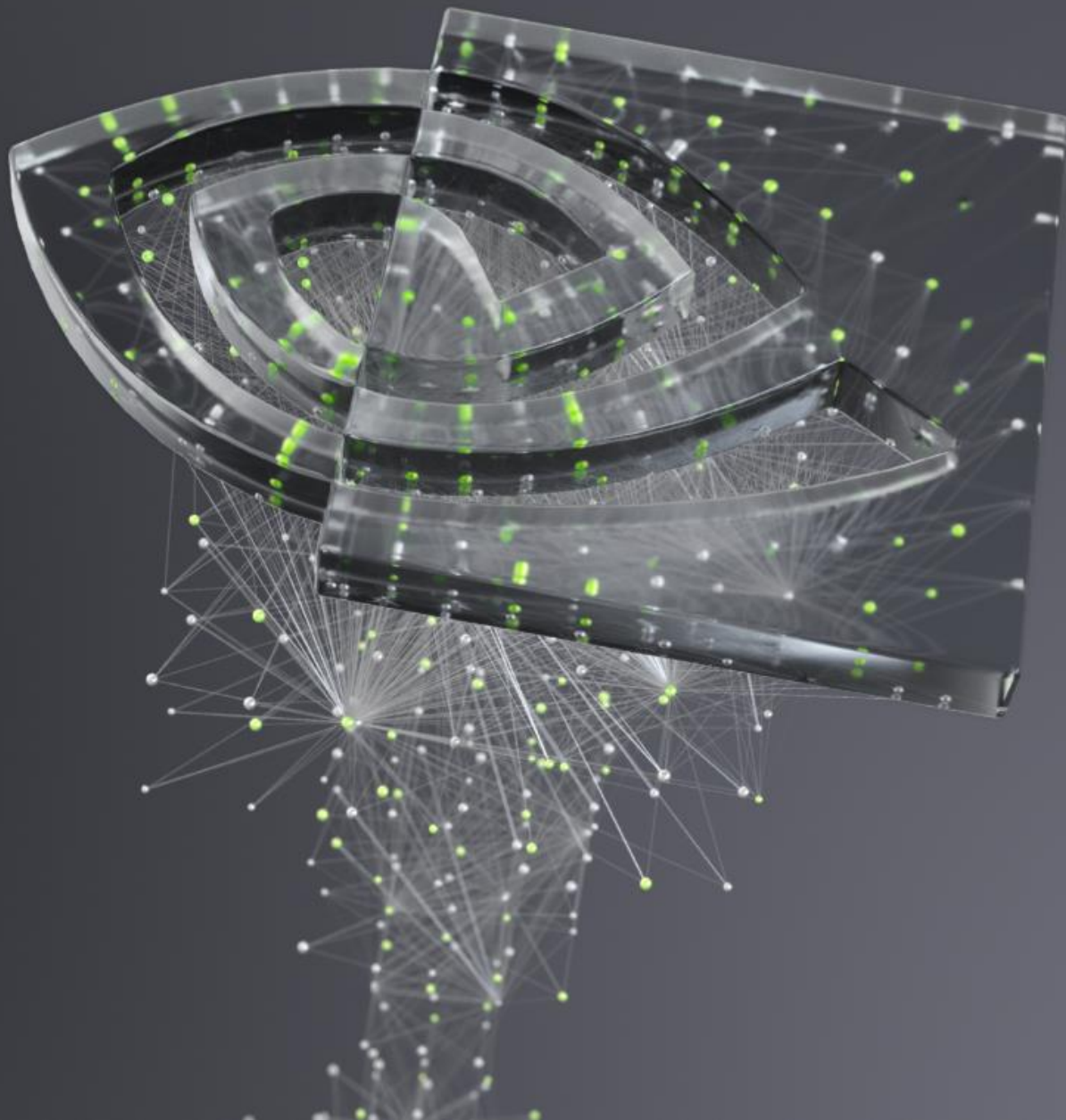




Nsight Graphics 方法及建议

梁佳晨 2019.12.18





概要

Nsight Graphics

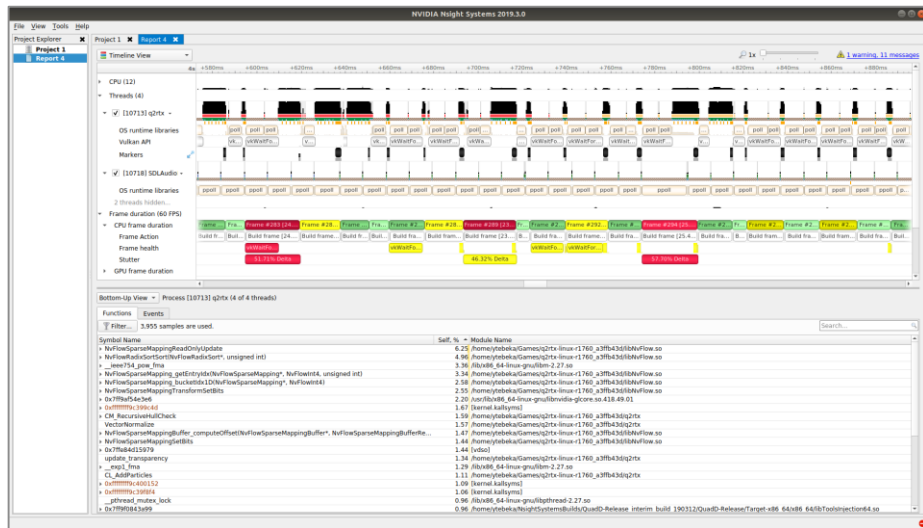
峰值性能百分比分析法

案例学习

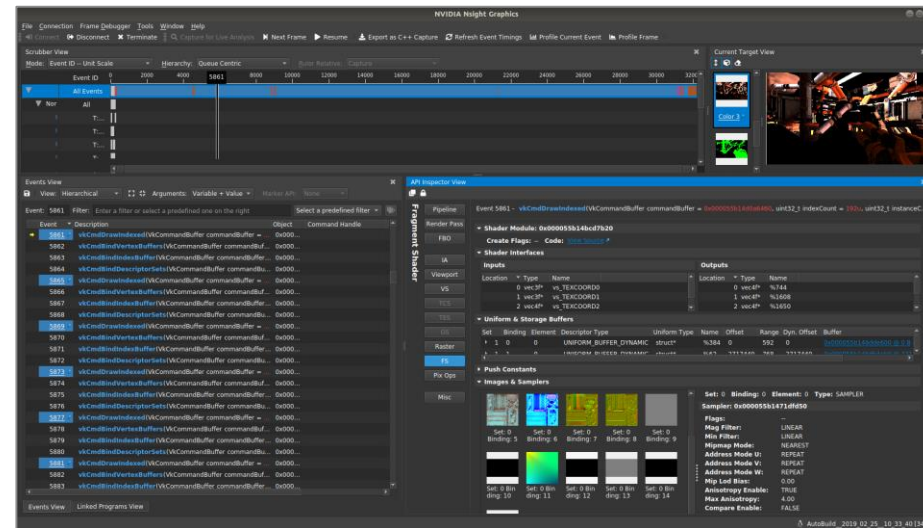


Nsight Graphics

NVIDIA 开发工具

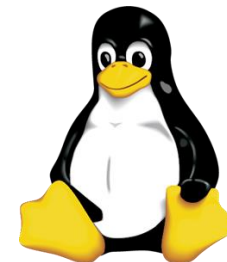
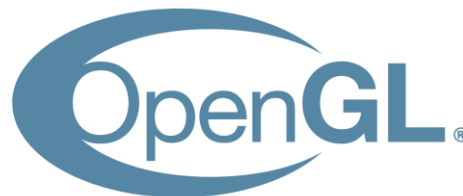
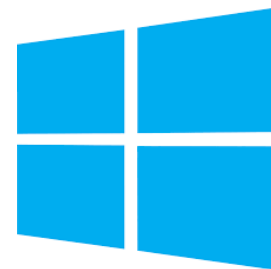


Nsight Systems



Nsight Graphics

Welcome To Nsight Graphics

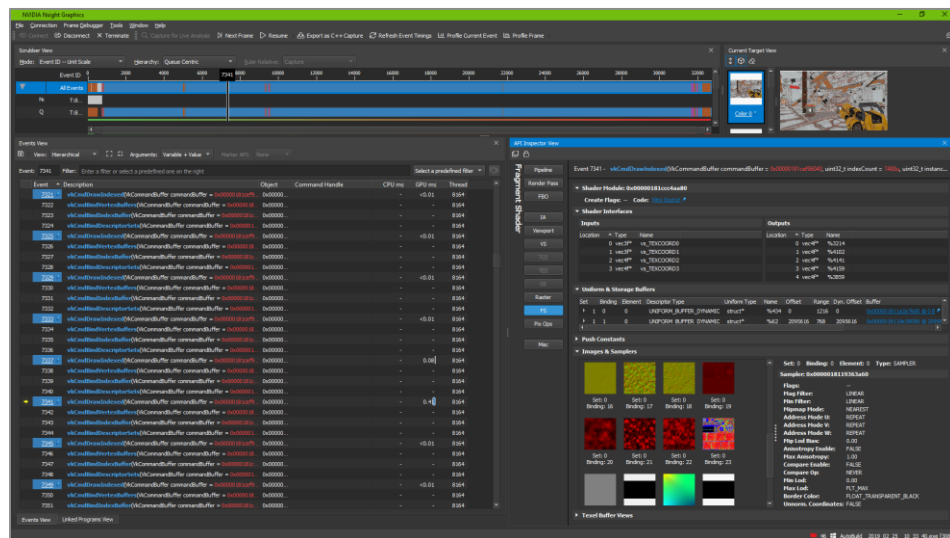


Work Flow

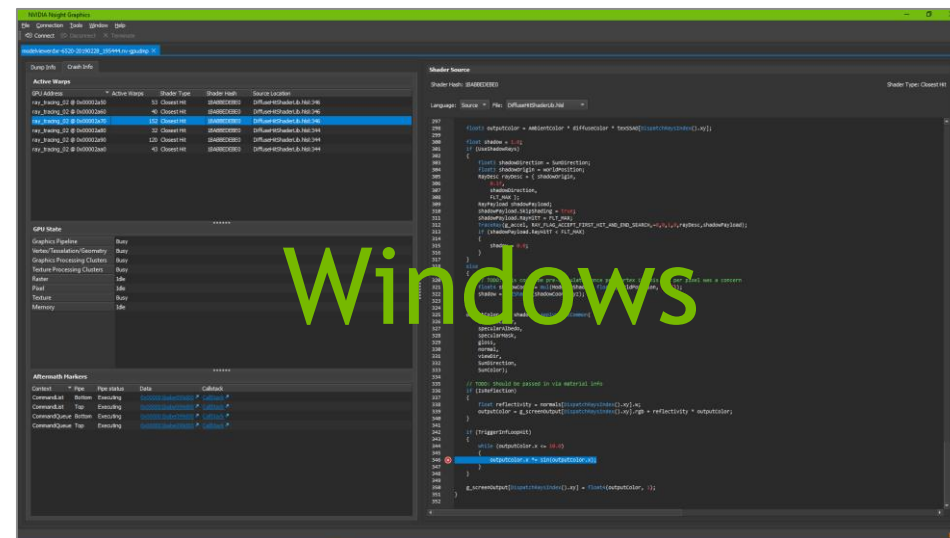




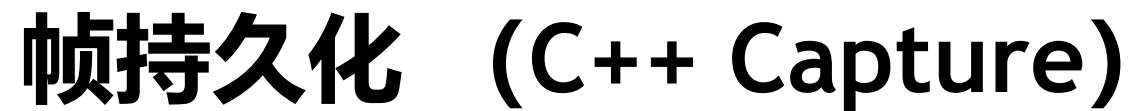
调试



Frame Debugger



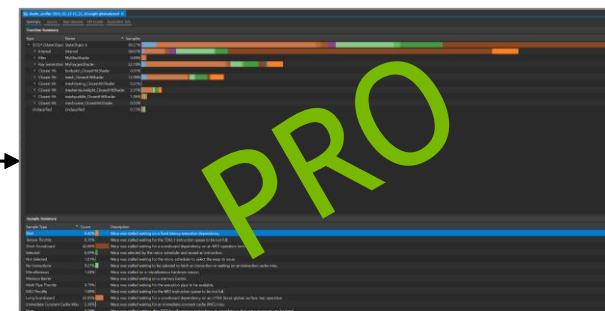
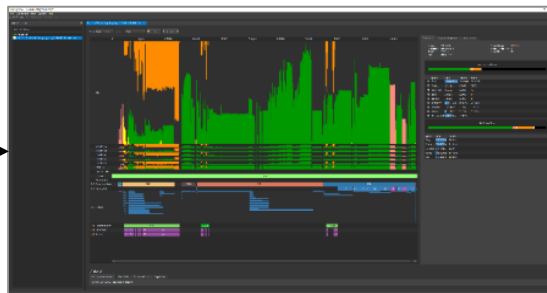
Aftermath





