GETTING STARTED WITH RAY TRACING AND NVIDIA'S RAY TRACING DEVELOPER TOOLS

CWE41887

Aurelio Reis
SWE Director, Graphics Developer Tools
NVIDIA

Jeff Kiel
Sr. Engineering Manager, Graphics Developer Tools
NVIDIA

Russ Kerschner
Sr. Software Engineer, Graphics Developer Tools
NVIDIA

Richard Meth
Sr. Software Engineer, PerfWorks
NVIDIA

Kyle Spagnoli
Sr. Software Engineer, Graphics Developer Tools
NVIDIA

Daniel Horowitz
SWE Director, Platform Developer Tools
NVIDIA

Chris Cottrell
Sr. Software Engineer, Graphics Developer Tools
NVIDIA

REAL TIME RAY TRACING

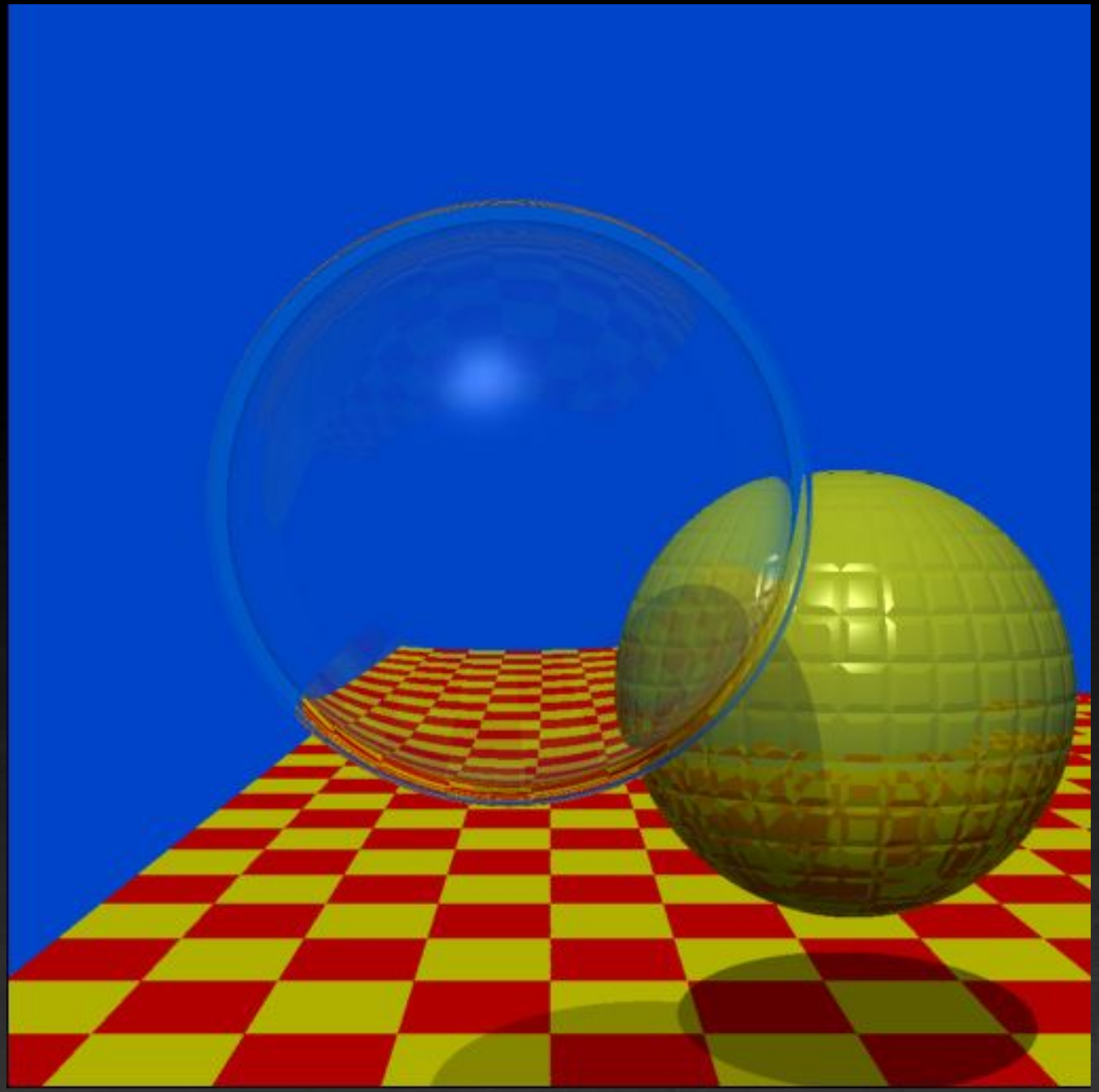
A Technological Revolution

Vulkan 1.2 FORGING AHEAD

DirectX 12 Raytracing

OptiX for CUDA

1979



2018

2021

NVIDIA

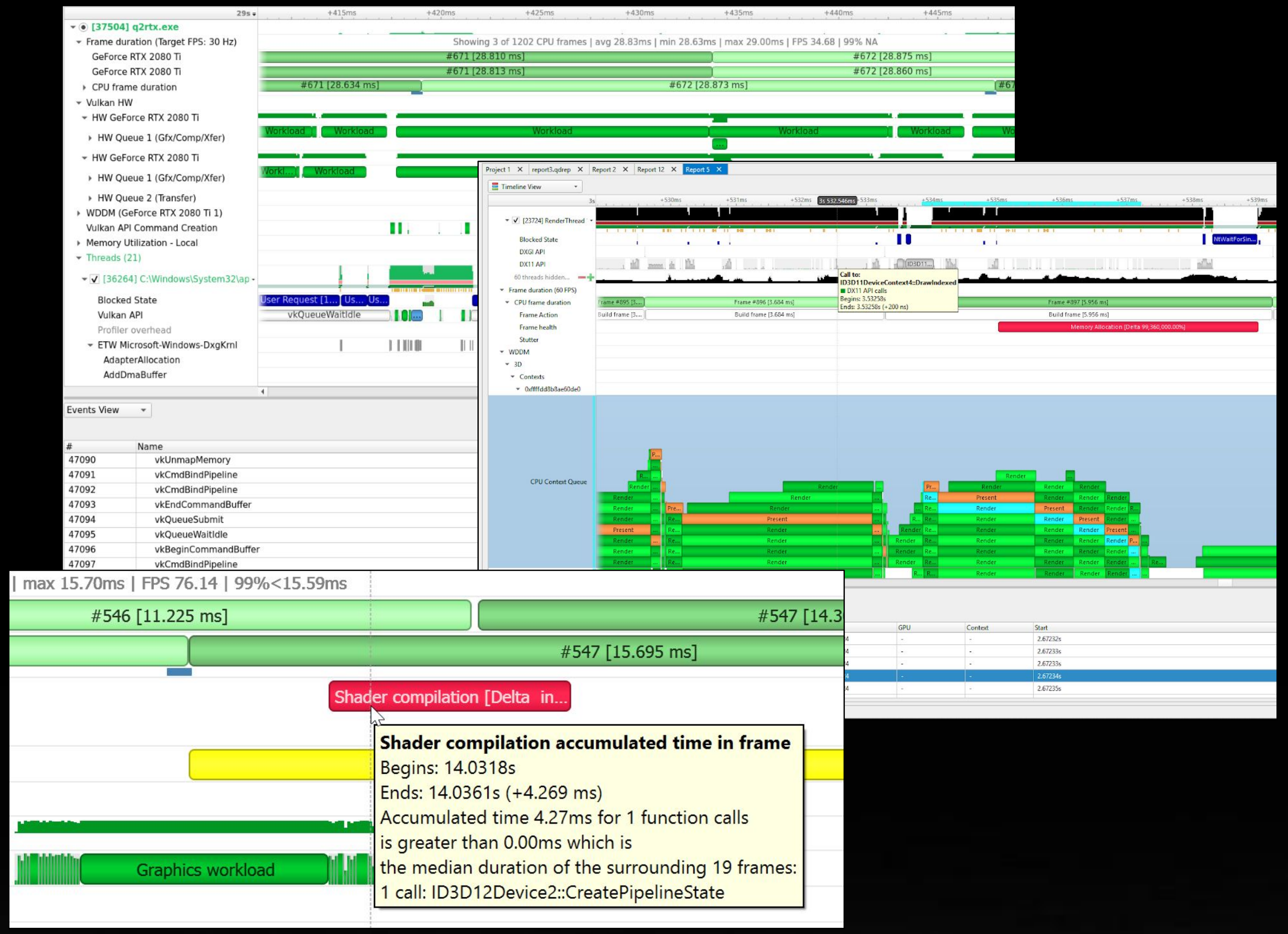
DEVELOPER TOOLS

We work to solve developer problems to make your lives better

PROFESSIONAL TOOLS FOR RAY TRACING

Improving development to drive innovation and create amazing graphics

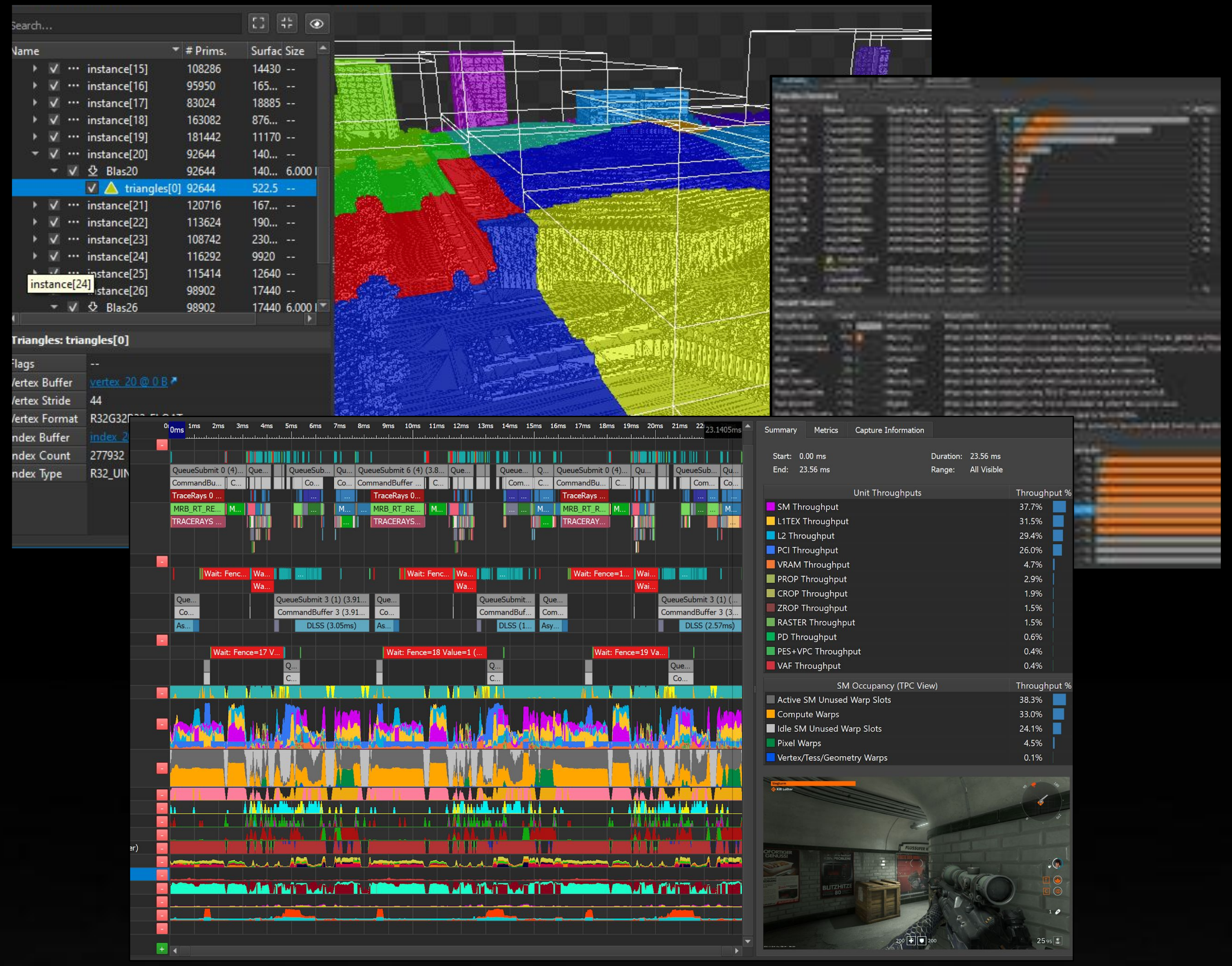
Nsight Systems



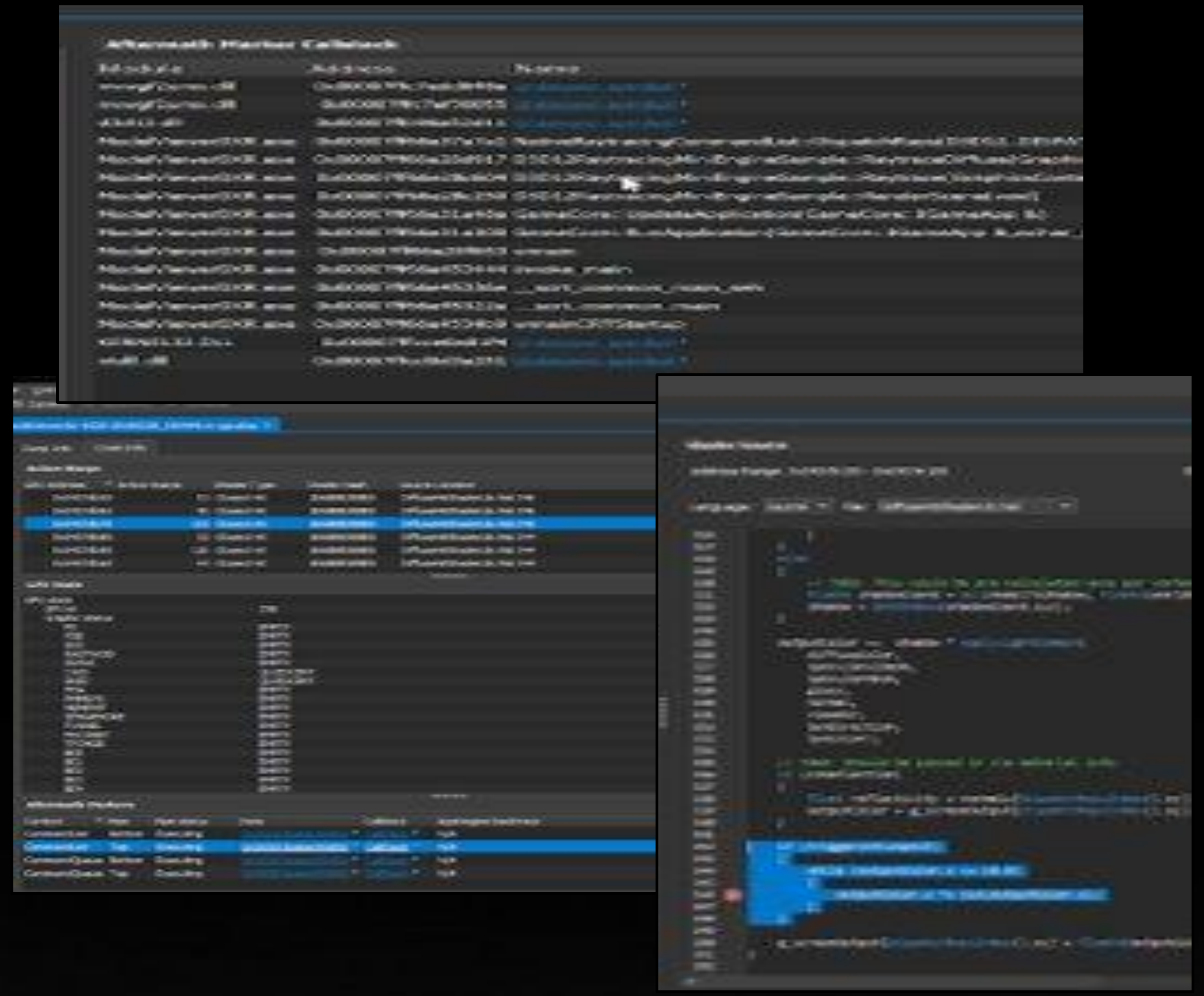
max 15.70ms | FPS 76.14 | 99%<15.59ms
#546 [11.225 ms]
#547 [14.3
#547 [15.695 ms]
Shader compilation [Delta in...
Graphics workload

Shader compilation accumulated time in frame
Begins: 14.0318s
Ends: 14.0361s (+4.269 ms)
Accumulated time 4.27ms for 1 function calls
is greater than 0.00ms which is
the median duration of the surrounding 19 frames:
1 call: ID3D12Device2::CreatePipelineState

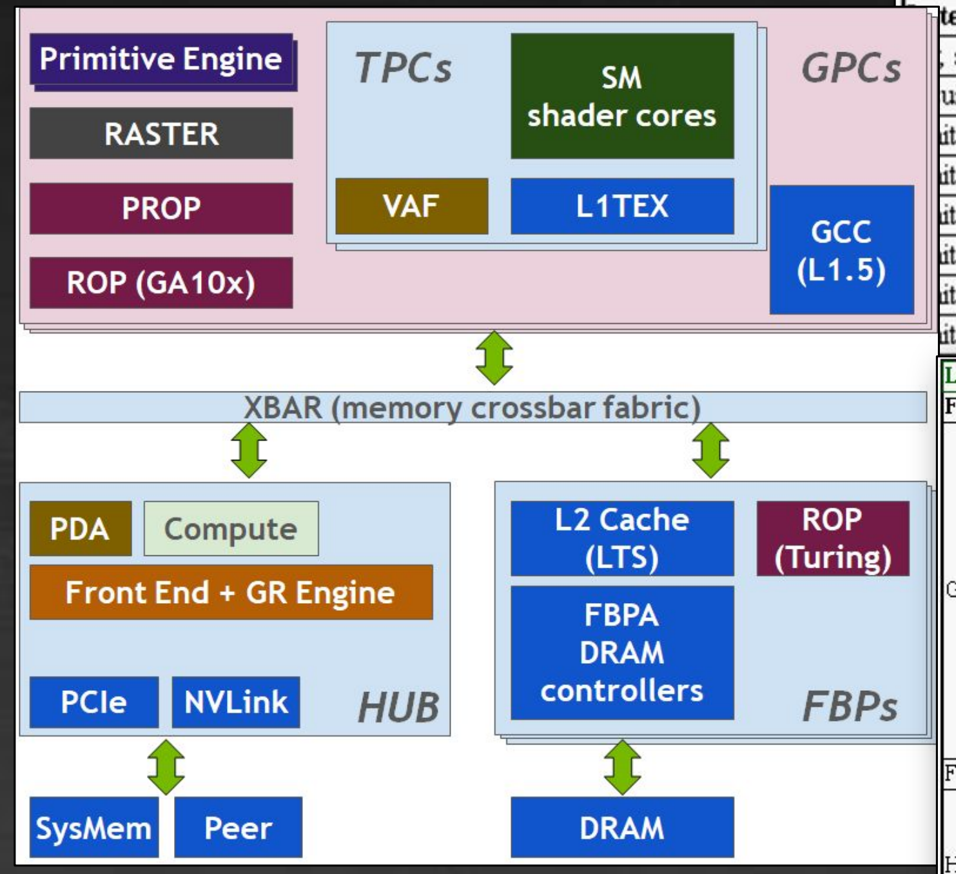
Nsight Graphics

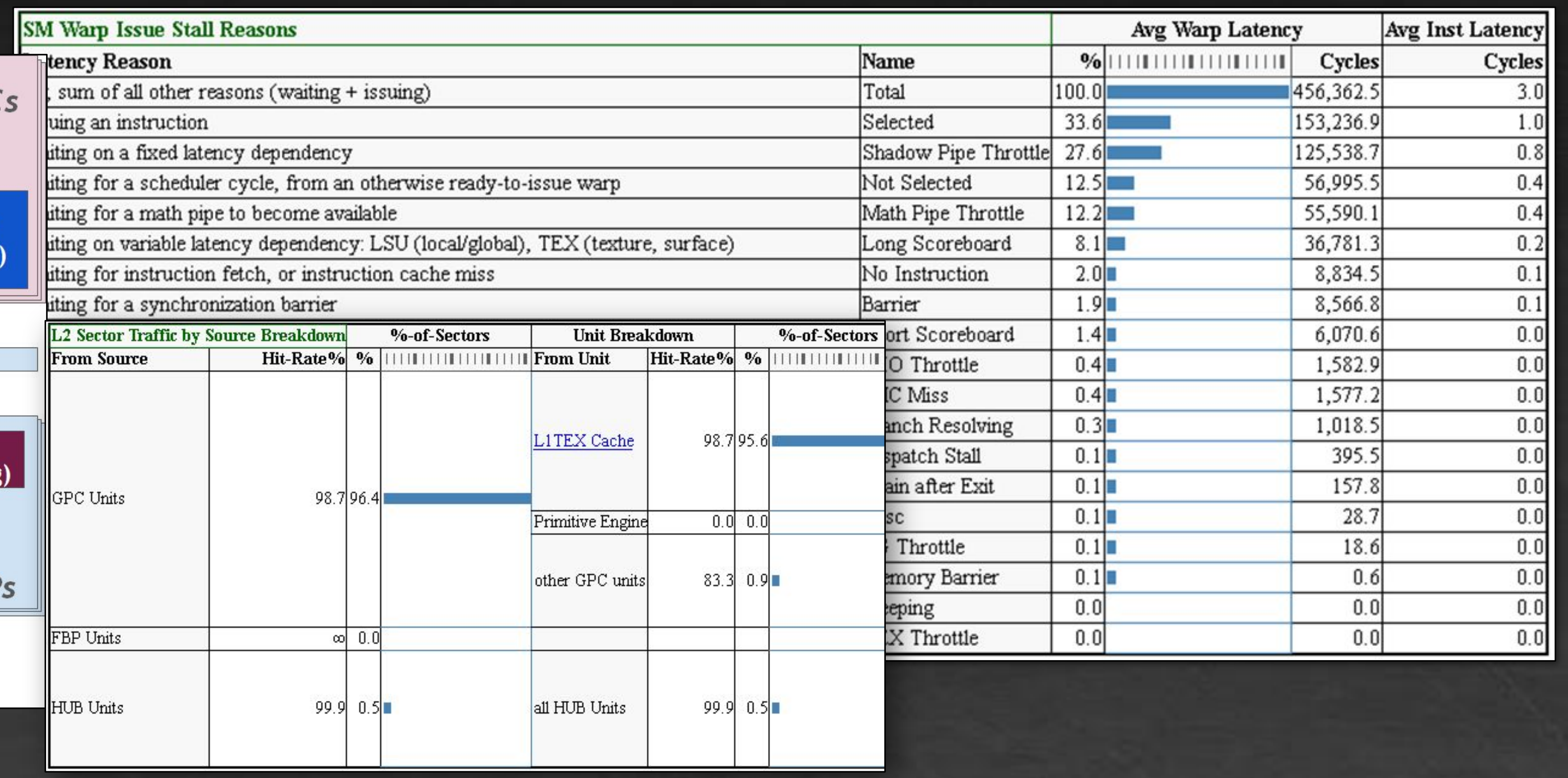


Nsight Aftermath



Nsight Perf SDK





SM Warp Issue Stall Reasons		Avg Warp Latency			Avg Inst Latency
