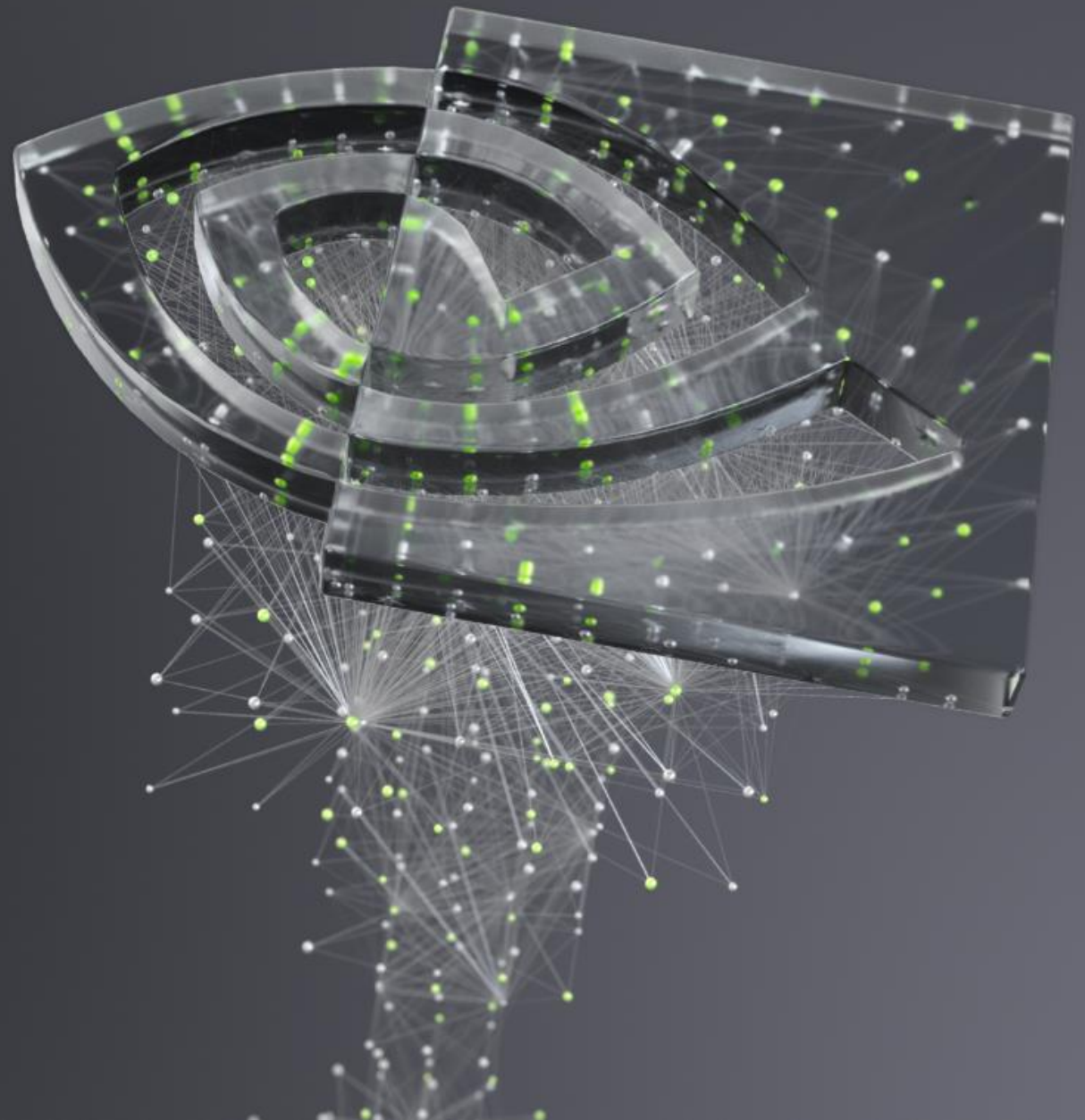


Nsight Graphics 方法及建议

梁佳晨 2019.12.18

概要

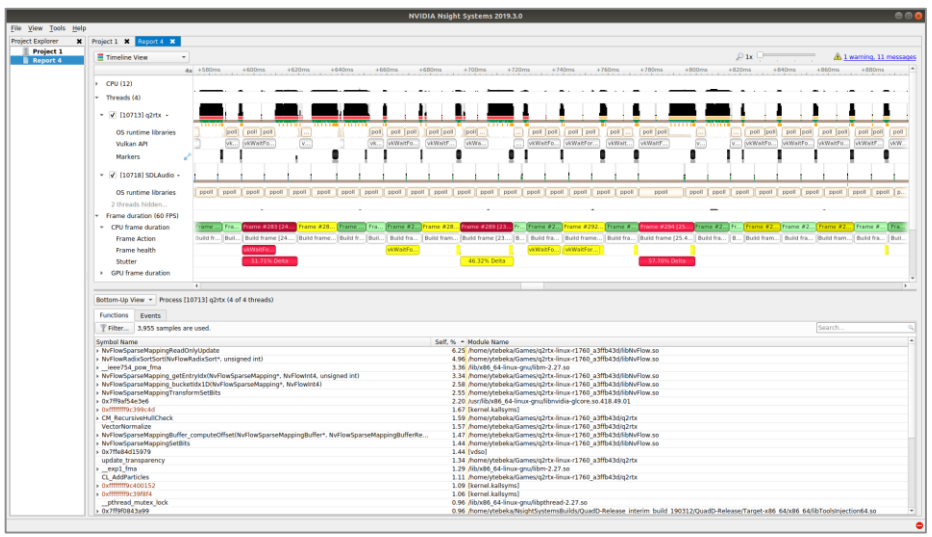
Nsight Graphics

峰值性能百分比分析法

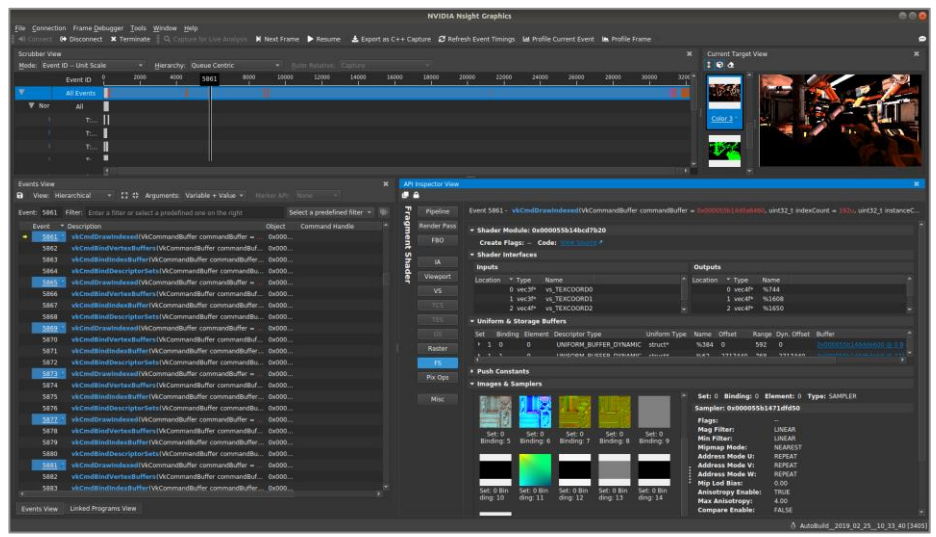
案例学习


