OPTIMIZING DX12/DXR GPU WORKLOADS

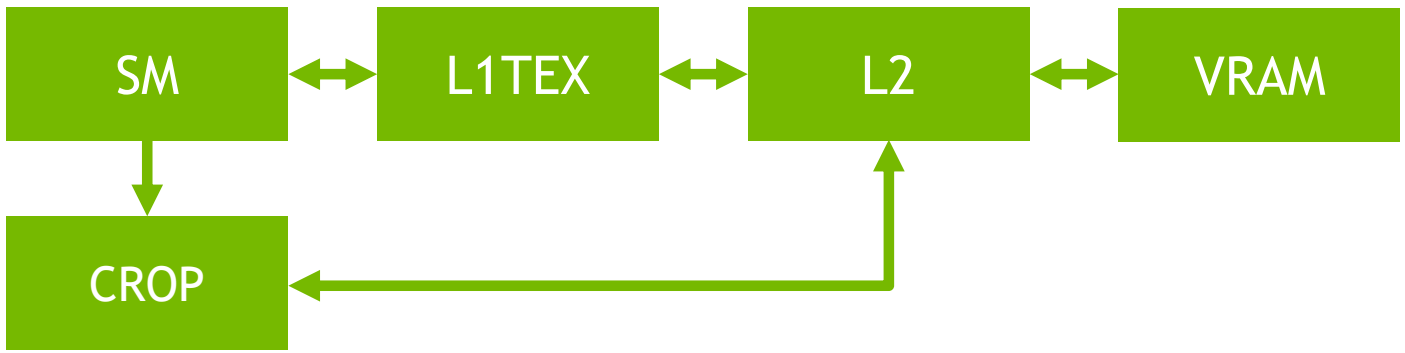
USING NSIGHT GRAPHICS: GPU TRACE AND THE PEAK-PERFORMANCE-PERCENTAGE (P3) METHOD

Louis Bavoil | Principal Engineer, Developer Technology

www.nvidia.com/GDC

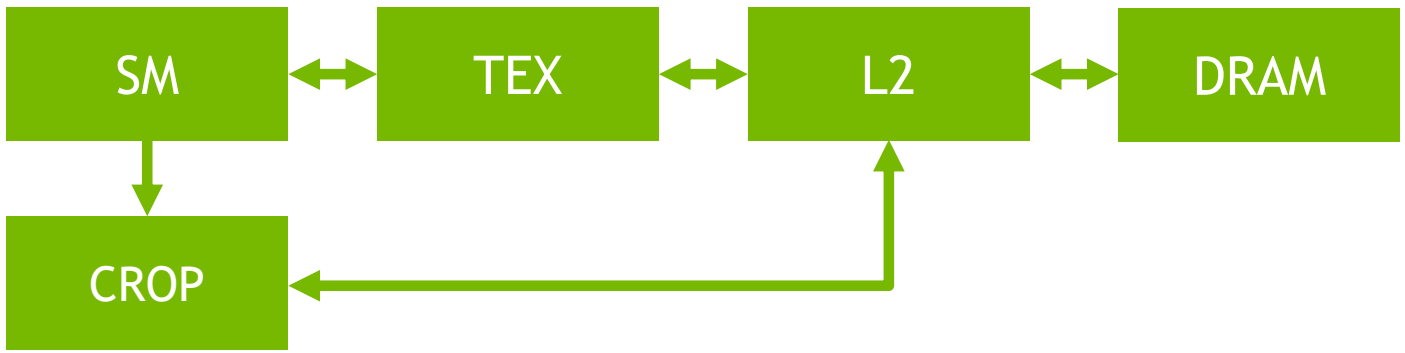

NVIDIA RTX and GameWorks Ray Tracing Demo (64-bit Development PCD3D_SM5)
RTAO DENOISER PIXEL SHADER

Full-Screen Pixel Shader



SM = Streaming Multiprocessor
L1TEX = Level 1 cache + Texture unit
L2 = Level 2 cache
VRAM = GDDR video-memory controller
CROP = Color ROP

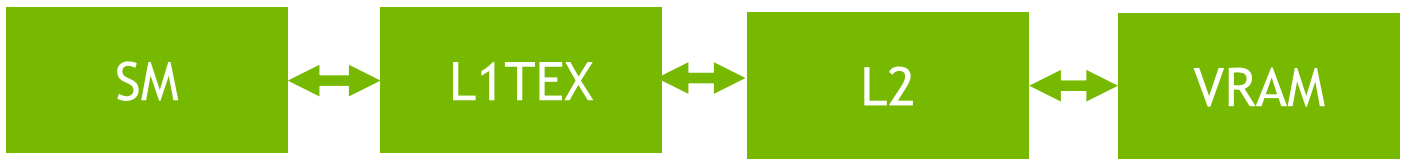
Unit Throughput% Metrics



"Throughput%" = % of max theoretical throughput

Also known as:
- Speed Of Light% (SOL%)
- Peak-Perf%

Top Throughput% Units

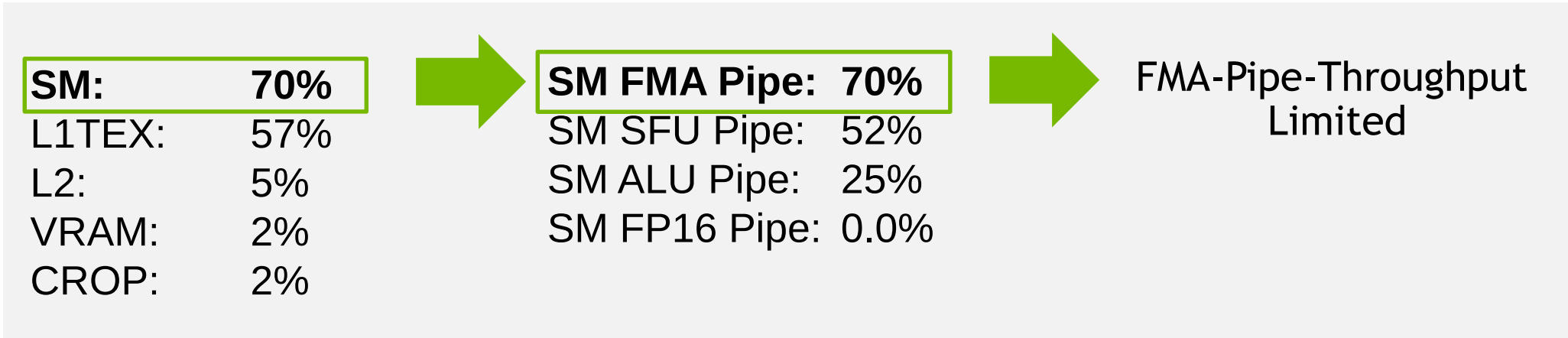


Example:

SM:	**70%**
L1TEX:	57%
L2:	5%
VRAM:	2%
CROP:	2%

SM-throughput limited

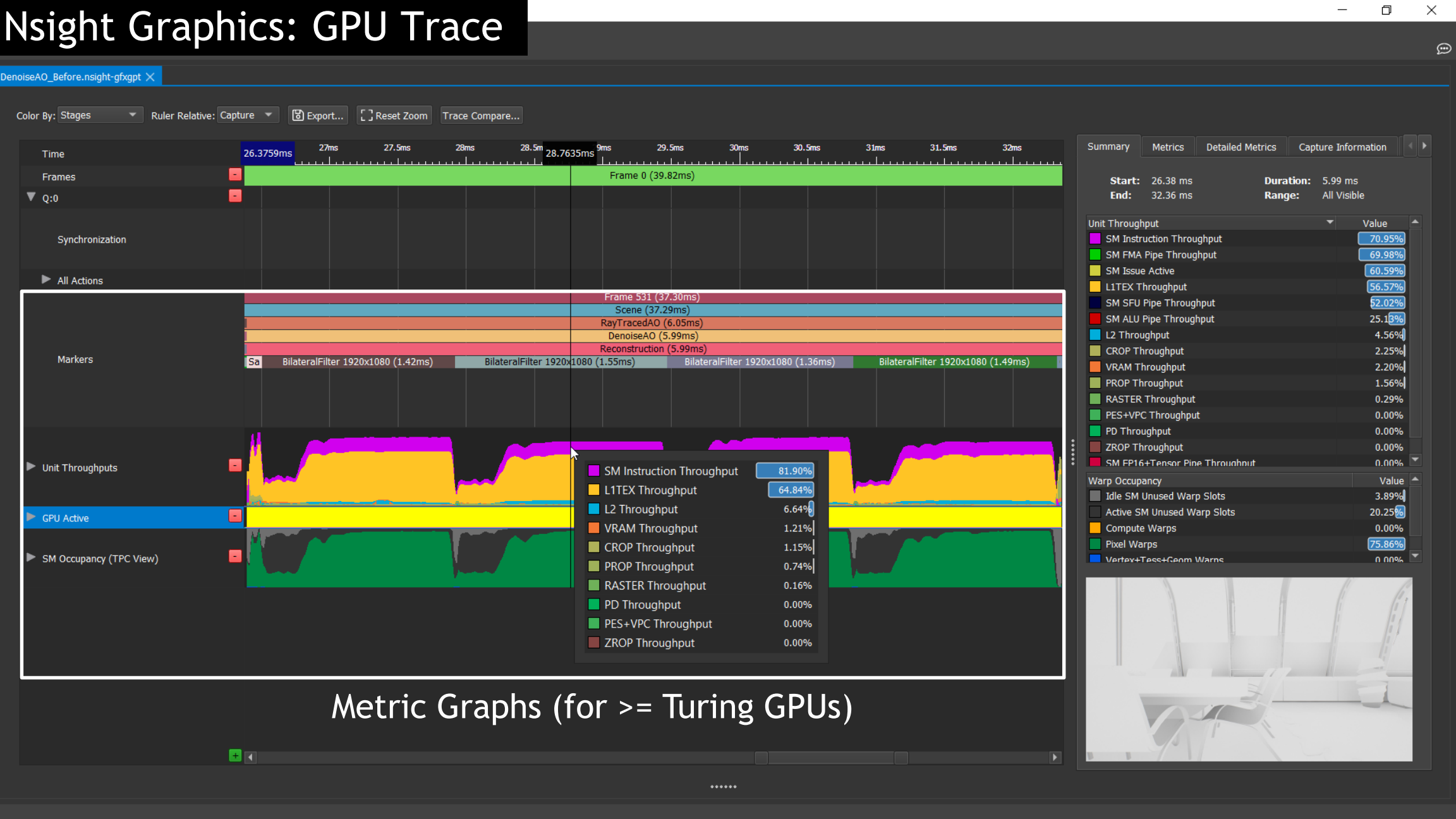
SM Sub-Throughput Metrics



SM pipes on Turing GPUs

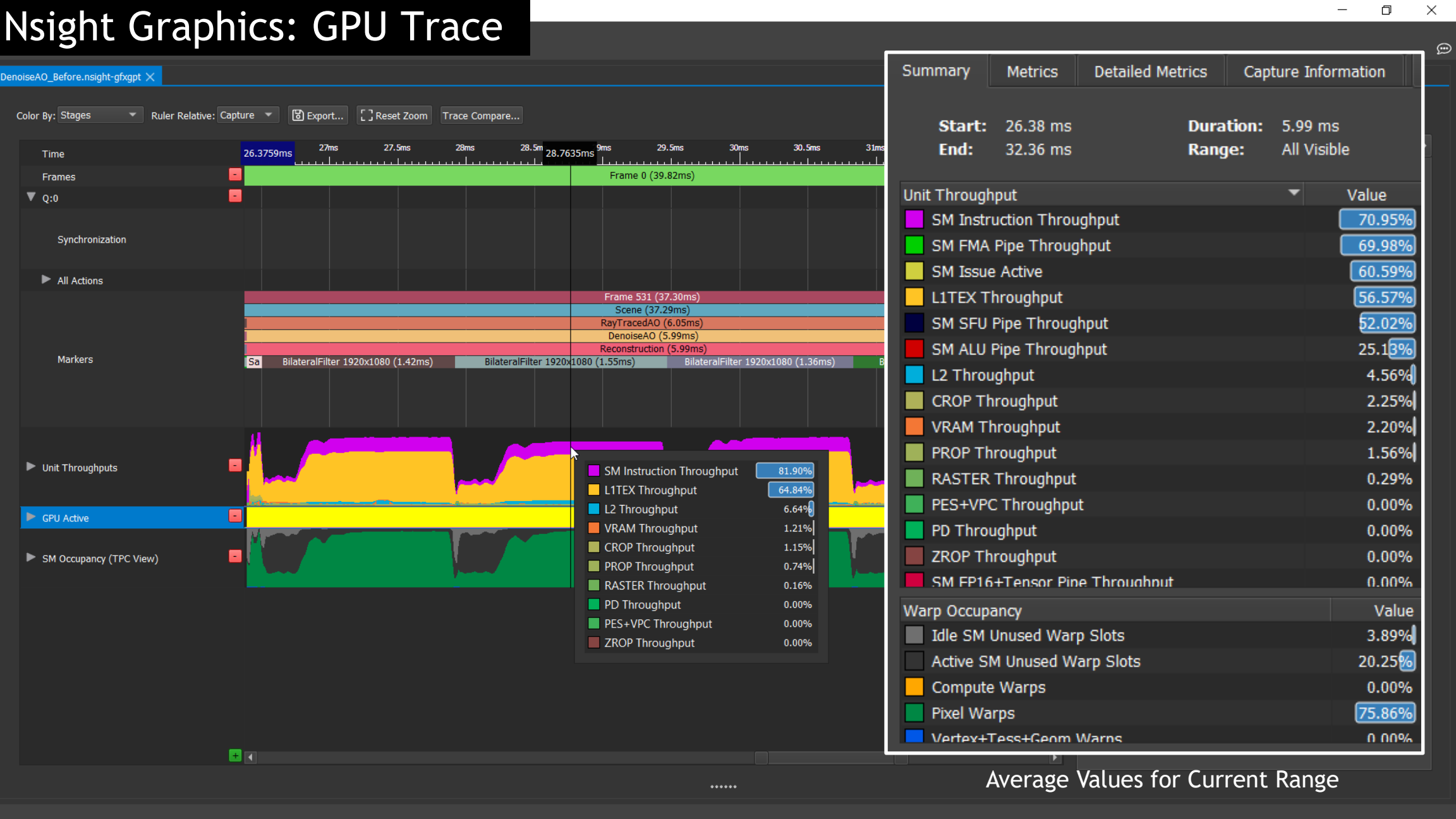
FMA: fp32 {FADD,FMUL,FMAD, …} ops + int {IMUL, IMAD} ops
ALU: integer & logic ops
FP16: FP16 ops executed in pairs
SFU: transcendental ops (rsqrt, cos/sin, etc.)

Nsight Graphics: GPU Trace



Metric Graphs (for >= Turing GPUs)

Nsight Graphics: GPU Trace



Average Values for Current Range

FMA-THROUGHPUT LIMITED

Why?

For each sample, shaders reconstruct a 3D world-space position with:

```
float2 SampleScreenPosition = (SampleScreenUV.xy - View.ScreenPositionScaleBias.wz) /
View.ScreenPositionScaleBias.xy;

float4 SampleHomogeneousWorldPosition = mul(float4(SampleScreenPosition *
SampleDepth, SampleDepth, 1), View.ScreenToWorld); ← 4x4 matrix mul

float3 SampleWorldPosition = SampleHomogeneousWorldPosition.xyz /
SampleHomogeneousWorldPosition.w;
```

FMA-REMOVAL EXPERIMENT:

```
#if 0

    float4 SampleHomogeneousWorldPosition =
      mul(float4(SampleScreenPosition * SampleZ, SampleZ, 1), View.ScreenToWorld);

#else

    float4 SampleHomogeneousWorldPosition =
      float4(SampleScreenPosition * SampleZ, SampleZ, 1);

#endif
```

FMA-REMOVAL EXPERIMENT:

4X4 MATRIX MUL -> NOP

	BEFORE	AFTER	RATIO
GPU Elapsed Time	5.99 ms	4.88 ms	**1.23x Gain**
Throughput: SM	71.0%	63.7%	0.90x
Throughput: L1TEX	56.6%	67.8%	1.20x
Throughput: L2	4.6%	5.5%	1.20x

On RTX 2080 with SetStablePowerState(TRUE)

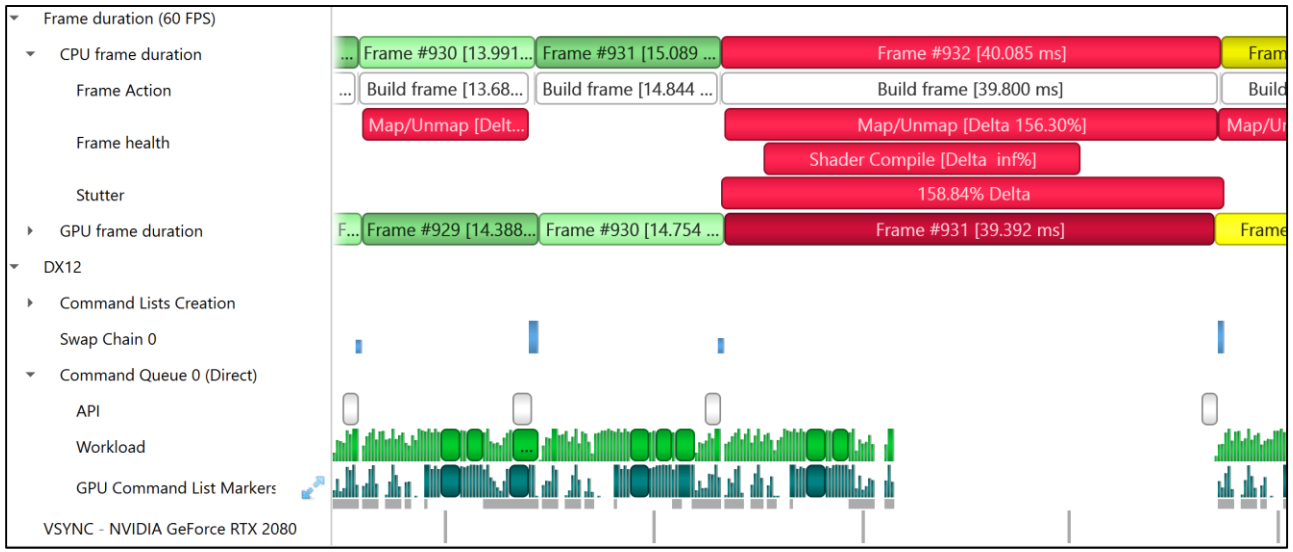
THE P3 (PEAK-PERF%) METHOD

START

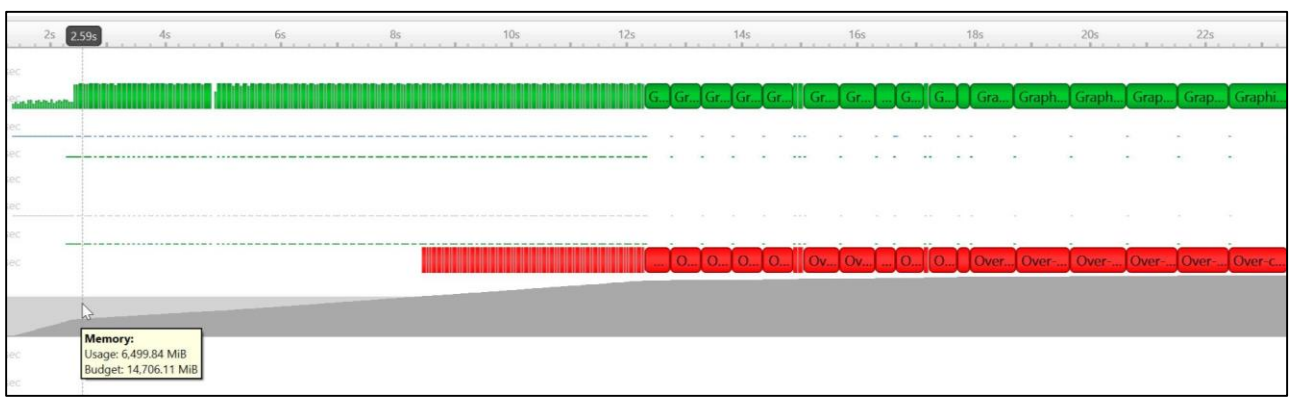
GPU Active%

< 95%

Use
Nsight Systems

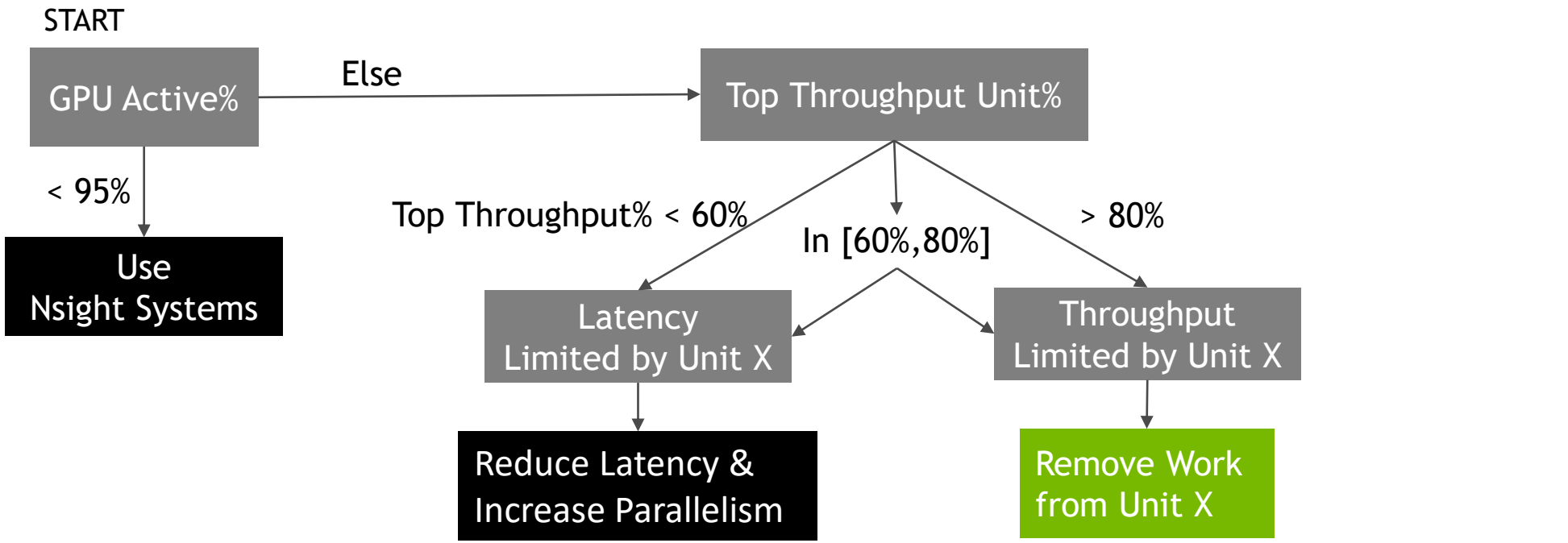


PSO Creation Stalls
on Critical Path



Video-Memory
Overcommitment

THE P3 (PEAK-PERF%) METHOD



Increase cache hit rates
Increase texture parallelism
Use async queues
…

If X==FMA: remove FMA-pipe ops

DEMO

dow Help

Terminate

DenoiseAO_Before.nsight-gfxgpt

Color By: Stages | Ruler Relative: Capture | Export... | Reset Zoom | Trace Compare...