tency Reason	Name	%		Cycles	Cycles
, sum of all other reasons (waiting + issuing)	Total	100.0		456,362.5	3.0
uing an instruction	Selected	33.6		153,236.9	1.0
iting on a fixed latency dependency	Shadow Pipe Throttle	27.6		125,538.7	0.8
iting for a scheduler cycle, from an otherwise ready-to-issue warp	Not Selected	12.5		56,995.5	0.4
iting for a math pipe to become available	Math Pipe Throttle	12.2		55,590.1	0.4
iting on variable latency dependency: LSU (local/global), TEX (texture, surface)	Long Scoreboard	8.1		36,781.3	0.2
iting for instruction fetch, or instruction cache miss	No Instruction	2.0		8,834.5	0.1
iting for a synchronization barrier	Barrier	1.9		8,566.8	0.1
	ort Scoreboard	1.4		6,070.6	0.0
	O Throttle	0.4		1,582.9	0.0
	C Miss	0.4		1,577.2	0.0
	anch Resolving	0.3		1,018.5	0.0
	spatch Stall	0.1		395.5	0.0
	ain after Exit	0.1		157.8	0.0
	sc	0.1		28.7	0.0
	Throttle	0.1		18.6	0.0
	emory Barrier	0.1		0.6	0.0
	eping	0.0		0.0	0.0
	X Throttle	0.0		0.0	0.0

L2 Sector Traffic by Source Breakdown		%-of-Sectors		Unit Breakdown		%-of-Sectors	
From Source	Hit-Rate%	%		From Unit	Hit-Rate%	%	
GPC Units	98.7	96.4		L1TEX Cache	98.7	95.6	
				Primitive Engine	0.0	0.0	
				other GPC units	83.3	0.9	
FBP Units	∞	0.0					
HUB Units	99.9	0.5		all HUB Units	99.9	0.5	

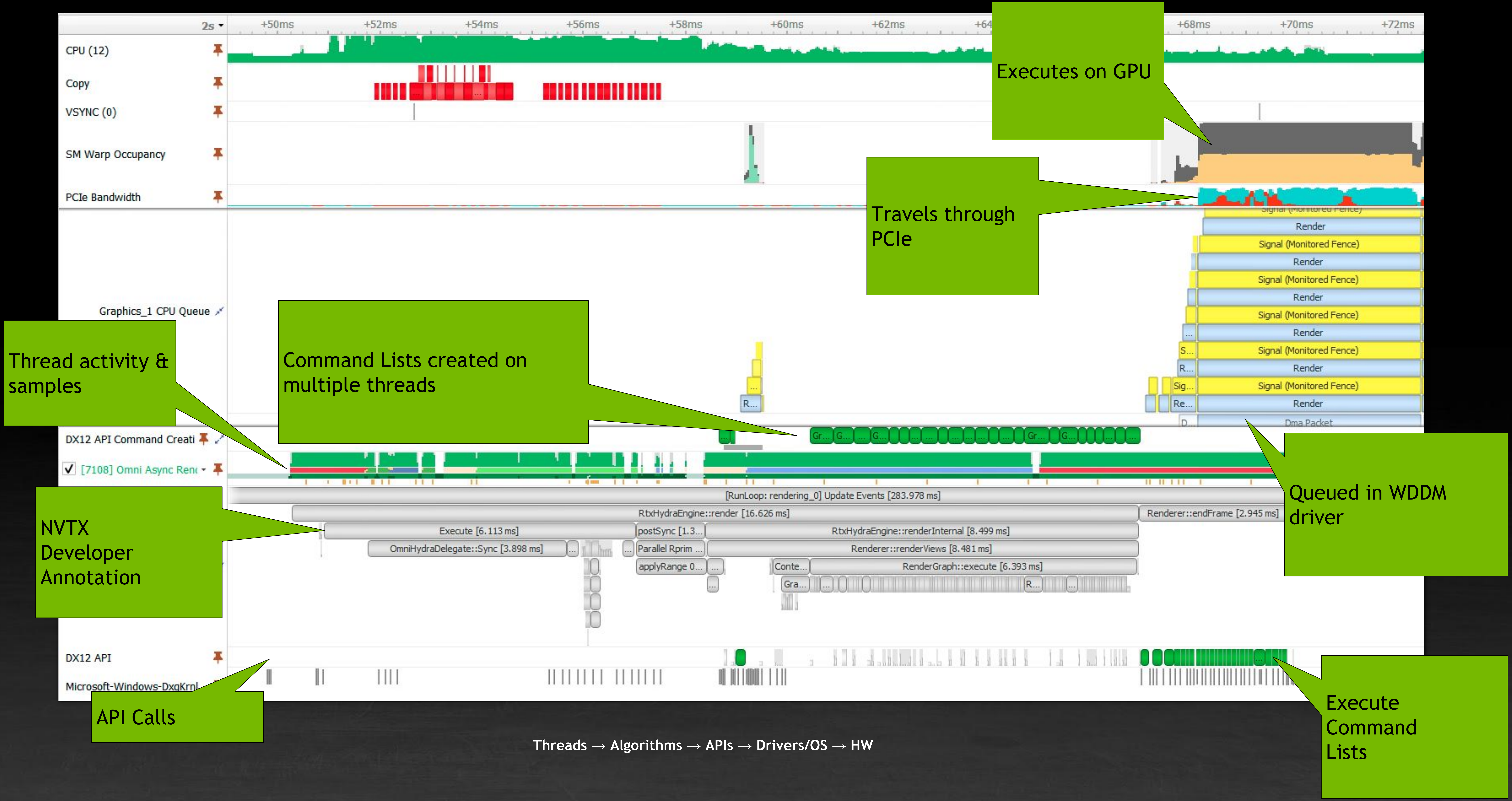
PERFORMANCE TRIAGE WORKFLOW

To help save you time and frustration

Game
Start Here
Performance Issue?
Yes
Rendering Issue?
Yes
Nsight Systems
Comprehensive workload-level analysis
GPU limited?
No
Yes
CPU limited?
Yes
No
Finished (for now)
Nsight Graphics
Detailed graphics debugging/profiling
Rendering Fixed