性能分析

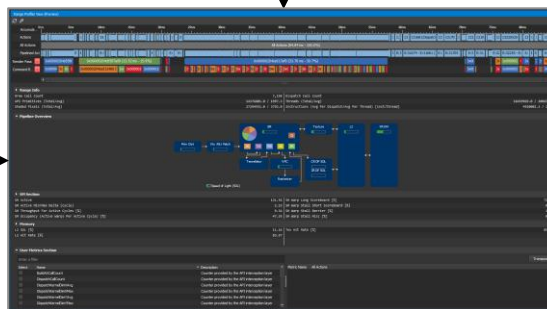
GPU Trace



Shader Profiler

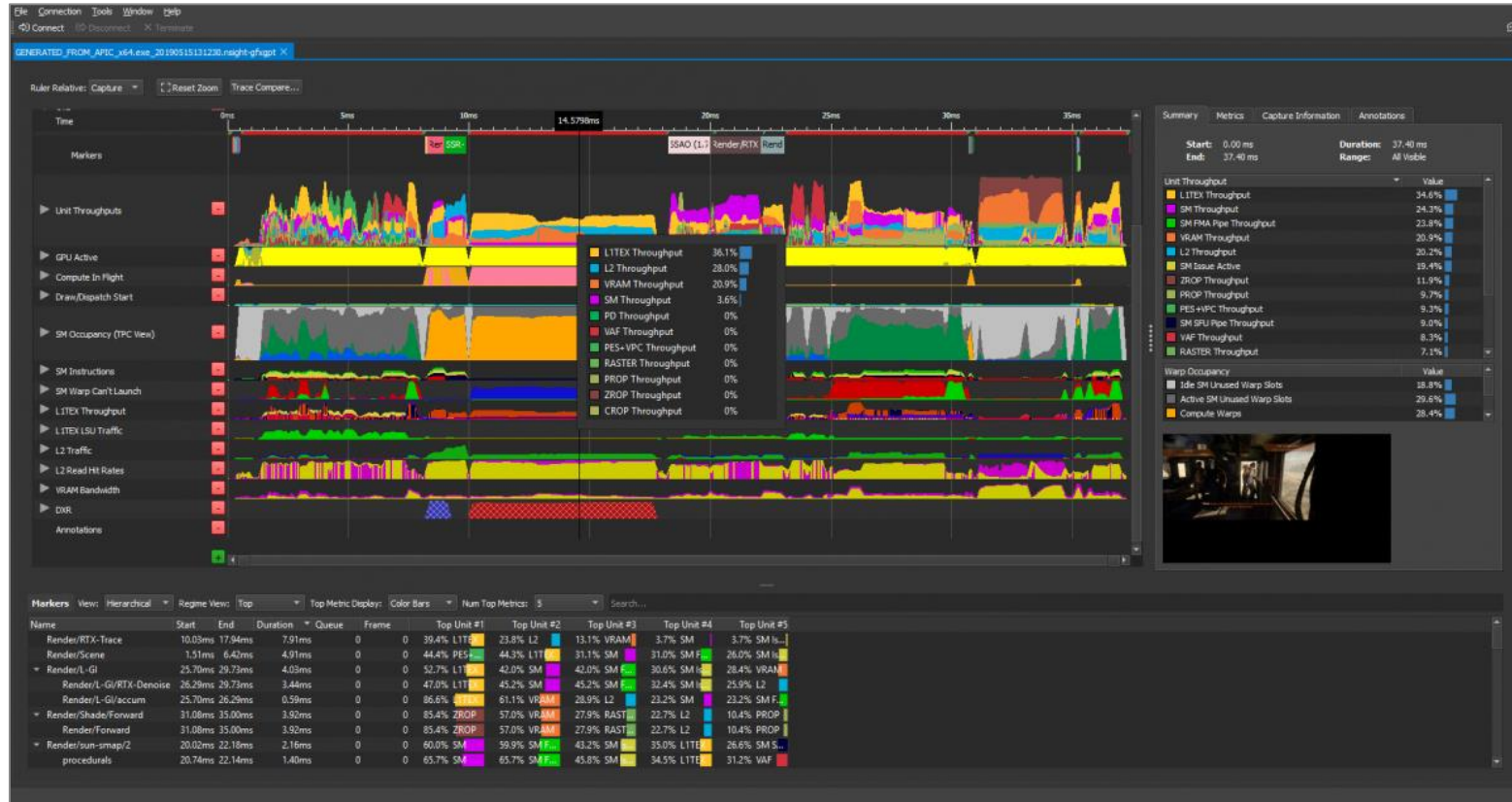


Nsight Systems

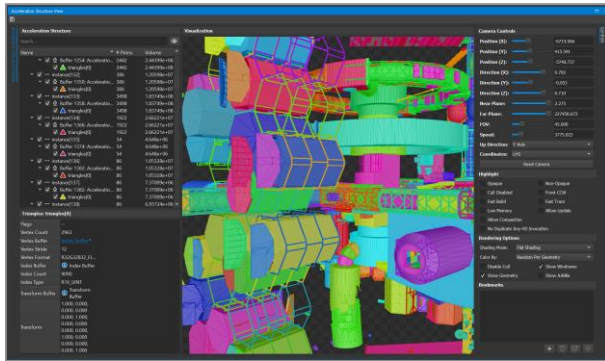


Range Profiler

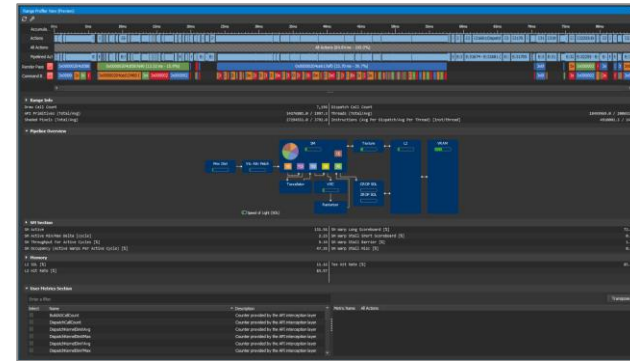
GPU Trace



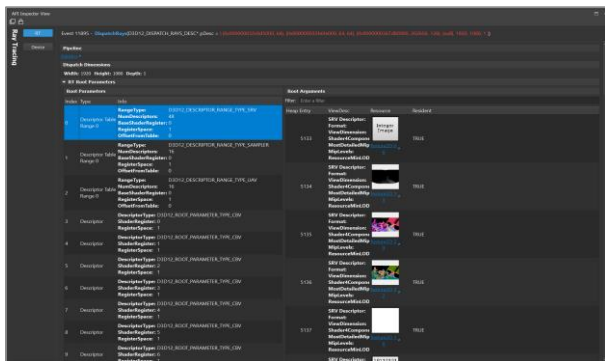
Ray Tracing Features



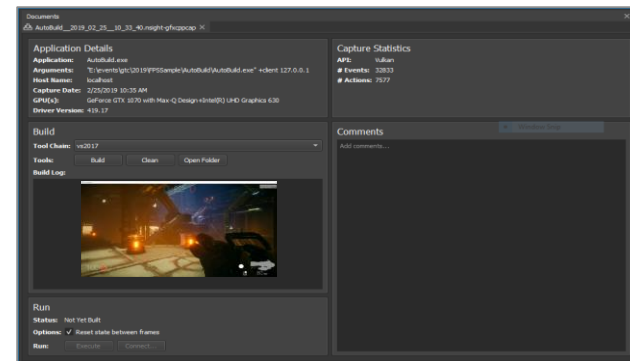
Acceleration Structure Viewer



Range Profiler



Resource Inspector



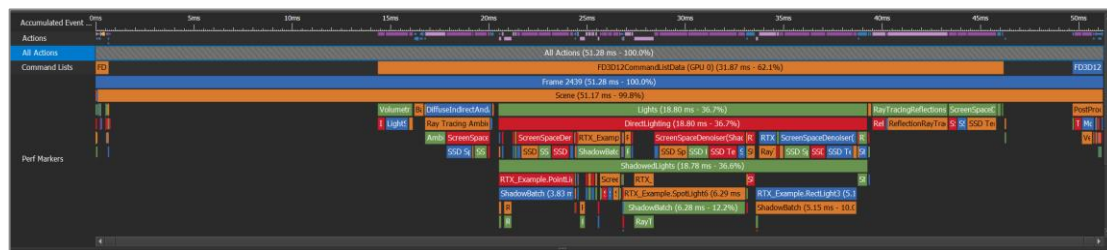
C++ Capture



峰值性能百分比分析法

Peak-Performance-Percentage (P3)