Time | 26.3759ms | 27ms | 27.5ms | 28.0829ms | 28.5ms | 29ms | 29.5ms | 30ms | 30.5ms | 31ms | 31.5ms | 32ms

Frames | Frame 0 (39.82ms)

Q:0

Synchronization

All Actions

Markers | Frame 531 (37.30ms) | Scene (37.29ms) | RayTracedAO (6.05ms) | DenoiseAO (5.99ms) | Reconstruction (5.99ms) | Sa | BilateralFilter 1920x1080 (1.42ms) | BilateralFilter 1920x1080 (1.55ms) | BilateralFilter 1920x

Unit Throughputs

GPU Active

SM Occupancy (TPC View)

Unit Throughputs:
SM_FMA: 70%
L1TEX: 57%
L2: 5%

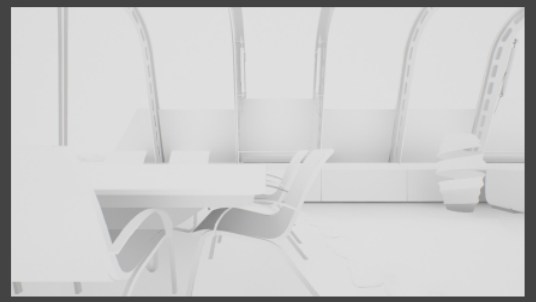
Active SM Unused Warp Slots	63.80%
Pixel Warps	25.45%
Idle SM Unused Warp Slots	10.75%
Vertex/Tess/Geometry Warps	0.00%
Compute Warps	0.00%

Summary | Metrics | Detailed Metrics | Capture Information

Start: 26.38 ms **Duration:** 5.99 ms
End: 32.36 ms **Range:** All Visible

Unit Throughput	Value
SM Instruction Throughput	70.95%
SM FMA Pipe Throughput	69.98%
SM Issue Active	60.59%
L1TEX Throughput	56.57%
SM SFU Pipe Throughput	52.02%
SM ALU Pipe Throughput	25.13%
L2 Throughput	4.56%
CROP Throughput	2.25%
VRAM Throughput	2.20%
PROP Throughput	1.56%
RASTER Throughput	0.29%
PES+VPC Throughput	0.00%
PD Throughput	0.00%
ZROP Throughput	0.00%
SM FP16+Tensor Pipe Throughput	0.00%

Warp Occupancy	Value
Idle SM Unused Warp Slots	3.89%
Active SM Unused Warp Slots	20.25%
Compute Warps	0.00%
Pixel Warps	75.86%
Vertex+Tess+Geom Warps	0.00%



NVIDIA Nsight Graphics

File Connection Tools Window Help

Connect Disconnect Terminate

DenoiseAO_Before.nsight-gfxgpt

Color By: Stages Ruler Relative: Capture Export... Reset Zoom Trace Compare...

Time 26.3759ms 27ms 27.5ms 28.0829ms 28.5ms 29ms 29.5ms 30ms 30.5ms 31ms 31.5ms 32ms

Frames Frame 0 (39.82ms)

Q:0

Synchronization

All Actions

Frame 531 (37.30ms)
Scene (37.29ms)
RayTracedAO (6.05ms)
DenoiseAO (5.99ms)
Reconstruction (5.99ms)

Markers Sa BilateralFilter 1920x1080 (1.42ms) BilateralFilter 1920x1080 (1.55ms) BilateralFilter 1920x1080 (1.36ms) BilateralFilter 1920x1080 (1.49ms)

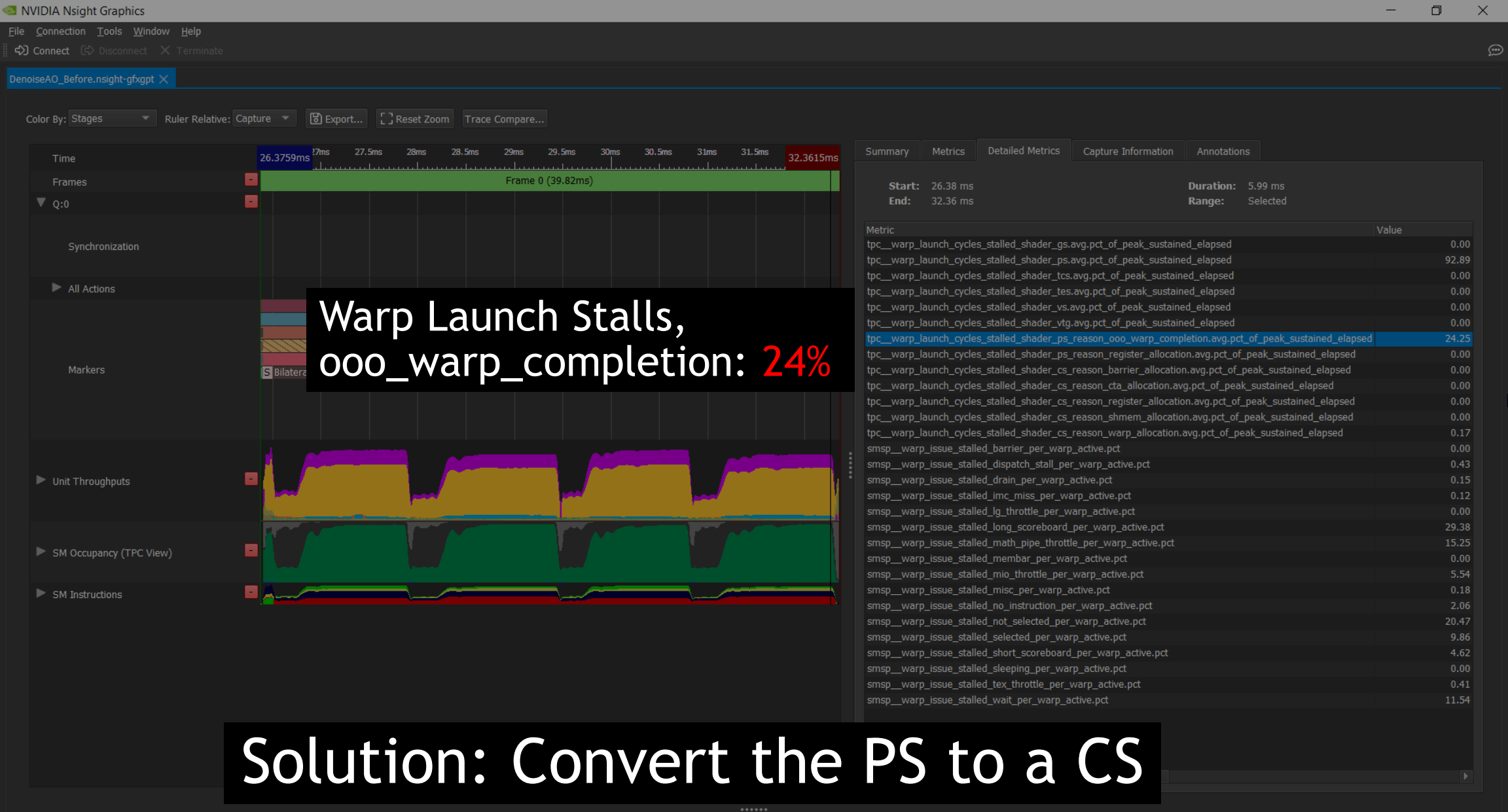
Summary Metrics Detailed Metrics Capture Information

Start: 26.38 ms **Duration:** 5.99 ms
End: 32.36 ms **Range:** All Visible

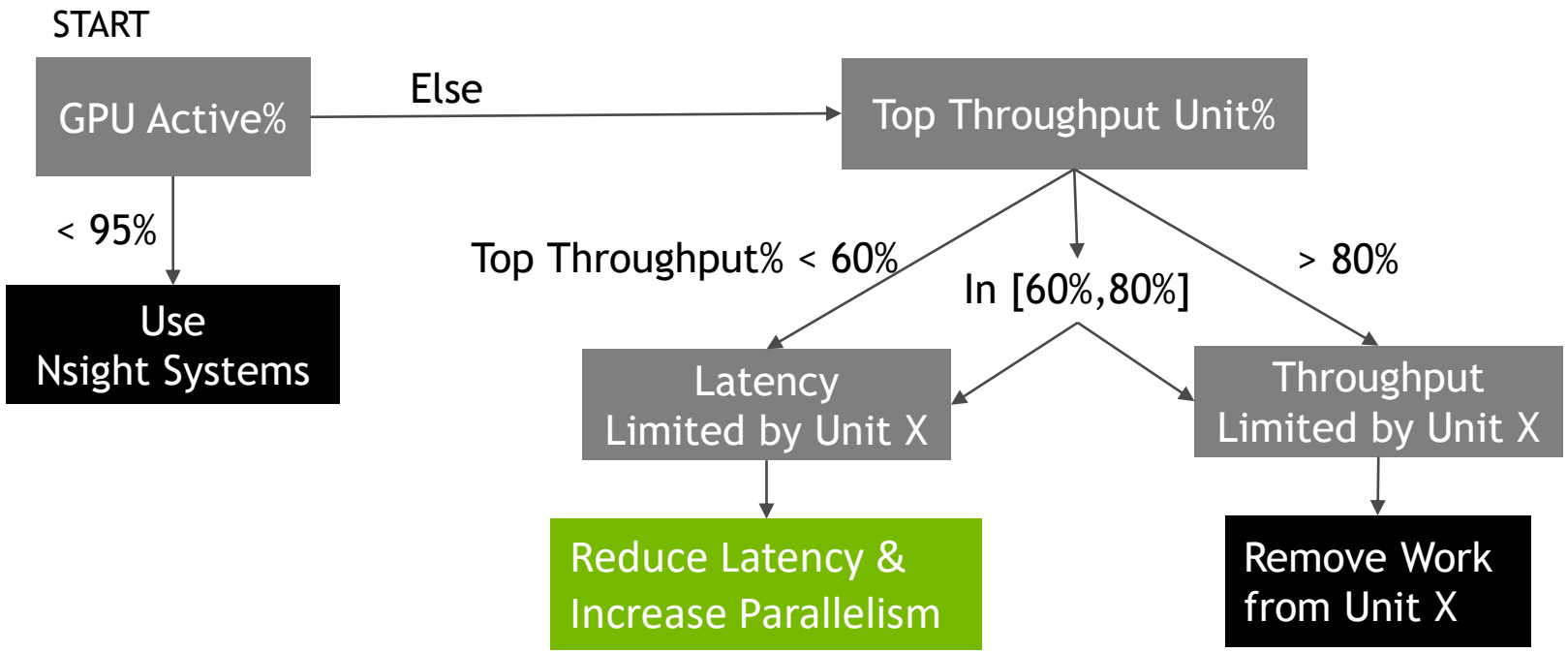
Unit Throughput	Value
SM Instruction Throughput	70.95%
SM FMA Pipe Throughput	69.98%
SM Issue Active	60.59%
L1TEX Throughput	56.57%
SM SFU Pipe Throughput	52.02%
SM ALU Pipe Throughput	25.13%
L2 Throughput	4.56%
CROP Throughput	2.25%
VRAM Throughput	2.20%
PROP Throughput	1.56%
RASTER Throughput	0.29%

SM Occupancy

Occupancy% = ActiveWarpsPerCycle / NumWarpSlots * 100


NVIDIA Nsight Graphics
File Connection Tools Window Help
Connect Disconnect Terminate
DenoiseAO_Before.nsight-gfxgpt
Color By: Stages
Ruler Relative: Capture
Export...
Reset Zoom
Trace Compare...
Time
Frames
Q:0
Synchronization
All Actions
Markers
Unit Throughputs
SM Occupancy (TPC View)
SM Instructions
26.3759ms
27.5ms
28ms
28.5ms
29ms
29.5ms
30ms
30.5ms
31ms
31.5ms
32.3615ms
Frame 0 (39.82ms)
Warp Launch Stalls,
ooo_warp_completion: 24%
Summary
Metrics
Detailed Metrics
Capture Information
Annotations
Start: 26.38 ms
End: 32.36 ms
Duration: 5.99 ms
Range: Selected
Metric
Value
tpc__warp_launch_cycles_stalled_shader_gs.avg.pct_of_peak_sustained_elapsed 0.00
tpc__warp_launch_cycles_stalled_shader_ps.avg.pct_of_peak_sustained_elapsed 92.89
tpc__warp_launch_cycles_stalled_shader_tcs.avg.pct_of_peak_sustained_elapsed 0.00
tpc__warp_launch_cycles_stalled_shader_tes.avg.pct_of_peak_sustained_elapsed 0.00
tpc__warp_launch_cycles_stalled_shader_vs.avg.pct_of_peak_sustained_elapsed 0.00
tpc__warp_launch_cycles_stalled_shader_vtg.avg.pct_of_peak_sustained_elapsed 0.00
tpc__warp_launch_cycles_stalled_shader_ps_reason_ooo_warp_completion.avg.pct_of_peak_sustained_elapsed 24.25
tpc__warp_launch_cycles_stalled_shader_ps_reason_register_allocation.avg.pct_of_peak_sustained_elapsed 0.00
tpc__warp_launch_cycles_stalled_shader_cs_reason_barrier_allocation.avg.pct_of_peak_sustained_elapsed 0.00
tpc__warp_launch_cycles_stalled_shader_cs_reason_cta_allocation.avg.pct_of_peak_sustained_elapsed 0.00
tpc__warp_launch_cycles_stalled_shader_cs_reason_register_allocation.avg.pct_of_peak_sustained_elapsed 0.00
tpc__warp_launch_cycles_stalled_shader_cs_reason_shmem_allocation.avg.pct_of_peak_sustained_elapsed 0.00
tpc__warp_launch_cycles_stalled_shader_cs_reason_warp_allocation.avg.pct_of_peak_sustained_elapsed 0.17
smsp__warp_issue_stalled_barrier_per_warp_active.pct 0.00
smsp__warp_issue_stalled_dispatch_stall_per_warp_active.pct 0.43
smsp__warp_issue_stalled_drain_per_warp_active.pct 0.15
smsp__warp_issue_stalled_imc_miss_per_warp_active.pct 0.12
smsp__warp_issue_stalled_lg_throttle_per_warp_active.pct 0.00
smsp__warp_issue_stalled_long_scoreboard_per_warp_active.pct 29.38
smsp__warp_issue_stalled_math_pipe_throttle_per_warp_active.pct 15.25
smsp__warp_issue_stalled_membar_per_warp_active.pct 0.00
smsp__warp_issue_stalled_mio_throttle_per_warp_active.pct 5.54
smsp__warp_issue_stalled_misc_per_warp_active.pct 0.18
smsp__warp_issue_stalled_no_instruction_per_warp_active.pct 2.06
smsp__warp_issue_stalled_not_selected_per_warp_active.pct 20.47
smsp__warp_issue_stalled_selected_per_warp_active.pct 9.86
smsp__warp_issue_stalled_short_scoreboard_per_warp_active.pct 4.62
smsp__warp_issue_stalled_sleeping_per_warp_active.pct 0.00
smsp__warp_issue_stalled_tex_throttle_per_warp_active.pct 0.41
smsp__warp_issue_stalled_wait_per_warp_active.pct 11.54
Solution: Convert the PS to a CS