NSIGHT SYSTEMS

Useful for investigating CPU/GPU Interactions



Threads → Algorithms → APIs → Drivers/OS → HW

Check out "ORCHESTRATING NEXT-GEN GRAPHICS WITH NSIGHT SYSTEMS" presented by Uri Shomroni

NSIGHT GRAPHICS

Powerful debugging and profiling for advanced 3d graphics

Debugging:

C++ Frame Serialization

Resource Viewer

Pixel History

API inspection

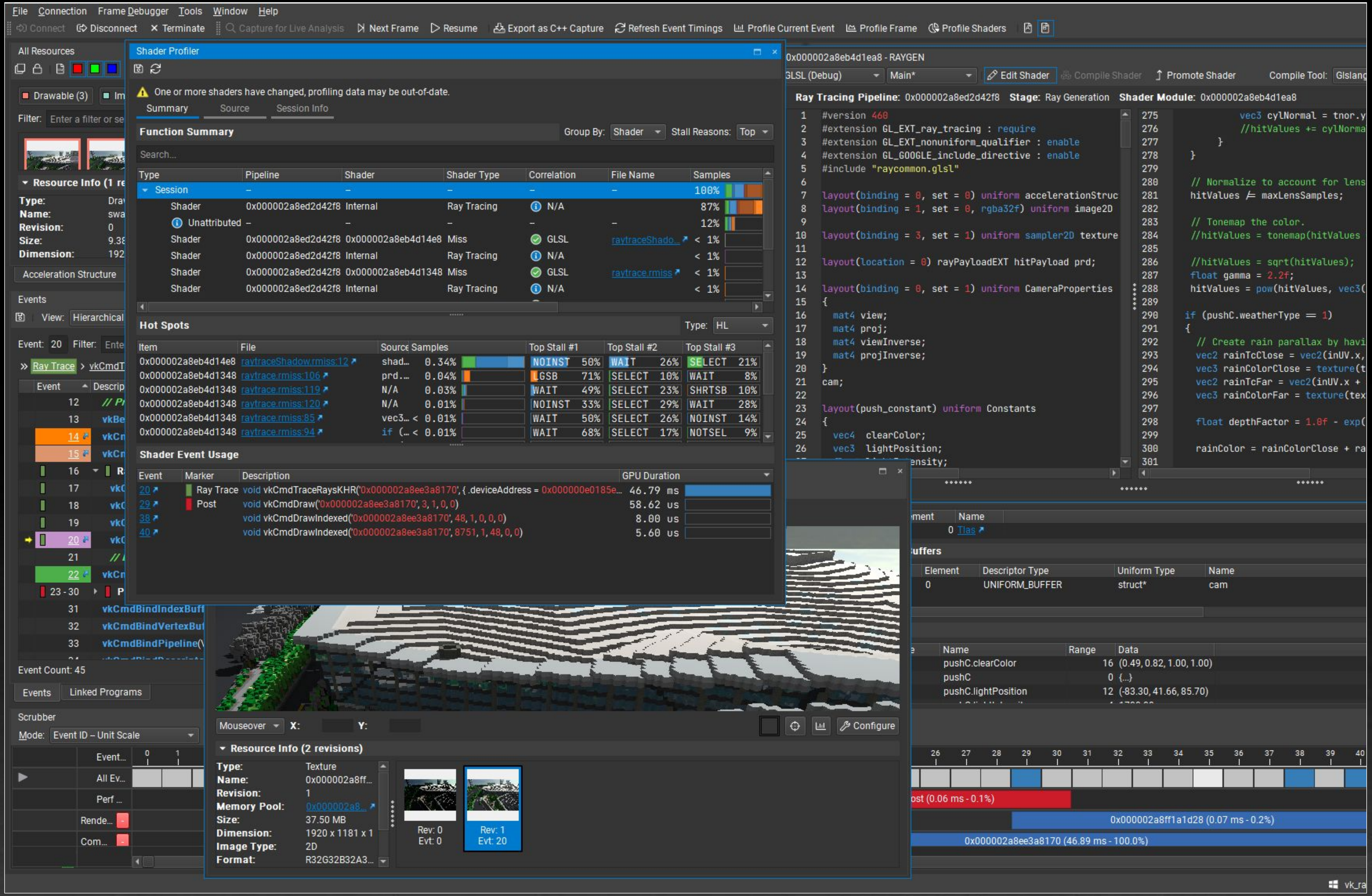
GPU Crashes

Profiling:

Range Profiler

GPU Trace

Shader Profiler


File Connection Frame Debugger Tools Window Help
Connect Disconnect Terminate Capture for Live Analysis Next Frame Resume Export as C++ Capture Refresh Event Timings Profile Current Event Profile Frame Profile Shaders
All Resources
Shader Profiler
Summary Source Session Info
Edit Shader Compile Shader Promote Shader
Mouseover Configure
Events Linked Programs
Scrubber

One or more shaders have changed, profiling data may be out-of-date.

Function Summary Group By: Shader Stall Reasons: Top

Type	Pipeline	Shader	Shader Type	Correlation	File Name	Samples
Session	–	–	–	–	–	100%
Shader	0x000002a8ed2d42f8	Internal	Ray Tracing	N/A		87%
Unattributed	–		–	–	–	12%
Shader	0x000002a8ed2d42f8	0x000002a8eb4d14e8	Miss	GLSL	raytraceShado...	< 1%
Shader	0x000002a8ed2d42f8	Internal	Ray Tracing	N/A		< 1%
Shader	0x000002a8ed2d42f8	0x000002a8eb4d1348	Miss	GLSL	raytrace.rmiss	< 1%
Shader	0x000002a8ed2d42f8	Internal	Ray Tracing	N/A		< 1%

Hot Spots Type: HL

Item	File	Source	Samples	Top Stall #1	Top Stall #2	Top Stall #3
0x000002a8eb4d14e8	raytraceShadow.rmiss:12	shad...	0.34%	NOINST 50%	WAIT 26%	SELECT 21%
0x000002a8eb4d1348	raytrace.rmiss:106	prd...	0.04%	LGSB 71%	SELECT 10%	WAIT 8%
0x000002a8eb4d1348	raytrace.rmiss:119	N/A	0.03%	WAIT 49%	SELECT 23%	SHRTSB 10%
0x000002a8eb4d1348	raytrace.rmiss:120	N/A	0.01%	NOINST 33%	SELECT 29%	WAIT 28%
0x000002a8eb4d1348	raytrace.rmiss:85	vec3...	< 0.01%	WAIT 50%	SELECT 26%	NOINST 14%
0x000002a8eb4d1348	raytrace.rmiss:94	if (...	< 0.01%	WAIT 68%	SELECT 17%	NOTSEL 9%

Shader Event Usage

Event	Marker	Description	GPU Duration
20	Ray Trace	void vkCmdTraceRaysKHR('0x000002a8ee3a8170', { .deviceAddress = 0x000000e0185e...	46.79 ms
29	Post	void vkCmdDraw('0x000002a8ee3a8170', 3, 1, 0, 0)	58.62 us
38		void vkCmdDrawIndexed('0x000002a8ee3a8170', 48, 1, 0, 0, 0)	8.00 us
40		void vkCmdDrawIndexed('0x000002a8ee3a8170', 8751, 1, 48, 0, 0)	5.60 us

0x000002a8eb4d1ea8 - RAYGEN

GLSL (Debug) Main* Compile Tool: Glslang

Ray Tracing Pipeline: 0x000002a8ed2d42f8 **Stage:** Ray Generation **Shader Module:** 0x000002a8eb4d1ea8

```
#version 460
#extension GL_EXT_ray_tracing : require
#extension GL_EXT_nonuniform_qualifier : enable
#extension GL_GOOGLE_include_directive : enable
#include "raycommon.glsl"

layout(binding = 0, set = 0) uniform accelerationStruc
layout(binding = 1, set = 0, rgba32f) uniform image2D

layout(binding = 3, set = 1) uniform sampler2D texture

layout(location = 0) rayPayloadEXT hitPayload prd;

layout(binding = 0, set = 1) uniform CameraProperties
{
    mat4 view;
    mat4 proj;
    mat4 viewInverse;
    mat4 projInverse;
}
cam;

layout(push_constant) uniform Constants
{
    vec4  clearColor;
    vec3  lightPosition;
```

```
            vec3 cylNormal = tnor.y
            //hitValues += cylNorma
        }
    }

    // Normalize to account for lens
    hitValues /= maxLensSamples;

    // Tonemap the color.
    //hitValues = tonemap(hitValues

    //hitValues = sqrt(hitValues);
    float gamma = 2.2f;
    hitValues = pow(hitValues, vec3(

    if (pushC.weatherType == 1)
    {
      // Create rain parallax by havi
      vec2 rainTcClose = vec2(inUV.x,
      vec3 rainColorClose = texture(t
      vec2 rainTcFar = vec2(inUV.x +
      vec3 rainColorFar = texture(tex

      float depthFactor = 1.8f - exp(

      rainColor = rainColorClose + ra
```

Element	Descriptor Type	Uniform Type	Name
0	UNIFORM_BUFFER	struct*	cam

Name	Range	Data
pushC.clearColor	16	(0.49, 0.82, 1.00, 1.00)
pushC	0	(...)
pushC.lightPosition	12	(-83.30, 41.66, 85.70)

Resource Info (2 revisions)

Type:	Texture
Name:	0x000002a8ff...
Revision:	1
Memory Pool:	0x000002a8...
Size:	37.50 MB
Dimension:	1920 x 1181 x 1
Image Type:	2D
Format:	R32G32B32A3...

Rev: 0 Evt: 0 · Rev: 1 Evt: 20

Post (0.06 ms - 0.1%) · 0x000002a8ff1a1d28 (0.07 ms - 0.2%) · 0x000002a8ee3a8170 (46.89 ms - 100.0%)