STEPS

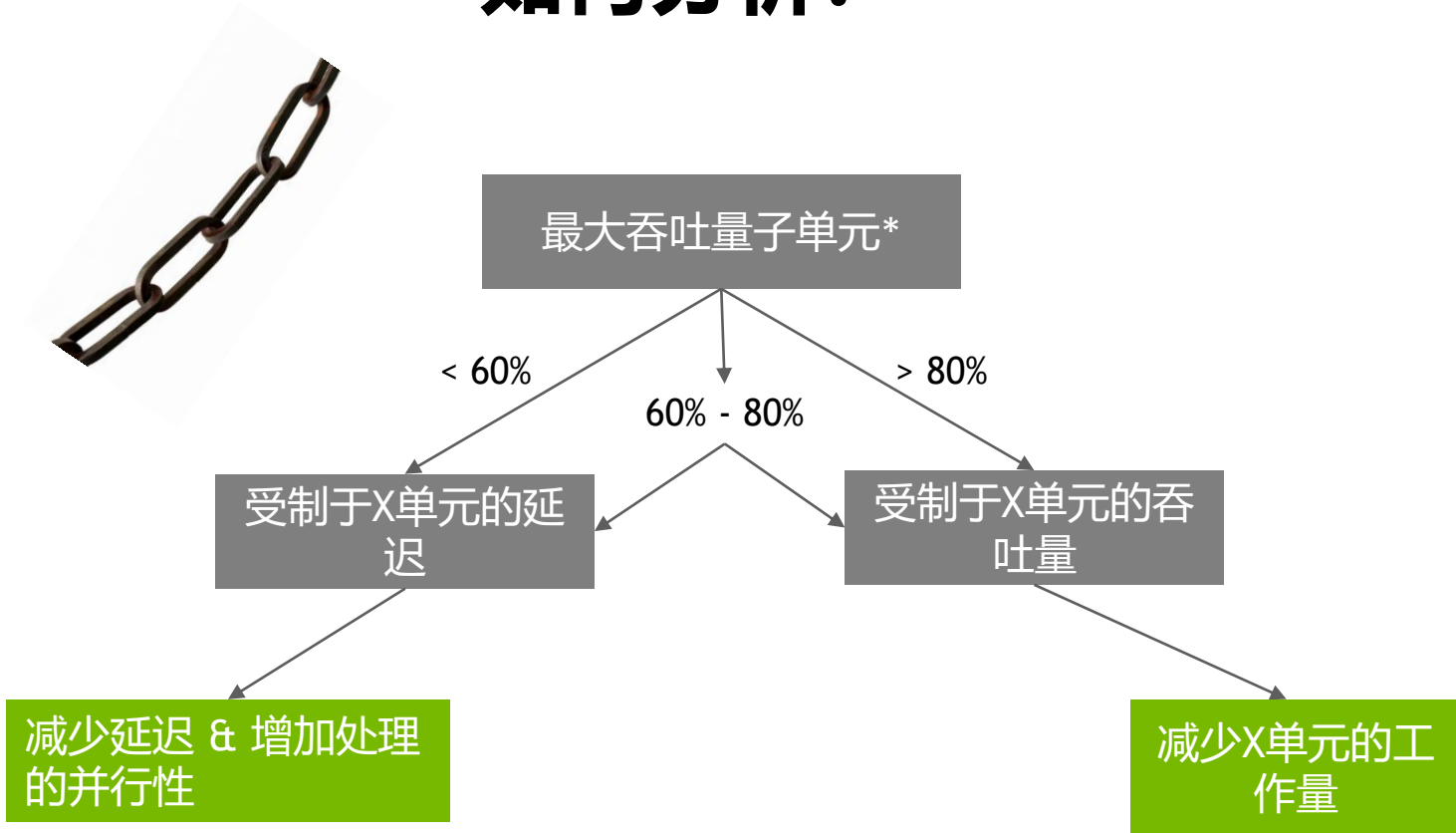


Draw Call Count	306	Dispatch Call Count	51
Input Primitives	269,655	Dispatch Total Threads	0
Shaded Fragments (thread)	25,343,651	CS Instruction Count [inst]	2,493,432,523.50
Early Z Passed	34,163,265	GPU Adapter	NVIDIA GeForce RTX 2060
Early Z Failed	7,905,915		
Input Primitives	242,644		
Output Primitives	135,487		
ZCull Samples Tested	91,176,448		
ZCull Samples Failed	3,478,696		

Recommendations

Range Info More draw calls than dispatches

如何分析?



*吞吐量(Throughput) == SOL%

MORE

NVIDIA Developer Blog

The Peak-Performance-Percentage Analysis Method for Optimizing Any GPU Workload

By Louis Bavoil

<https://devblogs.nvidia.com/the-peak-performance-analysis-method-for-optimizing-any-gpu-workload/>



案例学习一

RTAO DENOISER PIXEL SHADER

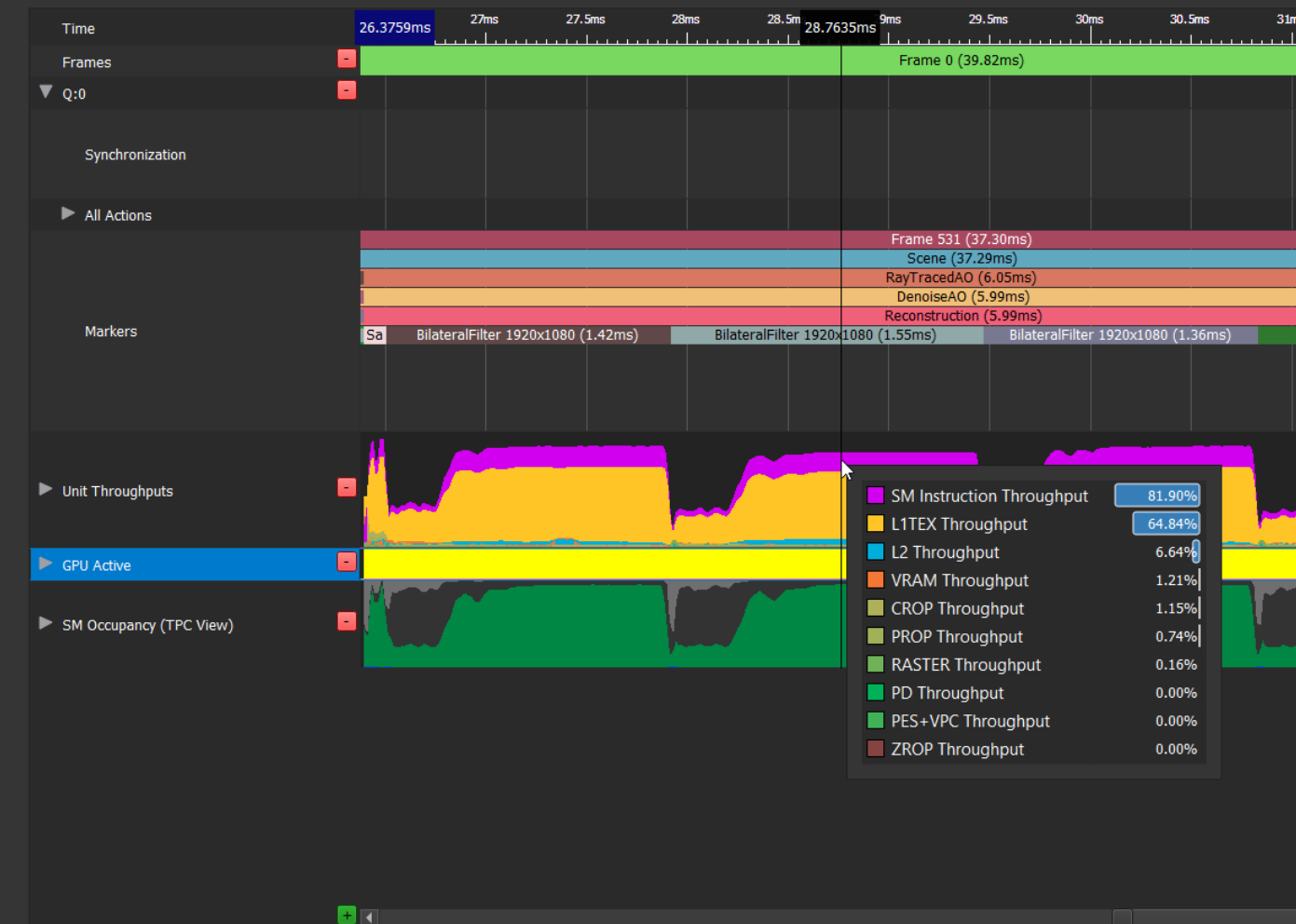
RTAO DENOISER PIXEL SHADER



Nsight Graphics: GPU Trace

DenoiseAO_Before.nsgt-gfxgpt X

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



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 Pine 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%	