Nsight Graphics

NVIDIA 开发工具

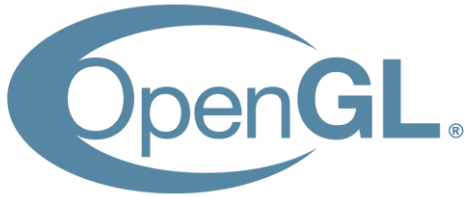


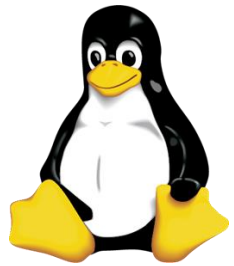
Nsight Systems



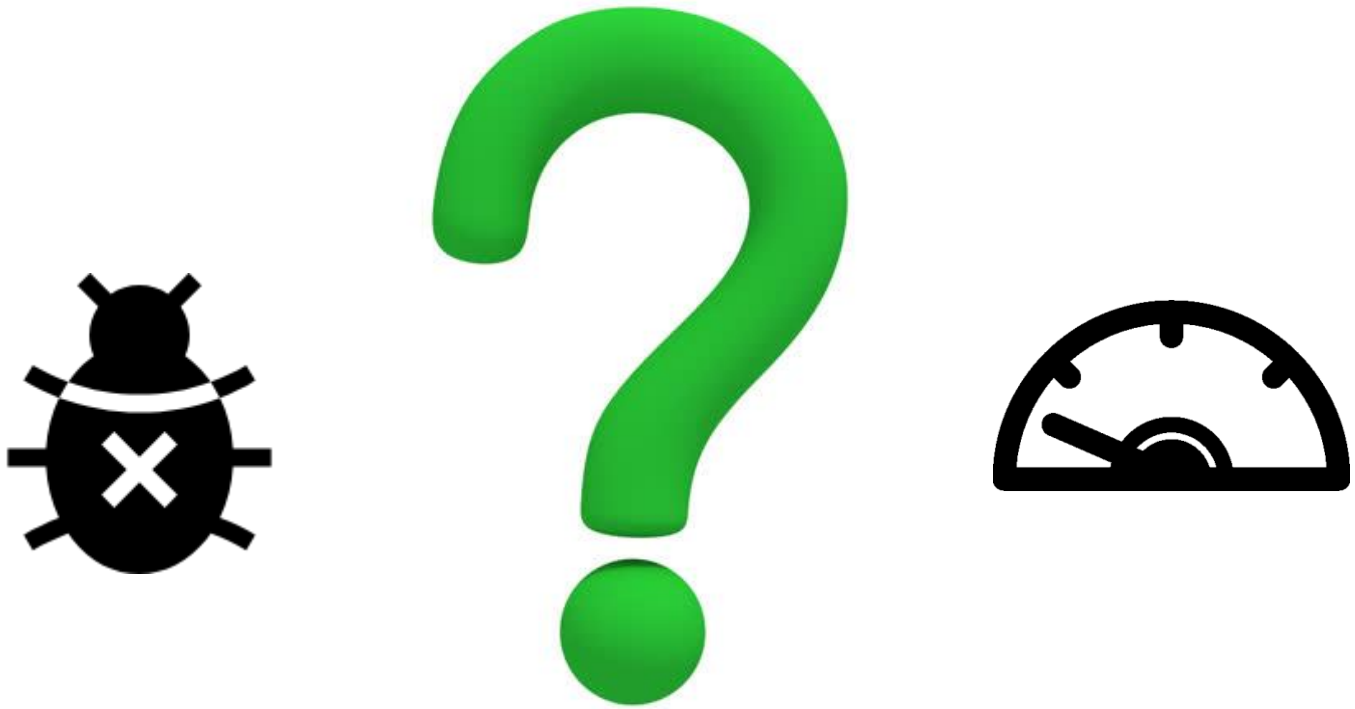
Nsight Graphics

Welcome To Nsight Graphics