THE P3 (PEAK-PERF%) METHOD

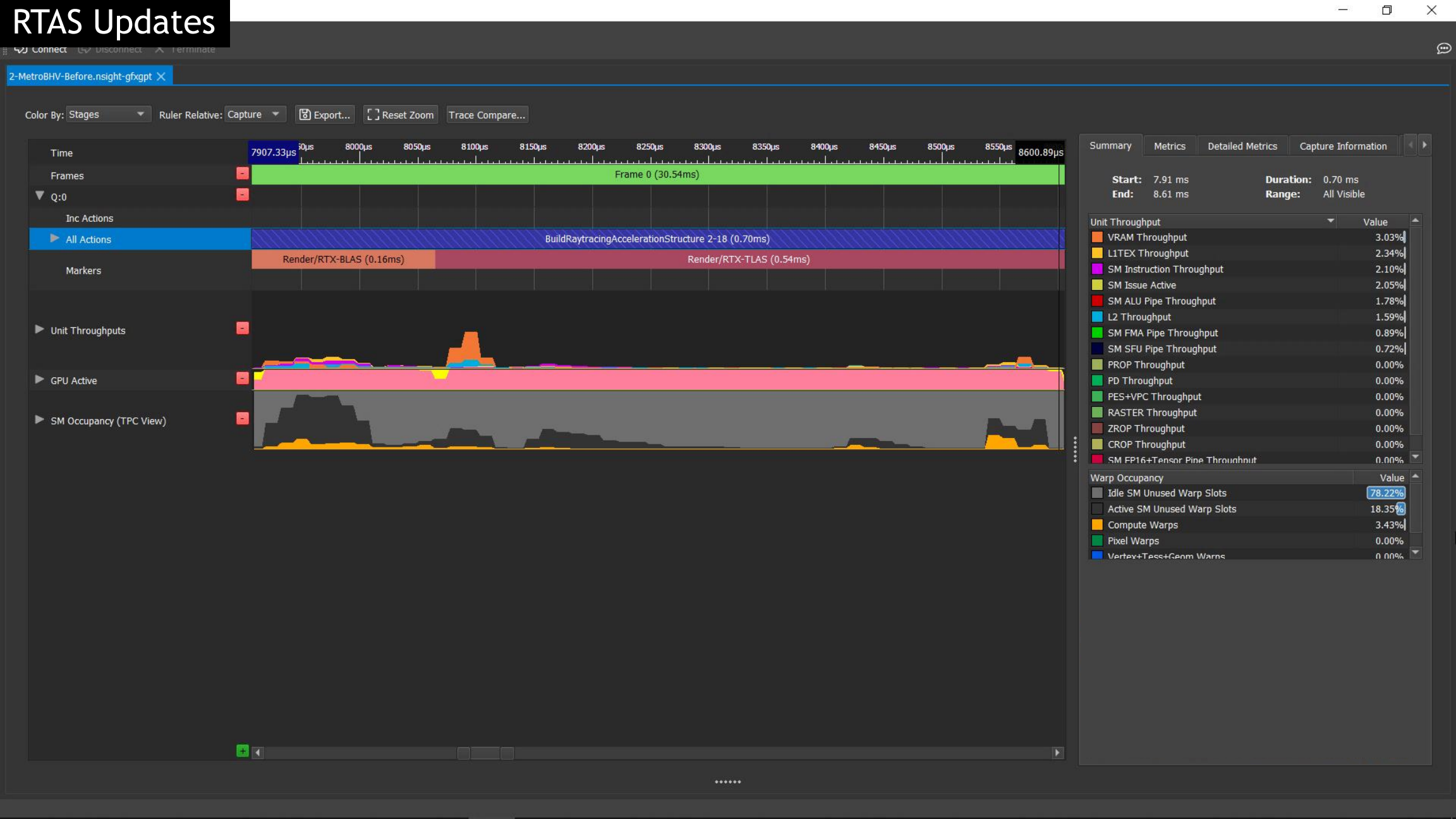


Increase cache hit rates
Increase texture parallelism
Use async queues
…

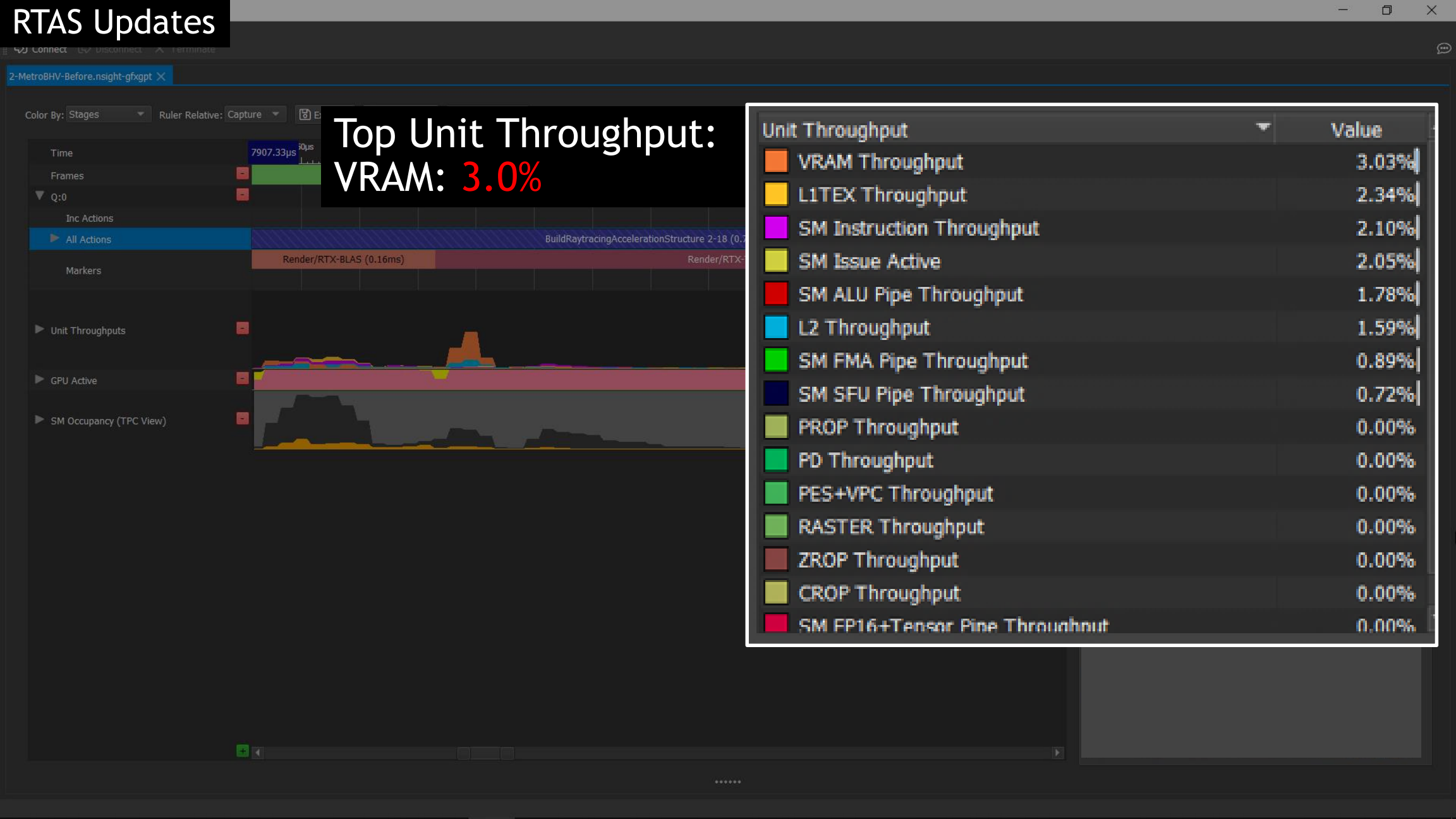
If X==FMA: remove FMA-pipe ops

Case Study #1:
RTAS Updates // Async Compute in Metro Exodus

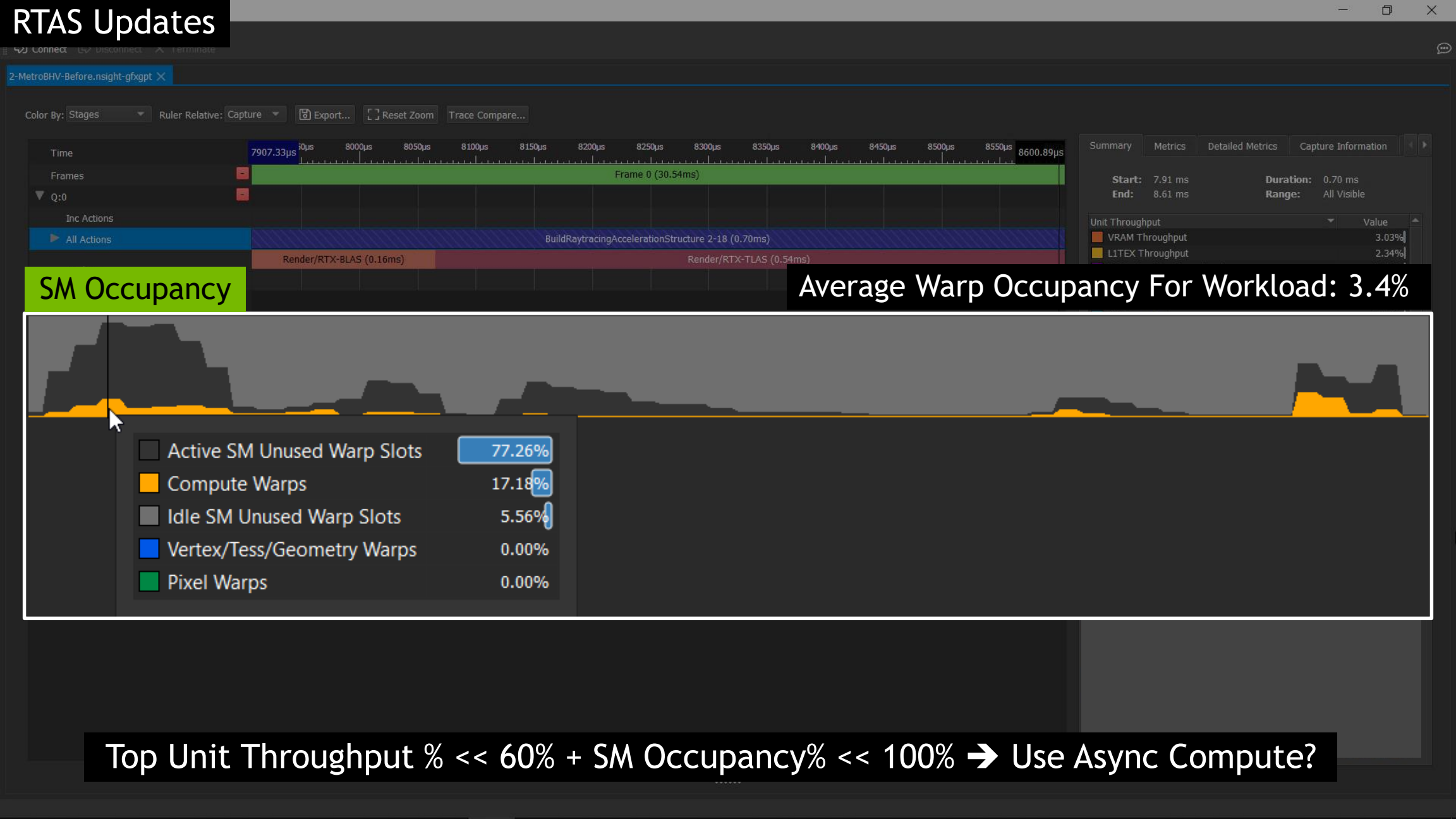
RTAS Updates



RTAS Updates



RTAS Updates

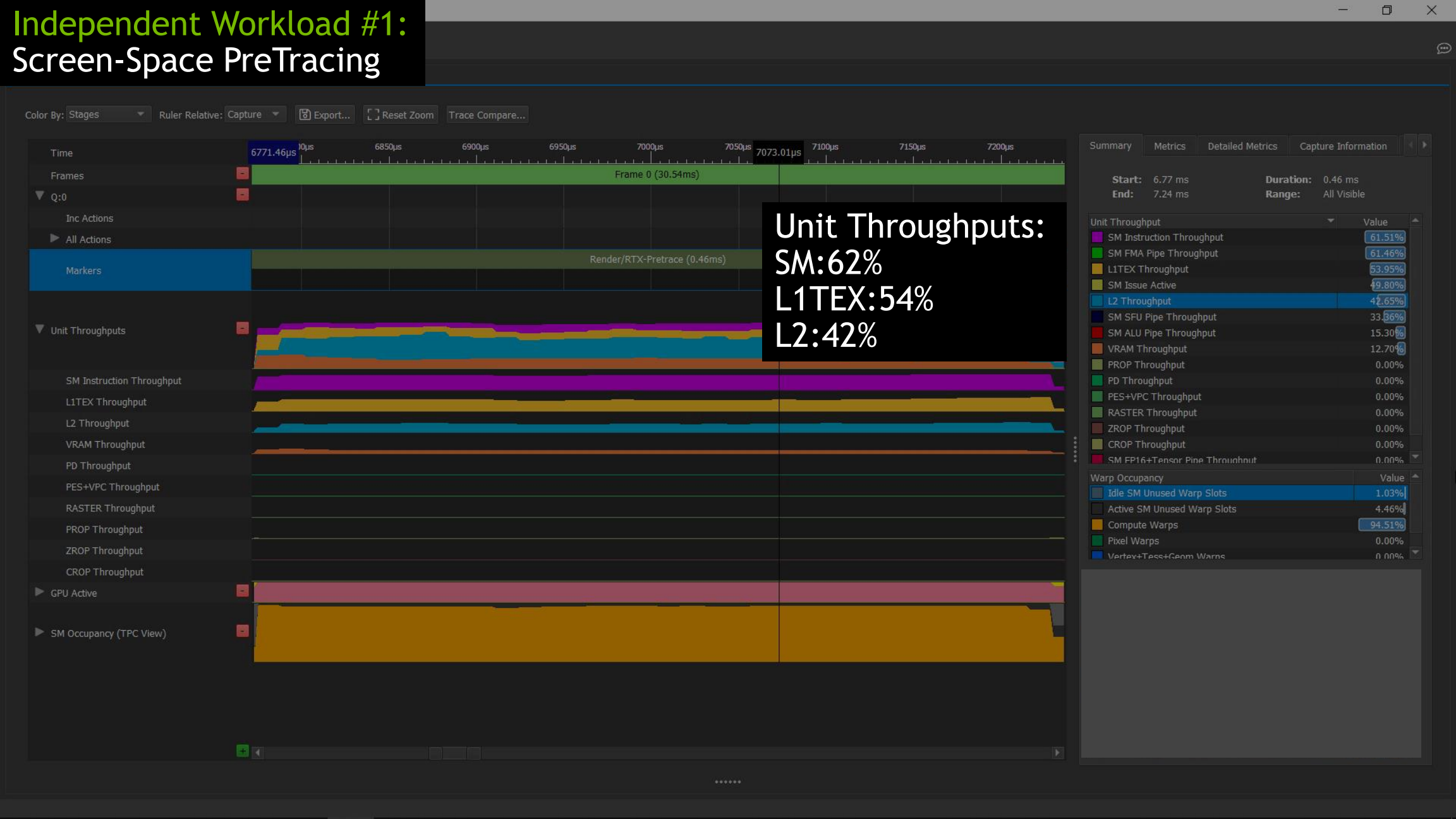


SM Occupancy

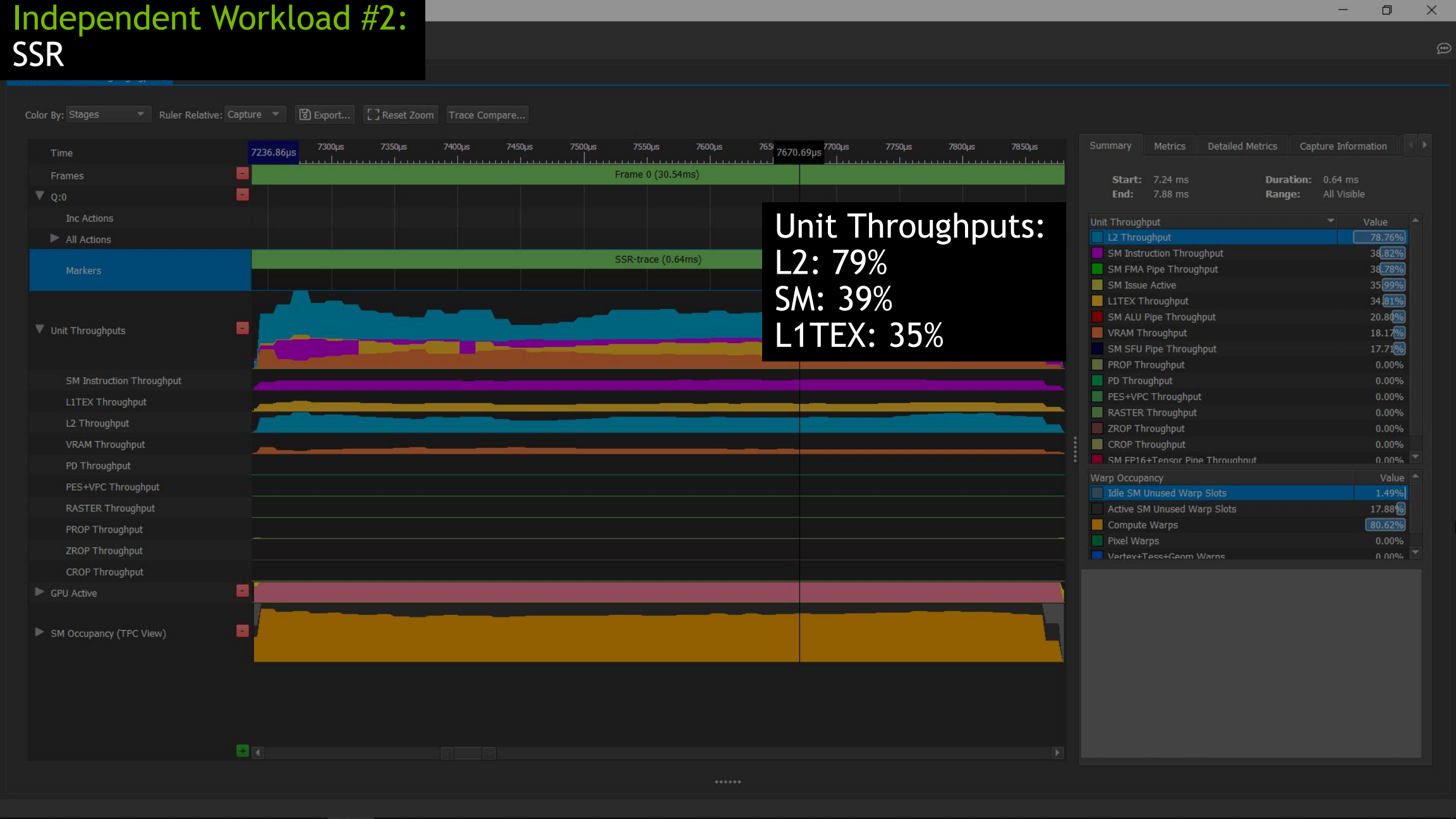
Average Warp Occupancy For Workload: 3.4%

Top Unit Throughput % << 60% + SM Occupancy% << 100% ➔ Use Async Compute?

Independent Workload #1:
Screen-Space PreTracing



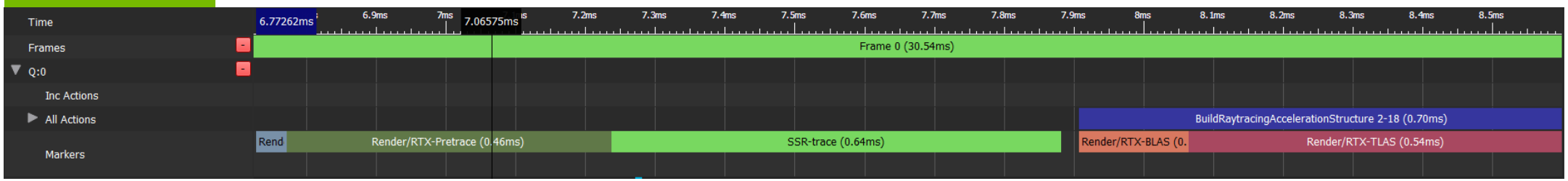
Independent Workload #2:
SSR



ASYNC COMPUTE DIFF

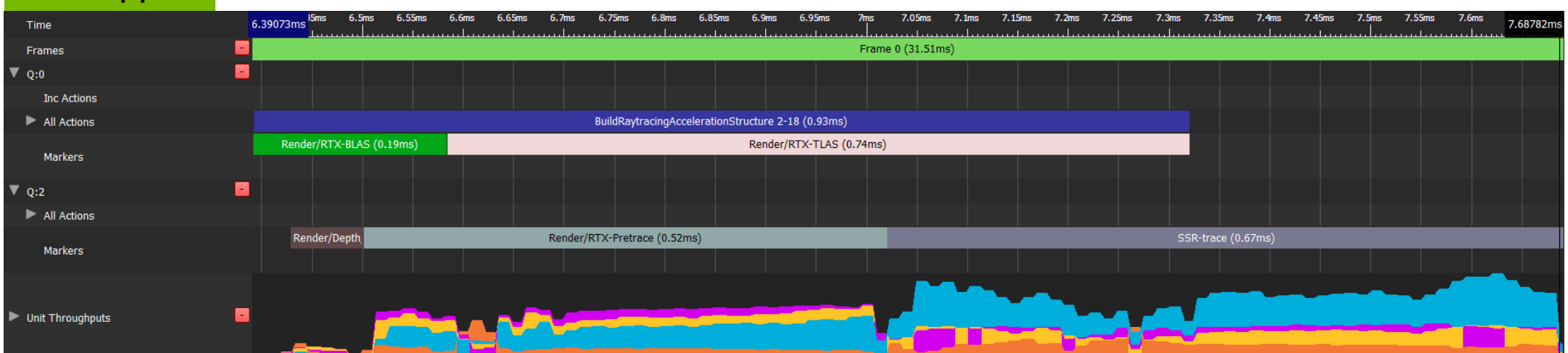
Serialized

1.83 ms



Overlapped

1.30 ms



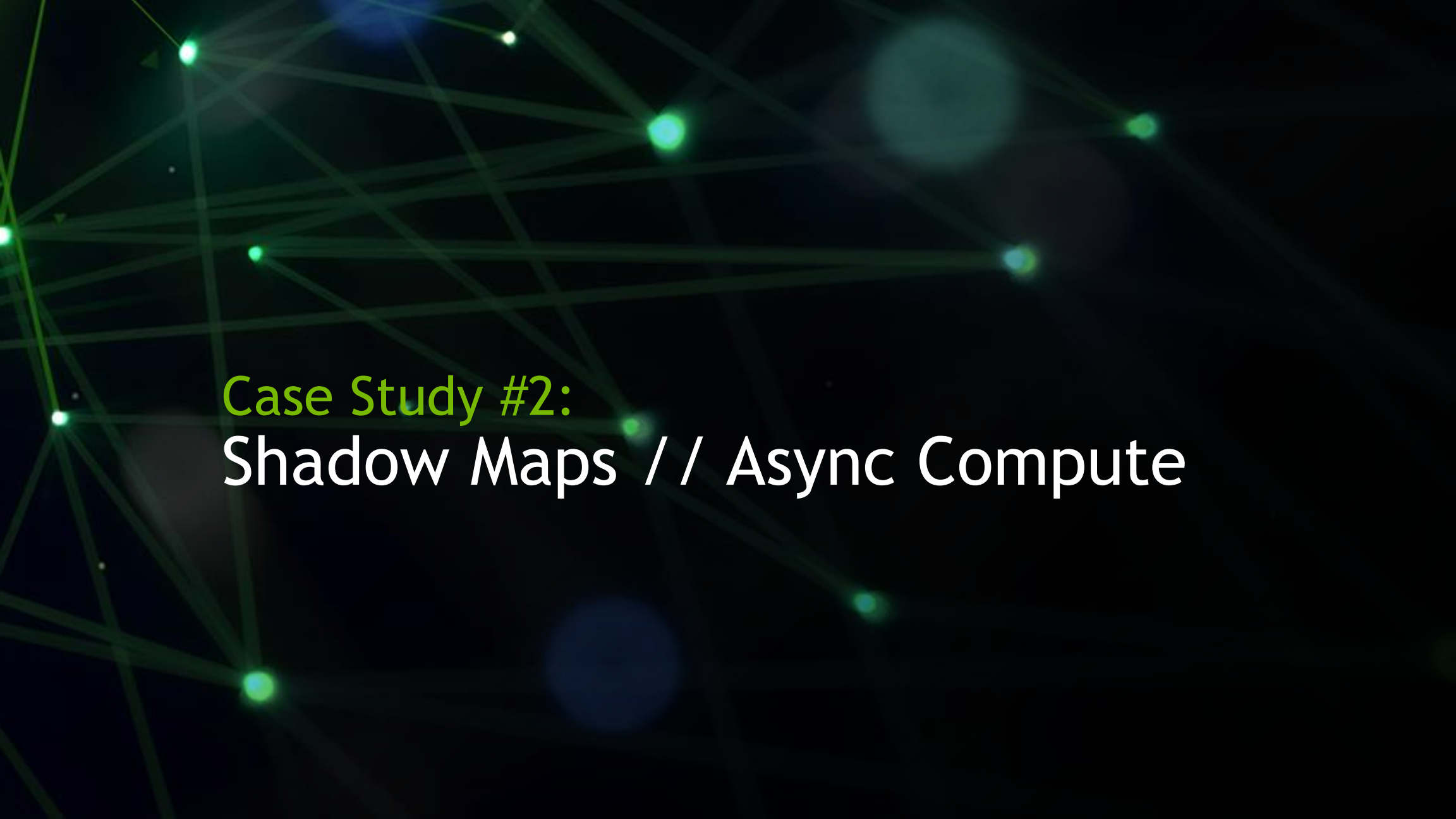
NVIDIA

ASYNC-COMPUTE OVERLAP

(RTAS Updates) // (Async Compute)

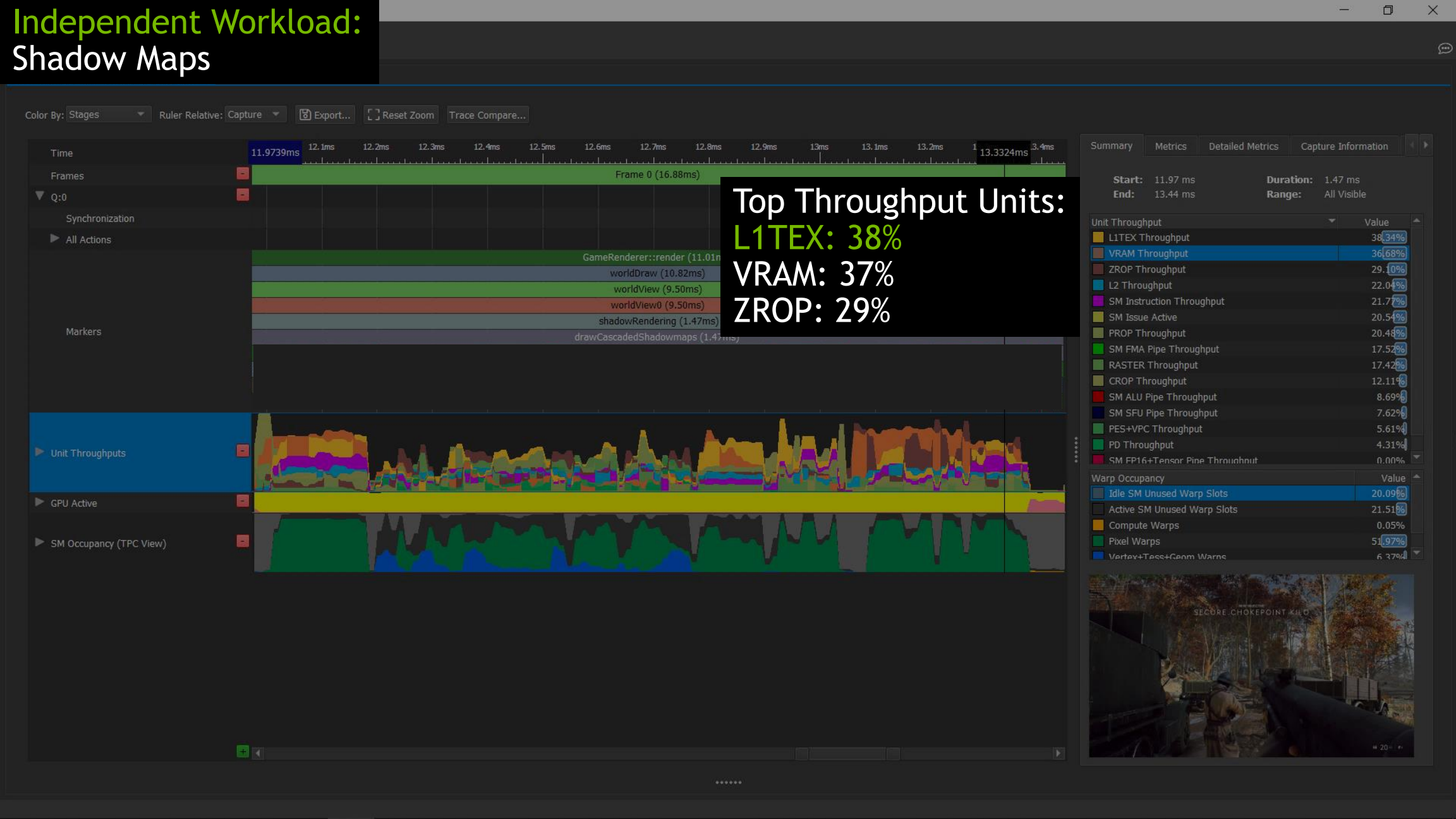
	BEFORE	AFTER	RATIO
GPU Elapsed Time	1.83 ms	1.30 ms	1.41x Gain
Throughput: L2	39.2%	54.8%	1.40x
Throughput: SM	30.1%	42.0%	1.40x
Throughput: L1TEX	26.9%	37.4%	1.39x
SM Occupancy	53.8%	78.2%	1.45x

On RTX 2080 with SetStablePowerState(TRUE)