Average Values for Current Range

子单元吞吐量

Turing GPU only

SM: 70%

L1TEX: 57%

L2: 5%

VRAM: 2%

CROP: 2%



SM FMA Pipe: 70%

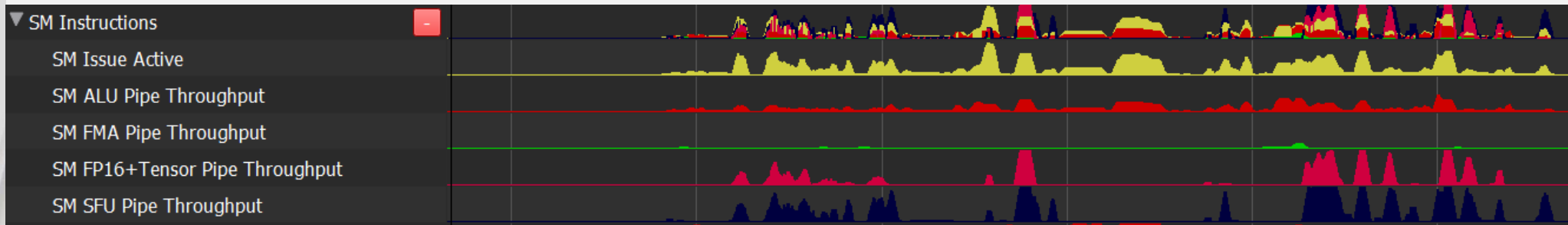
SM SFU Pipe: 52%

SM ALU Pipe: 25%

SM FP16 Pipe: 0.0%



FMA-Pipe-Throughput
Limited



移除FMA指令

```
#if 0
```

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

```
#else
```

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

```
#endif
```

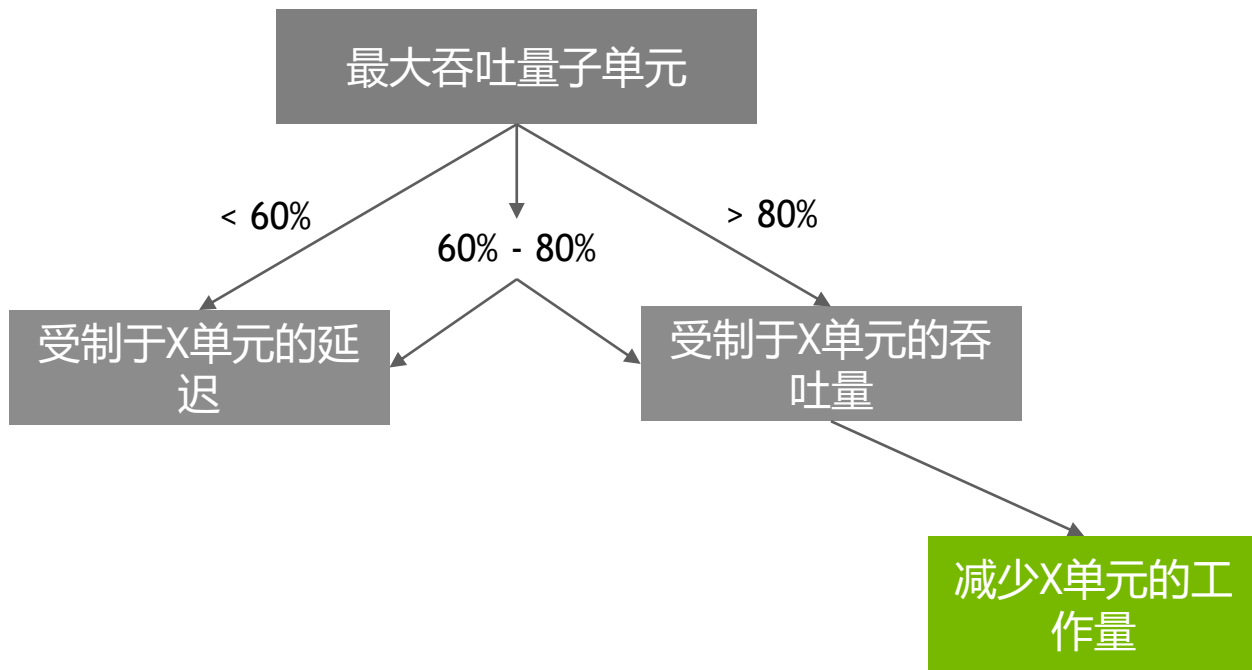
移除FMA指令

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 METHOD





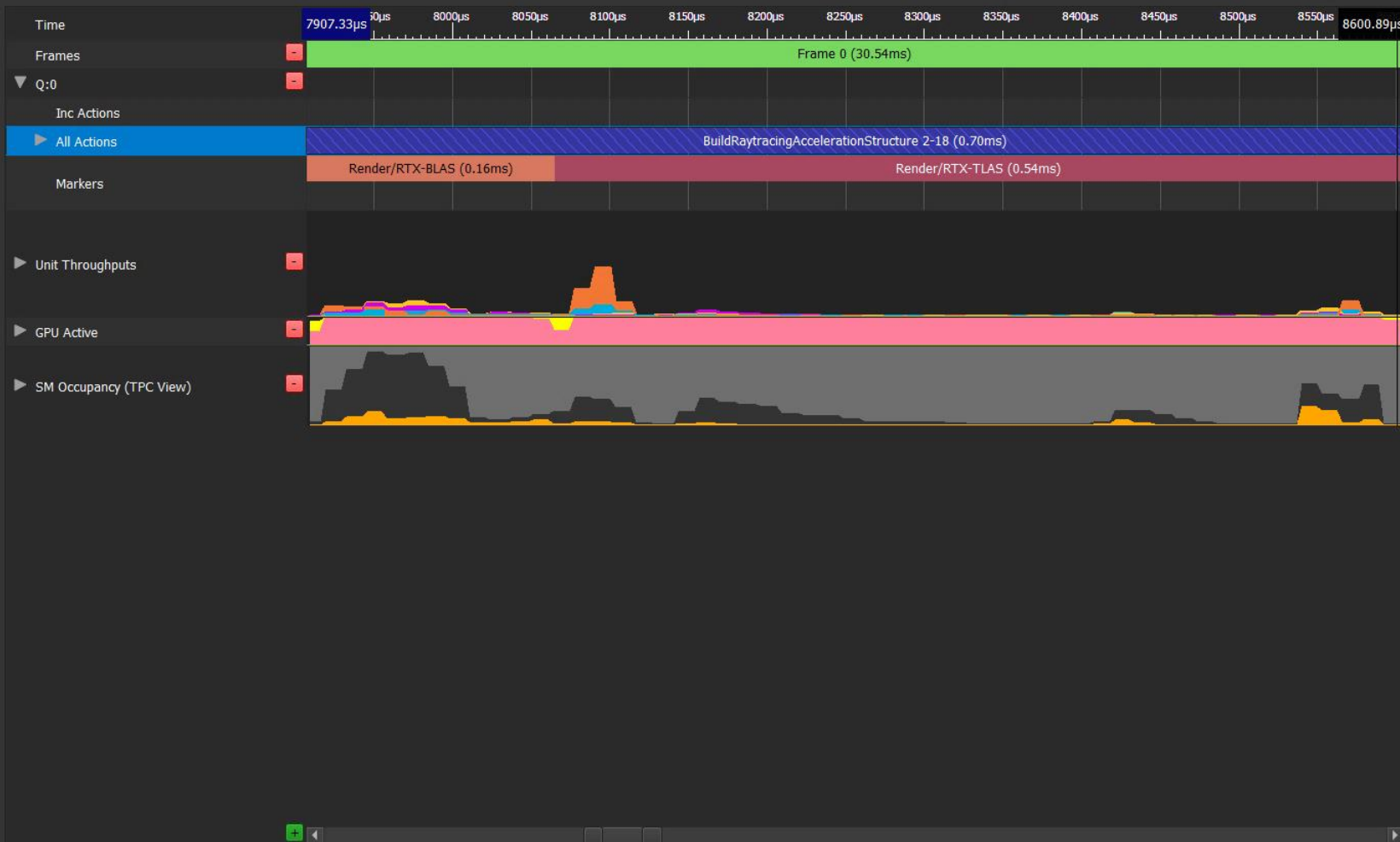
案例学习二 METRO EXODUS

RTAS Updates

Connect Disconnect Terminate

2-MetroBHV-Before.nsiht-gfxgpt X

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



Summary Metrics Detailed Metrics Capture Information

Start: 7.91 ms **Duration:** 0.70 ms
End: 8.61 ms **Range:** All Visible

Unit Throughput	Value
VRAM Throughput	3.03%
L1TEX Throughput	2.34%
SM Instruction Throughput	2.10%
SM Issue Active	2.05%
SM ALU Pipe Throughput	1.78%
L2 Throughput	1.59%
SM FMA Pipe Throughput	0.89%
SM SFU Pipe Throughput	0.72%
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 Pine Throughput	0.00%
Warp Occupancy	Value
Idle SM Unused Warp Slots	78.22%
Active SM Unused Warp Slots	18.35%
Compute Warps	3.43%
Pixel Warps	0.00%
Vertex+Tess+Geom Warps	0.00%

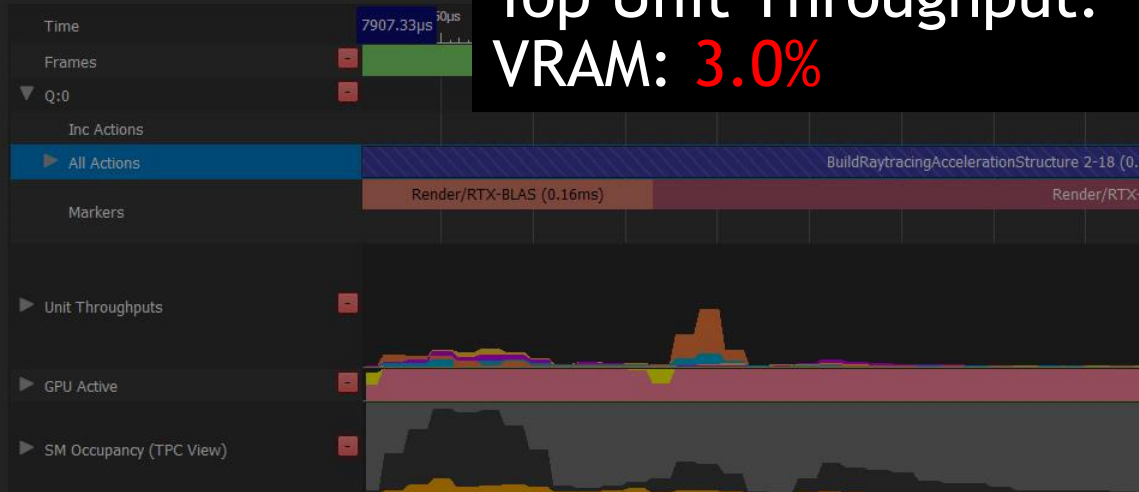
RTAS Updates

Connect Disconnect Terminate

2-MetroBHV-Before.nsisight-gfxgpt X

Color By: Stages Ruler Relative: Capture

Top Unit Throughput:
VRAM: 3.0%



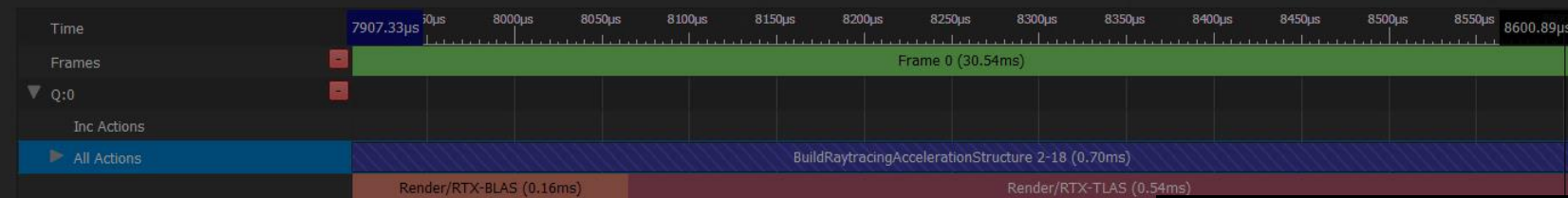
Unit Throughput	Value
VRAM Throughput	3.03%
L1TEX Throughput	2.34%
SM Instruction Throughput	2.10%
SM Issue Active	2.05%
SM ALU Pipe Throughput	1.78%
L2 Throughput	1.59%
SM FMA Pipe Throughput	0.89%
SM SFU Pipe Throughput	0.72%
PROP Throughput	0.00%
PD Throughput	0.00%
PES+VPC Throughput	0.00%
RASTER Throughput	0.00%
ZROP Throughput	0.00%
CROP Throughput	0.00%
SM FP16+Tensor Pipe Throughput	0.00%

RTAS Updates

Connect Disconnect Terminate

2-MetroBHV-Before.nsiht-gfxgpt X

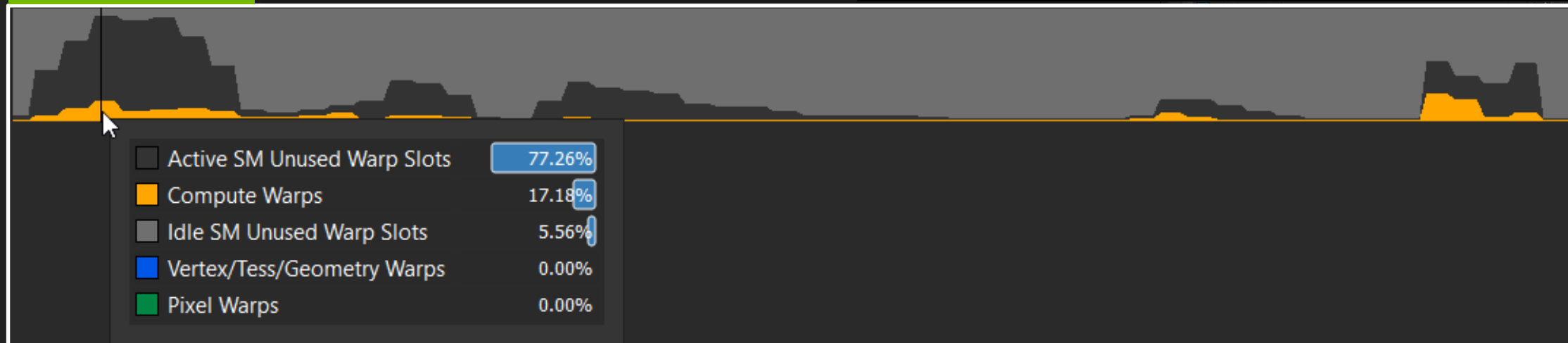
Color By: Stages Ruler Relative: Capture Export... Reset Zoom Trace Compare...