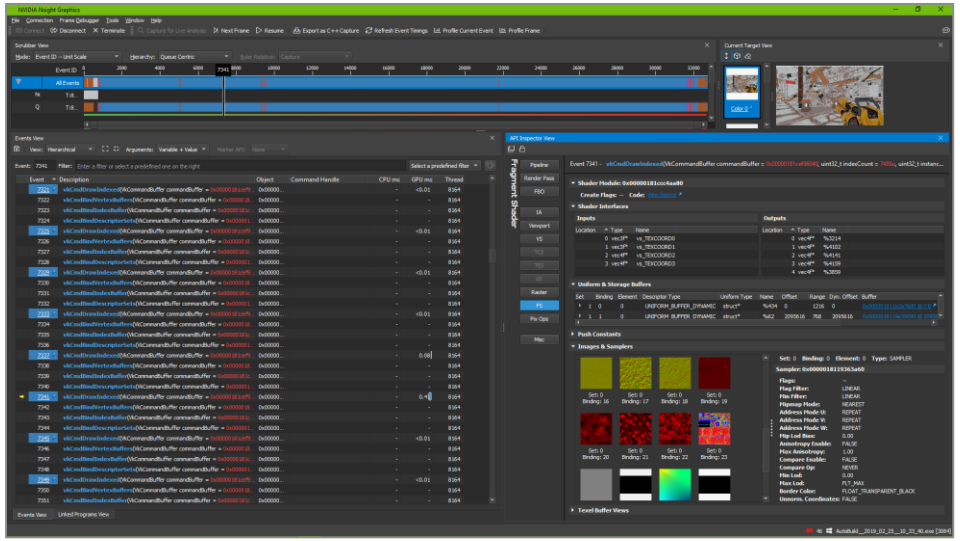




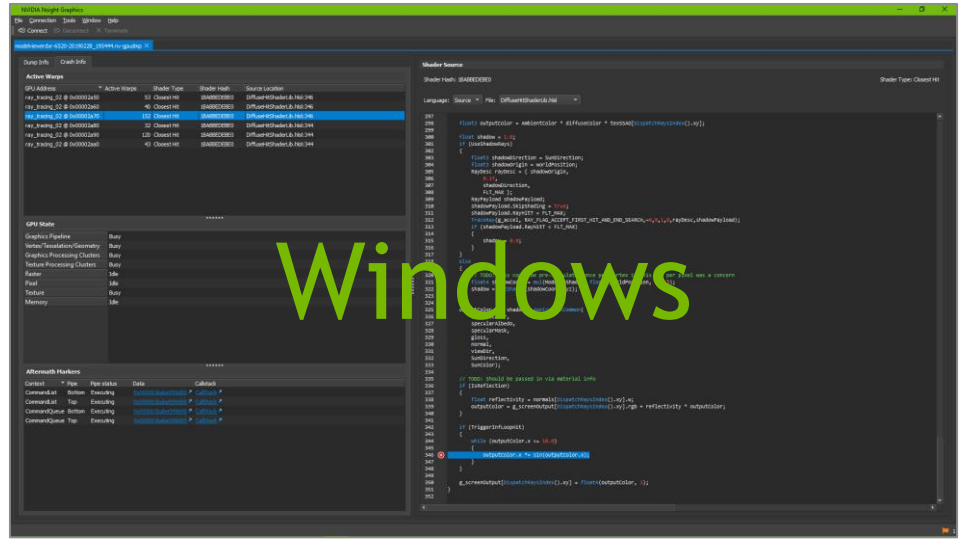
Work Flow



调试

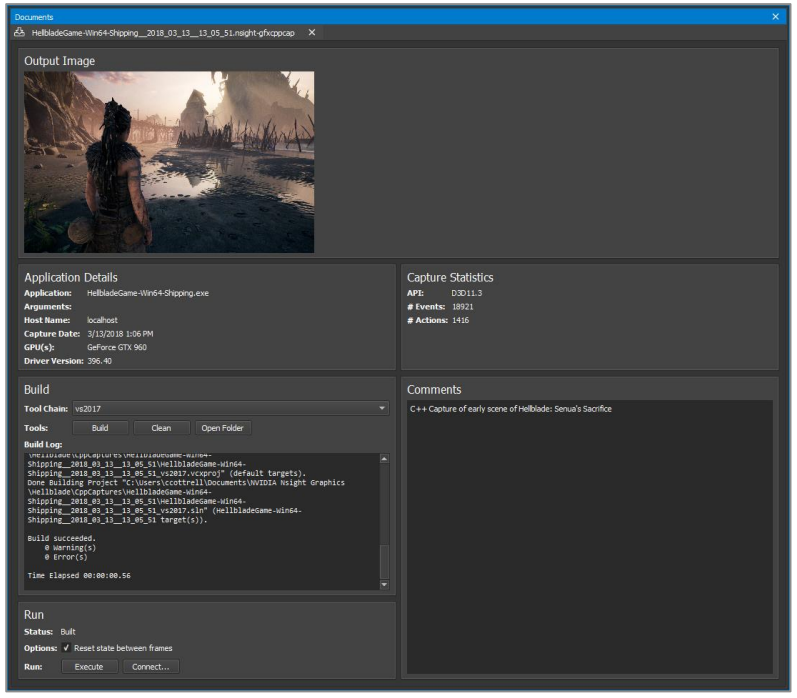


Frame Debugger



Aftermath

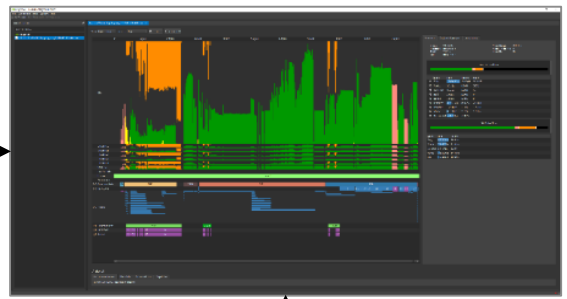