Case Study #2:
Shadow Maps // Async Compute

HBAO + SSR + Light Culling

Independent Workload:
Shadow Maps

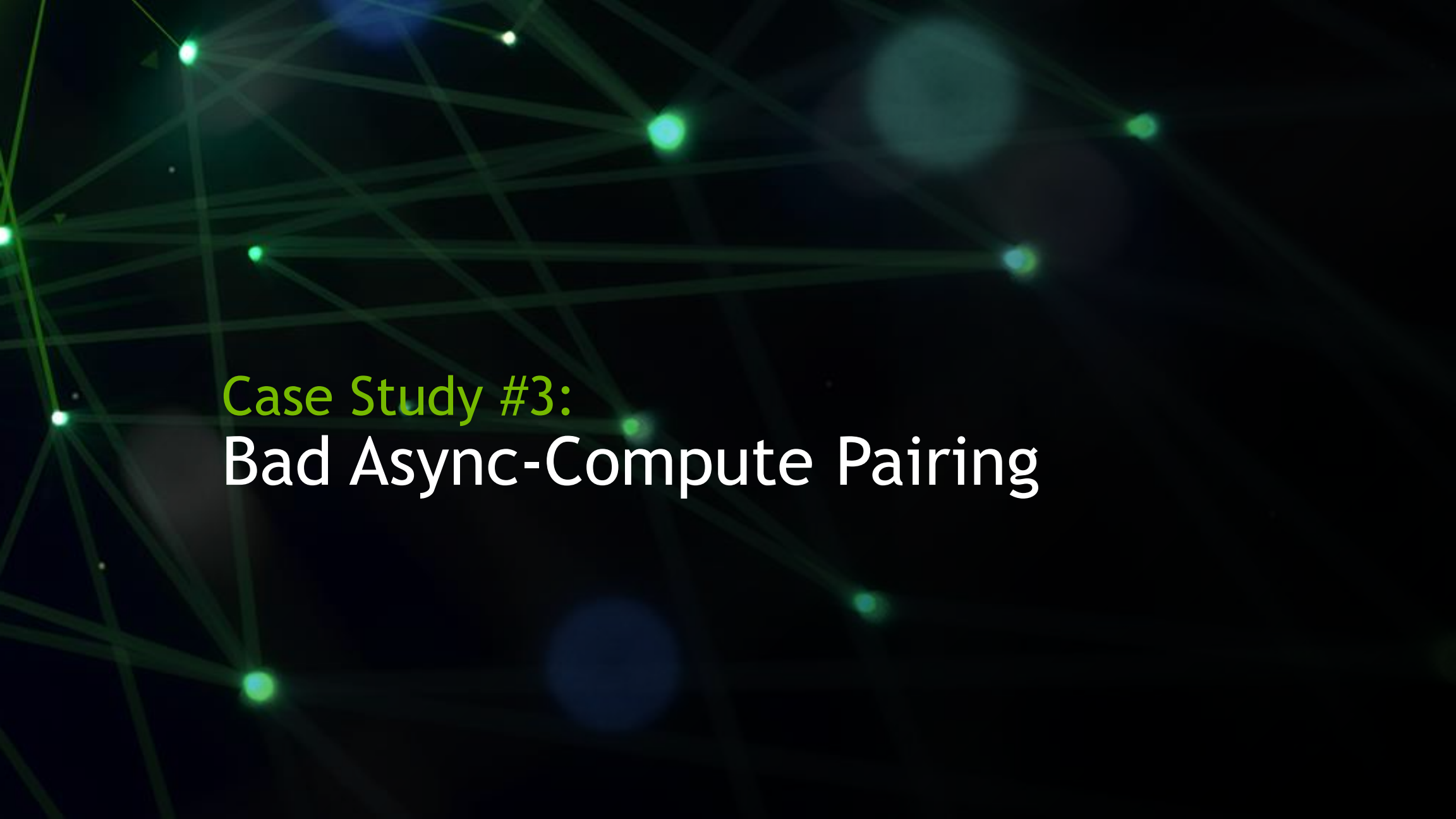


ASYNC-COMPUTE OVERLAP

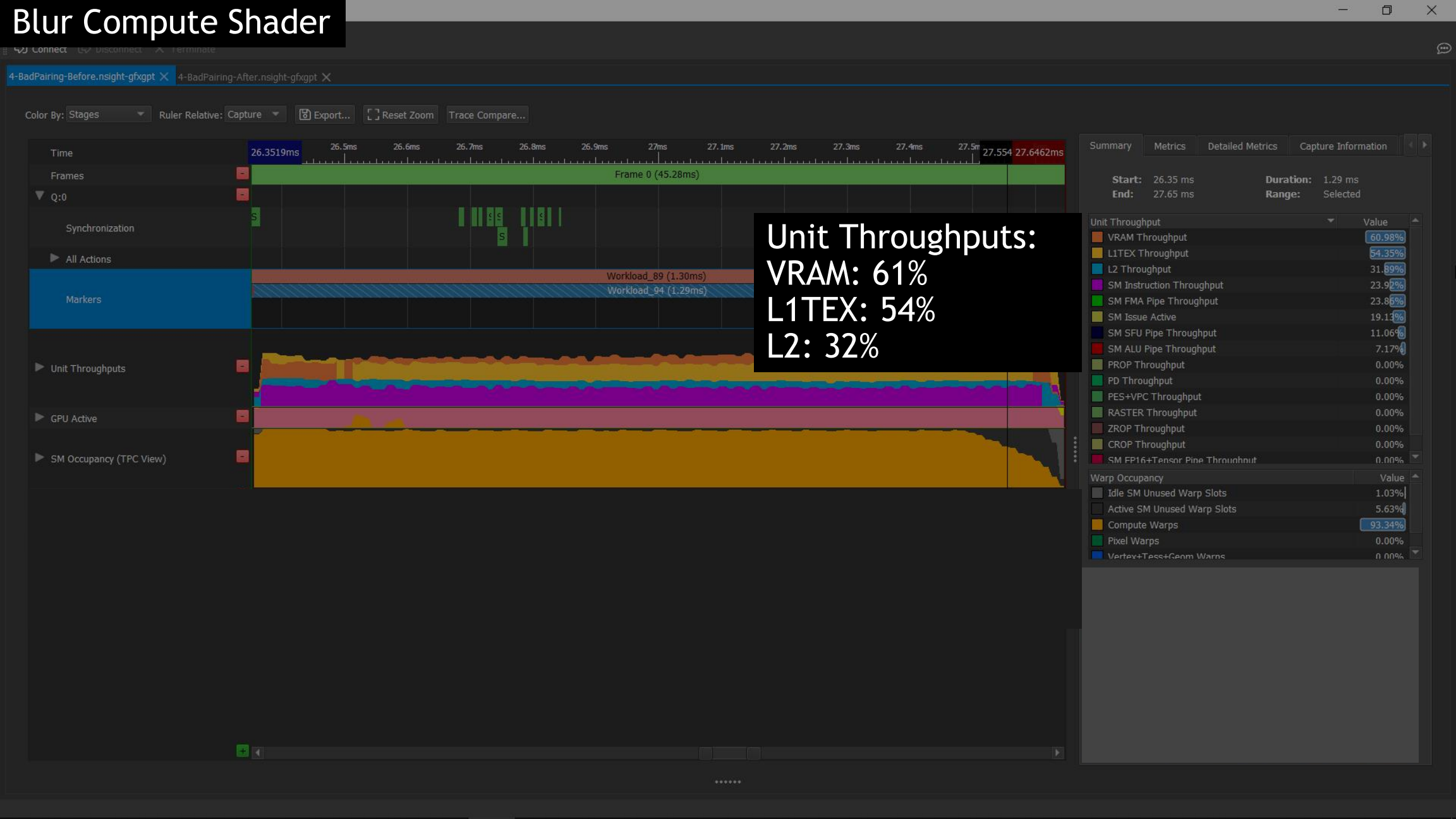
(SHADOW MAPS) // (HBAO+SSR+LIGHT-CULL)

	BEFORE	AFTER	RATIO
GPU Elapsed Time	2.45 ms	2.15 ms	1.14x Gain
Throughput: VRAM	34.2%	40.8%	1.19x
Throughput: L1TEX	31.0%	34.0%	1.10x
Throughput: SM	22.1%	24.3%	1.10x
SM Occupancy	59.5%	70.6%	1.19x

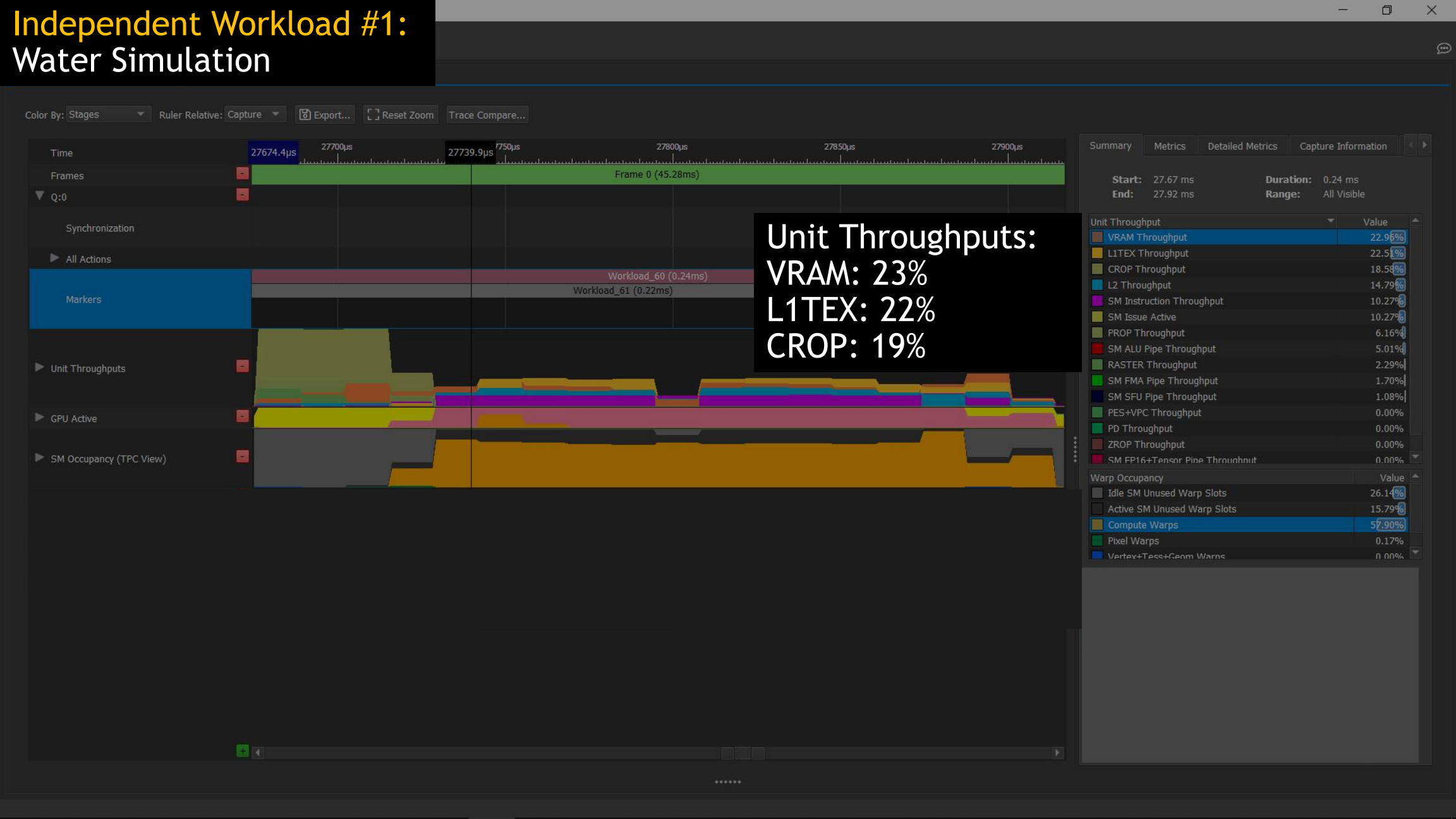
On RTX 2080 + SetStablePowerState(TRUE)


Case Study #3:
Bad Async-Compute Pairing

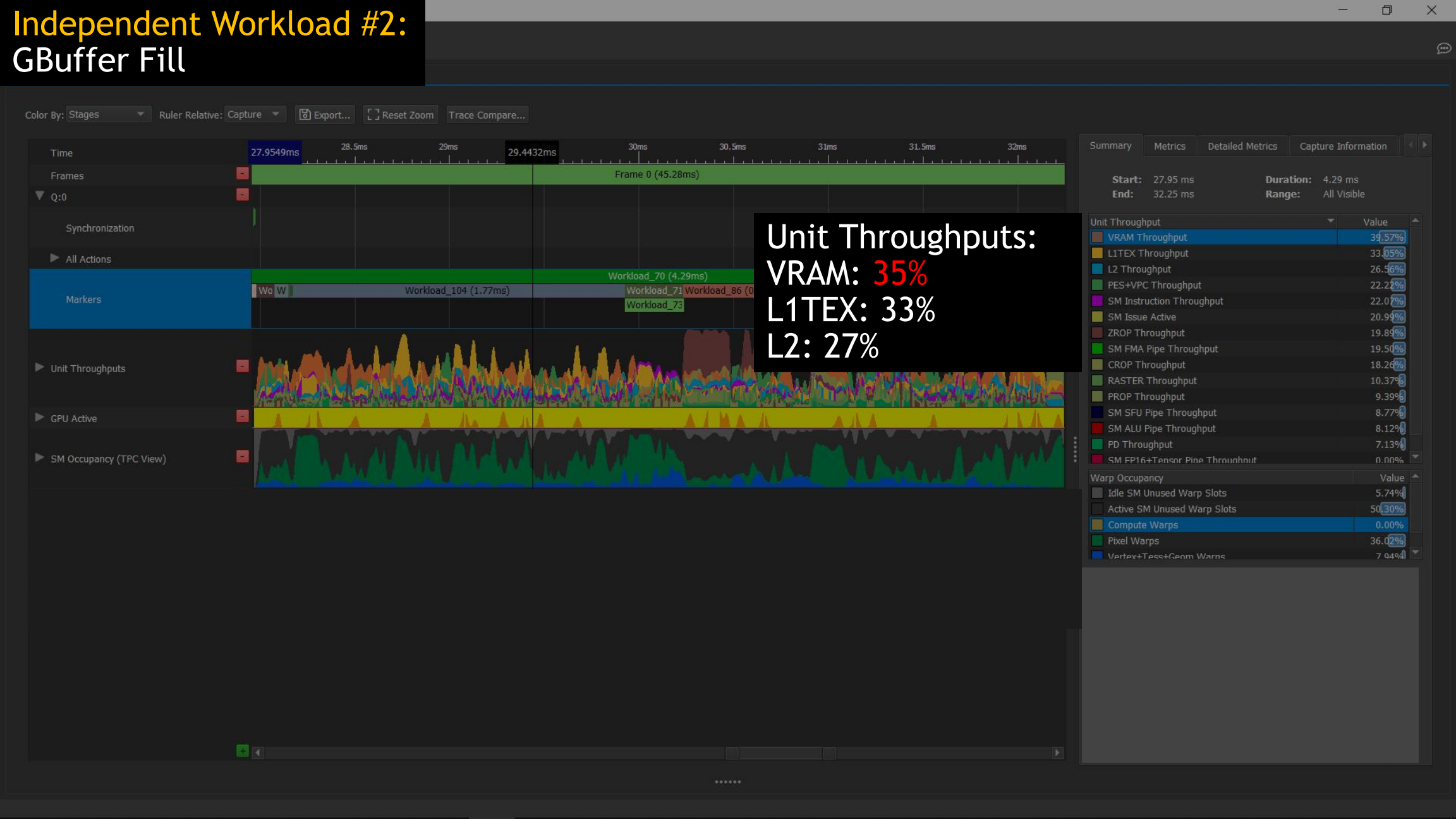
Blur Compute Shader



Independent Workload #1:
Water Simulation



Independent Workload #2:
GBuffer Fill



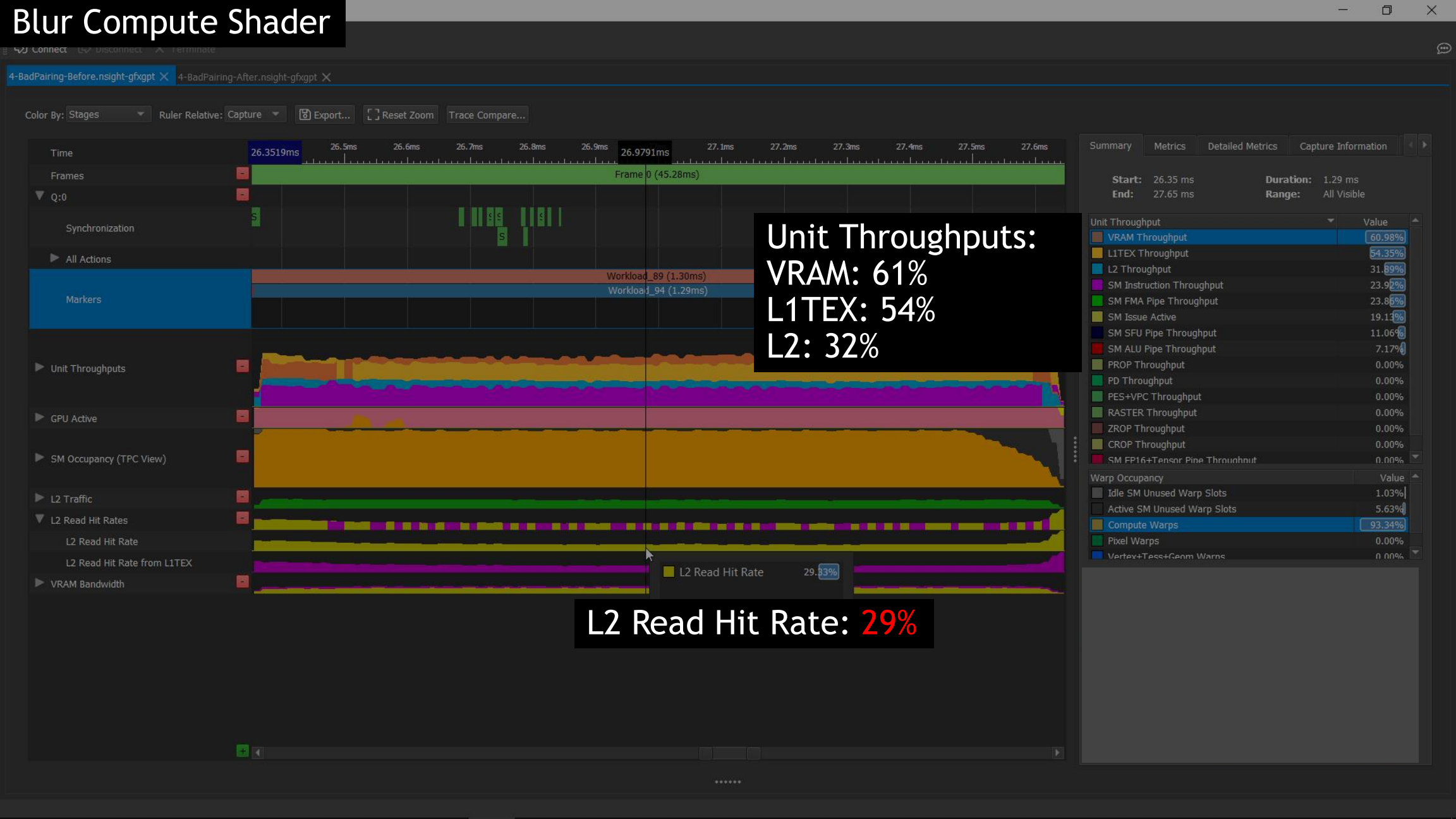
BAD ASYNC-COMPUTE PAIRING

(BLUR CS) // (GBUFFER + WATER SIM)

	BEFORE	AFTER	RATIO
GPU Elapsed Time	5.89 ms	6.12 ms	0.96x Loss
Throughput: VRAM	43.1%	47.1%	1.09x
Throughput: L1TEX	36.9%	35.4%	0.96x
Throughput: L2	27.0%	26.0%	0.96x
SM Occupancy	54.9%	57.5%	1.05x
L2 Read Hit Rate	52.3%	44.5%	0.85x

On RTX 2080 with SetStablePowerState(TRUE)

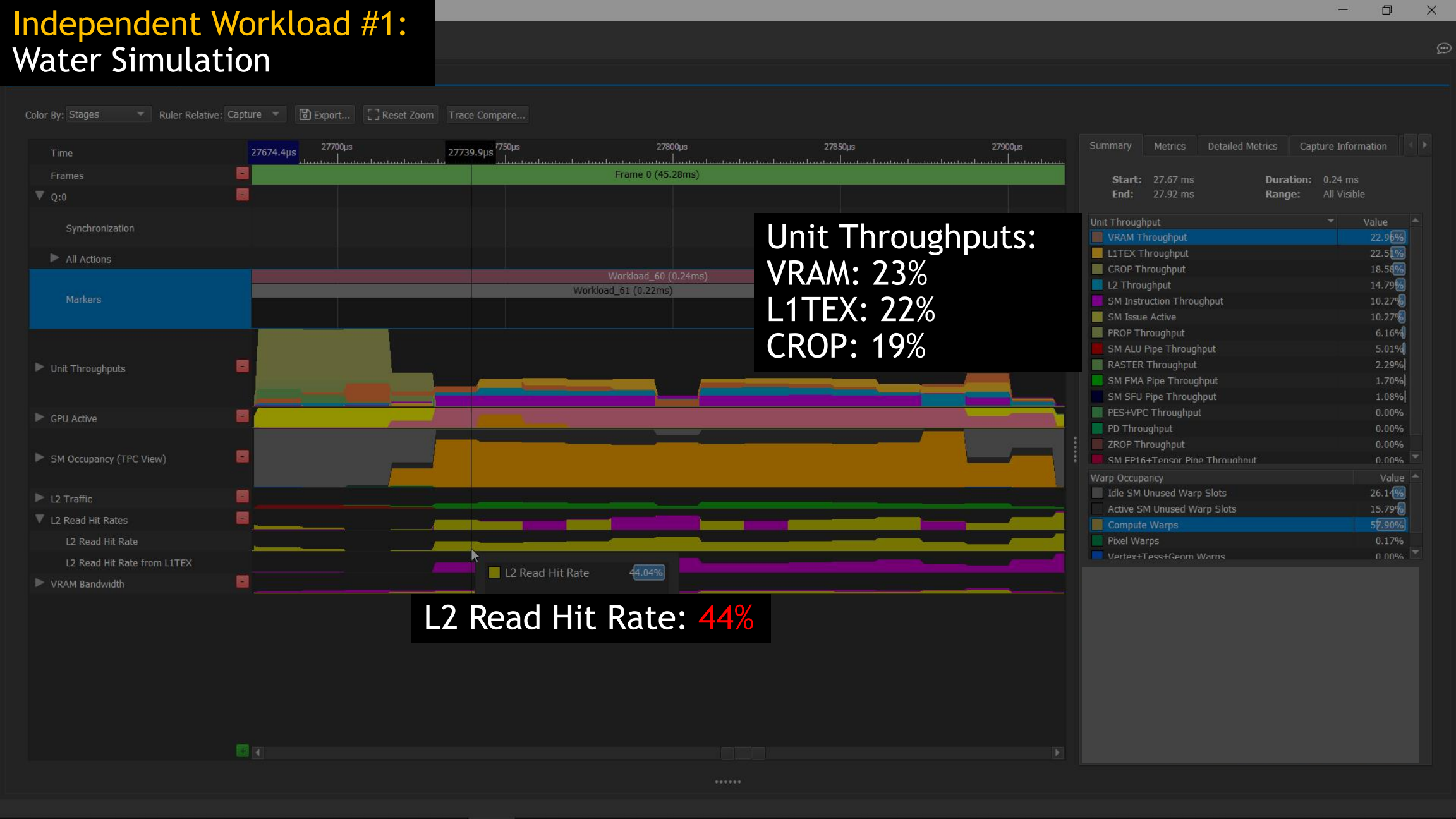
Blur Compute Shader



Unit Throughputs:
VRAM: 61%
L1TEX: 54%
L2: 32%

L2 Read Hit Rate: 29%

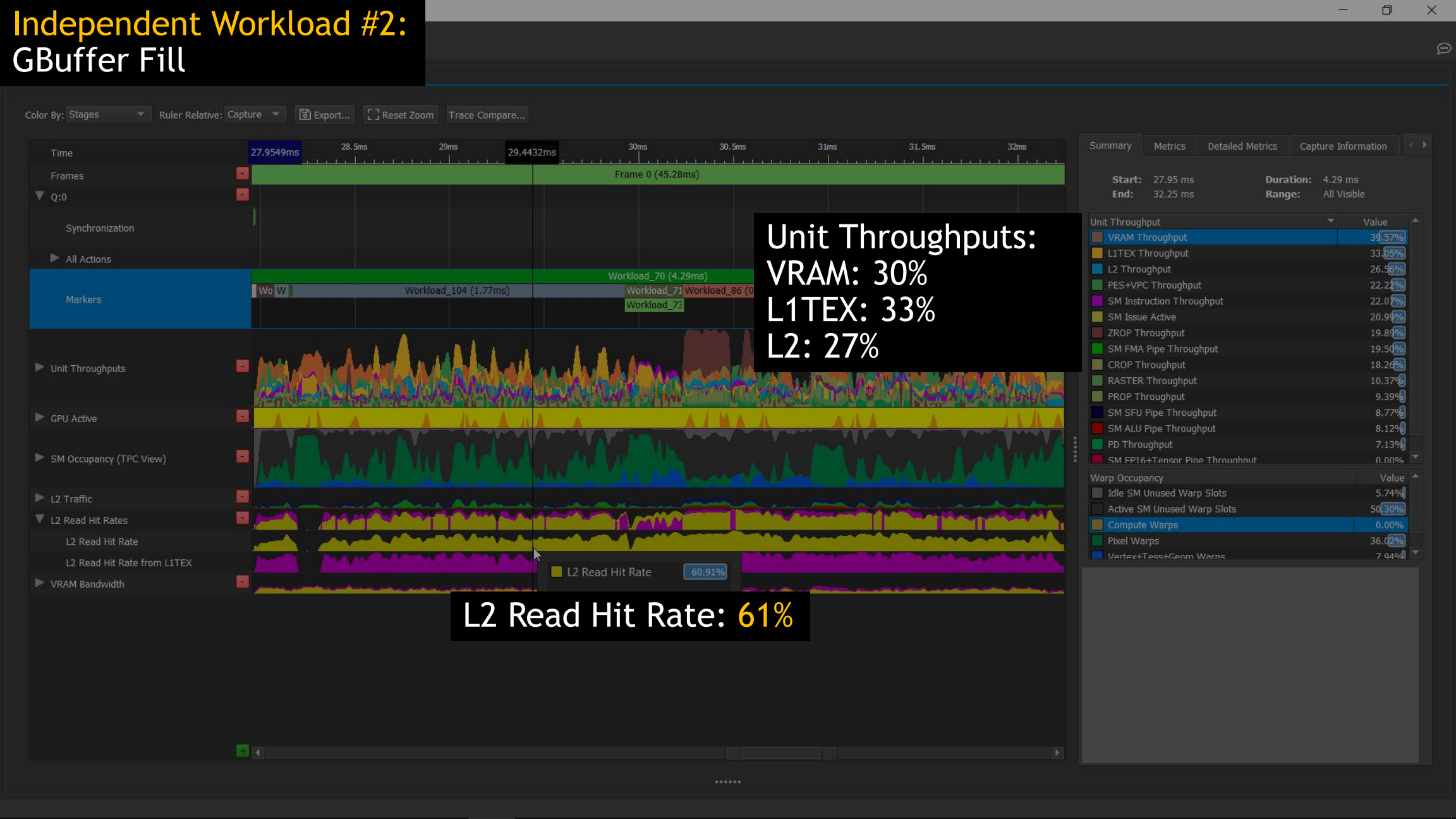
Independent Workload #1:
Water Simulation