Summary	Metrics	Detailed Metrics	Capture Information
Start: 7.91 ms	Duration: 0.70 ms		
End: 8.61 ms	Range: All Visible		
Unit Throughput		Value	
VRAM Throughput		3.03%	
L1TEX Throughput		2.34%	

SM Occupancy

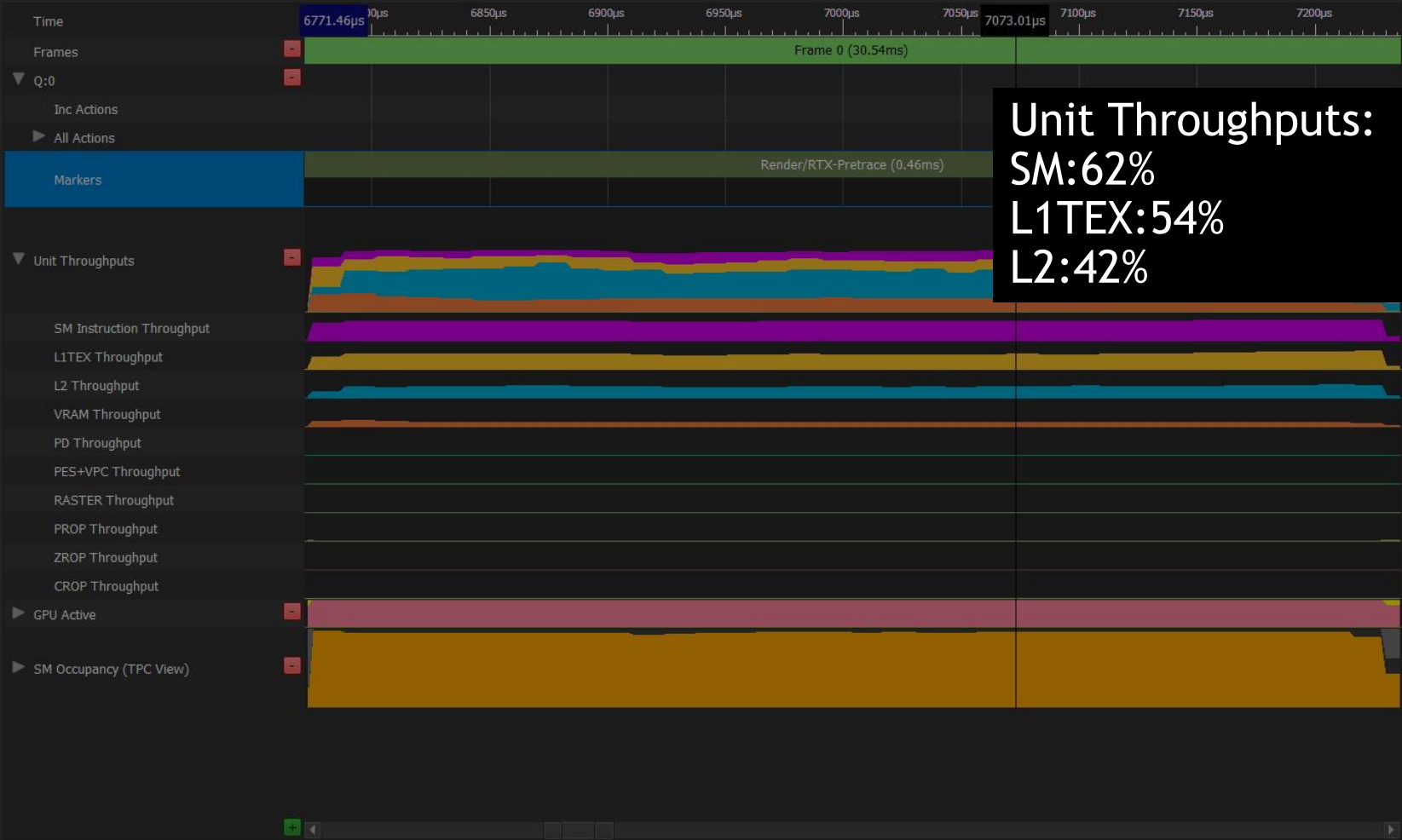
Average Warp Occupancy For Workload: 3.4%



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

Independent Workload #1: Screen-Space PreTracing

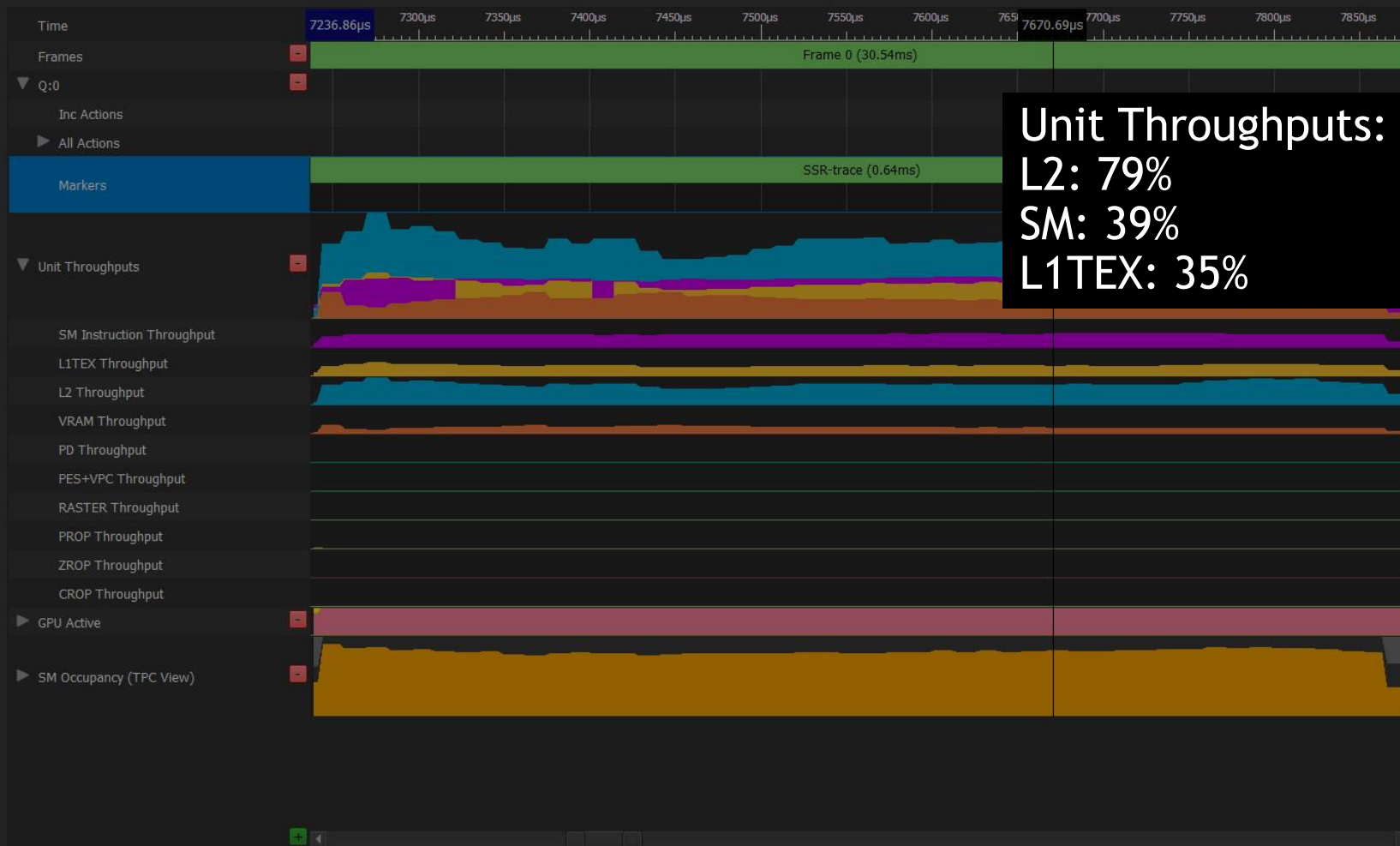
Color By: Stages Ruler Relative: Capture Export... Reset Zoom Trace Compare...



Summary	Metrics	Detailed Metrics	Capture Information
Start: 6.77 ms		Duration: 0.46 ms	
End: 7.24 ms		Range: All Visible	
Unit Throughput		Value	
SM Instruction Throughput		61.51%	
SM FMA Pipe Throughput		61.46%	
L1TEX Throughput		53.95%	
SM Issue Active		49.80%	
L2 Throughput		42.65%	
SM SFU Pipe Throughput		33.86%	
SM ALU Pipe Throughput		15.30%	
VRAM Throughput		12.70%	
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		1.03%	
Active SM Unused Warp Slots		4.46%	
Compute Warps		94.51%	
Pixel Warps		0.00%	
Vertex+Tess+Geom Warps		0.00%	

Independent Workload #2: SSR

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



Unit Throughputs:
L2: 79%
SM: 39%
L1TEX: 35%

Summary Metrics Detailed Metrics Capture Information

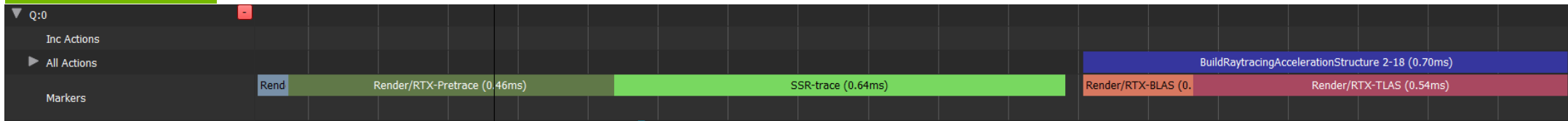
Start: 7.24 ms End: 7.88 ms Duration: 0.64 ms Range: All Visible

Unit Throughput	Value
L2 Throughput	78.76%
SM Instruction Throughput	38.82%
SM FMA Pipe Throughput	38.78%
SM Issue Active	35.99%
L1TEX Throughput	34.81%
SM ALU Pipe Throughput	20.84%
VRAM Throughput	18.17%
SM SFU Pipe Throughput	17.71%
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	1.49%
Active SM Unused Warp Slots	17.88%
Compute Warps	80.62%
Pixel Warps	0.00%
Vertex+Tess+Geom Warps	0.00%

ASync COMPUTE DIFF

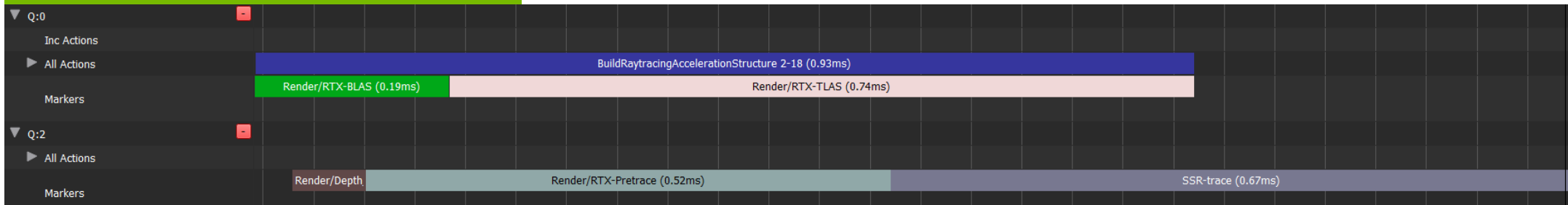
Serialized

1.83 ms



Overlapped (Async Compute)