帧持久化 （C++ Capture）



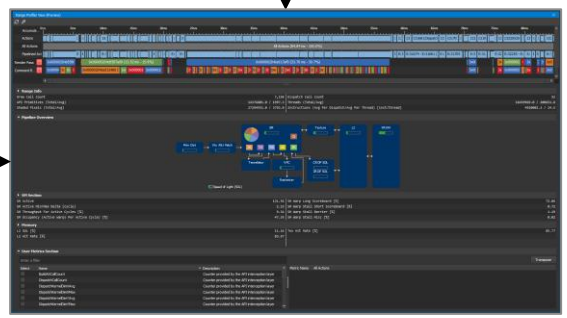
- snapshot
- 持久化
- 可重现
- 随意修改

性能分析

GPU Trace



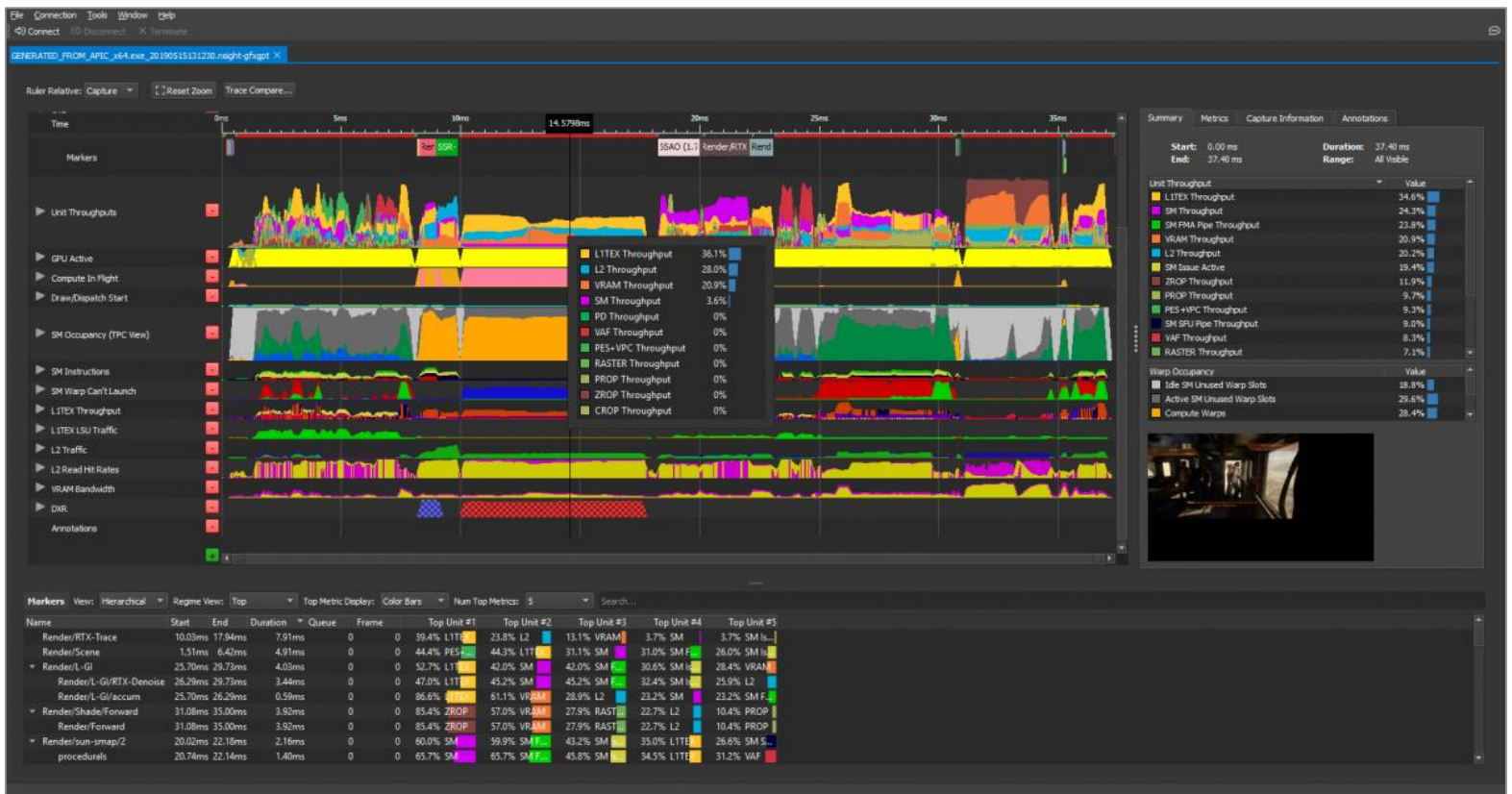
Nsight Systems



Range Profiler