Unit Throughputs:
VRAM: 23%
L1TEX: 22%
CROP: 19%

L2 Read Hit Rate: 44%

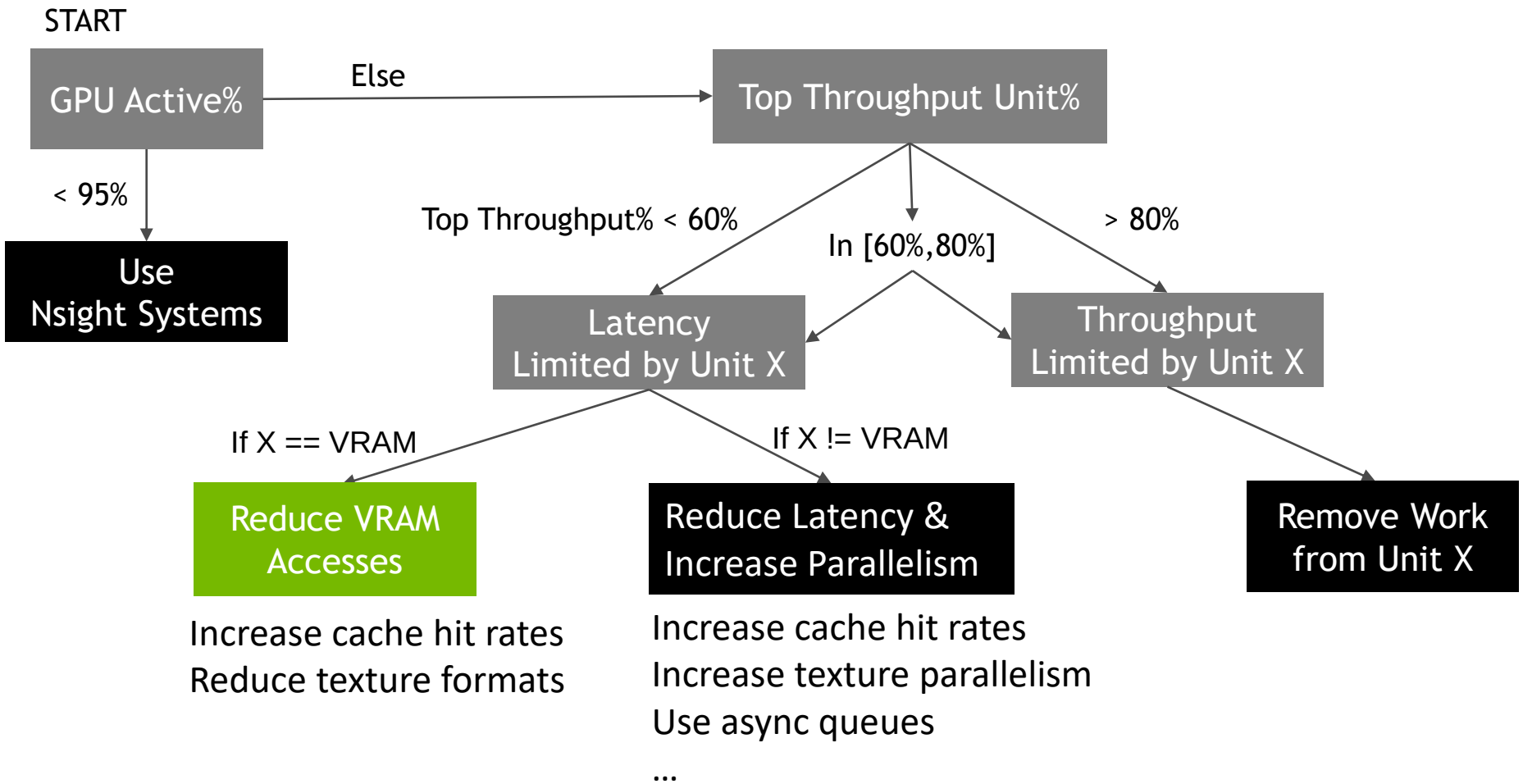
Independent Workload #2:
GBuffer Fill

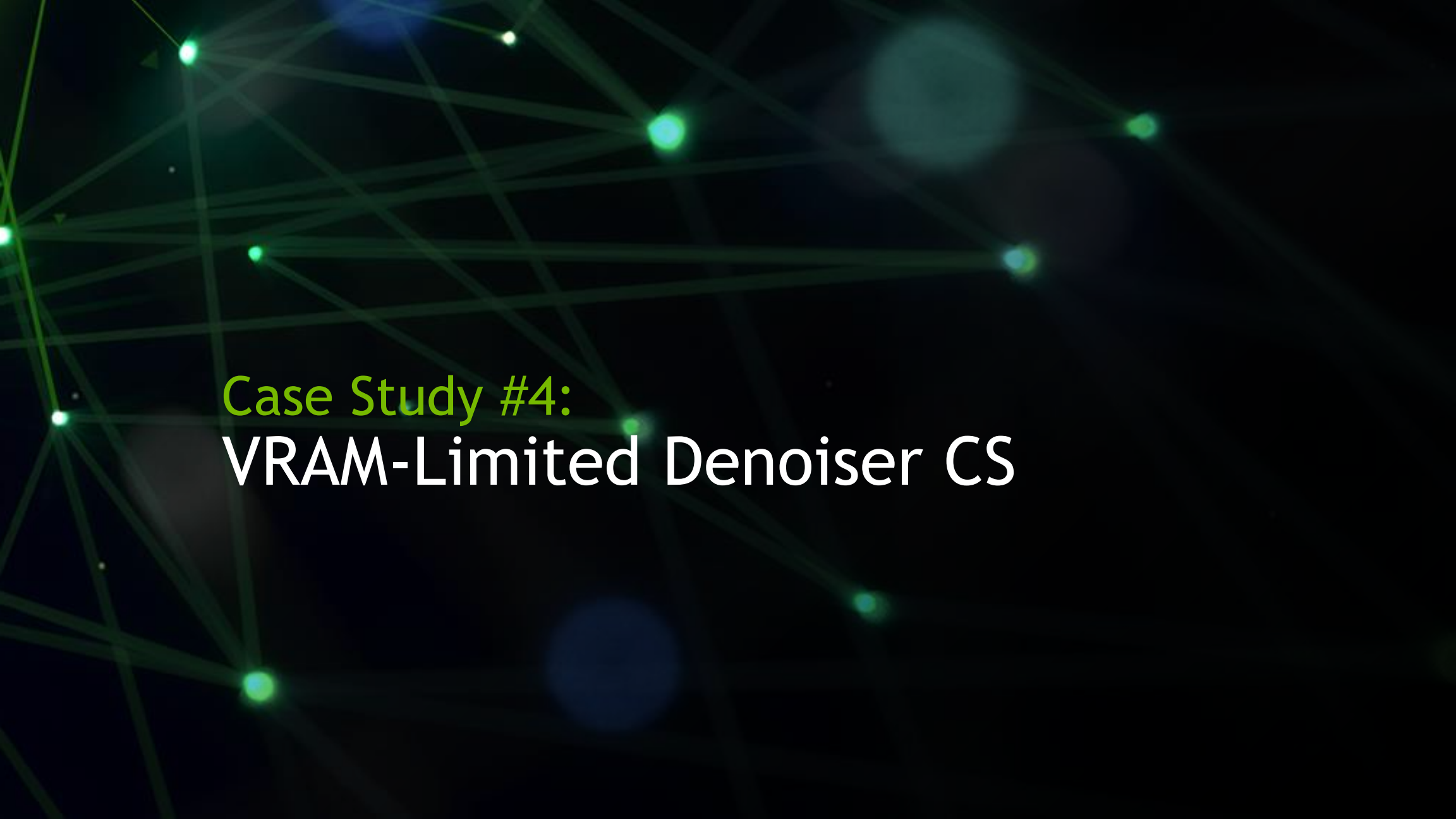


Unit Throughputs:
VRAM: 30%
L1TEX: 33%
L2: 27%

L2 Read Hit Rate: 61%

THE P3 (PEAK-PERF%) METHOD




Case Study #4:
VRAM-Limited Denoiser CS


Battlefield V, DXR, 1440p
W
149 m
SECURE CHOKEPOINT KILO
91
20 40
3

NVIDIA Nsight Graphics

File Connection Tools Window Help

Connect Disconnect Terminate

6-Denoising_CS-Before_2.nsight-gfxgpt

Color By: Stages Ruler Relative: Capture Export... Reset Zoom Trace Compare...

Time 21.0545ms 21.5ms 22ms 22.5ms 23ms 23.4185ms

Frames Frame 0 (27.66ms)

Q:0

Synchronization

Inc Actions

All Actions

Markers Dispatch[PSO_0_0_0_0_0_584F3225] (2.37ms)

Unit Throughputs

GPU Active

SM Occupancy (TPC View)

L2 Read Hit Rates

L2 Read Hit Rate

L2 Read Hit Rate from L1TEX

VRAM Bandwidth

VRAM Read Throughput

VRAM Write Throughput

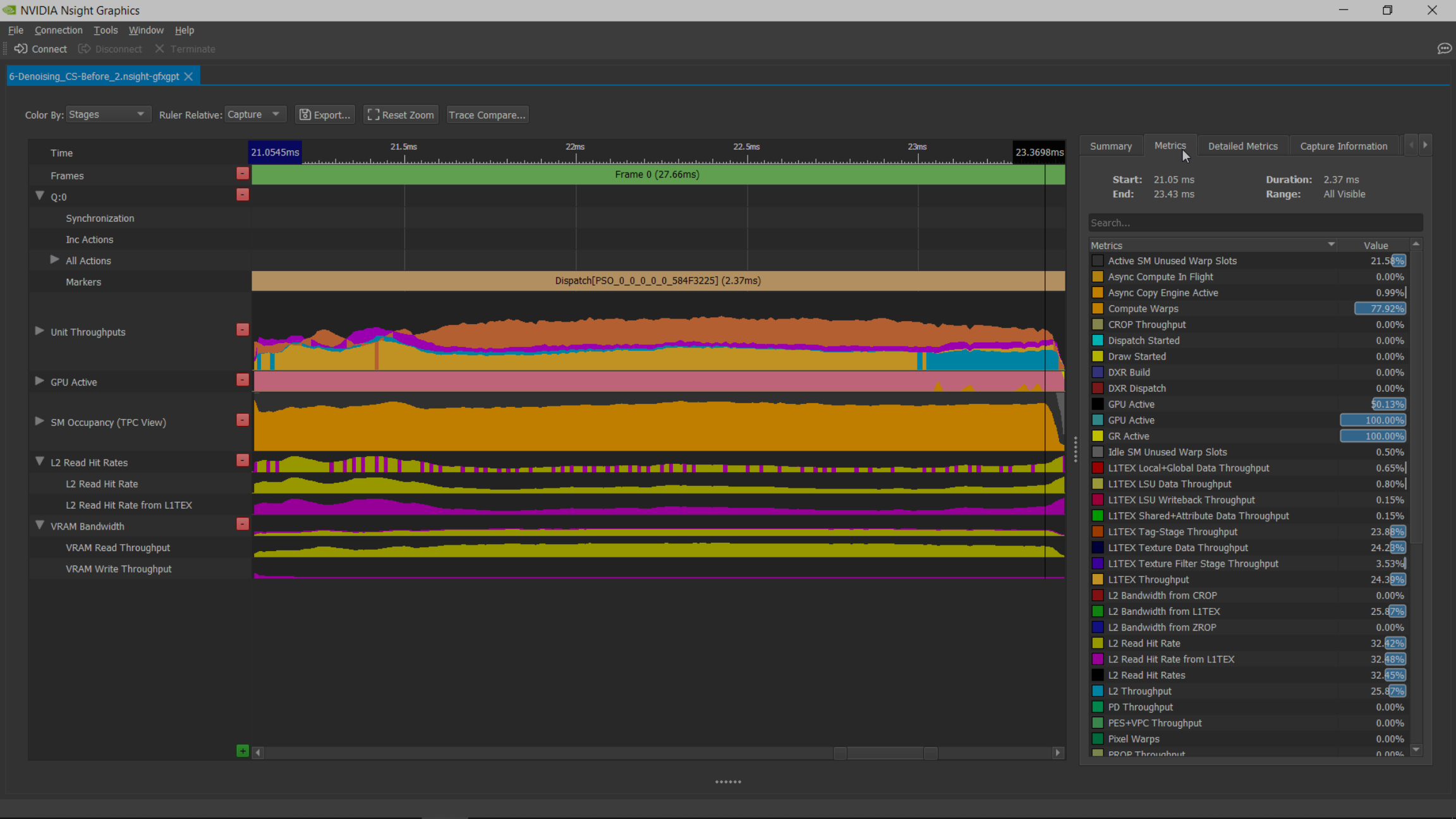
Unit Throughputs:
VRAM: 57%
SM: 33%
L2: 26%

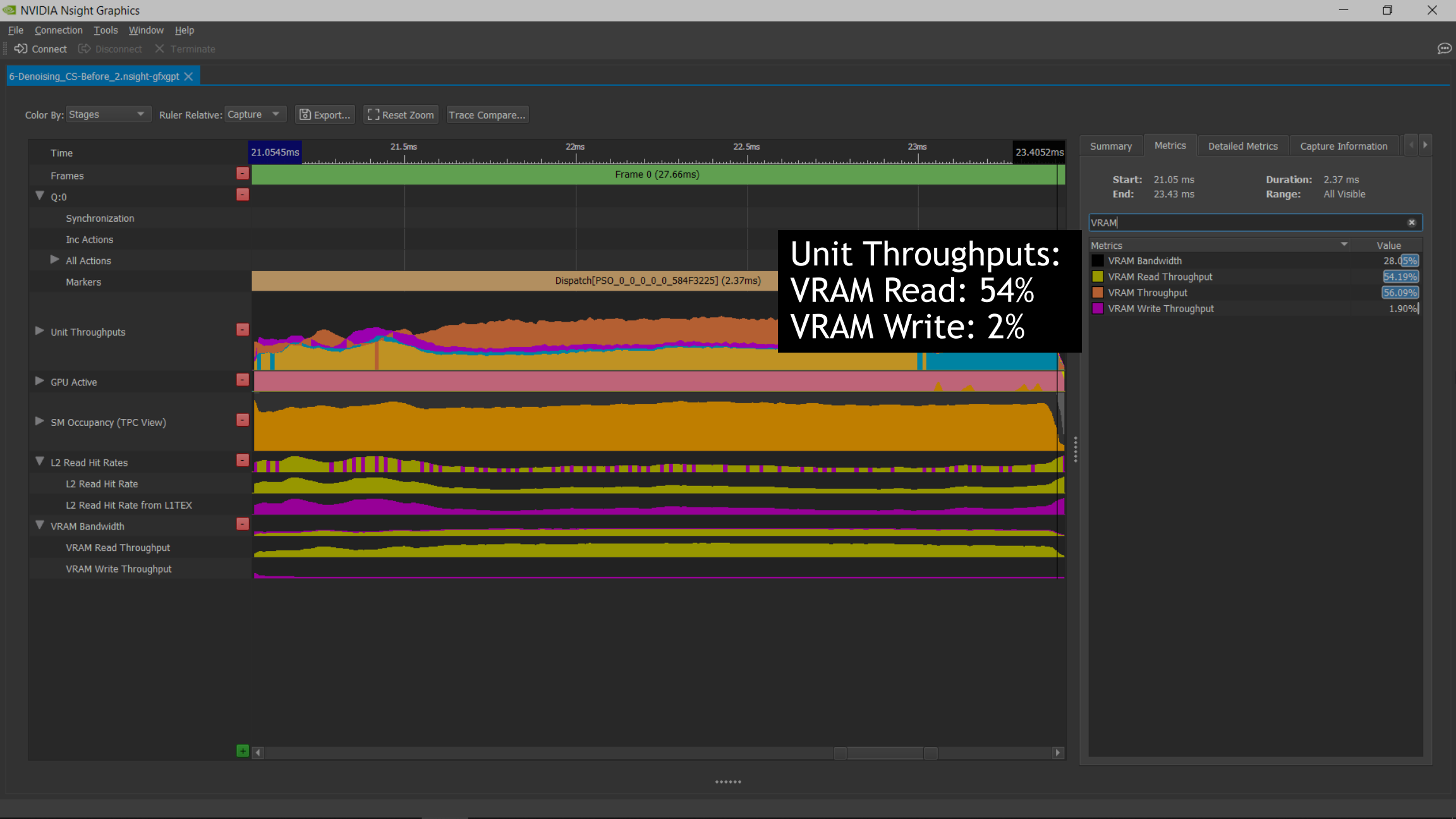
Summary Metrics Detailed Metrics Capture Information

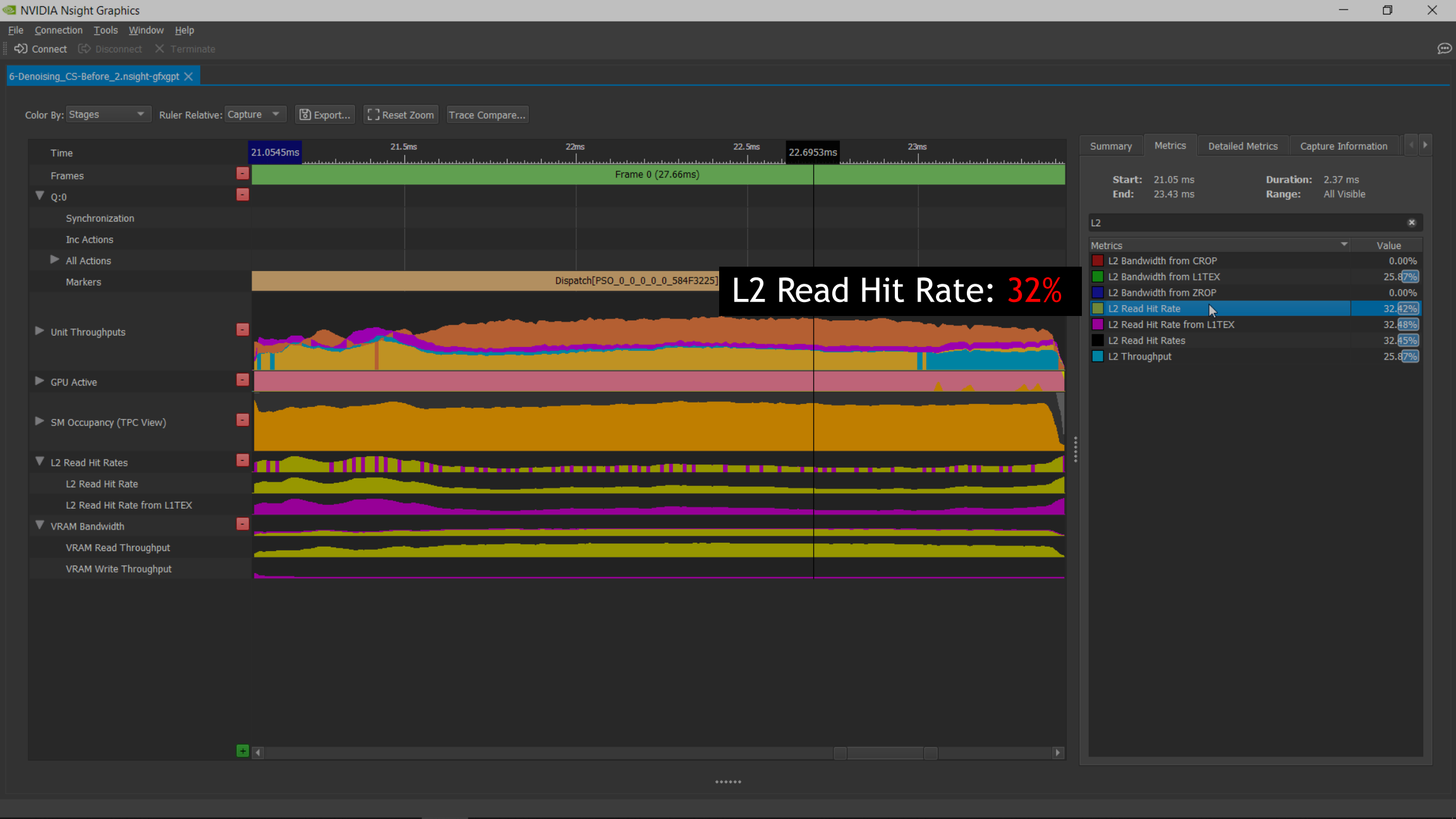
Start: 21.05 ms **Duration:** 2.37 ms
End: 23.43 ms **Range:** All Visible