Event Count: 45

Windows

Linux

Android

LuminOS

x64

ARM *(Coming Soon!)*

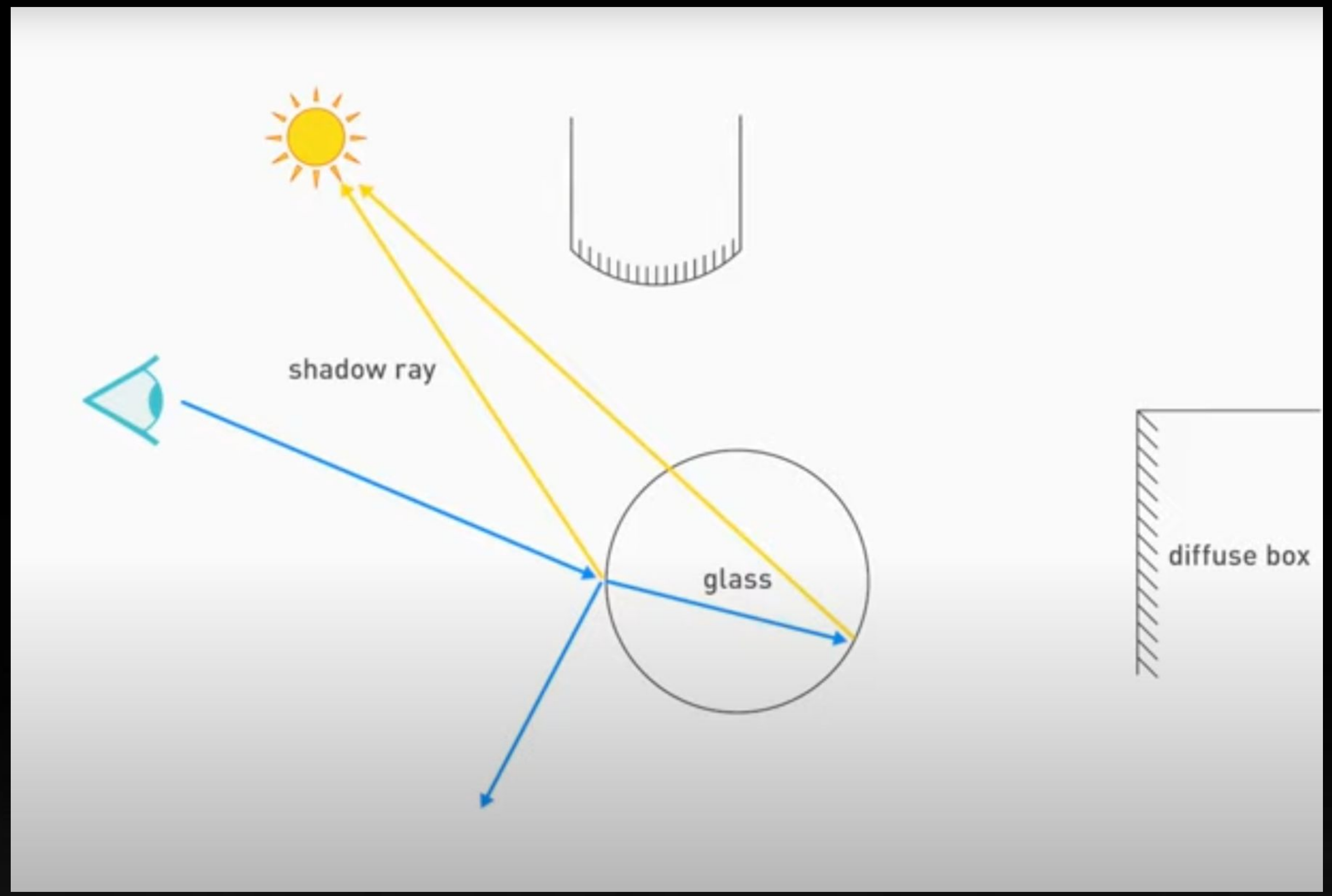
D3D+11/12, DXR

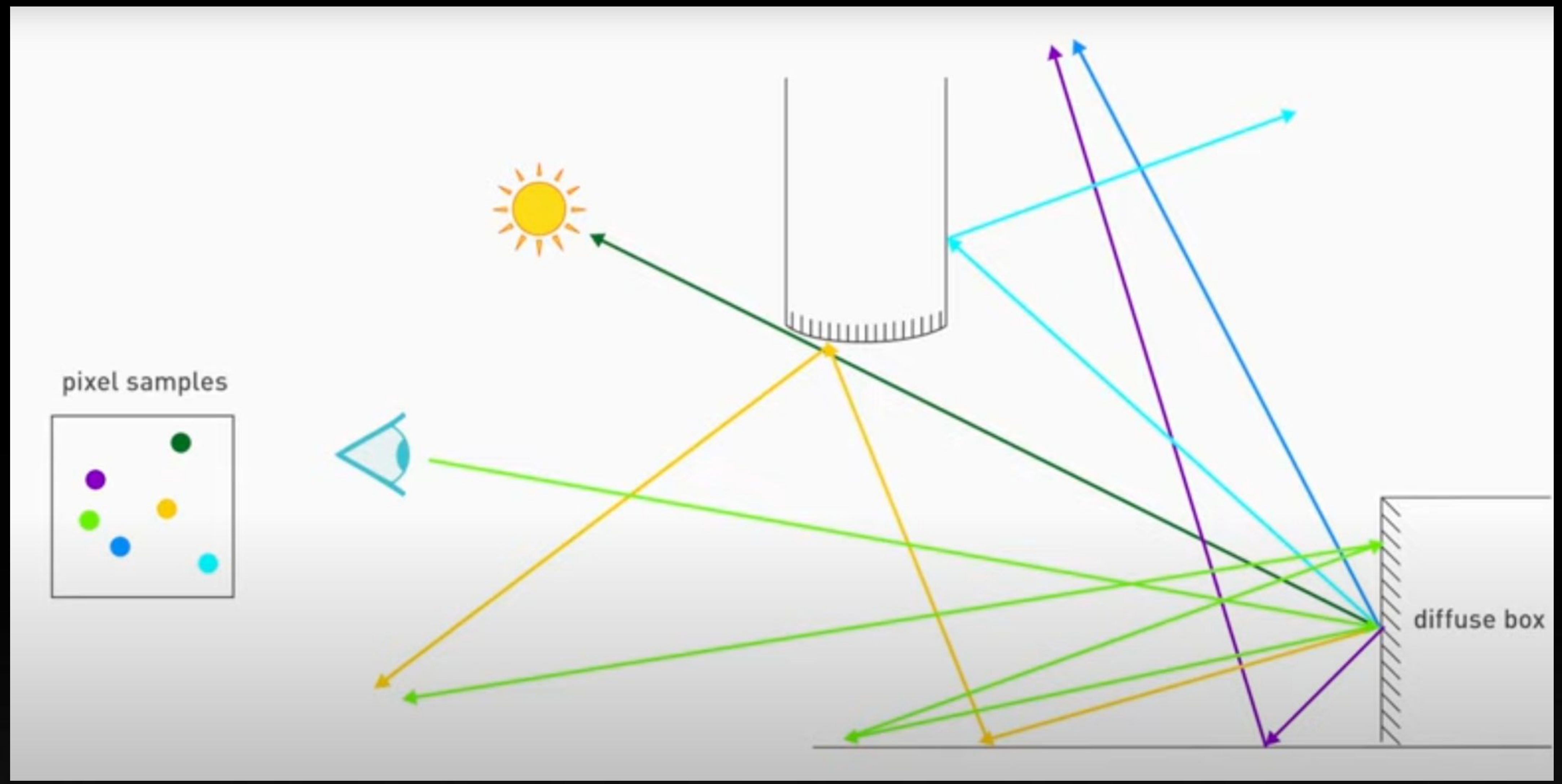
Vulkan 1.2, VRT

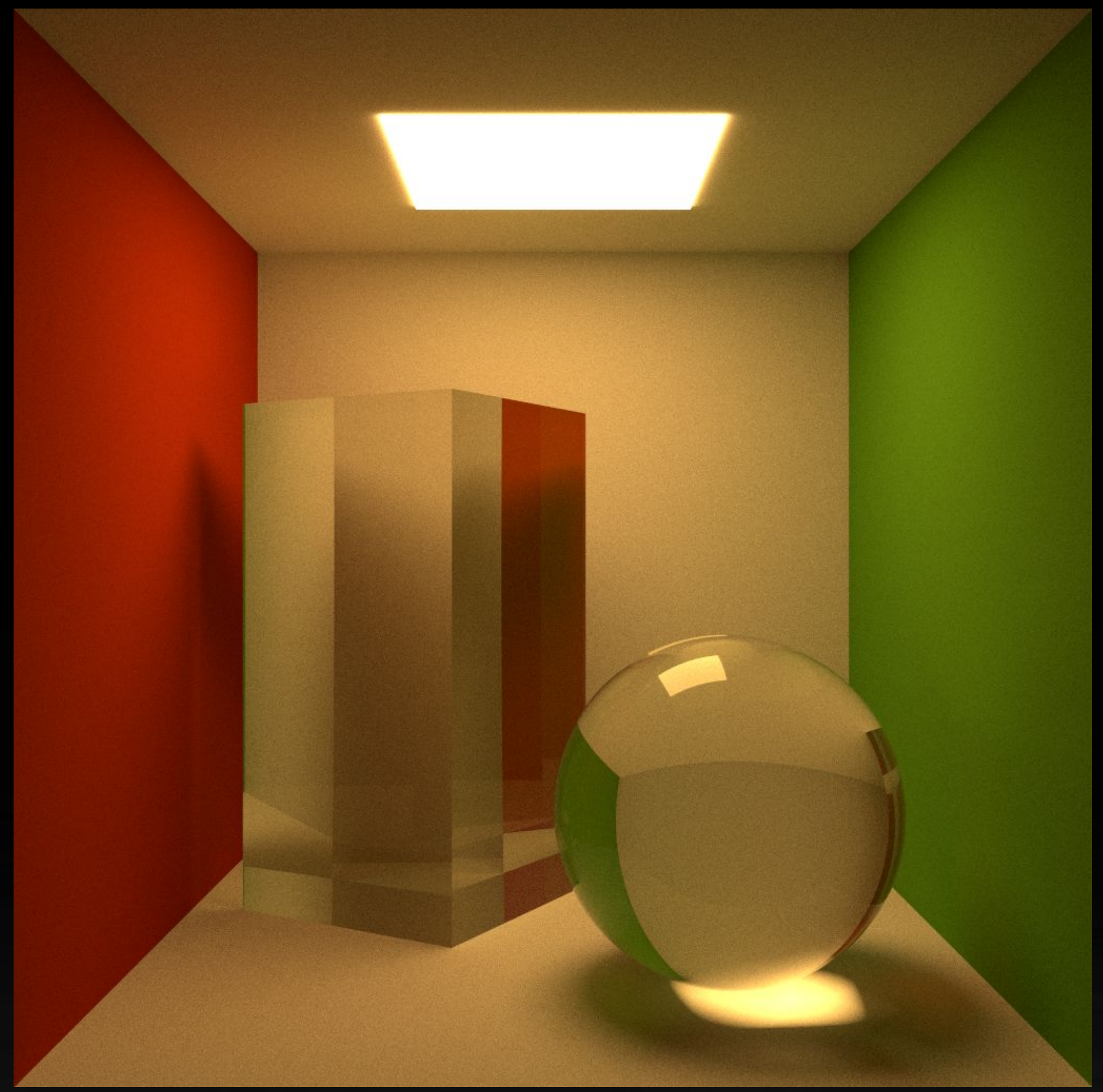
OpenGL 4.6

RAY TRACING CONCEPTS

From concept to tool





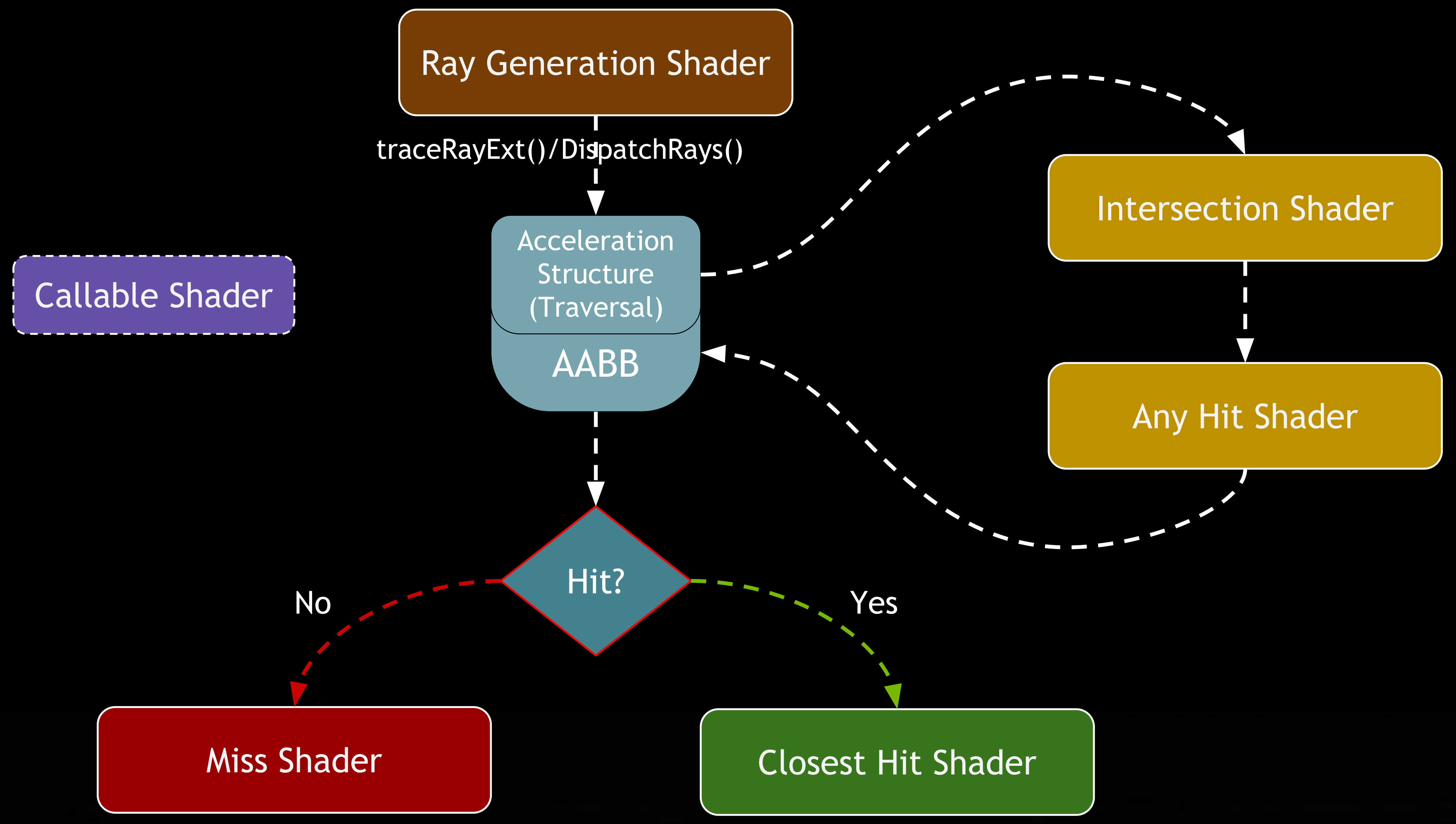


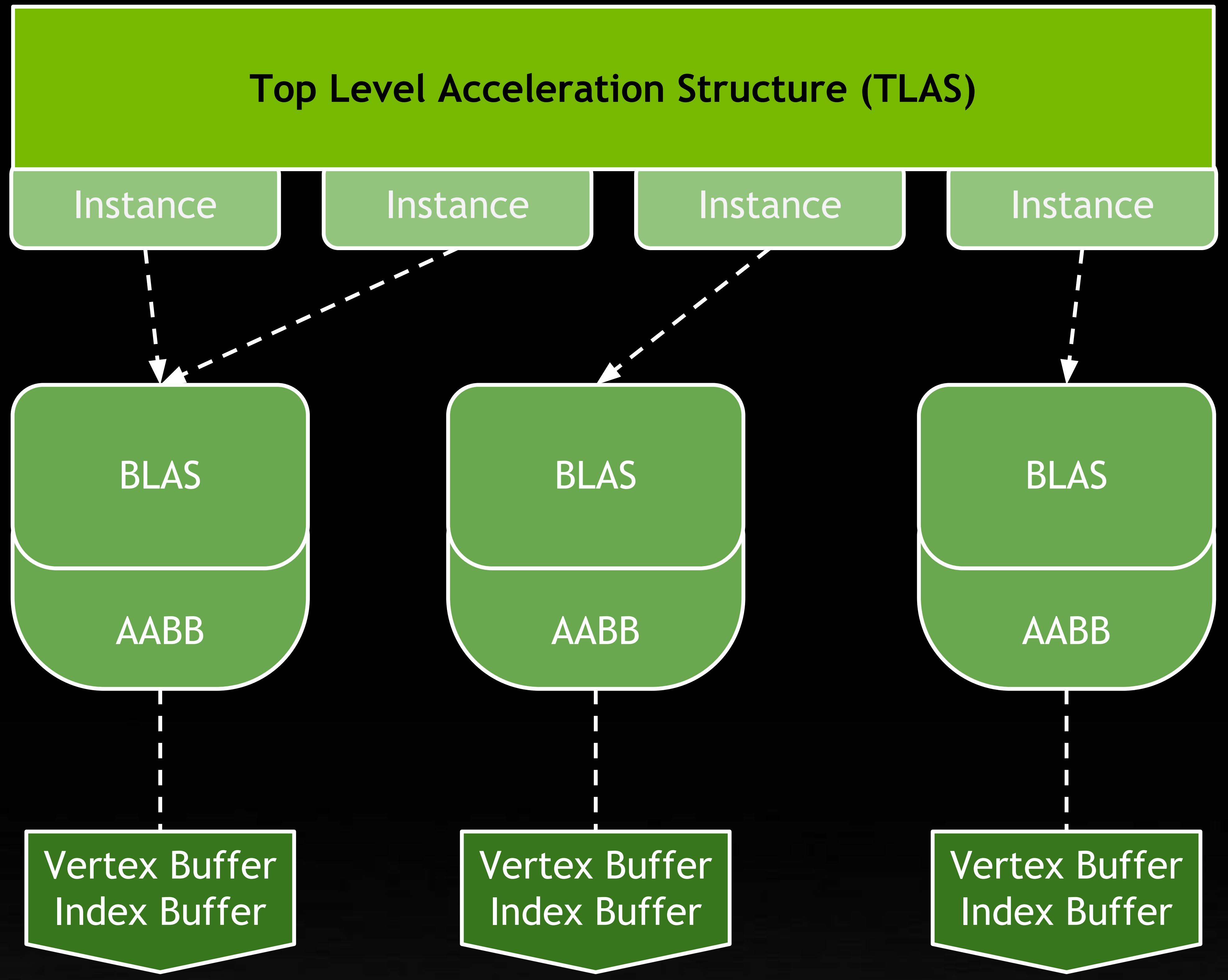
https://blogs.nvidia.com/blog/2018/03/19/whats-difference-between-ray-tracing-rasterization/

