1tex Read Sectors [%]	0.22
L1: Texin Sm2tex Req Cycles Active [%]	0.01
L2: D Sectors Fill Sysmem [%]	0.00
L1: Data Pipe Tex Wavefronts [%]	0

(Left panel items:)
- Compute (SM) Throughput [%]
- Memory Throughput [%]
- L1/TEX Cache Throughput [%]
- L2 Cache Throughput [%]
- DRAM Throughput [%]

⚠ Latency Issue — This kernel exhibits low compute throughput... Statistics and ▸ Warp State Statistics for...

ⓘ Roofline Analysis — The ratio of peak float (fp32) to dou...

Compute Workload Analysis
Detailed analysis of the compute resources of the streaming multi...
- Executed Ipc Elapsed [inst/cycle]
- Executed Ipc Active [inst/cycle]
- Issued Ipc Active [inst/cycle]

⚠ High Utilization — ALU is the highest-utilized pipeline...

Memory Workload Analysis
Detailed analysis of the memory resources of the GPU. Memory ca... maximum throughput of issuing memory instructions (Mem Pipes...
- Memory Throughput [Gbyte/second]
- L1/TEX Hit Rate [%]
- L2 Hit Rate [%]
- L2 Compression Success Rate [%]

⚠ L1TEX Global Store Access Pattern
⚠ L1TEX Local Load Access Pattern
⚠ L1TEX Local Store Access Pattern

3. Nsight Compu

Memory Chart

Show As: Transfer Size

Shared Memory

	Instructions	Requests	Wavefronts	% Peak	Bank Conflicts
Shared Load	0	0	0	0	0
Shared Load Matrix	0	0	0	0	0
Shared Store	0	0	0	0	0
Shared Store From Global Load	0	0	0	0	0
Shared Atomic	0	0	0	0	0
Other	-	-	81,922	0.03	0
Total	0	0	81,922	0.03	0

14

3. Nsight Compute 사용

3. Nsight Compute 사용

감 사 합 니 다