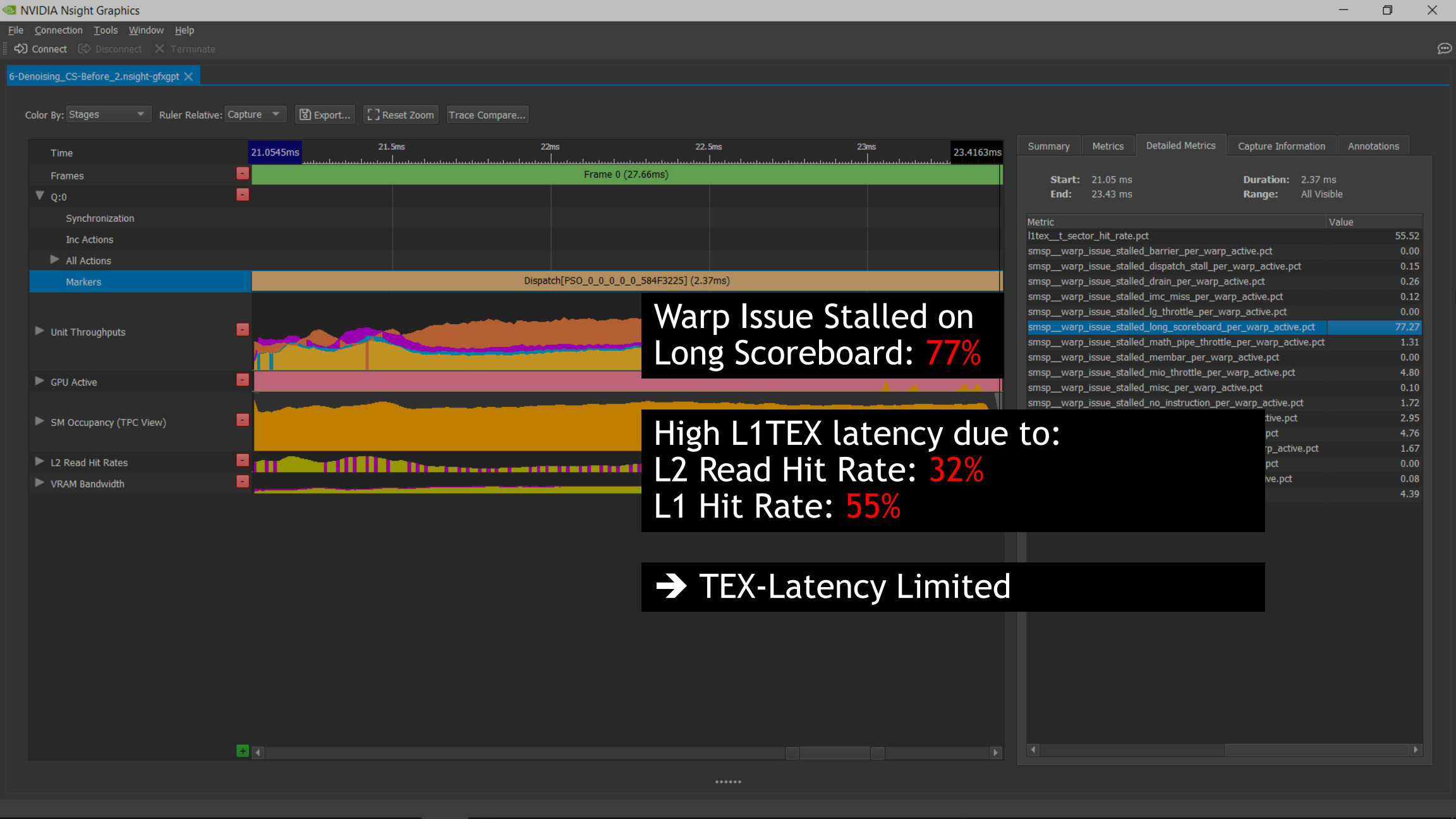
Unit Throughput	Value
VRAM Throughput	56.09%
SM Instruction Throughput	33.76%
SM SFU Pipe Throughput	33.58%
SM Issue Active	29.74%
L2 Throughput	25.87%
SM FMA Pipe Throughput	25.31%
L1TEX Throughput	24.39%
SM ALU Pipe Throughput	18.03%
PROP Throughput	0.00%
PD Throughput	0.00%
PES+VPC Throughput	0.00%
RASTER Throughput	0.00%
ZROP Throughput	0.00%
CROP Throughput	0.00%
SM FP16+Tensor Pipe Throughput	0.00%

Warp Occupancy	Value
Idle SM Unused Warp Slots	0.50%
Active SM Unused Warp Slots	21.58%
Compute Warps	77.92%
Pixel Warps	0.00%
Vertex+Tess+Geom Warps	0.00%

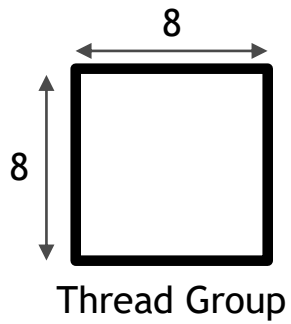

NVIDIA Nsight Graphics
File Connection Tools Window Help
Connect Disconnect Terminate
6-Denoising_CS-Before_2.nsight-gfxgpt
Color By: Stages
Ruler Relative: Capture
Export...
Reset Zoom
Trace Compare...
Time
21.0545ms
21.5ms
22ms
22.5ms
23ms
23.3698ms
Frames
Frame 0 (27.66ms)
Q:0
Synchronization
Inc Actions
All Actions
Markers
Dispatch[PSO_0_0_0_0_0_584F3225] (2.37ms)
Unit Throughputs
GPU Active
SM Occupancy (TPC View)
L2 Read Hit Rates
L2 Read Hit Rate
L2 Read Hit Rate from L1TEX
VRAM Bandwidth
VRAM Read Throughput
VRAM Write Throughput
Summary
Metrics
Detailed Metrics
Capture Information
Start: 21.05 ms
End: 23.43 ms
Duration: 2.37 ms
Range: All Visible
Search...
Metrics
Value
Active SM Unused Warp Slots 21.58%
Async Compute In Flight 0.00%
Async Copy Engine Active 0.99%
Compute Warps 77.92%
CROP Throughput 0.00%
Dispatch Started 0.00%
Draw Started 0.00%
DXR Build 0.00%
DXR Dispatch 0.00%
GPU Active 50.13%
GPU Active 100.00%
GR Active 100.00%
Idle SM Unused Warp Slots 0.50%
L1TEX Local+Global Data Throughput 0.65%
L1TEX LSU Data Throughput 0.80%
L1TEX LSU Writeback Throughput 0.15%
L1TEX Shared+Attribute Data Throughput 0.15%
L1TEX Tag-Stage Throughput 23.88%
L1TEX Texture Data Throughput 24.25%
L1TEX Texture Filter Stage Throughput 3.53%
L1TEX Throughput 24.39%
L2 Bandwidth from CROP 0.00%
L2 Bandwidth from L1TEX 25.87%
L2 Bandwidth from ZROP 0.00%
L2 Read Hit Rate 32.42%
L2 Read Hit Rate from L1TEX 32.48%
L2 Read Hit Rates 32.45%
L2 Throughput 25.87%
PD Throughput 0.00%
PES+VPC Throughput 0.00%
Pixel Warps 0.00%


NVIDIA Nsight Graphics
File Connection Tools Window Help
Connect Disconnect Terminate
6-Denoising_CS-Before_2.nsight-gfxgpt
Color By: Stages
Ruler Relative: Capture
Export...
Reset Zoom
Trace Compare...
Time
21.0545ms
21.5ms
22ms
22.5ms
23ms
23.4052ms
Frames
Frame 0 (27.66ms)
Q:0
Synchronization
Inc Actions
All Actions
Markers
Dispatch[PSO_0_0_0_0_0_584F3225] (2.37ms)
Unit Throughputs
GPU Active
SM Occupancy (TPC View)
L2 Read Hit Rates
L2 Read Hit Rate
L2 Read Hit Rate from L1TEX
VRAM Bandwidth
VRAM Read Throughput
VRAM Write Throughput
Unit Throughputs:
VRAM Read: 54%
VRAM Write: 2%
Summary
Metrics
Detailed Metrics
Capture Information
Start: 21.05 ms
End: 23.43 ms
Duration: 2.37 ms
Range: All Visible
VRAM
Metrics
Value
VRAM Bandwidth 28.05%
VRAM Read Throughput 54.19%
VRAM Throughput 56.09%
VRAM Write Throughput 1.90%

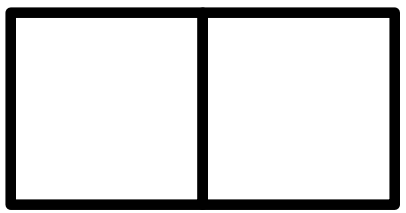

NVIDIA Nsight Graphics
File Connection Tools Window Help
Connect Disconnect Terminate
6-Denoising_CS-Before_2.nsight-gfxgpt
Color By: Stages
Ruler Relative: Capture
Export...
Reset Zoom
Trace Compare...
Time
21.0545ms
21.5ms
22ms
22.5ms
22.6953ms
23ms
Frames
Frame 0 (27.66ms)
Q:0
Synchronization
Inc Actions
All Actions
Markers
Dispatch[PSO_0_0_0_0_0_584F3225]
L2 Read Hit Rate: 32%
Unit Throughputs
GPU Active
SM Occupancy (TPC View)
L2 Read Hit Rates
L2 Read Hit Rate
L2 Read Hit Rate from L1TEX
VRAM Bandwidth
VRAM Read Throughput
VRAM Write Throughput
Summary
Metrics
Detailed Metrics
Capture Information
Start: 21.05 ms
End: 23.43 ms
Duration: 2.37 ms
Range: All Visible
L2
Metrics
Value
L2 Bandwidth from CROP 0.00%
L2 Bandwidth from L1TEX 25.87%
L2 Bandwidth from ZROP 0.00%
L2 Read Hit Rate 32.42%
L2 Read Hit Rate from L1TEX 32.48%
L2 Read Hit Rates 32.45%
L2 Throughput 25.87%


NVIDIA Nsight Graphics
File Connection Tools Window Help
Connect Disconnect Terminate
6-Denoising_CS-Before_2.nsight-gfxgpt
Color By: Stages
Ruler Relative: Capture
Export...
Reset Zoom
Trace Compare...
Time
21.0545ms
21.5ms
22ms
22.5ms
23ms
23.4163ms
Frames
Frame 0 (27.66ms)
Q:0
Synchronization
Inc Actions
All Actions
Markers
Dispatch[PSO_0_0_0_0_0_584F3225] (2.37ms)
Unit Throughputs
GPU Active
SM Occupancy (TPC View)
L2 Read Hit Rates
VRAM Bandwidth
Warp Issue Stalled on Long Scoreboard: 77%
High L1TEX latency due to:
L2 Read Hit Rate: 32%
L1 Hit Rate: 55%
➔ TEX-Latency Limited
Summary
Metrics
Detailed Metrics
Capture Information
Annotations
Start: 21.05 ms
End: 23.43 ms
Duration: 2.37 ms
Range: All Visible
Metric
Value
l1tex__t_sector_hit_rate.pct 55.52
smsp__warp_issue_stalled_barrier_per_warp_active.pct 0.00
smsp__warp_issue_stalled_dispatch_stall_per_warp_active.pct 0.15
smsp__warp_issue_stalled_drain_per_warp_active.pct 0.26
smsp__warp_issue_stalled_imc_miss_per_warp_active.pct 0.12
smsp__warp_issue_stalled_lg_throttle_per_warp_active.pct 0.00
smsp__warp_issue_stalled_long_scoreboard_per_warp_active.pct 77.27
smsp__warp_issue_stalled_math_pipe_throttle_per_warp_active.pct 1.31
smsp__warp_issue_stalled_membar_per_warp_active.pct 0.00
smsp__warp_issue_stalled_mio_throttle_per_warp_active.pct 4.80
smsp__warp_issue_stalled_misc_per_warp_active.pct 0.10
smsp__warp_issue_stalled_no_instruction_per_warp_active.pct 1.72
2.95
4.76
1.67
0.00
0.08
4.39

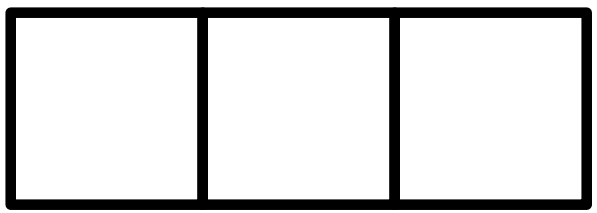
THREAD-GROUP LAUNCH ORDER



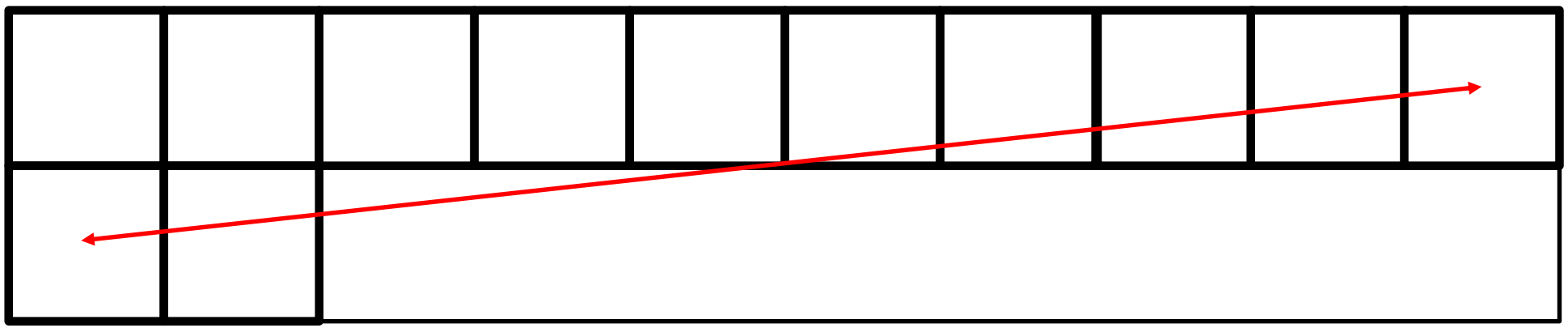
THREAD-GROUP LAUNCH ORDER



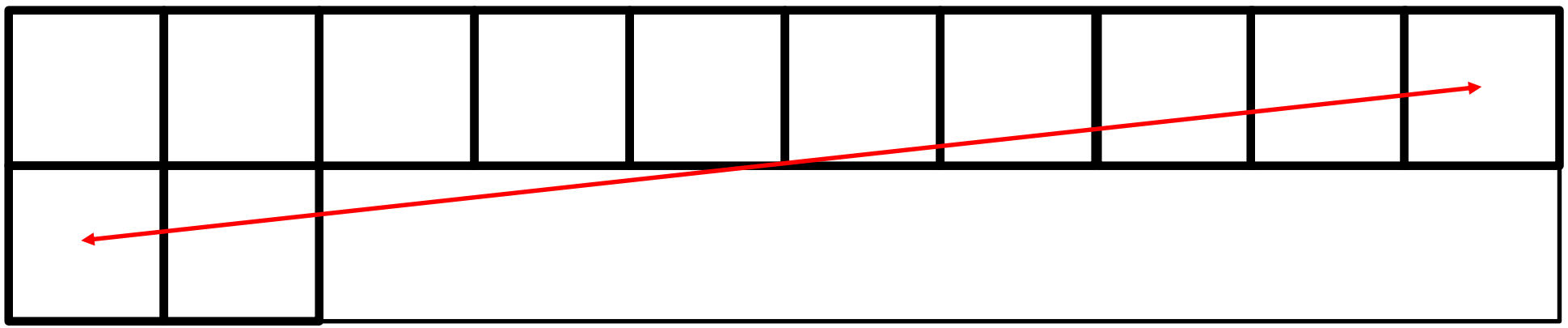
THREAD-GROUP LAUNCH ORDER



THREAD-GROUP LAUNCH ORDER

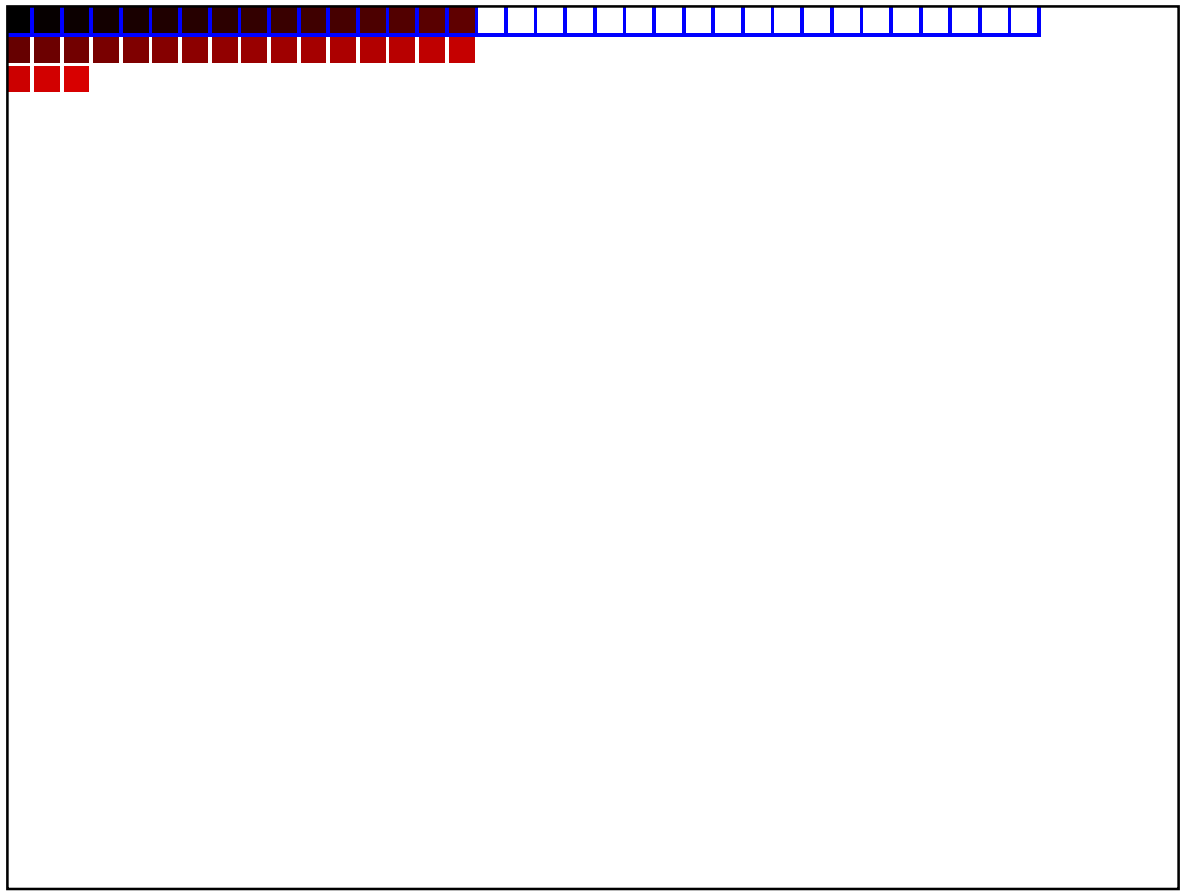


THREAD-GROUP LAUNCH ORDER



Thread groups launched sequentially have far-appart pixel coordinates
➔ Increases the size of the working set in L2
➔ Can cause poor L2 hit rate and long TEX miss latencies

THREAD-GROUP TILING



Divide the Dispatch grid into tiles of width=N

```
uint N = 16;
uint vThreadGroupIDFlattened = (Dispatch_Grid_Dim.x)*groupId.y + groupId.x;
uint Total_number_of_ThreadGroups_in_one_tile = N*(Dispatch_Grid_Dim.y);

uint Tile_ID_of_current_ThreadGroup = (vThreadGroupIDFlattened)/Total_number_of_ThreadGroups_in_one_tile;
uint Local_ThreadGroup_ID_flattened_within_current_tile =
(vThreadGroupIDFlattened)%Total_number_of_ThreadGroups_in_one_tile;
uint Local_ThreadGroup_ID_y_within_current_tile = (Local_ThreadGroup_ID_flattened_within_current_tile)/N;
uint Local_ThreadGroup_ID_x_within_current_tile = (Local_ThreadGroup_ID_flattened_within_current_tile)%N;

uint Tiled_vThreadGroupIDFlattened = Tile_ID_of_current_ThreadGroup*N +
Local_ThreadGroup_ID_y_within_current_tile*(Dispatch_Grid_Dim.x) + Local_ThreadGroup_ID_x_within_current_tile;

uint2 TiledvThreadGroupID;
TiledvThreadGroupID.y = Tiled_vThreadGroupIDFlattened/Dispatch_Grid_Dim.x;
TiledvThreadGroupID.x = Tiled_vThreadGroupIDFlattened%Dispatch_Grid_Dim.x;

uint2 TiledvThreadID;
TiledvThreadID.x = (ThreadGroup_Dim.x)*TiledvThreadGroupID.x + groupThreadIndex.x;
TiledvThreadID.y = (ThreadGroup_Dim.y)*TiledvThreadGroupID.y + groupThreadIndex.y;
```