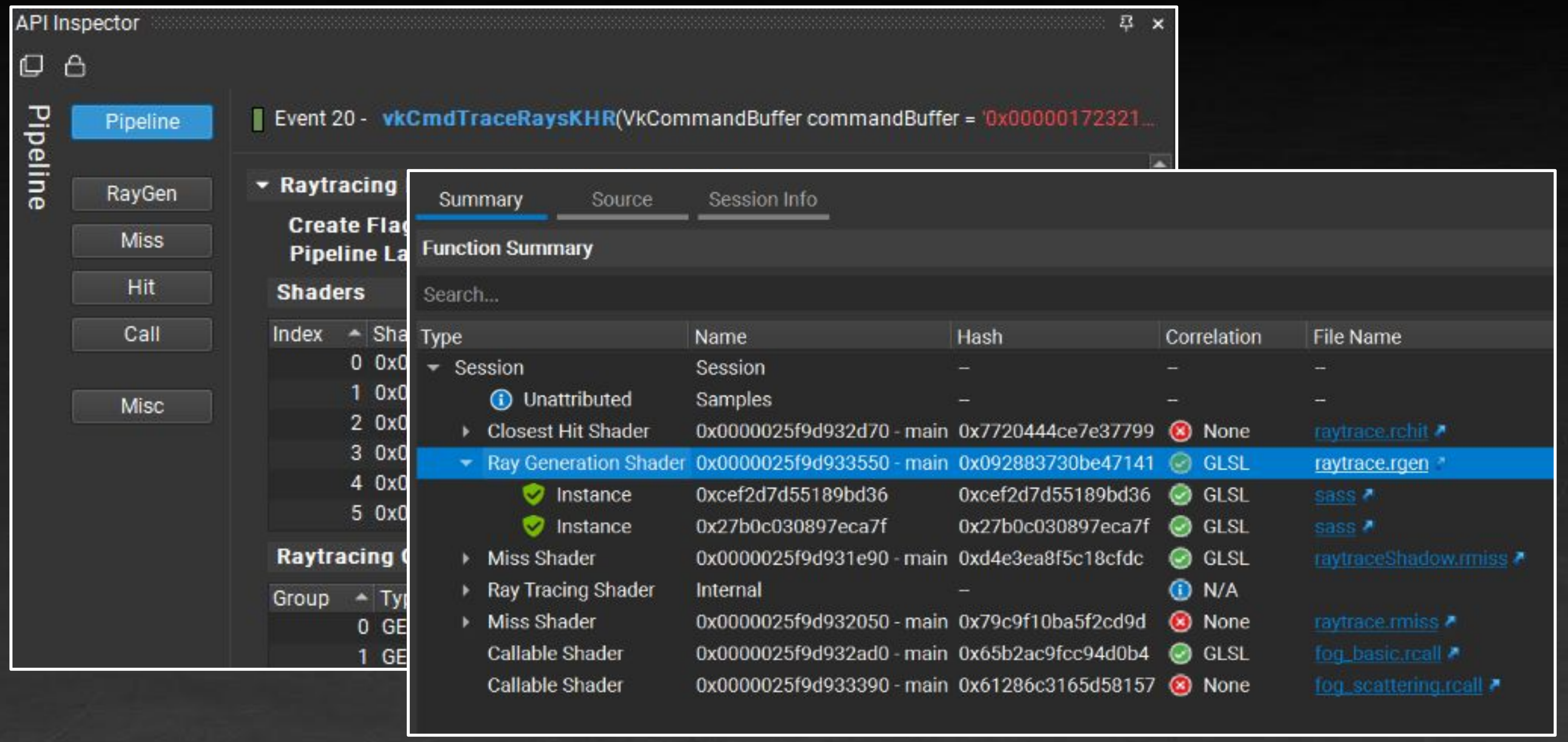
https://youtu.be/gBPNO6ruevk

RAY TRACING CONCEPTS

And How They Map To The Tools






API Inspector

Pipeline

Pipeline | RayGen | Miss | Hit | Call | Misc

Event 20 - vkCmdTraceRaysKHR(VkCommandBuffer commandBuffer = '0x00000172321...

Raytracing
Create Flag
Pipeline La
Shaders

Index	Sha
0	0x0
1	0x0
2	0x0
3	0x0
4	0x0
5	0x0

Raytracing (

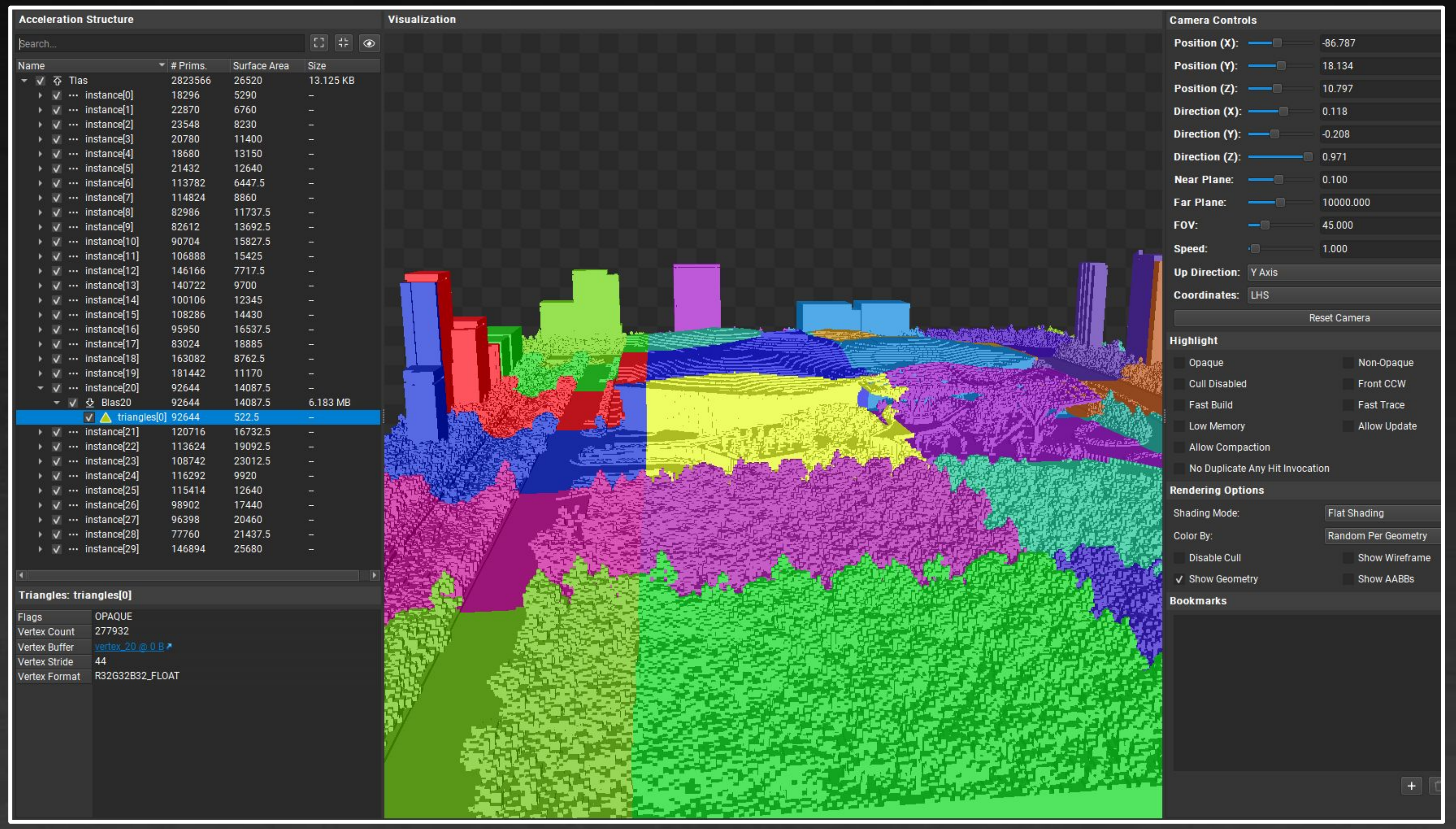
Group	Typ
0	GE
1	GE

Summary | Source | Session Info

Function Summary

Search...

Type	Name	Hash	Correlation	File Name
Session	Session	–	–	–
Unattributed	Samples	–	–	–
Closest Hit Shader	0x0000025f9d932d70 - main	0x7720444ce7e37799	None	raytrace.rchit
Ray Generation Shader	0x0000025f9d933550 - main	0x092883730be47141	GLSL	raytrace.rgen
Instance	0xcef2d7d55189bd36	0xcef2d7d55189bd36	GLSL	sass
Instance	0x27b0c030897eca7f	0x27b0c030897eca7f	GLSL	sass
Miss Shader	0x0000025f9d931e90 - main	0xd4e3ea8f5c18cfdc	GLSL	raytraceShadow.rmiss
Ray Tracing Shader	Internal	–	N/A	
Miss Shader	0x0000025f9d932050 - main	0x79c9f10ba5f2cd9d	None	raytrace.rmiss
Callable Shader	0x0000025f9d932ad0 - main	0x65b2ac9fcc94d0b4	GLSL	fog_basic.rcall
Callable Shader	0x0000025f9d933390 - main	0x61286c3165d58157	None	fog_scattering.rcall



BVH ANALYSIS

Instance Heatmap

ments

ellbladeGame-Win64-Shipping__TLAS_1bee660000__2022_02_22-08_19_17.ngfx-bvh ✕ HellbladeGame-Win64-Shipping.exe_20220222084316730.ngfx-gputrace ✕

Acceleration Structure

Search...

Name	# Prims.	Surface Area	Size
TLAS 1bee660000	4586494	2.05622E+12	1.11 MB
instance[2592]	85509	8.03994E+07	–
instance[19]	84194	59030.9	–
instance[2591]	35604	7.27741F+07	–
instance[3]	29541	374434	–
instance[878]	24297	126576	–
instance[879]	24297	123002	–
instance[901]	24297	121177	–
instance[902]	24297	114550	–
instance[922]	24279	175593	–
instance[923]	24279	164886	–
instance[939]	24279	165020	–
instance[1147]	24279	116876	–
instance[1148]	24279	157564	–
instance[1149]	24279	151079	–
instance[1150]	24279	113369	–
instance[1151]	24279	139323	–
instance[1152]	24279	143545	–
instance[1153]	24279	161917	–
instance[1154]	24279	135617	–
instance[1155]	24279	113363	–
instance[1156]	24279	148607	–
instance[1157]	24279	120212	–
instance[1158]	24279	125626	–
instance[1159]	24279	117945	–
instance[840]	23512	135594	–
instance[841]	23512	137154	–
instance[891]	23512	108474	–
instance[910]	23512	131999	–
instance[911]	23512	159755	–
instance[548]	23446	129344	–
instance[549]	23446	110316	–
instance[550]	23446	136514	–
instance[854]	23446	103565	–
instance[855]	23446	110351	–

Instance: instance[19]

Instance Custom Index	1401
Mask	0b00001011
Instance Offset	56
Flags	FRONT_CCW
BLAS	BLAS 1c01b70000
Transform	1.000, 0.000, 0.000, 0.000 0.000, 1.000, 0.000, 0.000 0.000, 0.000, 1.000, 0.000

Performance Analysis

Instance Overlaps... Instance Heatmap...