1.30 ms



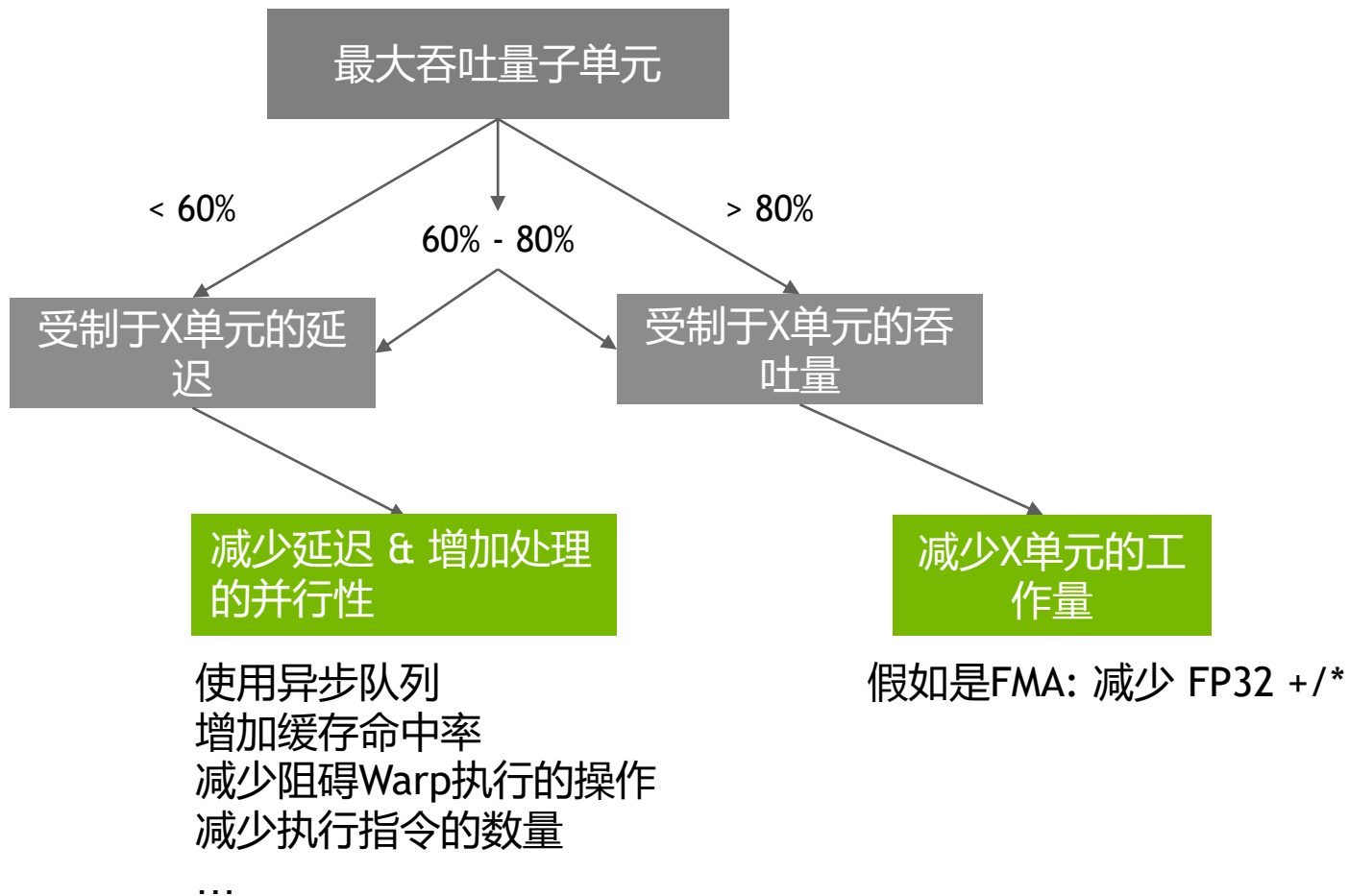
ASYNC-COMPUTE OVERLAP

(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)

THE P3 METHOD





案例学习三 阴影贴图//异步计算

HBAO + SSR + Light Culling

Connect Disconnect Terminate

1-BF5_AsyncCompute-Before.nsisight-gfxgpt X

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



Summary Metrics Detailed Metrics Capture Information

Start: 11.00 ms End: 11.97 ms Duration: 0.98 ms Range: All Visible

Unit Throughput	Value
VRAM Throughput	30.63%
SM Instruction Throughput	22.61%
L1TEX Throughput	19.96%
SM Issue Active	19.90%
L2 Throughput	18.30%
SM FMA Pipe Throughput	15.42%
SM SFU Pipe Throughput	12.62%
SM ALU Pipe Throughput	11.95%
CROP Throughput	0.30%
RASTER Throughput	0.15%
PROP Throughput	0.12%
ZROP Throughput	0.00%
PES+VPC Throughput	0.00%
PD Throughput	0.00%
SM FP16+Tensor Pipe Throughput	0.00%
Warp Occupancy	Value
Idle SM Unused Warp Slots	28.24%
Active SM Unused Warp Slots	10.56%
Compute Warps	61.24%
Pixel Warps	0.00%
Vertex+Tess+Geom Warps	0.00%



Independent Workload: Shadow Maps

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



Summary Metrics Detailed Metrics Capture Information

Start: 11.97 ms End: 13.44 ms Duration: 1.47 ms Range: All Visible

Unit Throughput	Value
L1TEX Throughput	38.34%
VRAM Throughput	36.68%
ZROP Throughput	29.10%
L2 Throughput	22.04%
SM Instruction Throughput	21.77%
SM Issue Active	20.54%
PROP Throughput	20.48%
SM FMA Pipe Throughput	17.52%
RASTER Throughput	17.42%
CROP Throughput	12.11%
SM ALU Pipe Throughput	8.69%
SM SFU Pipe Throughput	7.62%
PES+VPC Throughput	5.61%
PD Throughput	4.31%
SM FP16+Tensor Pipe Throughput	0.00%
Warp Occupancy	Value
Idle SM Unused Warp Slots	20.09%
Active SM Unused Warp Slots	21.51%
Compute Warps	0.05%
Pixel Warps	51.97%
Vertex+Tess+Geom Warps	6.37%



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)



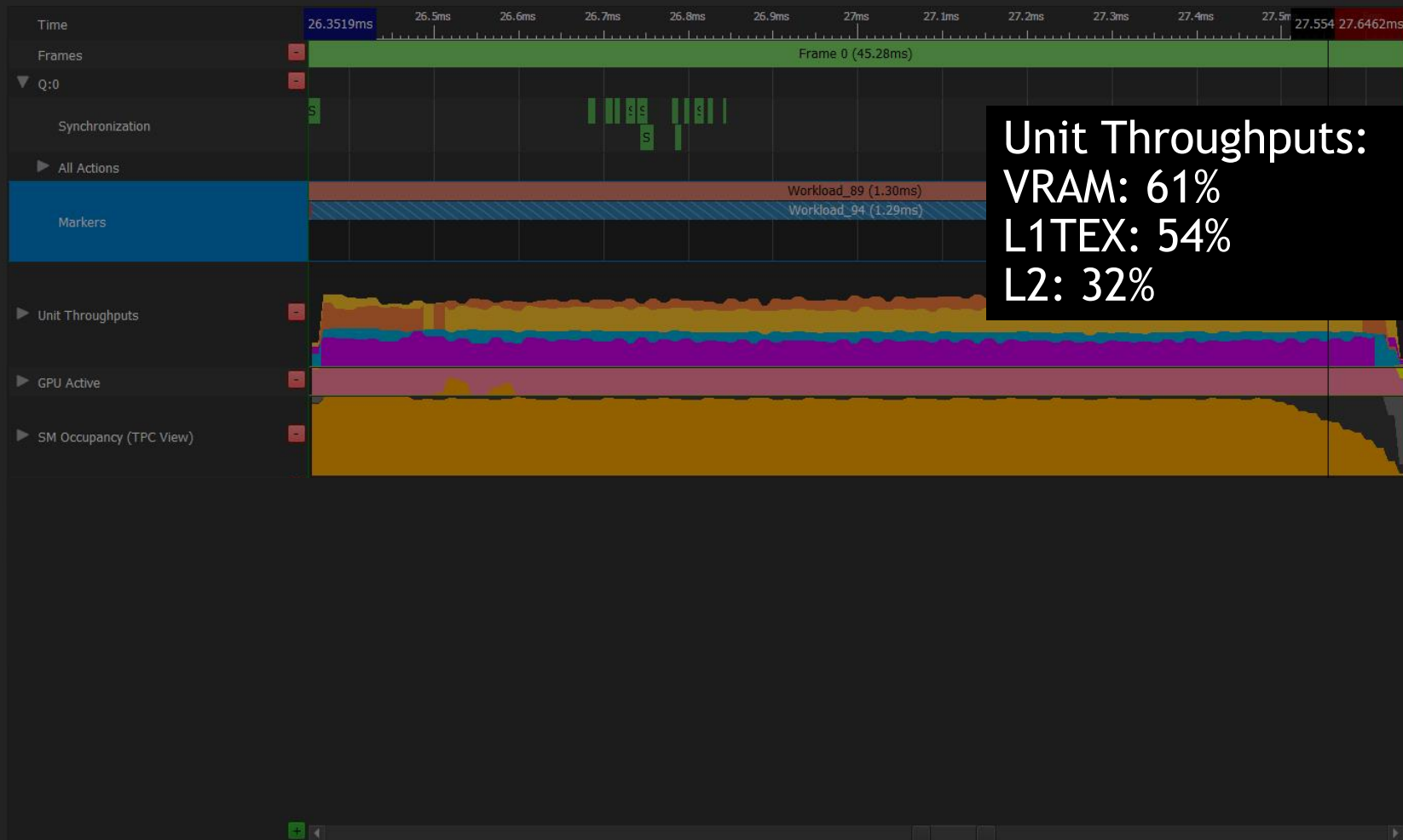
案例学习四 不宜使用异步计算的情况

Blur Compute Shader

Connect Disconnect Terminate

4-BadPairing-Before.nsisight-gfxgpt 4-BadPairing-After.nsisight-gfxgpt

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



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

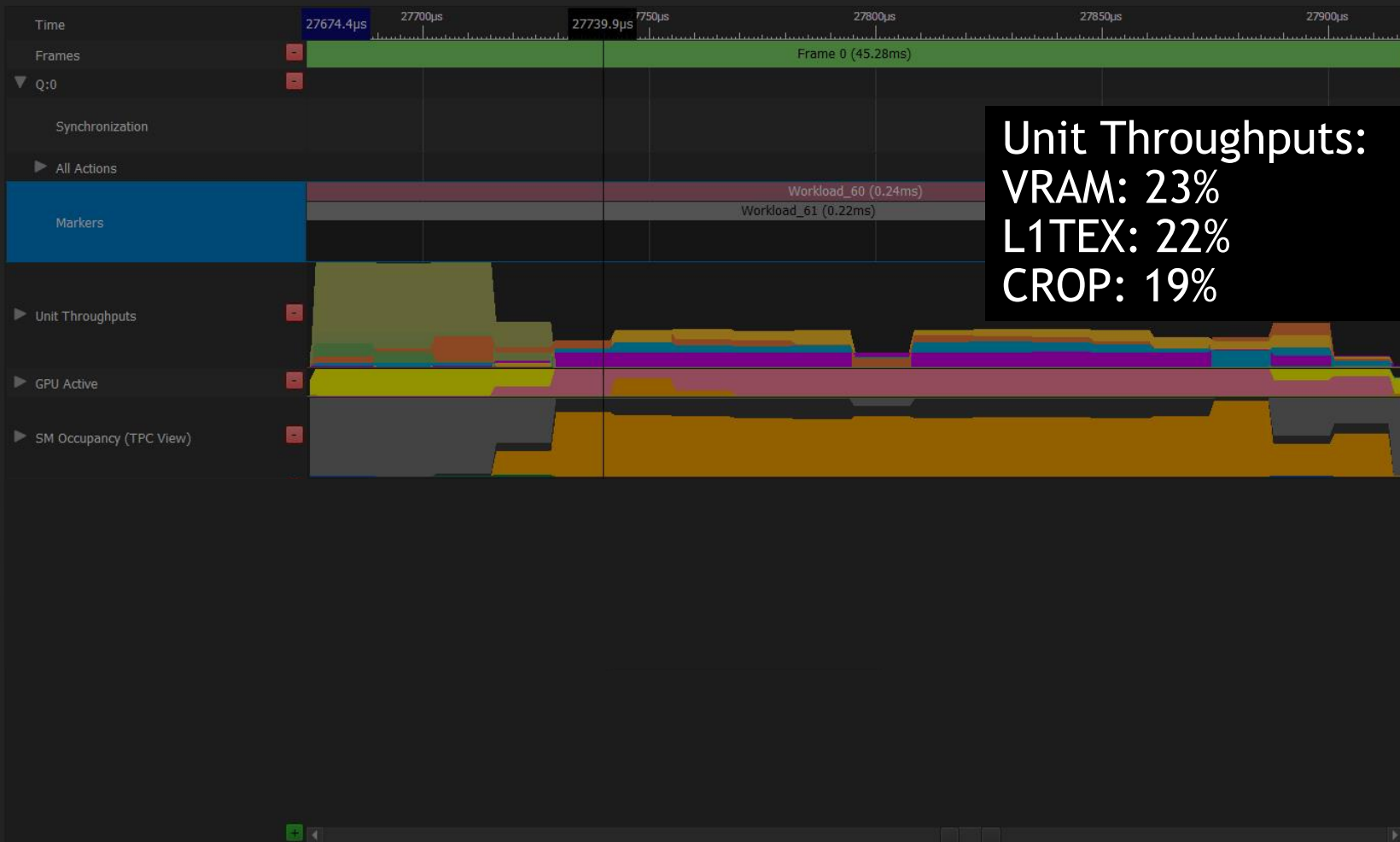
Summary Metrics Detailed Metrics Capture Information