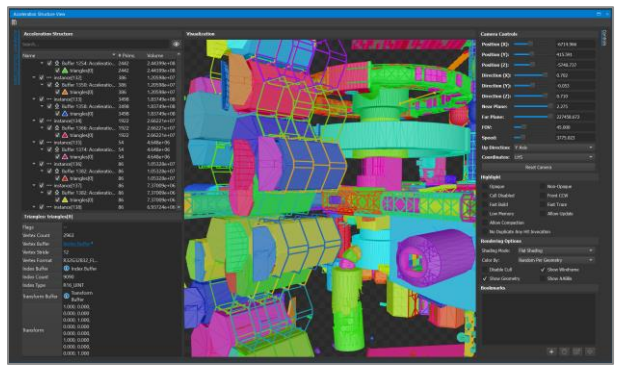
Shader Profiler

GPU Trace

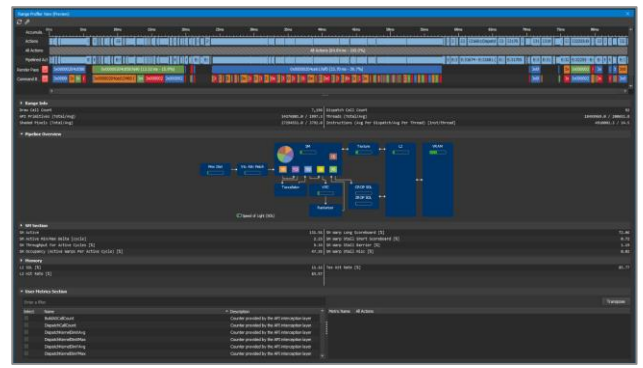


NVIDIA.

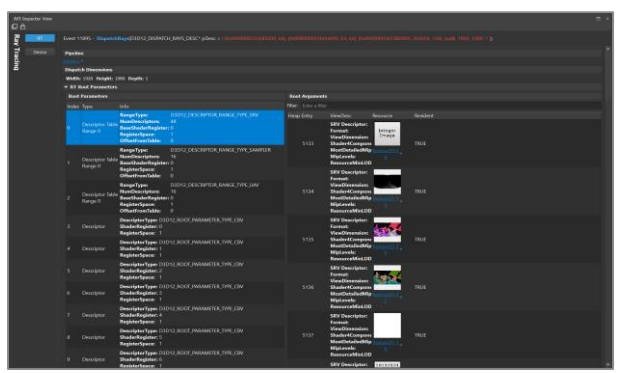
Ray Tracing Features



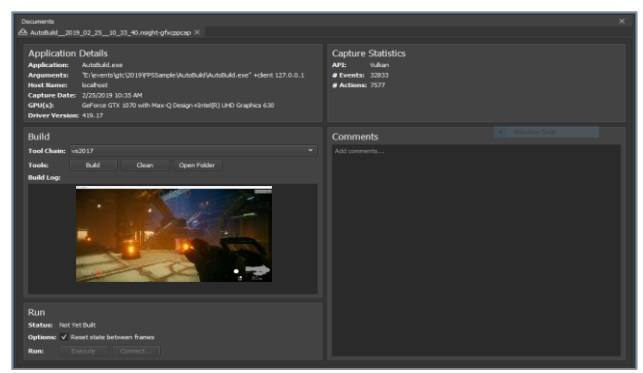