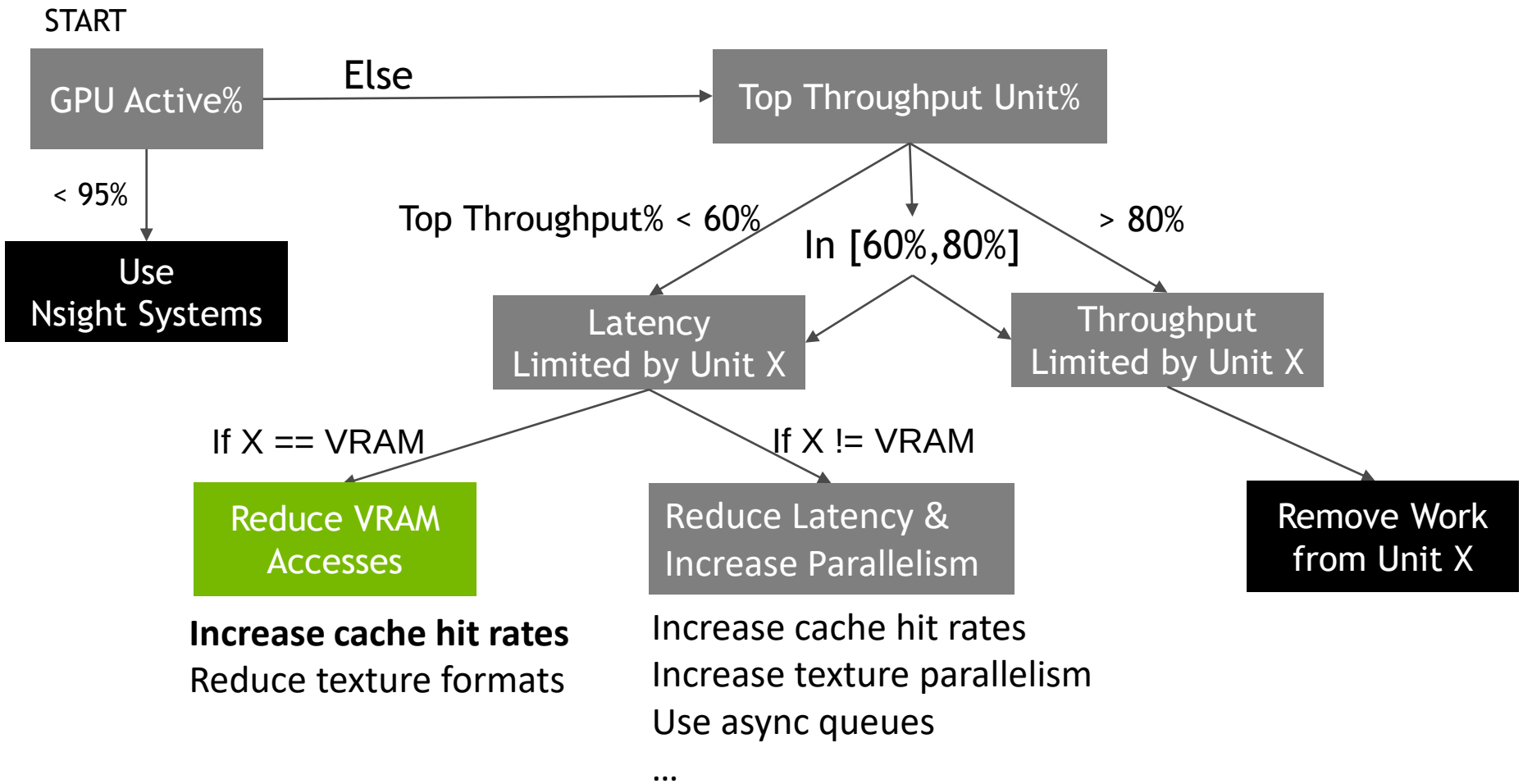
THREAD-GROUP TILING

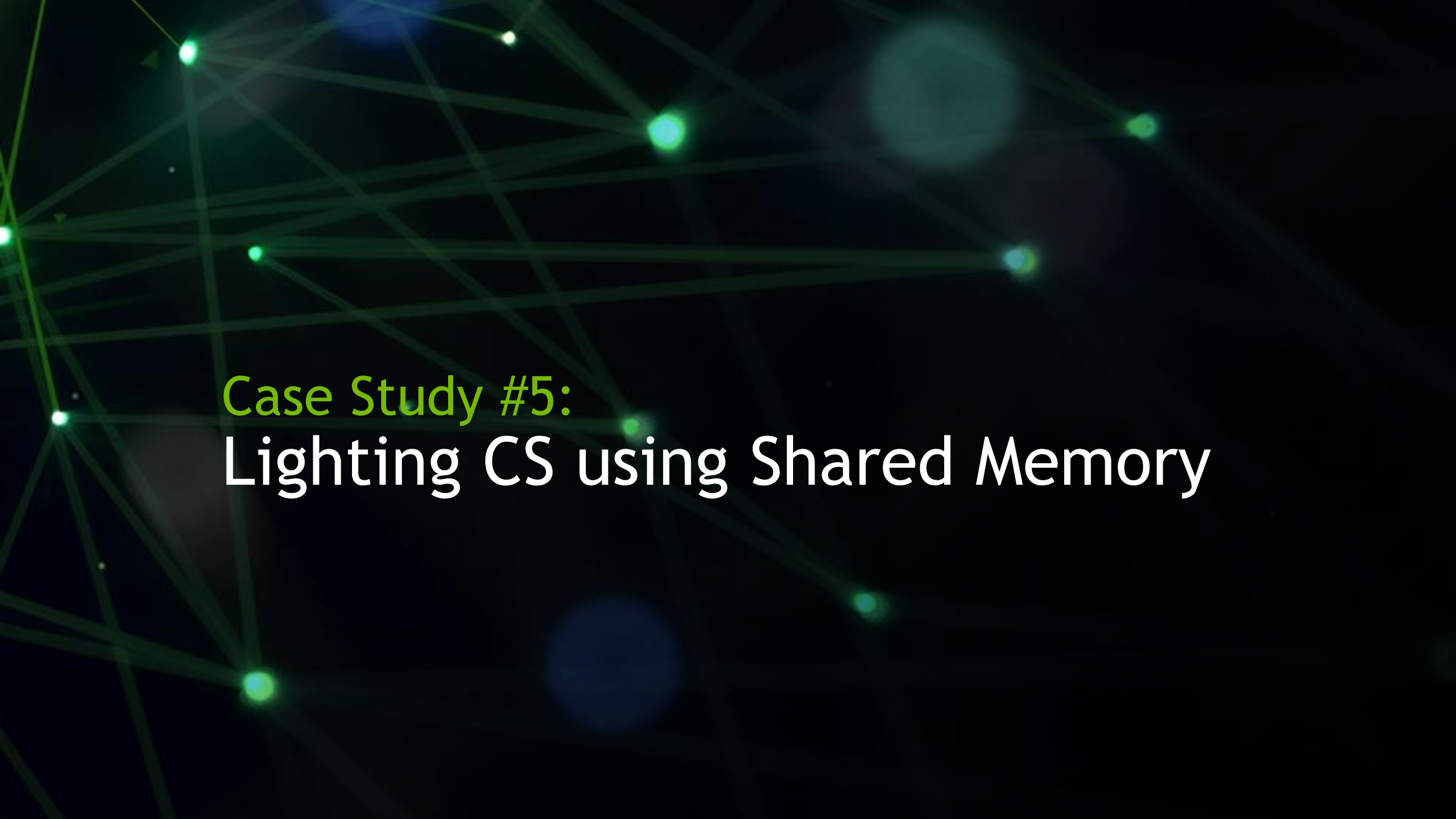
Tile Size: [16, Dispatch_Grid_Dim.y]

	BEFORE	AFTER	RATIO
GPU Elapsed Time	2.36 ms	1.61 ms	**1.47x Gain**
Throughput: VRAM	56.3%	30.4%	0.54x
Throughput: SM	33.8%	51.3%	1.52x
L2 Read Hit Rate	62.5%	85.8%	1.37x
SM Warp Stalls on long_scoreboard	77.8%	58.6%	0.75x

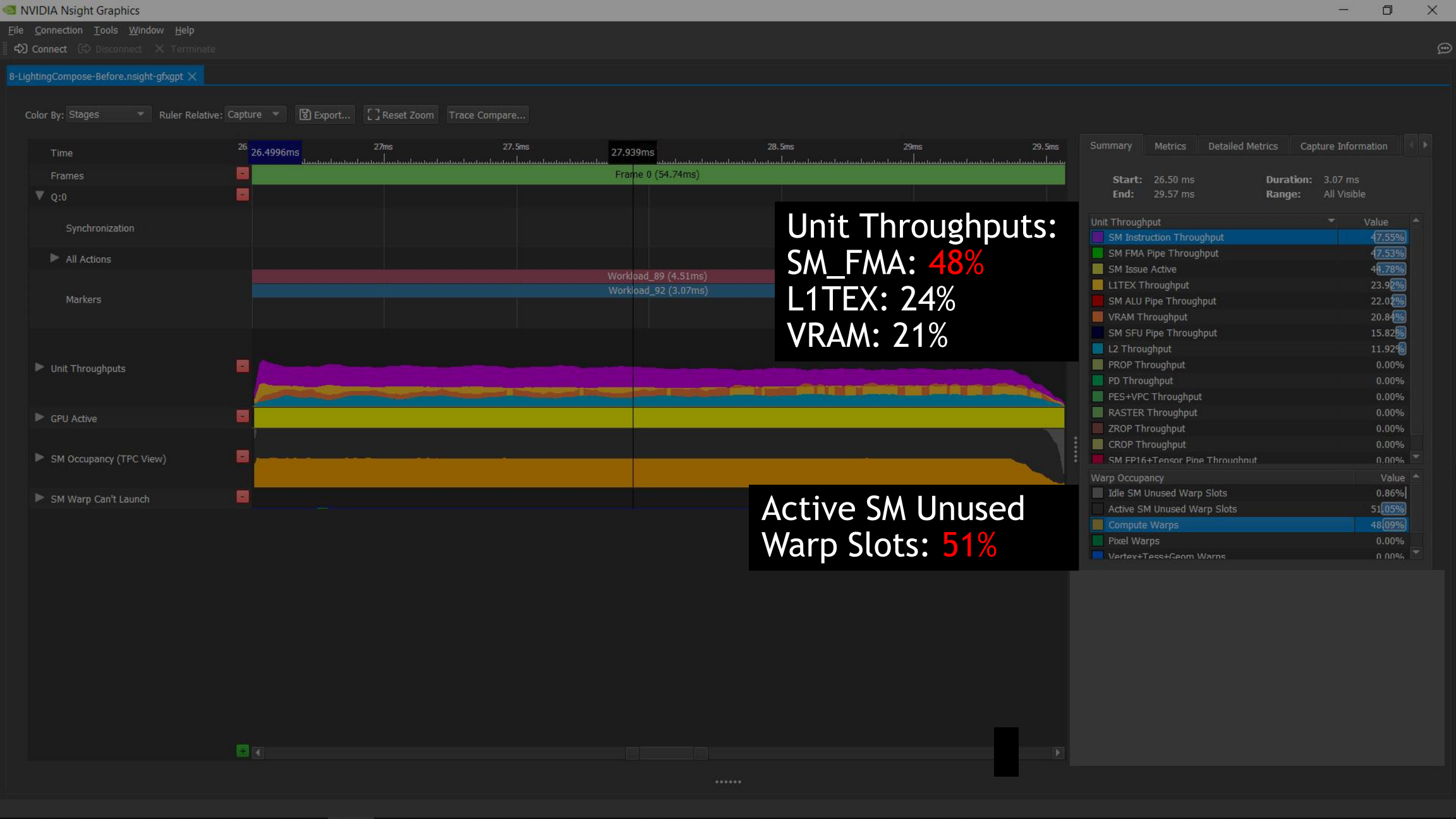
On RTX 2080 with SetStablePowerState(TRUE)

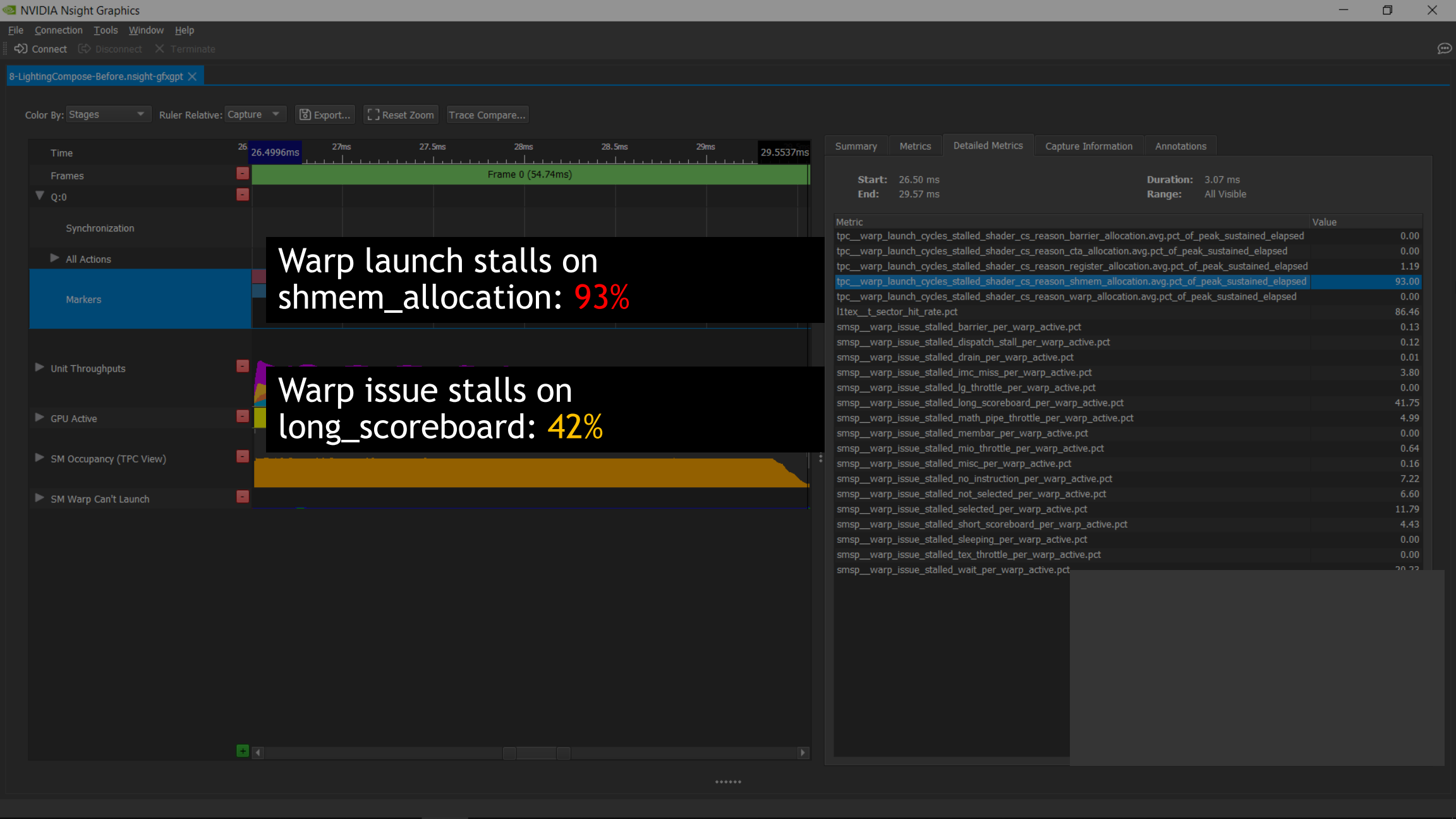
THE P3 (PEAK-PERF%) METHOD





Case Study #5:
Lighting CS using Shared Memory


NVIDIA Nsight Graphics
File Connection Tools Window Help
Connect Disconnect Terminate
8-LightingCompose-Before.nsight-gfxgpt
Color By: Stages
Ruler Relative: Capture
Export...
Reset Zoom
Trace Compare...
Time
26.4996ms
27.939ms
Frames
Frame 0 (54.74ms)
Q:0
Synchronization
All Actions
Workload_89 (4.51ms)
Workload_92 (3.07ms)
Markers
Unit Throughputs
GPU Active
SM Occupancy (TPC View)
SM Warp Can't Launch
Unit Throughputs:
SM_FMA: 48%
L1TEX: 24%
VRAM: 21%
Active SM Unused
Warp Slots: 51%
Summary
Metrics
Detailed Metrics
Capture Information
Start: 26.50 ms
End: 29.57 ms
Duration: 3.07 ms
Range: All Visible
Unit Throughput
Value
SM Instruction Throughput 47.55%
SM FMA Pipe Throughput 47.53%
SM Issue Active 48.78%
L1TEX Throughput 23.92%
SM ALU Pipe Throughput 22.02%
VRAM Throughput 20.84%
SM SFU Pipe Throughput 15.82%
L2 Throughput 11.92%
PROP Throughput 0.00%
PD Throughput 0.00%
PES+VPC Throughput 0.00%
RASTER Throughput 0.00%
ZROP Throughput 0.00%
CROP Throughput 0.00%
Warp Occupancy
Value
Idle SM Unused Warp Slots 0.86%
Active SM Unused Warp Slots 51.05%
Compute Warps 48.09%
Pixel Warps 0.00%


NVIDIA Nsight Graphics
File Connection Tools Window Help
Connect Disconnect Terminate
8-LightingCompose-Before.nsight-gfxgpt
Color By: Stages
Ruler Relative: Capture
Export...
Reset Zoom
Trace Compare...
Time
26.4996ms
27ms
27.5ms
28ms
28.5ms
29ms
29.5537ms
Frames
Frame 0 (54.74ms)
Q:0
Synchronization
All Actions
Markers
Unit Throughputs
GPU Active
SM Occupancy (TPC View)
SM Warp Can't Launch
Warp launch stalls on shmem_allocation: 93%
Warp issue stalls on long_scoreboard: 42%
Summary
Metrics
Detailed Metrics
Capture Information
Annotations
Start: 26.50 ms
End: 29.57 ms
Duration: 3.07 ms
Range: All Visible
Metric
Value
tpc__warp_launch_cycles_stalled_shader_cs_reason_barrier_allocation.avg.pct_of_peak_sustained_elapsed 0.00
tpc__warp_launch_cycles_stalled_shader_cs_reason_cta_allocation.avg.pct_of_peak_sustained_elapsed 0.00
tpc__warp_launch_cycles_stalled_shader_cs_reason_register_allocation.avg.pct_of_peak_sustained_elapsed 1.19
tpc__warp_launch_cycles_stalled_shader_cs_reason_shmem_allocation.avg.pct_of_peak_sustained_elapsed 93.00
tpc__warp_launch_cycles_stalled_shader_cs_reason_warp_allocation.avg.pct_of_peak_sustained_elapsed 0.00
l1tex__t_sector_hit_rate.pct 86.46
smsp__warp_issue_stalled_barrier_per_warp_active.pct 0.13
smsp__warp_issue_stalled_dispatch_stall_per_warp_active.pct 0.12
smsp__warp_issue_stalled_drain_per_warp_active.pct 0.01
smsp__warp_issue_stalled_imc_miss_per_warp_active.pct 3.80
smsp__warp_issue_stalled_lg_throttle_per_warp_active.pct 0.00
smsp__warp_issue_stalled_long_scoreboard_per_warp_active.pct 41.75
smsp__warp_issue_stalled_math_pipe_throttle_per_warp_active.pct 4.99
smsp__warp_issue_stalled_membar_per_warp_active.pct 0.00
smsp__warp_issue_stalled_mio_throttle_per_warp_active.pct 0.64
smsp__warp_issue_stalled_misc_per_warp_active.pct 0.16
smsp__warp_issue_stalled_no_instruction_per_warp_active.pct 7.22
smsp__warp_issue_stalled_not_selected_per_warp_active.pct 6.60
smsp__warp_issue_stalled_selected_per_warp_active.pct 11.79
smsp__warp_issue_stalled_short_scoreboard_per_warp_active.pct 4.43
smsp__warp_issue_stalled_sleeping_per_warp_active.pct 0.00
smsp__warp_issue_stalled_tex_throttle_per_warp_active.pct 0.00
smsp__warp_issue_stalled_wait_per_warp_active.pct

SHARED-MEM-SIZE REDUCTION

Before: store light data into shared mem

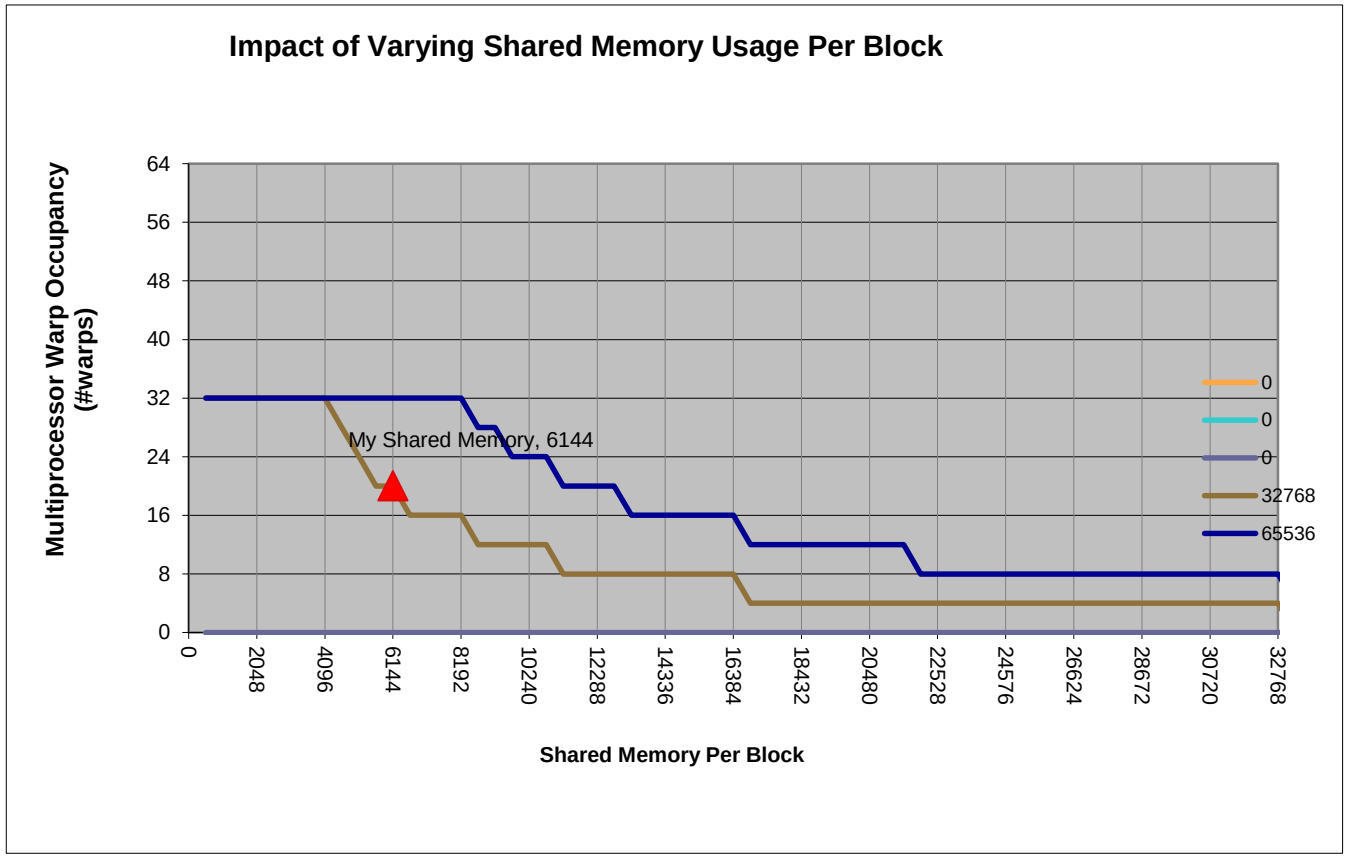
```
static groupshared LightData SharedMem[MaxLightCount];

...
```

After: store light indices into shared mem

```
static groupshared uint SharedMem[MaxLightCount];
```

And load light data via non-divergent indexed constant-buffer loads

CUDA OCCUPANCY CALCULATOR



For Turing RTX GPUs (Compute Capability 7.5)