Visualization

Camera Controls

Mode:	Fly Camera
Position (X):	2176.253
Position (Y):	725.979
Position (Z):	20234.133
Direction (X):	0.329
Direction (Y):	0.717
Direction (Z):	-0.615
Near Plane:	17.733
Far Plane:	7733176.000
FOV:	45.000
Speed:	57.817
Up Direction:	7 Axis
Coordinates:	LHS

Reset Camera

Highlight

- Opaque
- Non-Opaque
- Cull Disabled
- Front CCW
- Fast Build
- Fast Trace
- Low Memory
- Allow Update
- Allow Compaction
- Compacted
- No Duplicate Any Hit Invocation

Rendering Options

Color By:	Instance Complexity
Instance AABB Mode:	Heatmap
AABB Heatmap Threshold:	16.000
Anti-aliasing Mode:	MSAA (4x)

- ✓ Disable Back Face Cull
- Show Wireframes
- ✓ Show Geometry
- ✓ Show Axes
- ✓ Show Non-Opaque Transparency

Bookmarks

SHADER TIMING HEATMAP

Ray Tracing Hotspot Analysis

NVIDIA Nsight Graphics Pro
File
Connection
Tools
Window
Help
Connect
Disconnect
Terminate
Documents
API
Heatmap
Overlay
Enable Overlay
No Image
Opacity:

Shader: View Source Profile Shaders

Dimensions: 1920 x 1080 x 1

X: 0 Y: 0 W: 1920 H: 1080 Z: 0 Pixels: 2073600

RayGen	Time
234, 38, 0	1.66 ms
1137, 904, 0	1.65 ms
212, 18, 0	1.63 ms
220, 4, 0	1.63 ms
220, 5, 0	1.63 ms
222, 1, 0	1.63 ms
927, 985, 0	1.61 ms
216, 31, 0	1.60 ms
251, 1019, 0	1.60 ms
280, 1002, 0	1.60 ms
231, 35, 0	1.59 ms
231, 38, 0	1.59 ms
231, 33, 0	1.59 ms
258, 1016, 0	1.59 ms
1170, 928, 0	1.59 ms
230, 25, 0	1.58 ms
1039, 937, 0	1.58 ms
1162, 927, 0	1.57 ms
1069, 911, 0	1.56 ms
232, 37, 0	1.55 ms
219, 0, 0	1.55 ms
189, 1027, 0	1.55 ms
181, 1028, 0	1.55 ms
31, 988, 0	1.54 ms
234, 40, 0	1.54 ms
1163, 929, 0	1.54 ms
237, 60, 0	1.54 ms
1919, 848, 0	1.54 ms
225, 16, 0	1.54 ms
1279, 37, 0	1.54 ms
227, 19, 0	1.54 ms
226, 18, 0	1.54 ms
1277, 37, 0	1.54 ms
1276, 37, 0	1.54 ms
281, 976, 0	1.53 ms
1276, 34, 0	1.53 ms
1276, 36, 0	1.53 ms
1277, 32, 0	1.53 ms
1279, 34, 0	1.53 ms
1276, 35, 0	1.53 ms
1277, 35, 0	1.53 ms
1278, 35, 0	1.53 ms
1276, 32, 0	1.53 ms
1278, 37, 0	1.53 ms
1278, 36, 0	1.53 ms
1278, 32, 0	1.53 ms
1279, 36, 0	1.53 ms
1277, 36, 0	1.53 ms
1279, 32, 0	1.53 ms
1279, 33, 0	1.53 ms
1278, 33, 0	1.53 ms
1278, 34, 0	1.53 ms
1276, 33, 0	1.53 ms
1277, 34, 0	1.53 ms
1279, 35, 0	1.53 ms
272, 1003, 0	1.53 ms

GPU TRACE

Low-level Metrics Graph Profiler

Ruler Relative: Capture
Export...
Trace Analysis...
Trace Compare...
Overlay Barriers
Flat Queue Rows
Aggregate Frames
Start: 1.34ms
End: 6.52ms
Reset Zoom
Zoom To Selection
Time
1.5ms
2ms
2.5ms
2.66172ms
3ms
3.38515ms Δt=0.723436ms
4ms
4.5ms
5ms
5.26038ms
5.5ms
6ms
6.5ms
Frames
Matched Frame (7.53ms) (active=7.44ms)
queueGC1 (Vulkan Graphics Q:0)
Synchronization
Actions
QueueSubmit 0 (1) (7.44ms)
CommandBuffer 0 (7.44ms)
TraceRays 0 (7.37ms)
Markers
Ray Trace (7.37ms)
GPU Active
Unit Throughputs
SM Throughput
L1 Throughput
RTCORE Throughput (Pro)
VAF Throughput
MMU Throughput (Pro)
L2 Throughput
VRAM Throughput
PCIe Throughput
PD Throughput
PES+VPC Throughput
RASTER Throughput
PROP Throughput
ZROP Throughput
CROP Throughput
Compute In Flight
Draw/Dispatch Start
SM Occupancy (TPC View)
Warps Eligible
FE Stalls
FE Pipeline Stalling Commands
SM Register Usage (Occupancy Limiter)
Summary
Metrics
Capture Information
Type: Marker
Top Metrics Only
Visible Only
Search... or select a predefined one on the right

Start: 2.66 ms Duration: 0.72 ms
End: 3.39 ms Range: User Range

	Unit Throughputs	Throughput %
🟦	PCIe Throughput	14.2%
🟧	VRAM Throughput	13.7%
🟩	PROP Throughput	12.3%
🟨	CROP Throughput	12.0%
🟦	L2 Throughput	7.9%
✓	RTCORE Throughput (Pro)	7.0%
🟨	L1 Throughput	3.6%
🟪	SM Throughput	2.7%
✓	MMU Throughput (Pro)	2.2%
🟩	RASTER Throughput	1.5%
🟩	PES+VPC Throughput	< 0.1%
🟥	VAF Throughput	< 0.1%
🟩	PD Throughput	< 0.1%
🟫	ZROP Throughput	0%

	SM Occupancy (TPC View)	Throughput %	Avg	Avg Warps per Cycle
⬜	Idle SM Unused Warp Slots	82.2%	79486019.1	78.9
⬛	Active SM Unused Warp Slots	8.3%	8017538.0	8.0
🟩	Pixel Warps	5.4%	5212915.5	5.2
🟧	Compute Warps	4.1%	3975725.3	3.9
🟦	Vertex/Tess/Geometry Warps	< 0.1%	2619.6	0.0

Name	Path	Type	Start	End	Duration	Frame	SM	L1	RTCORE	VAF	MMU	L2	VRAM	PCIe	PD	PES+VPC	RASTER	PROP	ZROP	CROP	SM Issue	SM ALU	SM FMAL	SM FMAH	SM SFU
Post		Marker	7.43ms	7.47ms	0.04ms (0%)	0	23.7%	21.3%	0.0%	0.0%	0.9%	38.7%	91.3%	14.2%	0.0%	0.3%	8.3%	31.7%	2.9%	31.3%	6.9%	0.0%	3.5%	0.8%	23.7%
Ray Trace		Marker	0.04ms	7.41ms	7.37ms (97%)	0	20.4%	9.7%	81.9%	0.0%	1.2%	15.2%	11.5%	14.0%	0.0%	0.0%	0.1%	1.2%	0.0%	1.2%	20.4%	17.5%	3.8%	9.1%	2.9%

GPU TRACE

Trace Analysis

Skip GPU-idle gaps for frame gain
Show analysis colors only for leaf markers
Export...
Frames
Other Processes
Direct3D Direct Q:0
Markers
Unit Throughputs
SM Occupancy (TPC Vie...
User Ranges
RTX Rendering (16.34ms)
Frame 0 (42.68ms) (active: 10.74ms)
Ruler Relative: Capture
Trace Analysis...
Trace Compare...
queueGCT (Vulkan Graphics Q:0)
Synchronization
Actions
GPU Active

	ms	%	Frame gain
ne 0	42.68	100.0%	
Queue 0	42.68	100.0%	
Upload Buffer Data	0.00	0%	N/A
RTX Rendering	16.34	38.3%	42.8%
SceneDb: Skinning	0.00	0%	N/A
SceneDb: GPU Normals	0.00	0%	N/A
SceneDb: flushBufferUpdates	0.03	< 0.1%	0.3%
Update SBT-Instance-BlockData	0.03	< 0.1%	0.3%
Prepare GPU Scene Instance Data	0.02	< 0.1%	0.0%
Update GPU SceneDb Instances	0.01	< 0.1%	0.1%
Translate Instance position	0.01	< 0.1%	0.1%
Update GPU SceneDb AccelStruct	0.47	1.1%	3.9%
TLAS rebuild	0.46	1.1%	3.8%
Cached PT ClearNew	0.10	0.2%	0.1%
Flow	0.09	0.2%	0.4%
SparseCS	0.01	< 0.1%	0.1%
RayMarchUpdateColormapCS	0.01	< 0.1%	0.1%
SparseCS	0.01	< 0.1%	0.1%
GBuffer RT	0.31	0.7%	0.2%
Shadows	0.90	2.1%	5.9%
Shadow RT	0.16	0.4%	0.4%
Shadow Spatial Filter	0.08	0.2%	0.2%
Shadow Denoiser	0.65	1.5%	3.6%
NVRTD PrepareInput	0.02	< 0.1%	0.1%
NVRTD ComplexMask	0.01	< 0.1%	0.1%
NVRTD PenumbraMask	0.03	< 0.1%	0.1%
NVRTD PenumbraMask	0.02	< 0.1%	0.1%
NVRTD AvgMinDist	<0.01	< 0.1%	0.0%
NVRTD AvgMinDist	<0.01	< 0.1%	0.0%
NVRTD Denoise	0.04	< 0.1%	0.2%
NVRTD Denoise	0.03	< 0.1%	0.2%
NVRTD Denoise	0.03	< 0.1%	0.2%
NVRTD Denoise	0.02	< 0.1%	0.2%
NVRTD UpSample	0.04	< 0.1%	0.3%
NVRTD UpSampleSimple	<0.01	< 0.1%	0.1%
Deferred LTC Lighting RT	1.60	3.8%	0.4%
Texture To Buffer Copy	0.02	< 0.1%	0.1%
Texture To Buffer Copy	0.03	< 0.1%	0.0%
Optix-7 Denoising	4.65	10.9%	N/A

Range: "RTX Rendering/Update GPU SceneDb AccelStruct/TLAS rebuild" Duration: 0.46ms

Overview [0.0%] | SM Warp Occupancy [3.7%] | SM Warp Latency [0.0%] | Graphics/Compute Idle [0.0%] | L1 L2 Traffic [0.0%] | Copy Engines | L1 L2 Hit Rates | L1 L2 Sectors | L1 Latency

Analysis Results (click on each entry for details): Copy to Clipboard

Name	Value	Description
SM Fully Idle		
Idle SM Unused Warp Slots [%]	88.1%	cumulative number of idle warp slots on idle SMs
Engine Active Copy Sync	0	Number of elapsed cycles where a synchronous Copy Engine is active (Copy in a DIRECT or COMPUTE queue)
Engine Active Copy Async	0	Number of elapsed cycles where an asynchronous Copy Engine is active (Copy in a COPY queue)
Projected range speedup factor	8.40	Estimation of the ratio between the current GPU time for the range divided by the projected GPU time for the range (2.0 means the range may get 2x faste
Projected range gain	0.40ms	The potential reduction in time for the range (based on the projected range speedup factor)
Projected frame gain	3.7%	Projected frame Gain in % based on the speedup factor of the current range, and the % of the GPU frame time that this range is taking (1.0% means the fr
Warp Launch Stalled by CS Register Count		
Warp Launch Stalled by CS Warp Slots		
Warp Launch Stalled by CS Shared-Memory Size		
Warp Launch Stalled by CS Thread-Group Slots		
Warp Launch Stalled by Reasons [Cycles]		

Percentage of the cycles where an SM has all of its warp slots empty (SM is fully idle).
This may be due to the "Tail Effect", that is the warp occupancy going down at the end of a SM workload before a Wait For Idle.

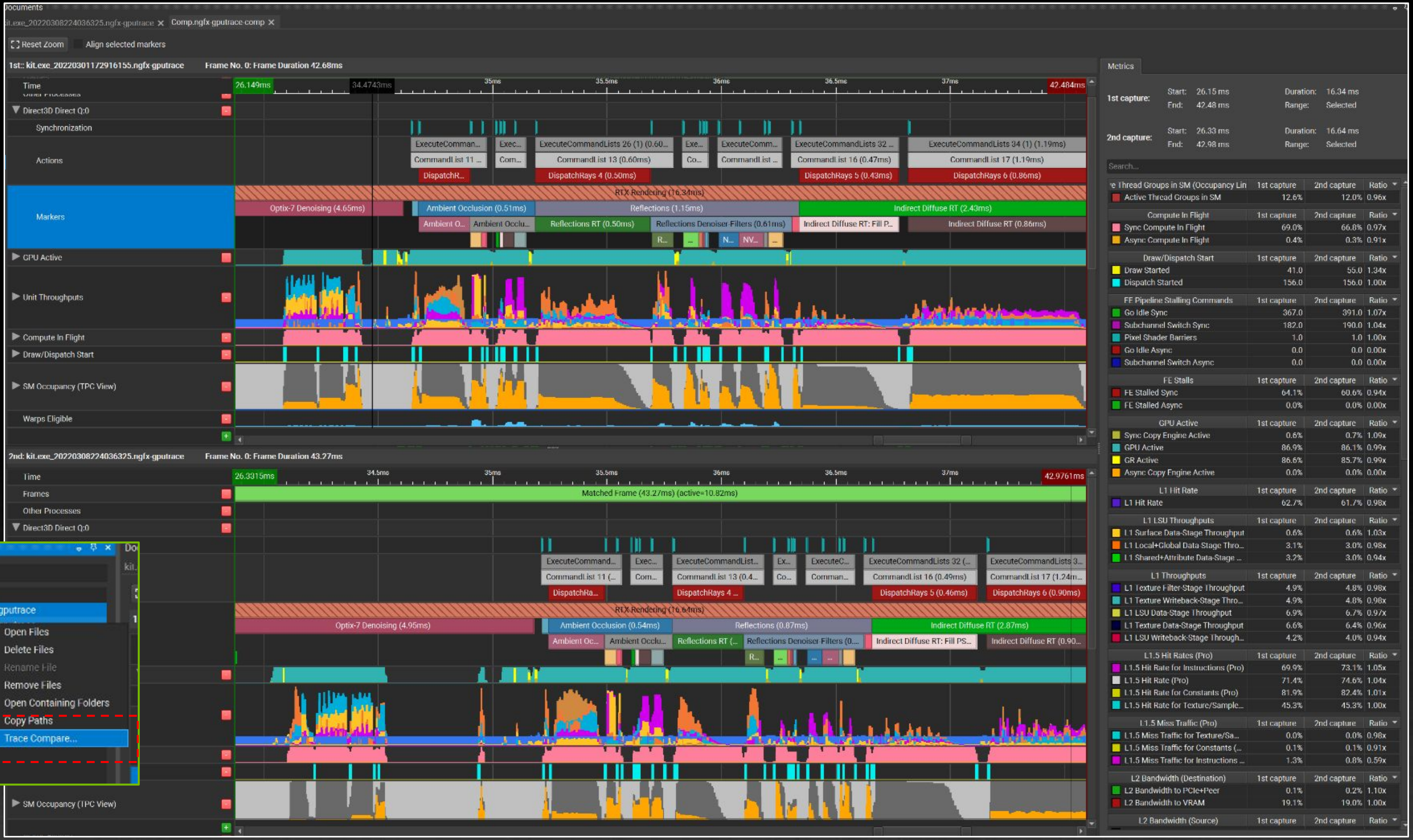
Suggestion

Look at the SM Warp Occupancy metric graph in GPU Trace. If a significant tail is there, then:

- See if a DX12 ResourceBarrier command can be removed at the end of the current workload, to let it execute concurrently with the next one,
- assuming that the current and next workloads are in the same subchannel (Graphics or Compute).
- If high, and resource dependencies permit, then:
- If this workload is a Compute Shader, then it may be a good candidate for being moved to an async-compute queue.
- If this workload is Graphics, then it may be a good candidate to be overlapped with async-compute work.
- See also: https://devblogs.nvidia.com/the-peak-performance-analysis-method-for-optimizing-any-gpu-workload/#5.1.2

GPU TRACE

Trace Compare

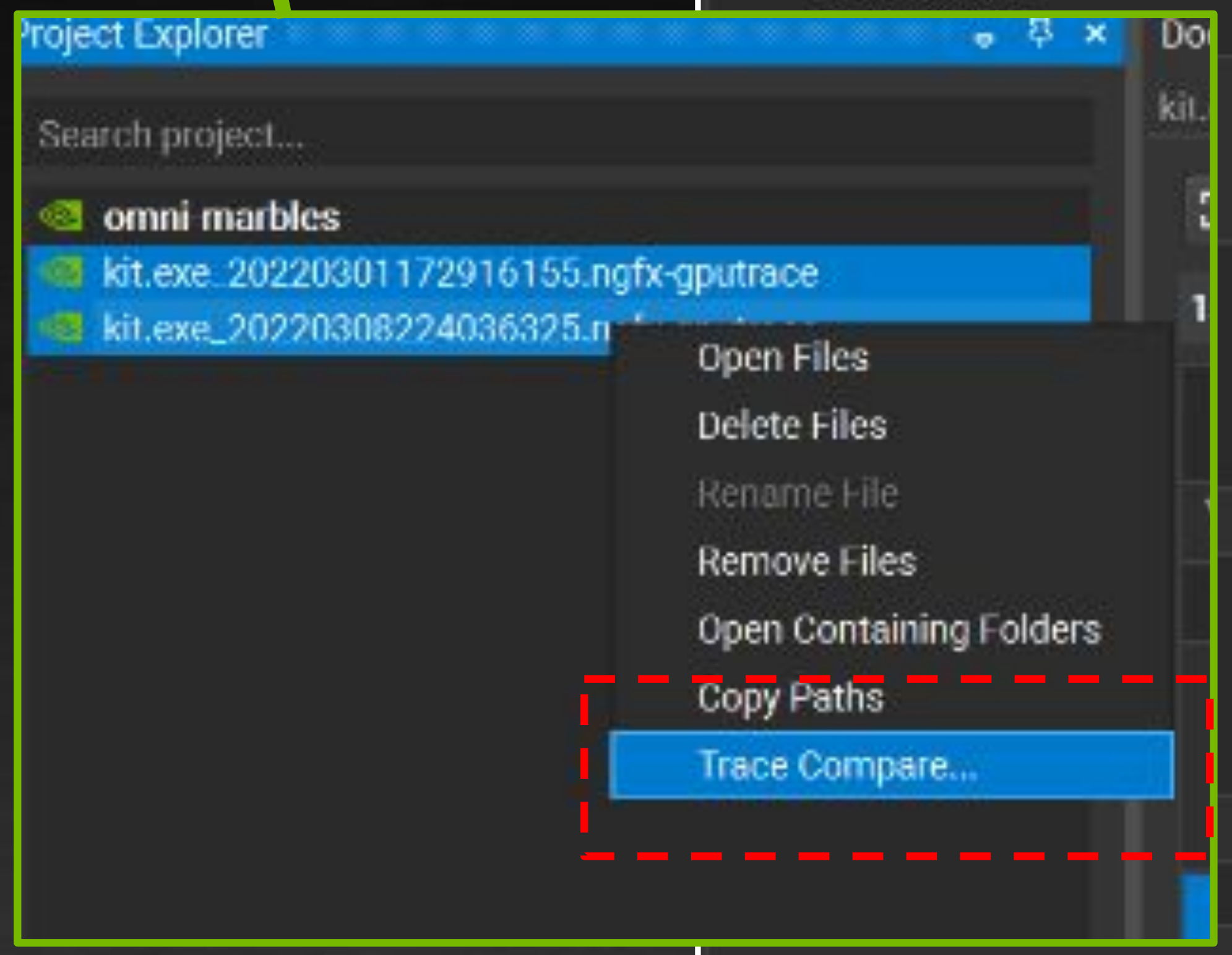


1st: kit.exe_20220301172916155.ngfx-gputrace   Frame No. 0: Frame Duration 42.68ms

2nd: kit.exe_20220308224036325.ngfx-gputrace   Frame No. 0: Frame Duration 43.27ms

1st capture:	Start: 26.15 ms	End: 42.48 ms	Duration: 16.34 ms	Range: Selected
2nd capture:	Start: 26.33 ms	End: 42.98 ms	Duration: 16.64 ms	Range: Selected

e Thread Groups in SM (Occupancy Lin	1st capture	2nd capture	Ratio
🟥 Active Thread Groups in SM	12.6%	12.0%	0.96x
Compute In Flight	**1st capture**	**2nd capture**	**Ratio**
🟥 Sync Compute In Flight	69.0%	66.8%	0.97x
🟧 Async Compute In Flight	0.4%	0.3%	0.91x
Draw/Dispatch Start	**1st capture**	**2nd capture**	**Ratio**
🟨 Draw Started	41.0	55.0	1.34x
🟦 Dispatch Started	156.0	156.0	1.00x
FE Pipeline Stalling Commands	**1st capture**	**2nd capture**	**Ratio**
🟩 Go Idle Sync	367.0	391.0	1.07x
🟪 Subchannel Switch Sync	182.0	190.0	1.04x
🟦 Pixel Shader Barriers	1.0	1.0	1.00x
🟥 Go Idle Async	0.0	0.0	0.00x
🟦 Subchannel Switch Async	0.0	0.0	0.00x
FE Stalls	**1st capture**	**2nd capture**	**Ratio**
🟥 FE Stalled Sync	64.1%	60.6%	0.94x
🟩 FE Stalled Async	0.0%	0.0%	0.00x
GPU Active	**1st capture**	**2nd capture**	**Ratio**
🟨 Sync Copy Engine Active	0.6%	0.7%	1.09x
🟦 GPU Active	86.9%	86.1%	0.99x
🟨 GR Active	86.6%	85.7%	0.99x
🟧 Async Copy Engine Active	0.0%	0.0%	0.00x
L1 Hit Rate	**1st capture**	**2nd capture**	**Ratio**
🟪 L1 Hit Rate	62.7%	61.7%	0.98x
L1 LSU Throughputs	**1st capture**	**2nd capture**	**Ratio**
🟨 L1 Surface Data-Stage Throughput	0.6%	0.6%	1.03x
🟧 L1 Local+Global Data-Stage Thro...	3.1%	3.0%	0.98x
🟩 L1 Shared+Attribute Data-Stage ...	3.2%	3.0%	0.94x
L1 Throughputs	**1st capture**	**2nd capture**	**Ratio**
🟦 L1 Texture Filter-Stage Throughput	4.9%	4.8%	0.98x
🟦 L1 Texture Writeback-Stage Thro...	4.9%	4.8%	0.98x
🟨 L1 LSU Data-Stage Throughput	6.9%	6.7%	0.97x
🟦 L1 Texture Data-Stage Throughput	6.6%	6.4%	0.96x
🟥 L1 LSU Writeback-Stage Through...	4.2%	4.0%	0.94x
L1.5 Hit Rates (Pro)	**1st capture**	**2nd capture**	**Ratio**
🟪 L1.5 Hit Rate for Instructions (Pro)	69.9%	73.1%	1.05x
⬜ L1.5 Hit Rate (Pro)	71.4%	74.6%	1.04x
🟨 L1.5 Hit Rate for Constants (Pro)	81.9%	82.4%	1.01x
🟦 L1.5 Hit Rate for Texture/Sample...	45.3%	45.3%	1.00x
L1.5 Miss Traffic (Pro)	**1st capture**	**2nd capture**	**Ratio**
🟦 L1.5 Miss Traffic for Texture/Sa...	0.0%	0.0%	0.98x
🟨 L1.5 Miss Traffic for Constants (...	0.1%	0.1%	0.91x
🟪 L1.5 Miss Traffic for Instructions ...	1.3%	0.8%	0.59x
L2 Bandwidth (Destination)	**1st capture**	**2nd capture**	**Ratio**
🟩 L2 Bandwidth to PCIe+Peer	0.1%	0.2%	1.10x
🟥 L2 Bandwidth to VRAM	19.1%	19.0%	1.00x
L2 Bandwidth (Source)	**1st capture**	**2nd capture**	**Ratio**



SHADER PROFILER

Instruction Stalls as an indicator to optimal shader efficiency

Source Code
Stalls
SASS
Stall Dependencies
Stalls
Live Registers
Register Dependencies

```
  for(uint n = 0; n < 16; n++)
  {
    s0 += 0x9e3779b9;
    v0 += ((v1 << 4) + 0xa341316c) ^ (v1 + s0) ^ ((v1 >> 5) + 0xc801      0.23%
    v1 += ((v0 << 4) + 0xad90777d) ^ (v0 + s0) ^ ((v0 >> 5) + 0x7e95      0.26%
  }

  return v0;
}

// Generate a random unsigned int in [0, 2^24) given the previous RNG sta
// using the Numerical Recipes linear congruential generator
uint lcg(inout uint prev)
{
  uint LCG_A = 1664525u;
  uint LCG_C = 1013904223u;
  prev       = (LCG_A * prev + LCG_C);                                  < 0.01%
  return prev & 0x00FFFFFF;
}

// Generate a random float in [0, 1) given the previous RNG state
float rnd(inout uint prev)
{
  return (float(lcg(prev)) / float(0x01000000));
}
```

1 …	SHF.R.U32.HI R3, RZ, c[0x3][0x74], R13	0.07%	NOINST	99%	MATHTH	< 1%	NOTSEL	< 1%	29
2 …	ULDC.64 UR6, c[0x3][0x1d0]	0.09%	IMCMIS	94%	NOINST	4%	SELECT	< 1%	29
3 …	LOP3.LUT R2, R23, c[0x3][0x70], RZ, 0xc0, !F	< 0.01%	SELECT	91%	NOTSEL	4%	MATHTH	3%	30
4 …	I2FP.F32.U32 R12, c[0x3][0x1c]	< 0.01%	NOTSEL	33%	MATHTH	29%	WAIT	19%	31
5 …	I2FP.F32.U32 R60, R2	< 0.01%	NOTSEL	44%	MATHTH	40%	SELECT	10%	32
6 …	IMAD R0, R3, c[0x3][0x1c], R2	< 0.01%	SELECT	100%					33
7 …	MUFU.RCP R15, R12	< 0.01%	SELECT	99%	MIOT	< 1%			34
8 …	I2FP.F32.U32 R61, R3	< 0.01%	SELECT	100%					34
9 …	MOV R16, 0x40000000	< 0.01%	WAIT	47%	SELECT	44%	NOINST	4%	35
10 …	IADD3 R6, R0, -0xa88898f, RZ	0.03%	NOTSEL	50%	MATHTH	47%	SELECT	2%	36
11 …	IADD3 R7, R0, -0x6c510fd6, RZ	< 0.01%	SELECT	52%	WAIT	45%	NOTSEL	2%	37
12 …	LEA R0, R6, 0xad90777d, 0x4	< 0.01%	WAIT	48%	SELECT	46%	NOTSEL	3%	37
13 …	LEA.HI R10, R6, 0x7e95761e, RZ, 0x1b	< 0.01%	WAIT	53%	SELECT	44%	NOTSEL	1%	38
14 …	LOP3.LUT R0, R10, R0, R7, 0x96, !PT	0.02%	MATHTH	44%	NOTSEL	44%	WAIT	9%	38
15 …	FMUL.FTZ R60, R60, R15	< 0.01%	SELECT	92%	SHRTSB	6%	NOTSEL	3%	36
16 …	LEA R7, R0, 0xa341316c, 0x4	< 0.01%	NOTSEL	40%	MATHTH	39%	WAIT	14%	36
17 …	IADD3 R10, R0, 0x3c6ef372, RZ	< 0.01%	SELECT	50%	WAIT	46%	NOTSEL	2%	37
18 …	LEA.HI R11, R0, 0xc8013ea4, RZ, 0x1b	< 0.01%	WAIT	48%	SELECT	46%	NOTSEL	3%	38
19 …	LOP3.LUT R7, R10, R11, R7, 0x96, !PT	< 0.01%	NOTSEL	38%	MATHTH	38%	WAIT	18%	38
20 …	IMAD.IADD R7, R6, 0x1, R7	< 0.01%	WAIT	74%	SELECT	17%	NOTSEL	6%	36
21 …	LEA R6, R7, 0xad90777d, 0x4	0.01%	MATHTH	40%	NOTSEL	39%	WAIT	17%	36
22 …	IADD3 R11, R7, 0x3c6ef372, RZ	< 0.01%	SELECT	49%	WAIT	48%	MATHTH	2%	37
23 …	LEA.HI R10, R7, 0x7e95761e, RZ, 0x1b	< 0.01%	SELECT	50%	WAIT	47%	NOTSEL	2%	38
24 …	LOP3.LUT R11, R10, R6, R11, 0x96, !PT	< 0.01%	NOTSEL	40%	MATHTH	39%	WAIT	15%	38
25 …	IMAD.IADD R11, R0, 0x1, R11	< 0.01%	WAIT	64%	SELECT	16%	NOINST	8%	36
26 …	LEA R0, R11, 0xa341316c, 0x4	< 0.01%	NOTSEL	38%	MATHTH	36%	WAIT	21%	36
27 …	IADD3 R13, R11, -0x255992d5, RZ	< 0.01%	WAIT	51%	SELECT	44%	NOTSEL	2%	37
28 …	LEA.HI R6, R11, 0xc8013ea4, RZ, 0x1b	< 0.01%	WAIT	49%	SELECT	46%	NOTSEL	5%	38
29 …	LOP3.LUT R0, R13, R6, R0, 0x96, !PT	< 0.01%	NOTSEL	36%	MATHTH	35%	WAIT	23%	38
30 …	IMAD.IADD R0, R7, 0x1, R0	< 0.01%	WAIT	71%	SELECT	18%	NOTSEL	6%	36