Acceleration Structure Viewer



Range Profiler



Resource Inspector

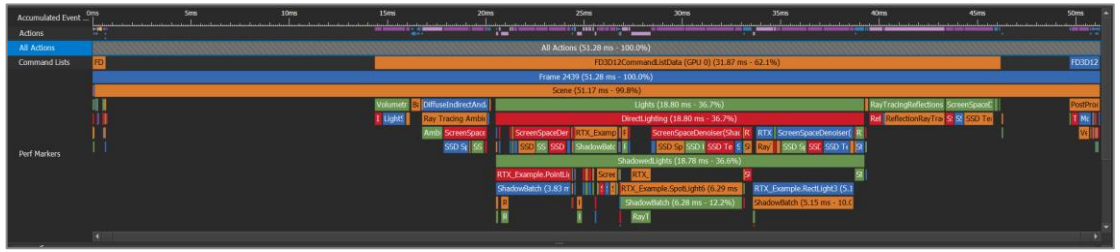


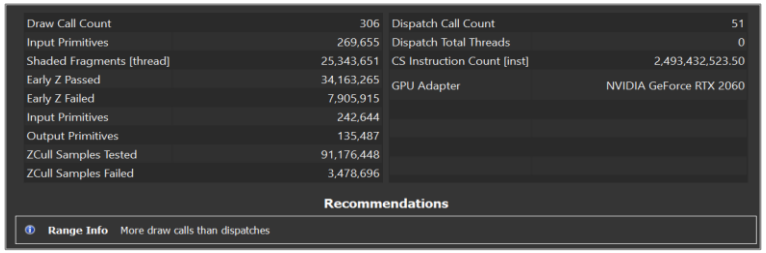
C++ Capture

峰值性能百分比分析法
Peak-Performance-Percentage (P3)