Start: 26.35 ms End: 27.65 ms
Duration: 1.29 ms Range: Selected

Unit Throughput	Value
VRAM Throughput	60.98%
L1TEX Throughput	54.35%
L2 Throughput	31.89%
SM Instruction Throughput	23.92%
SM FMA Pipe Throughput	23.86%
SM Issue Active	19.11%
SM SFU Pipe Throughput	11.06%
SM ALU Pipe Throughput	7.17%
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	1.03%
Active SM Unused Warp Slots	5.63%
Compute Warps	93.34%
Pixel Warps	0.00%
Vertex+Tess+Geom Warps	0.00%

Independent Workload #1: Water Simulation

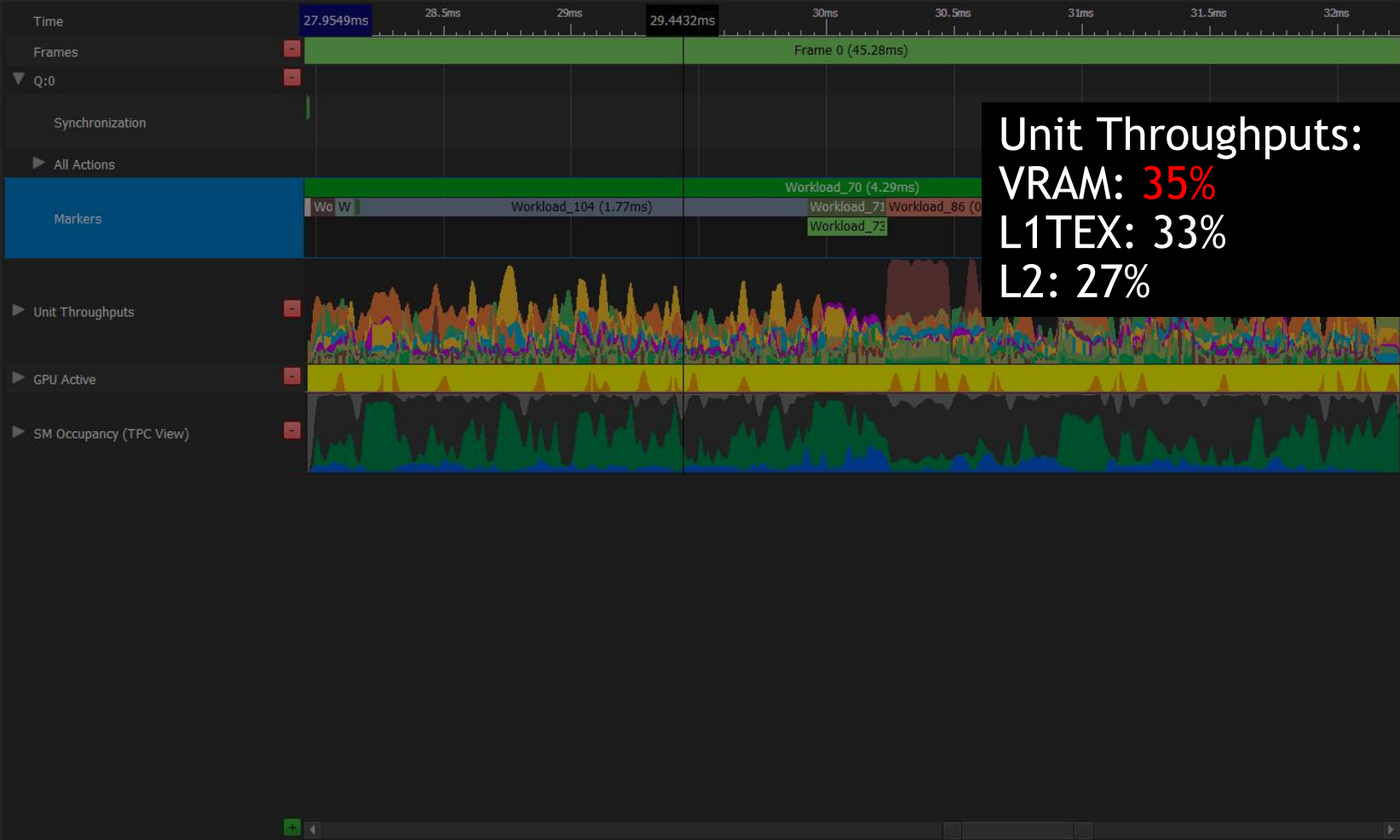
Color By: Stages Ruler Relative: Capture Export... Reset Zoom Trace Compare...



Summary	Metrics	Detailed Metrics	Capture Information
Start: 27.67 ms		Duration: 0.24 ms	
End: 27.92 ms		Range: All Visible	
Unit Throughput		Value	
VRAM Throughput		22.96%	
L1TEX Throughput		22.51%	
CROP Throughput		18.58%	
L2 Throughput		14.79%	
SM Instruction Throughput		10.27%	
SM Issue Active		10.27%	
PROP Throughput		6.16%	
SM ALU Pipe Throughput		5.01%	
RASTER Throughput		2.29%	
SM FMA Pipe Throughput		1.70%	
SM SFU Pipe Throughput		1.08%	
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		26.14%	
Active SM Unused Warp Slots		15.79%	
Compute Warps		57.90%	
Pixel Warps		0.17%	
Vertex+Tess+Geom Warps		0.00%	

Independent Workload #2: GBuffer Fill

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



Unit Throughputs:
VRAM: 35%
L1TEX: 33%
L2: 27%

Summary	Metrics	Detailed Metrics	Capture Information
Start: 27.95 ms		Duration: 4.29 ms	
End: 32.25 ms		Range: All Visible	
Unit Throughput		Value	
VRAM Throughput		34.57%	
L1TEX Throughput		33.05%	
L2 Throughput		26.56%	
PES+VPC Throughput		22.22%	
SM Instruction Throughput		22.07%	
SM Issue Active		20.99%	
ZROP Throughput		19.89%	
SM FMA Pipe Throughput		19.50%	
CROP Throughput		18.26%	
RASTER Throughput		10.37%	
PROP Throughput		9.39%	
SM SFU Pipe Throughput		8.77%	
SM ALU Pipe Throughput		8.12%	
PD Throughput		7.13%	
SM FP16+Tensor Pipe Throughput		0.00%	
Warp Occupancy		Value	
Idle SM Unused Warp Slots		5.74%	
Active SM Unused Warp Slots		50.30%	
Compute Warps		0.00%	
Pixel Warps		36.02%	
Vertex+Tess+Geom Warps		7.94%	