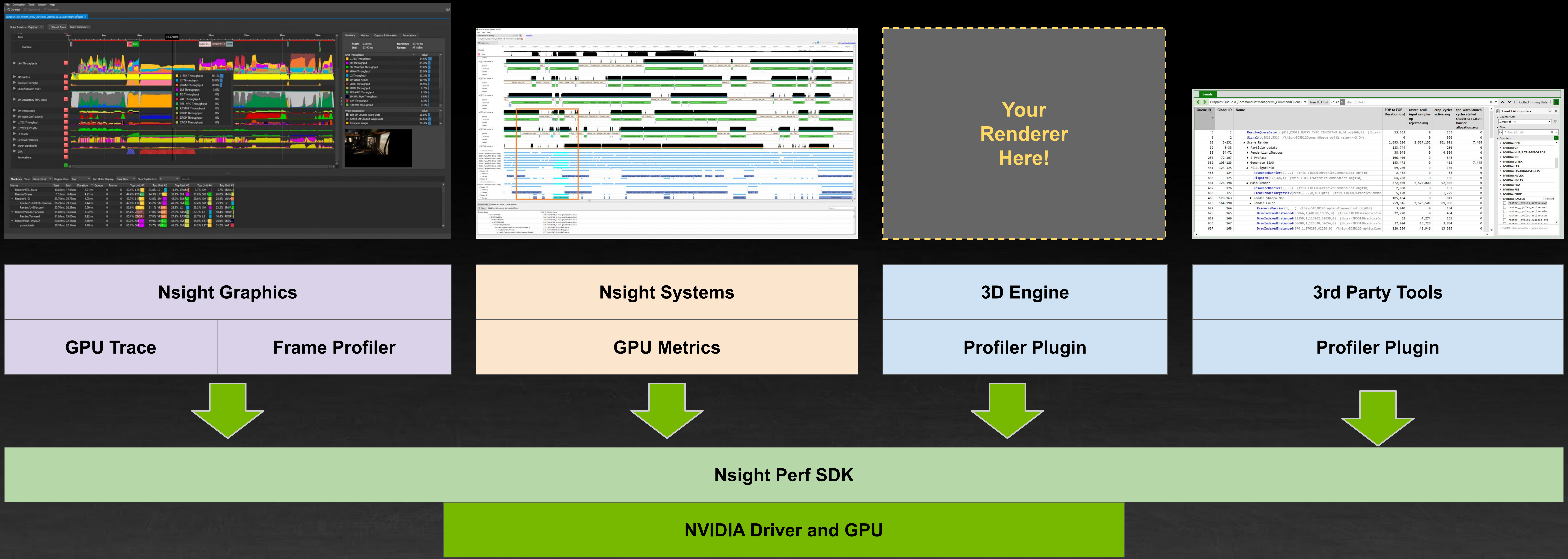
NSIGHT PERF SDK

Understand performance more effectively

It's a library (& API) that enables games to Collect GPU Performance Counters

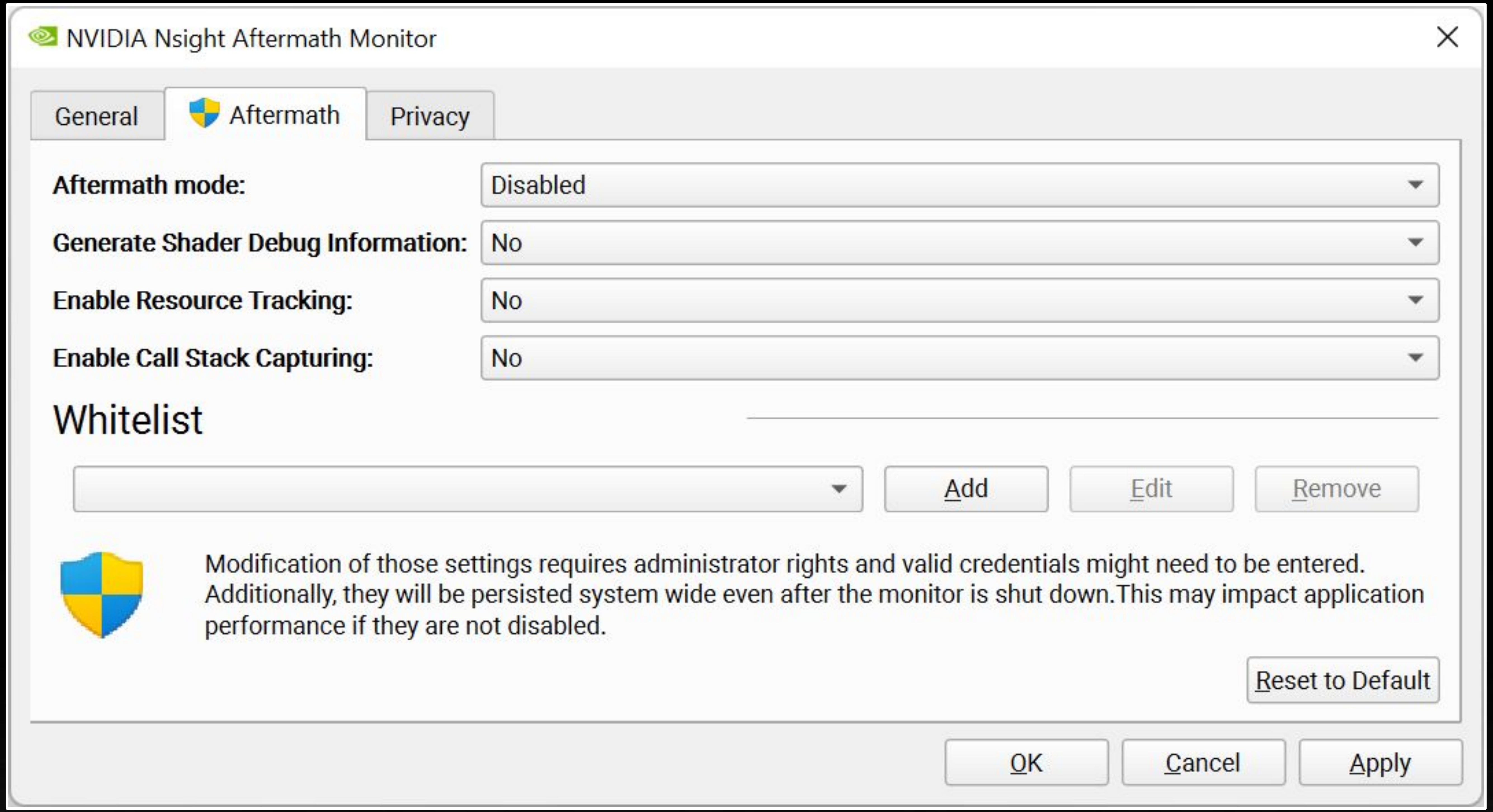
Very Low CPU and GPU Overhead

In-Application Integration via an intuitive API



NSIGHT AFTERMATH

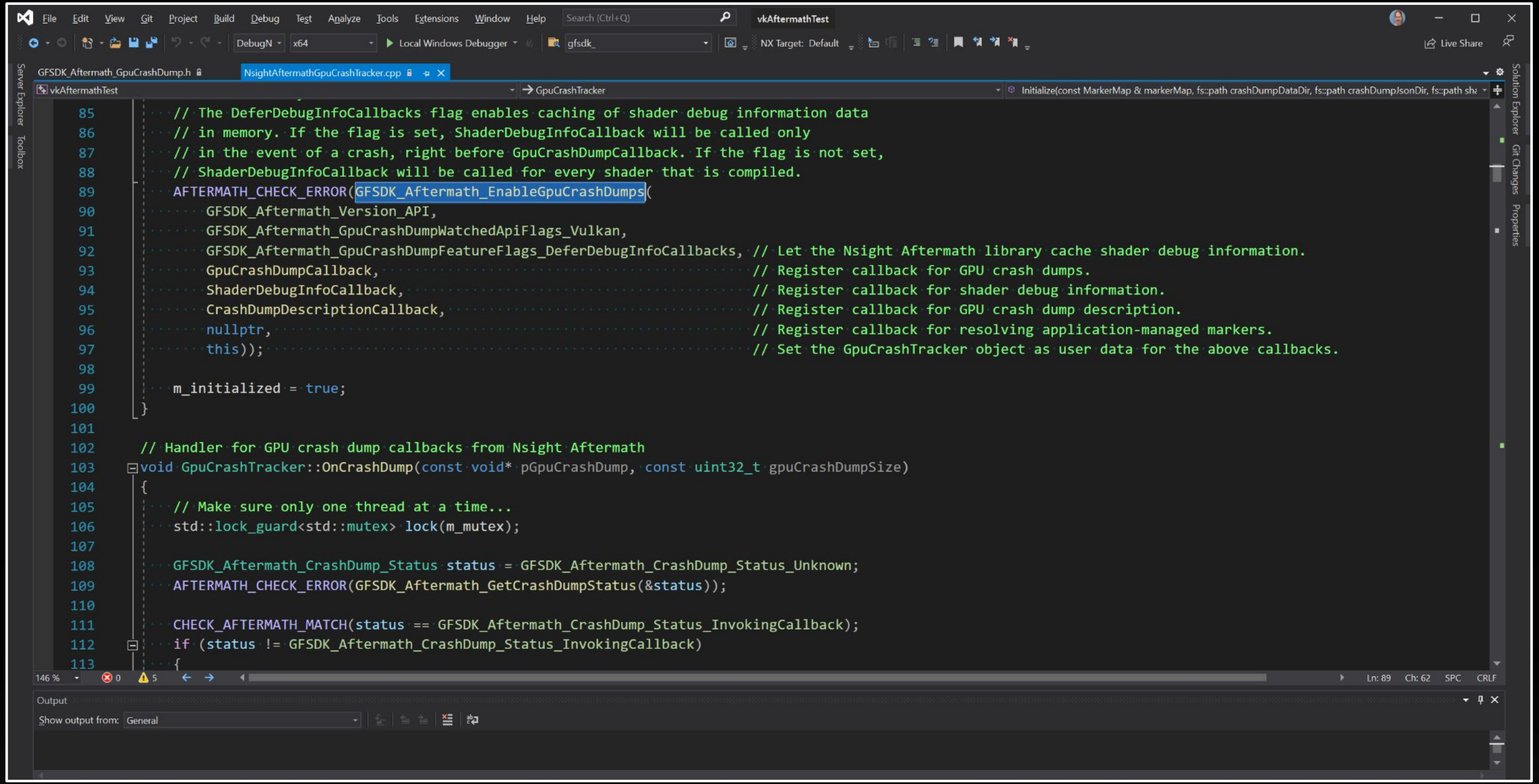
Multiple ways to use it



Aftermath mode: Disabled
Generate Shader Debug Information: No
Enable Resource Tracking: No
Enable Call Stack Capturing: No

Whitelist

Modification of those settings requires administrator rights and valid credentials might need to be entered. Additionally, they will be persisted system wide even after the monitor is shut down.This may impact application performance if they are not disabled.



```
    // The DeferDebugInfoCallbacks flag enables caching of shader debug information data
    // in memory. If the flag is set, ShaderDebugInfoCallback will be called only
    // in the event of a crash, right before GpuCrashDumpCallback. If the flag is not set,
    // ShaderDebugInfoCallback will be called for every shader that is compiled.
    AFTERMATH_CHECK_ERROR(GFSDK_Aftermath_EnableGpuCrashDumps(
        GFSDK_Aftermath_Version_API,
        GFSDK_Aftermath_GpuCrashDumpWatchedApiFlags_Vulkan,
        GFSDK_Aftermath_GpuCrashDumpFeatureFlags_DeferDebugInfoCallbacks, // Let the Nsight Aftermath library cache shader debug information.
        GpuCrashDumpCallback,                                             // Register callback for GPU crash dumps.
        ShaderDebugInfoCallback,                                          // Register callback for shader debug information.
        CrashDumpDescriptionCallback,                                     // Register callback for GPU crash dump description.
        nullptr,                                                          // Register callback for resolving application-managed markers.
        this));                                                           // Set the GpuCrashTracker object as user data for the above callbacks.

    m_initialized = true;
}

// Handler for GPU crash dump callbacks from Nsight Aftermath
void GpuCrashTracker::OnCrashDump(const void* pGpuCrashDump, const uint32_t gpuCrashDumpSize)
{
    // Make sure only one thread at a time...
    std::lock_guard<std::mutex> lock(m_mutex);

    GFSDK_Aftermath_CrashDump_Status status = GFSDK_Aftermath_CrashDump_Status_Unknown;
    AFTERMATH_CHECK_ERROR(GFSDK_Aftermath_GetCrashDumpStatus(&status));

    CHECK_AFTERMATH_MATCH(status == GFSDK_Aftermath_CrashDump_Status_InvokingCallback);
    if (status != GFSDK_Aftermath_CrashDump_Status_InvokingCallback)
    {
```

Nsight Aftermath Monitor

Automatically save GPU crash dumps

No code changes necessary

Control over settings, paths, etc.

Nsight Aftermath SDK

Full control over GPU crash dump serialization

Fine grained setup

User markers to narrow down faulting workload

WE'RE HIRING!

www.nvidia.com/en-us/about-nvidia/careers

GPU Debugging Tools Engineering Manager
GPU Debugging Tools Engineer
Game Console Developer Tools Engineer
GPU Profiling Tools Engineer
Shader Tools Engineer
Senior UX Designer
Product Manager

...and more!

THANK YOU!

developer.nvidia.com/tools-overview

Contact us at DevTools@nvidia.com

NVIDIA®