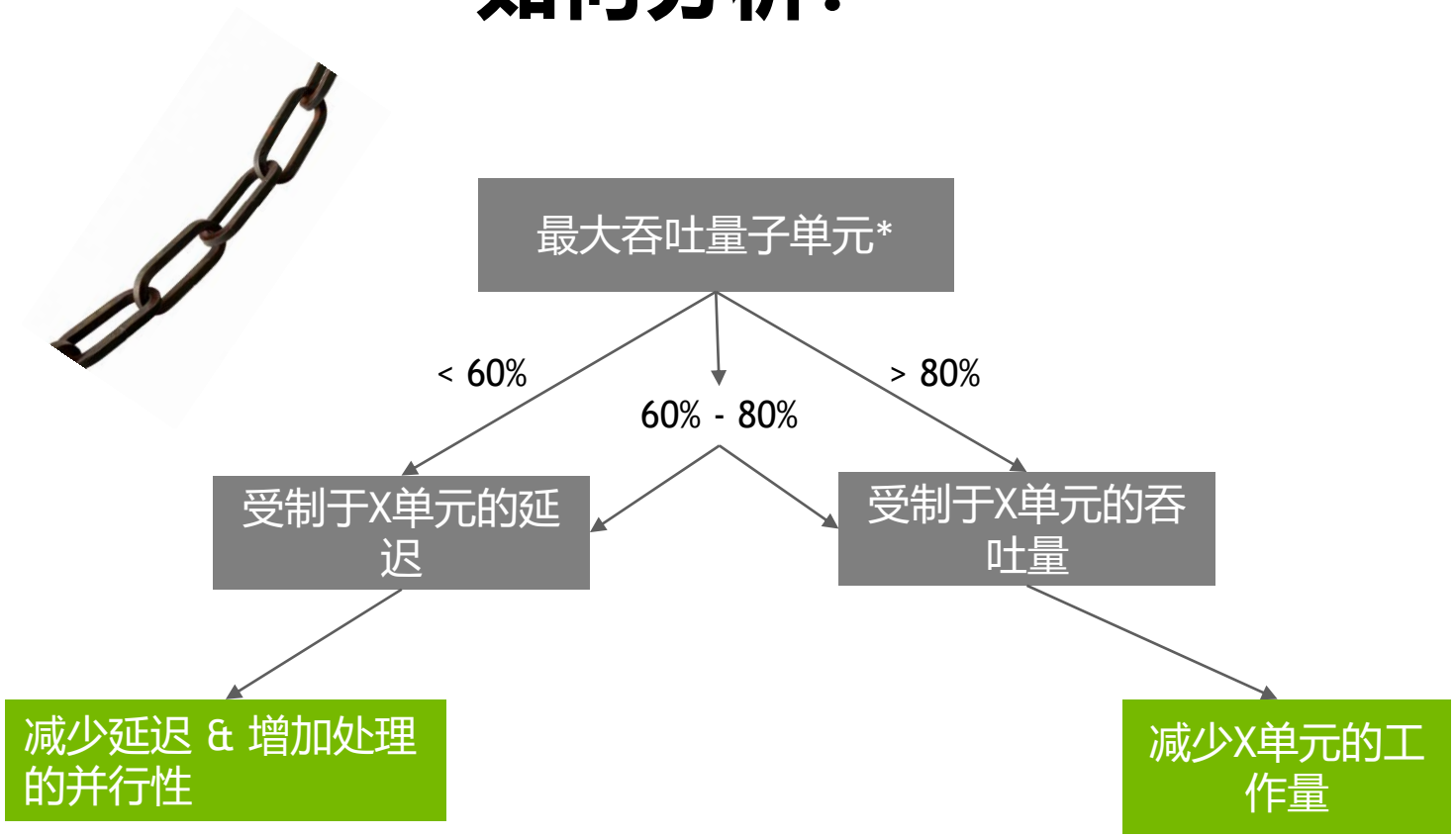
STEPS


捕捉帧 (C++)
性能分析
吞吐量(Throughput)
分析GPU负载
解决 验证


Accumulated Event ...
Actions
All Actions
Command Lists
Perf Markers
All Actions (51.28 ms - 100.0%)
Frame 2439 (51.28 ms - 100.0%)
Scene (51.17 ms - 99.8%)
Lights (18.80 ms - 36.7%)
DirectLighting (18.80 ms - 36.7%)
ShadowedLights (18.78 ms - 36.6%)