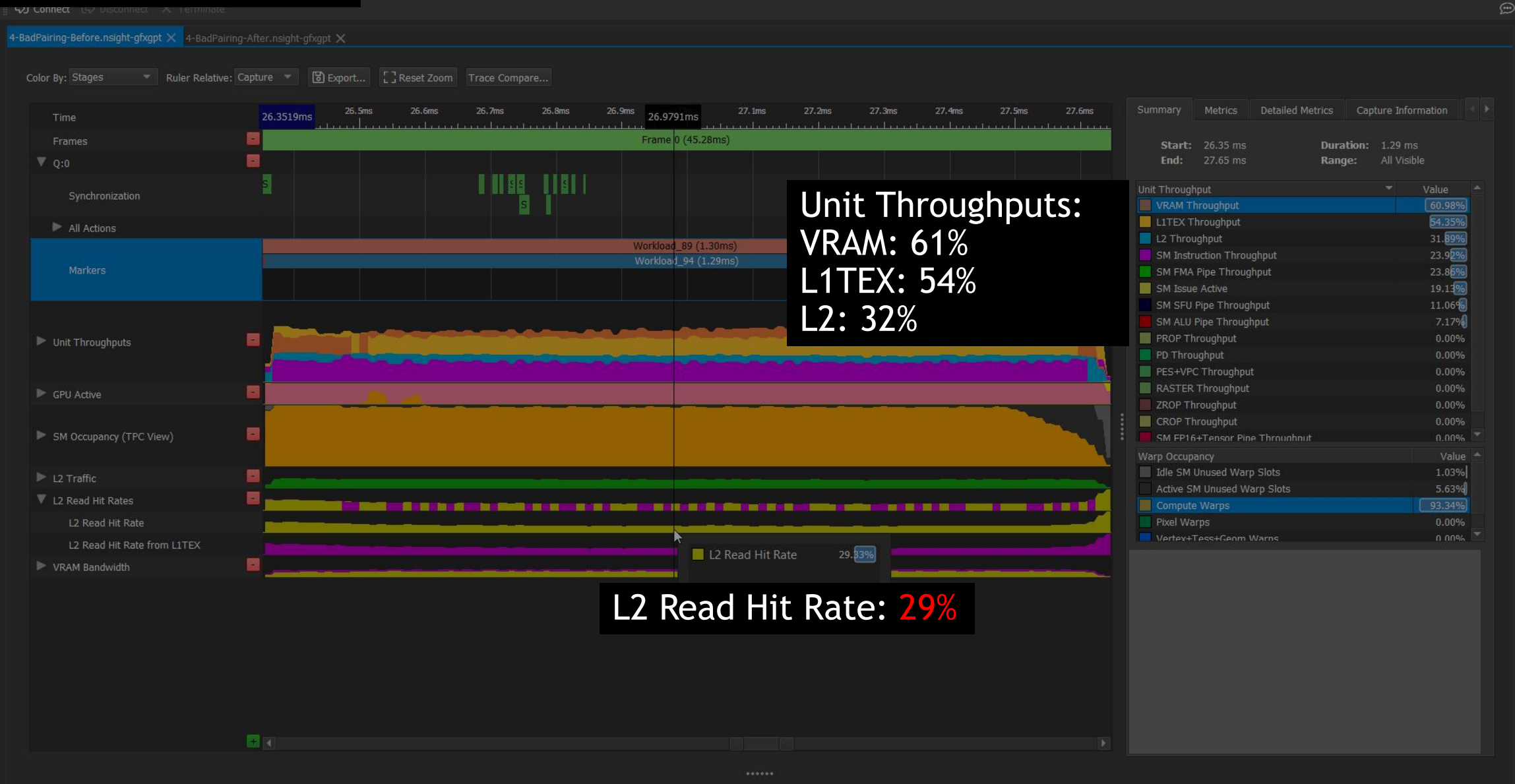
BAD ASYNC-COMPUTE PAIRING

(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

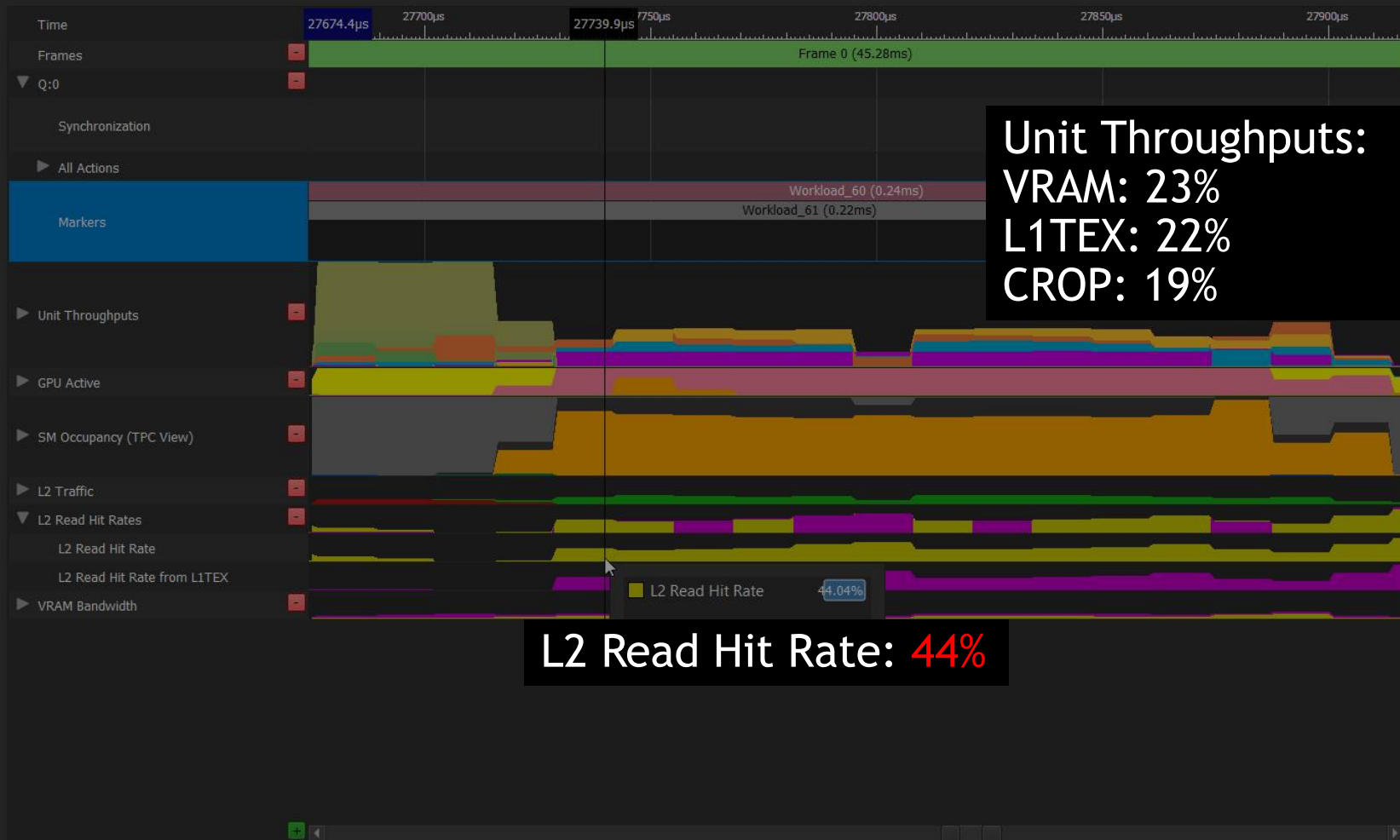
On RTX 2080 with SetStablePowerState(TRUE)

Blur Compute Shader



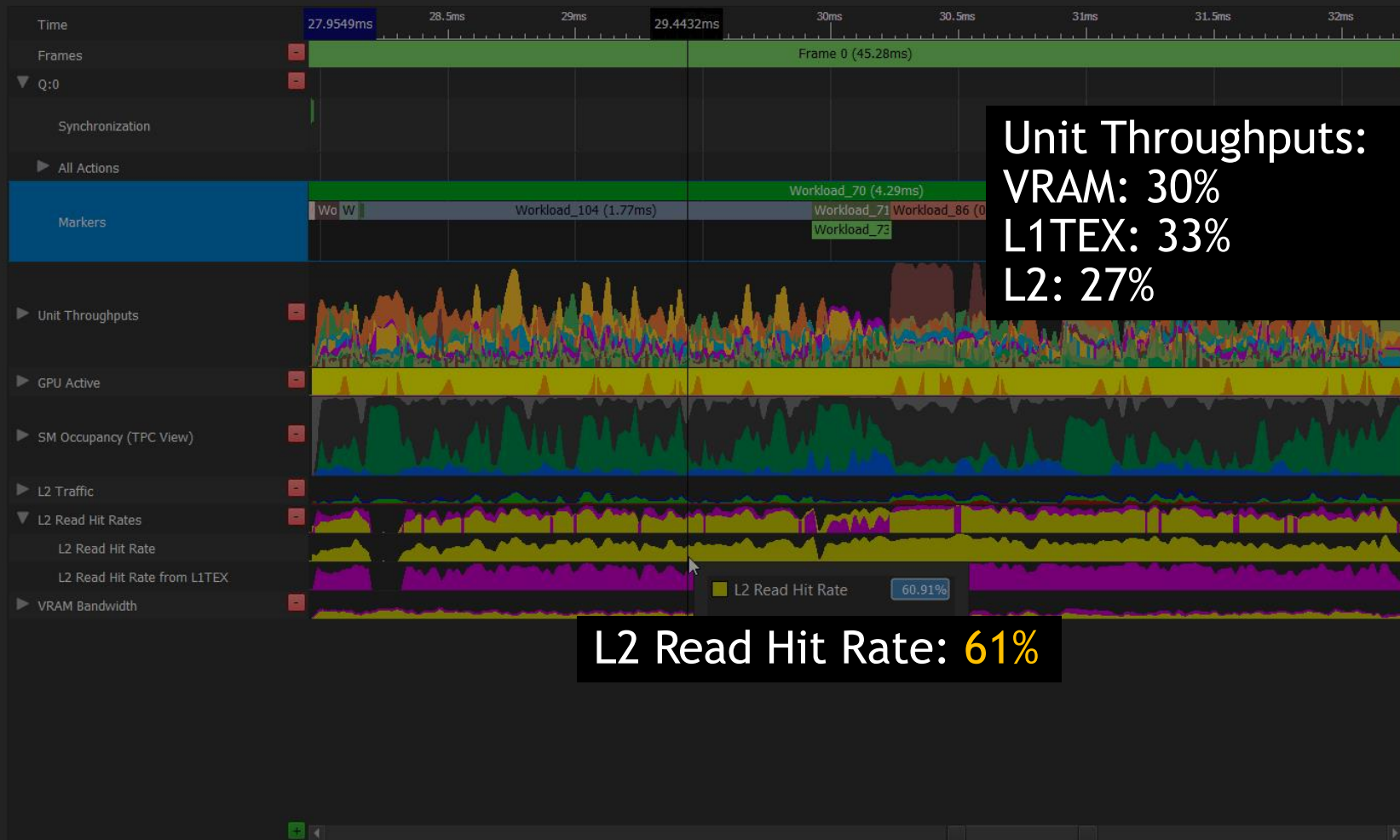
Independent Workload #1: Water Simulation

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



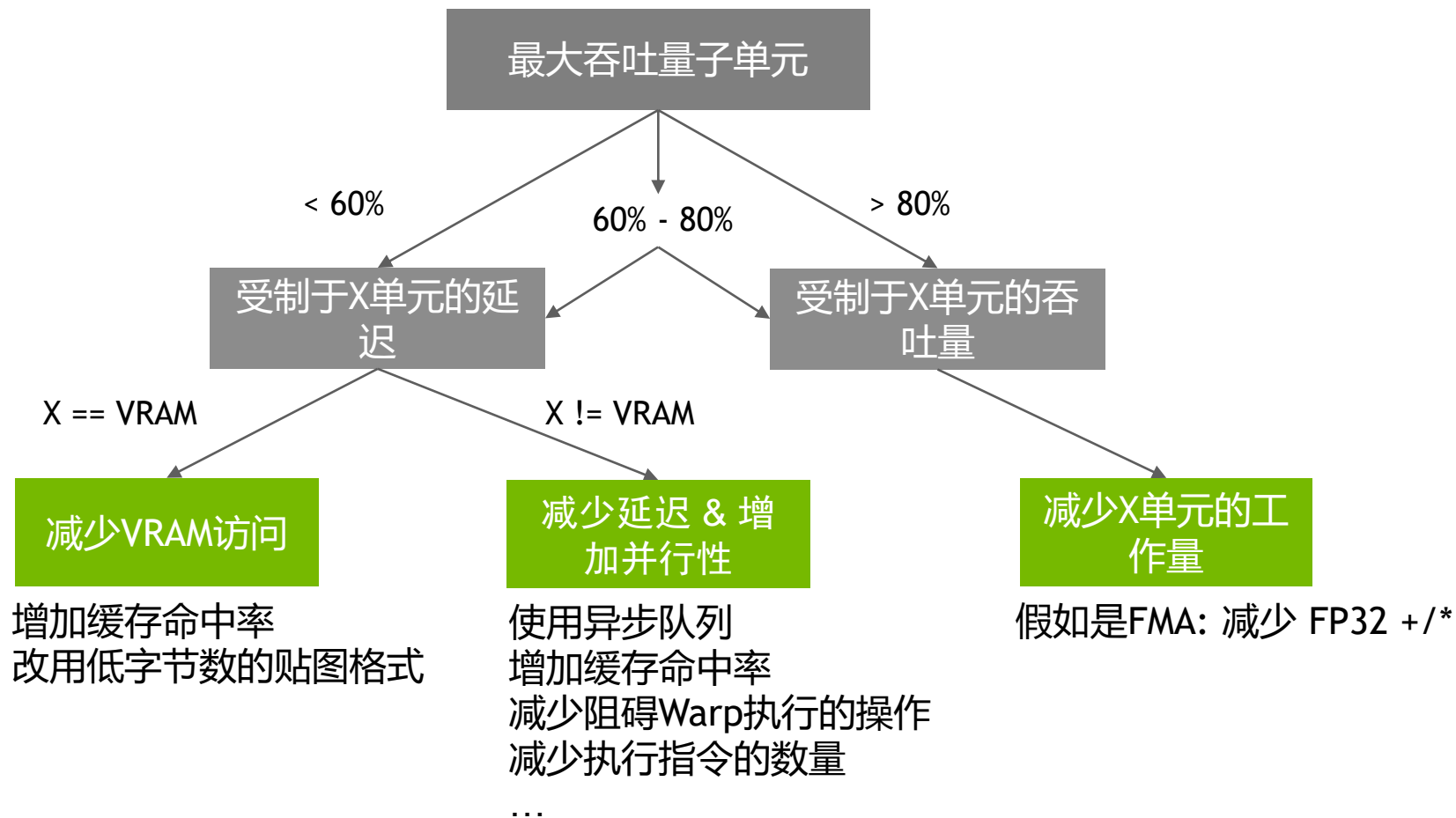
Independent Workload #2: GBuffer Fill

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



Summary	Metrics	Detailed Metrics	Capture Information
Start: 27.95 ms	Duration: 4.29 ms	End: 32.25 ms	Range: All Visible
Unit Throughput		Value	
VRAM Throughput		39.57%	
L1TEX Throughput		33.05%	
L2 Throughput		26.56%	
PES+VPC Throughput		22.22%	
SM Instruction Throughput		22.07%	
SM Issue Active		20.99%	
ZROP Throughput		19.89%	
SM FMA Pipe Throughput		19.50%	
CROP Throughput		18.26%	
RASTER Throughput		10.37%	
PROP Throughput		9.39%	
SM SFU Pipe Throughput		8.77%	
SM ALU Pipe Throughput		8.12%	
PD Throughput		7.13%	
SM FP16+Tensor Pipe Throughput		0.00%	
Warp Occupancy		Value	
Idle SM Unused Warp Slots		5.74%	
Active SM Unused Warp Slots		50.30%	
Compute Warps		0.00%	
Pixel Warps		36.02%	
Vertex+Tess+Geom Warps		7.94%	

THE P3 METHOD





总结

总结

- ▶ 工具 – Nsight Graphics
 - ▶ Modern GPU / API
- ▶ 方法论 – P3
 - ▶ 从最大吞吐量 (Throughput / SOL%) 开始
 - ▶ 减少子单元工作量 / 减少延迟 / 增加并行度
 - ▶ 不要重叠2个受VRAM延迟限制的工作负载



THANK YOU
JCL@NVIDIA.COM



nvidia.