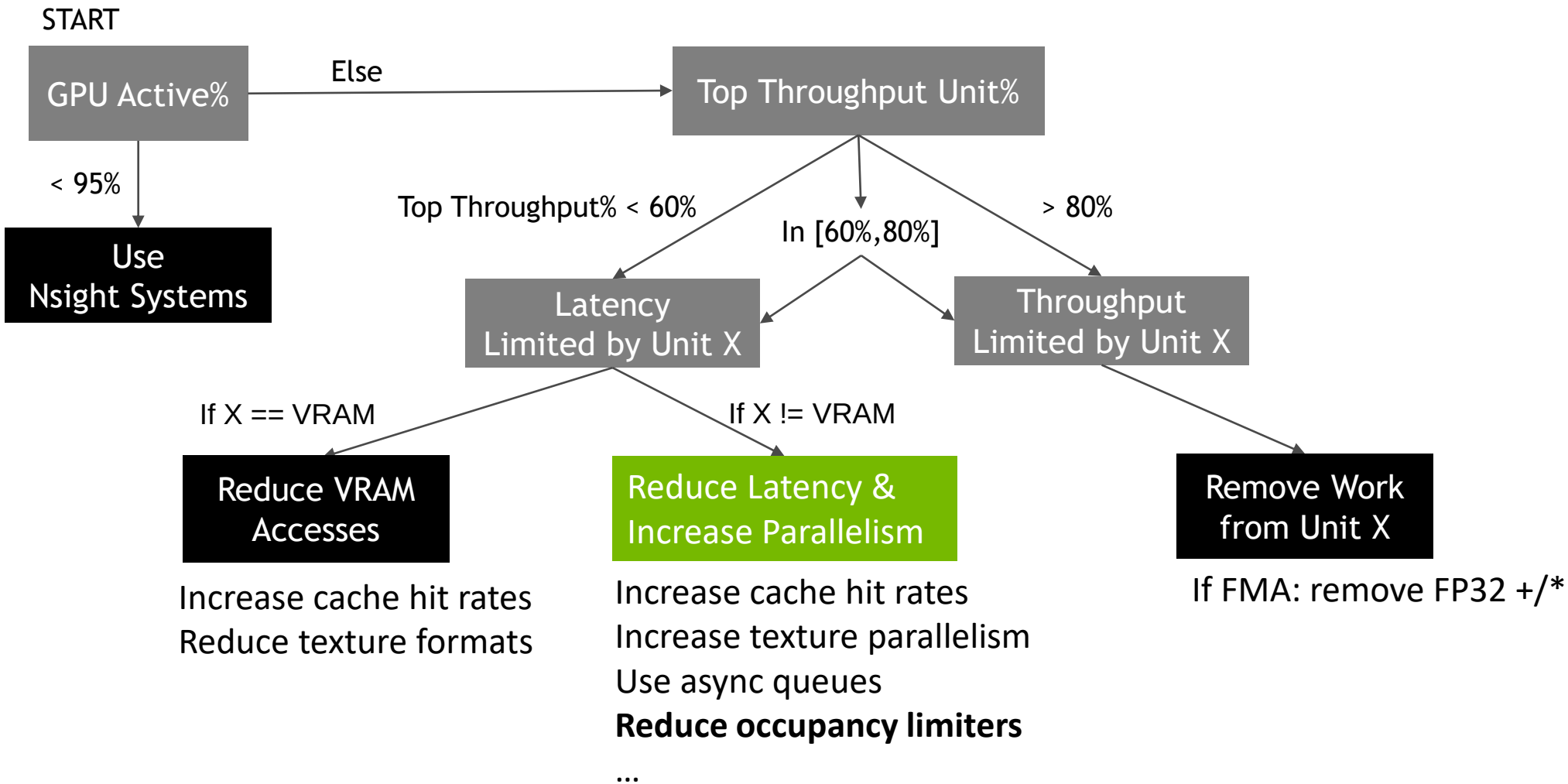
SHARED-MEM-SIZE REDUCTION:

6,144 -> 256 bytes per thread group

	BEFORE	AFTER	RATIO
GPU Elapsed Time	3.06 ms	2.21 ms	**1.38x Gain**
SM Throughput	47.5%	66.1%	1.39x
Warp launch stalls on shmem_allocation	94%	0%	+INF
Warp issue stalls on long_scoreboard	41.7%	36.7%	1.14x

On RTX 2080 with SetStablePowerState(TRUE)

THE P3 (PEAK-PERF%) METHOD




Case Study #6:
Thread-Divergence-Limited DispatchRays in Battlefield V


Battlefield V DXR Ultra
SECURE CHOKEPOINT KILO
20

NVIDIA Nsight Graphics

File Connection Tools Window Help

Connect Disconnect Terminate

Original.nsight-gfxgpt

Color By: Stages Ruler Relative: Capture Export... Reset Zoom Trace Compare...

Time 12.4854ms 12.6ms 12.7ms 12.8ms 12.9ms 13ms 13.1ms

Frames Frame 0 (20.39ms)

Q:0

Synchronization

Inc Actions

All Actions DispatchRays 2 (1.13ms

Markers

Unit Throughputs

GPU Active

SM Occupancy (TPC View)

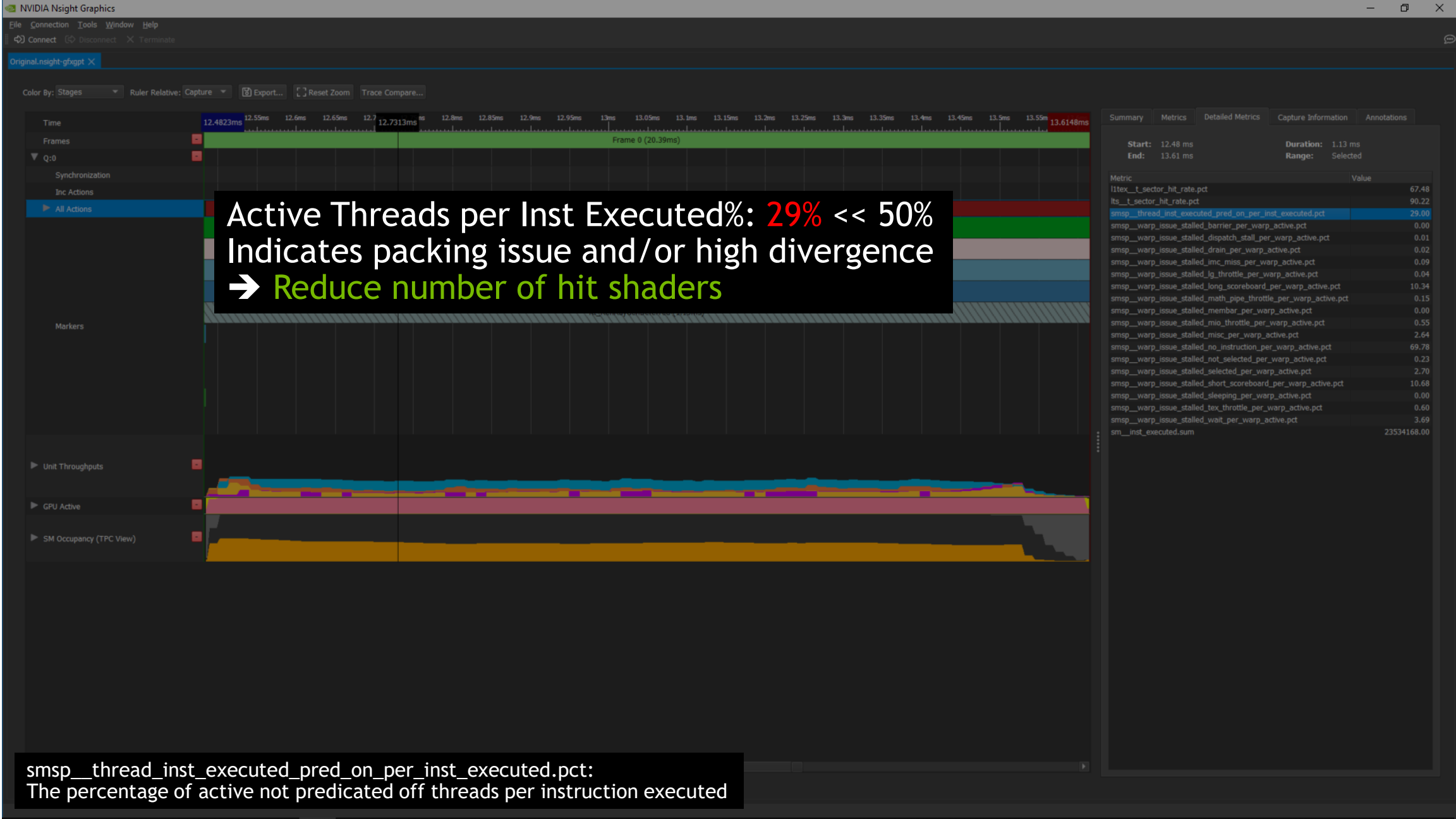
Unit Throughputs:
L2: 24%
VRAM: 12%
SM: 8%

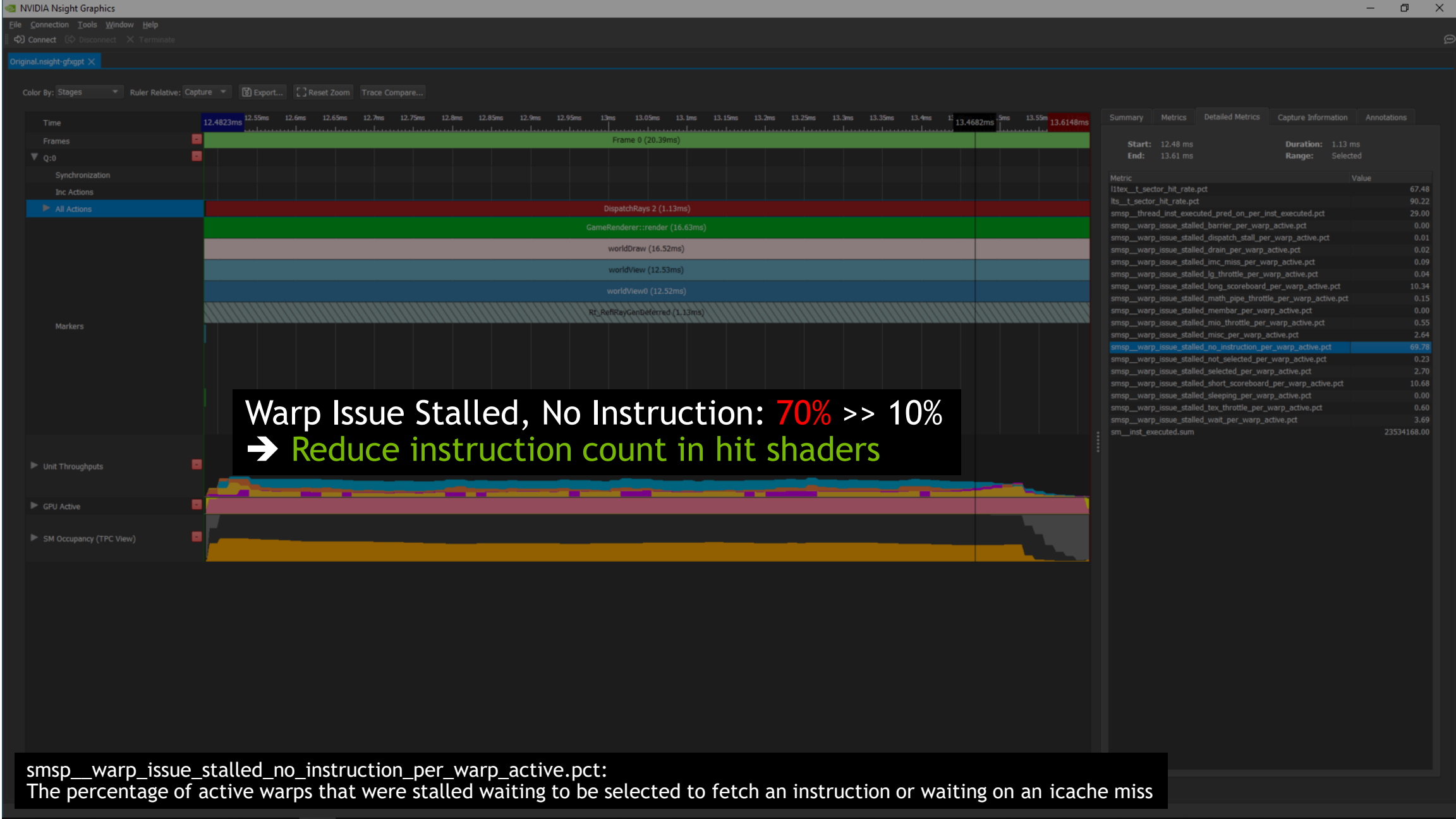
Summary Metrics Detailed Metrics Capture Information Annotations

Start: 12.48 ms **Duration:** 1.13 ms
End: 13.61 ms **Range:** Selected

Unit Throughput	Value
L2 Throughput	23.78%
VRAM Throughput	12.16%
SM Instruction Throughput	7.63%
SM Issue Active	7.63%
L1TEX Throughput	7.42%
SM FMA Pipe Throughput	5.81%
SM ALU Pipe Throughput	5.33%
SM SFU Pipe Throughput	2.42%
PROP Throughput	0.00%
PD Throughput	0.00%
PES+VPC Throughput	0.00%
RASTER Throughput	0.00%
ZROP Throughput	0.00%

Warp Occupancy	Value
Idle SM Unused Warp Slots	5.80%
Active SM Unused Warp Slots	57.11%
Compute Warps	37.09%


NVIDIA Nsight Graphics
File Connection Tools Window Help
Connect Disconnect Terminate
Original.nsight-gfxgpt
Color By: Stages
Ruler Relative: Capture
Export...
Reset Zoom
Trace Compare...
Time
Frames
Q:0
Synchronization
Inc Actions
All Actions
Markers
Unit Throughputs
GPU Active
SM Occupancy (TPC View)
12.4823ms
12.7313ms
13.6148ms
Frame 0 (20.39ms)
Active Threads per Inst Executed%: 29% << 50%
Indicates packing issue and/or high divergence
➔ Reduce number of hit shaders
Summary
Metrics
Detailed Metrics
Capture Information
Annotations
Start: 12.48 ms
End: 13.61 ms
Duration: 1.13 ms
Range: Selected
Metric
Value
l1tex__t_sector_hit_rate.pct 67.48
lts__t_sector_hit_rate.pct 90.22
smsp__thread_inst_executed_pred_on_per_inst_executed.pct 29.00
smsp__warp_issue_stalled_barrier_per_warp_active.pct 0.00
smsp__warp_issue_stalled_dispatch_stall_per_warp_active.pct 0.01
smsp__warp_issue_stalled_drain_per_warp_active.pct 0.02
smsp__warp_issue_stalled_imc_miss_per_warp_active.pct 0.09
smsp__warp_issue_stalled_lg_throttle_per_warp_active.pct 0.04
smsp__warp_issue_stalled_long_scoreboard_per_warp_active.pct 10.34
smsp__warp_issue_stalled_math_pipe_throttle_per_warp_active.pct 0.15
smsp__warp_issue_stalled_membar_per_warp_active.pct 0.00
smsp__warp_issue_stalled_mio_throttle_per_warp_active.pct 0.55
smsp__warp_issue_stalled_misc_per_warp_active.pct 2.64
smsp__warp_issue_stalled_no_instruction_per_warp_active.pct 69.78
smsp__warp_issue_stalled_not_selected_per_warp_active.pct 0.23
smsp__warp_issue_stalled_selected_per_warp_active.pct 2.70
smsp__warp_issue_stalled_short_scoreboard_per_warp_active.pct 10.68
smsp__warp_issue_stalled_sleeping_per_warp_active.pct 0.00
smsp__warp_issue_stalled_tex_throttle_per_warp_active.pct 0.60
smsp__warp_issue_stalled_wait_per_warp_active.pct 3.69
sm__inst_executed.sum 23534168.00
smsp__thread_inst_executed_pred_on_per_inst_executed.pct:
The percentage of active not predicated off threads per instruction executed


NVIDIA Nsight Graphics
File Connection Tools Window Help
Connect Disconnect Terminate
Original.nsight-gfxgpt
Color By: Stages
Ruler Relative: Capture
Export...
Reset Zoom
Trace Compare...
Time
Frames
Q:0
Synchronization
Inc Actions
All Actions
Markers
Unit Throughputs
GPU Active
SM Occupancy (TPC View)
12.4823ms
13.4682ms
13.6148ms
Frame 0 (20.39ms)
DispatchRays 2 (1.13ms)
GameRenderer::render (16.63ms)
worldDraw (16.52ms)
worldView (12.53ms)
worldView0 (12.52ms)
Rt_RefiRayGenDeferred (1.13ms)
Warp Issue Stalled, No Instruction: 70% >> 10%
➔ Reduce instruction count in hit shaders
Summary
Metrics
Detailed Metrics
Capture Information
Annotations
Start: 12.48 ms
End: 13.61 ms
Duration: 1.13 ms
Range: Selected
Metric
Value
l1tex__t_sector_hit_rate.pct 67.48
lts__t_sector_hit_rate.pct 90.22
smsp__thread_inst_executed_pred_on_per_inst_executed.pct 29.00
smsp__warp_issue_stalled_barrier_per_warp_active.pct 0.00
smsp__warp_issue_stalled_dispatch_stall_per_warp_active.pct 0.01
smsp__warp_issue_stalled_drain_per_warp_active.pct 0.02
smsp__warp_issue_stalled_imc_miss_per_warp_active.pct 0.09
smsp__warp_issue_stalled_lg_throttle_per_warp_active.pct 0.04
smsp__warp_issue_stalled_long_scoreboard_per_warp_active.pct 10.34
smsp__warp_issue_stalled_math_pipe_throttle_per_warp_active.pct 0.15
smsp__warp_issue_stalled_membar_per_warp_active.pct 0.00
smsp__warp_issue_stalled_mio_throttle_per_warp_active.pct 0.55
smsp__warp_issue_stalled_misc_per_warp_active.pct 2.64
smsp__warp_issue_stalled_no_instruction_per_warp_active.pct 69.78
smsp__warp_issue_stalled_not_selected_per_warp_active.pct 0.23
smsp__warp_issue_stalled_selected_per_warp_active.pct 2.70
smsp__warp_issue_stalled_short_scoreboard_per_warp_active.pct 10.68
smsp__warp_issue_stalled_sleeping_per_warp_active.pct 0.00
smsp__warp_issue_stalled_tex_throttle_per_warp_active.pct 0.60
smsp__warp_issue_stalled_wait_per_warp_active.pct 3.69
sm__inst_executed.sum 23534168.00
smsp__warp_issue_stalled_no_instruction_per_warp_active.pct:
The percentage of active warps that were stalled waiting to be selected to fetch an instruction or waiting on an icache miss

THE OPAQUE FLAGS

- D3D12_RAYTRACING_GEOMETRY_FLAG_OPAQUE
  - D3D12_RAYTRACING_INSTANCE_FLAG_FORCE_OPAQUE
    - RAY_FLAG_FORCE_OPAQUE

- Indicates whether triangles are fully opaque or not
  - Stop rays earlier
  - Helps reduce the number of evaluated materials


Battlefield V DXR
Original OPAQUE flags
DispatchRays: 1.13 ms
GeForce RTX 2080 +
SetStablePowerState
SECURE CHOKEPOINT KILO
20

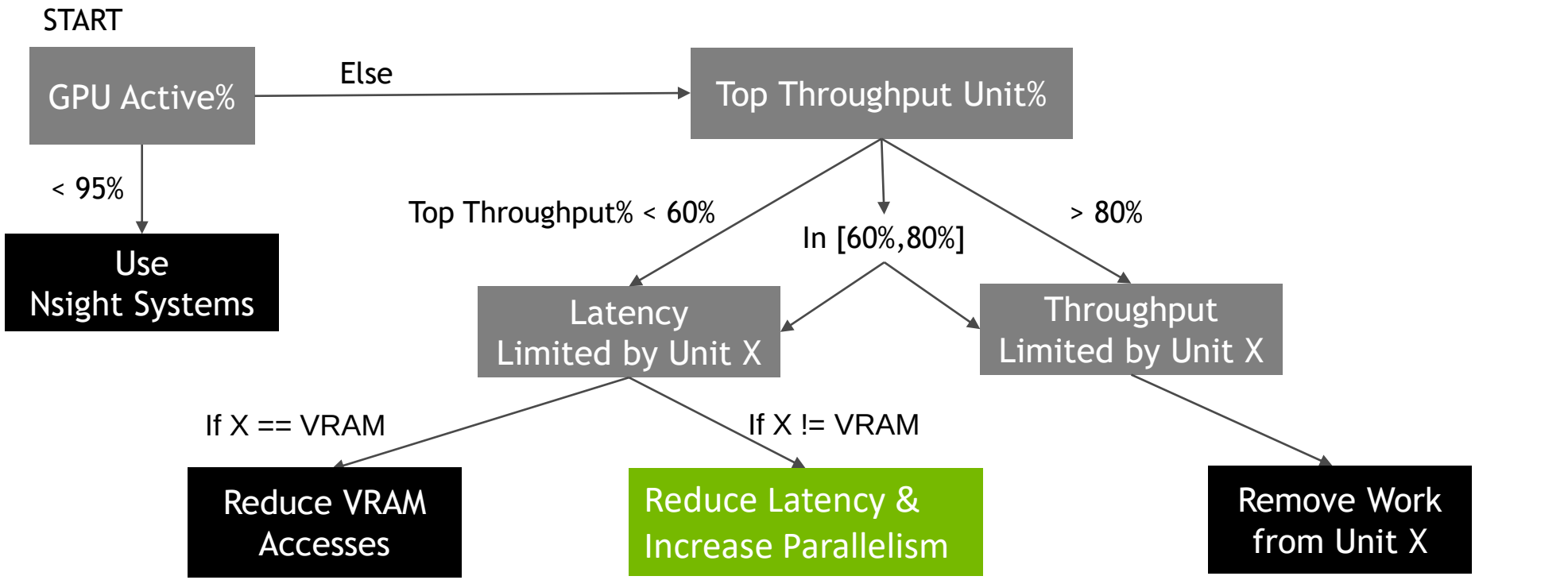

Battlefield V DXR
+OPAQUE flag for ALL geometries
DispatchRays: 0.58 ms (1.95x)
GeForce RTX 2080 +
SetStablePowerState
SECURE CHOKEPOINT KILO
20

FORCE_OPAQUE EXPERIMENT

	BEFORE	AFTER	RATIO
GPU Elapsed Time	1.13 ms	0.58 ms	**1.95x Gain**
Top Throughput: L2	23.8%	24.0%	1.01x
SM Active Threads Per Instruction Executed	29.0%	40.5%	**1.40x**
SM Instructions Executed Per Warp	725.1	425.2	0.59x
SM Warp Issue Stalls, No Instruction	69.8%	68.2%	0.98x

On RTX 2080 with SetStablePowerState(TRUE)

THE P3 (PEAK-PERF%) METHOD



Increase cache hit rates
Reduce texture formats

Increase cache hit rates
Use async queues
Reduce warp issue stall reasons
Reduce warp launch stall reasons
Reduce instructions executed
…

If FMA: remove FP32 +/*

NVIDIA TOOLS FOR GPU PROFILING

Two Nsight: Graphics modules

	GPU TRACE METRICS	RANGE PROFILER
Graphs over time	Y	N
Async queue support	**Unbiased**	Serialized
APIs	DX12 only so far	DX12, DX11 and Vk
GPUs	>= Turing	>= Kepler

CONCLUSION

The P3 Method (Peak-Perf-Percentage)

- A **method** to triage the performance of **any** GPU workload:
  - Start from the « Top Throughput% » Metrics (aka SOL% or Peak-Perf%)
  - Do NOT start from SM Warp Occupancy
  - We have been keeping a blog post up-to-date as tools and GPUs evolve.
- Async Compute rule of thumb:
  - Do not overlap 2 VRAM-latency-limited workloads

QUESTIONS?

Louis Bavoil | lbavoil@nvidia.com

www.nvidia.com/GDC