Draw Call Count 306
Input Primitives 269,655
Shaded Fragments [thread] 25,343,651
Early Z Passed 34,163,265
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
Dispatch Call Count 51
Dispatch Total Threads 0
CS Instruction Count [inst] 2,493,432,523.50
GPU Adapter NVIDIA GeForce RTX 2060
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

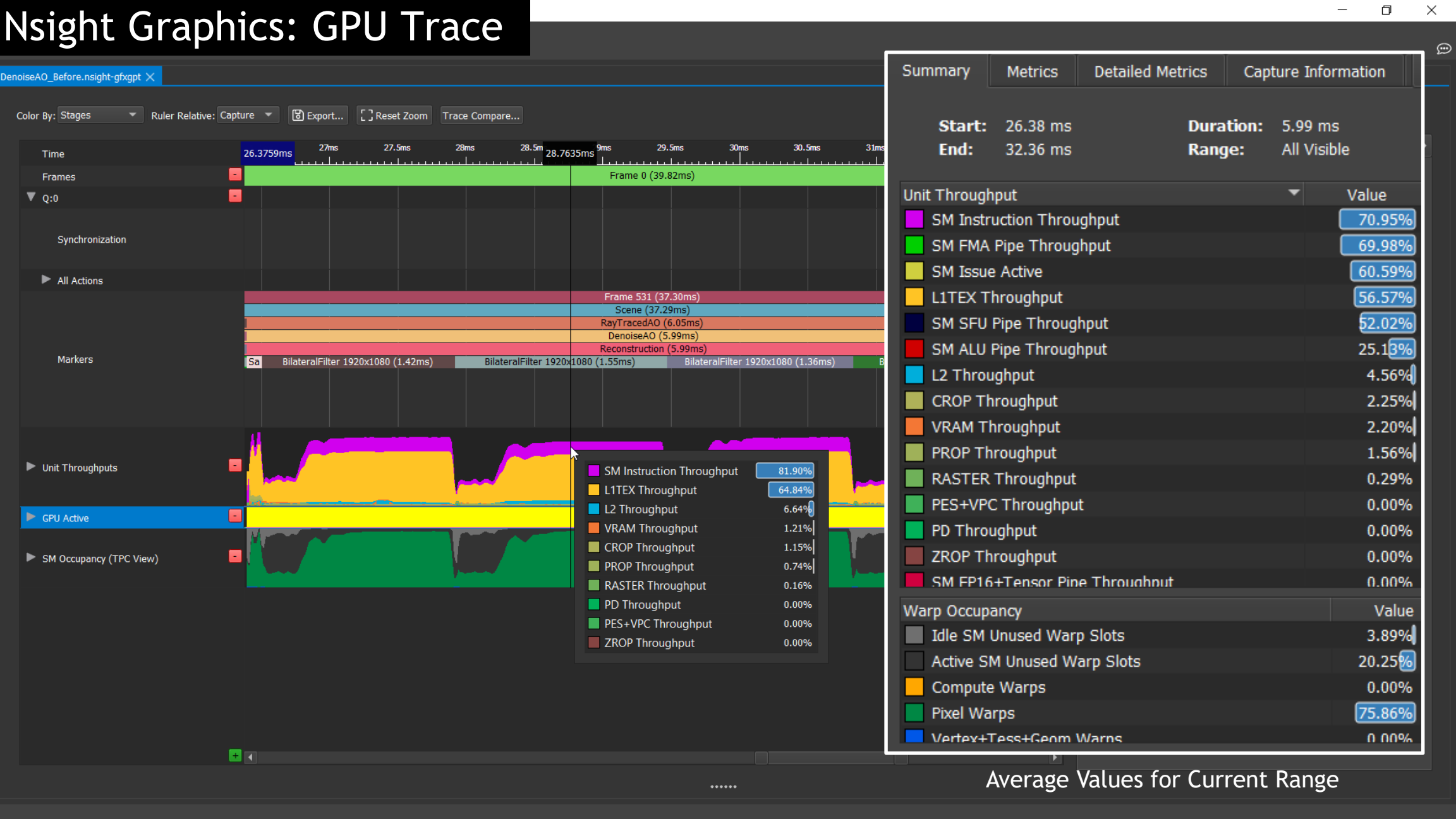
By Louis Bavoil

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

案例学习一
RTAO DENOISER PIXEL SHADER


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

Nsight Graphics: GPU Trace



Average Values for Current Range

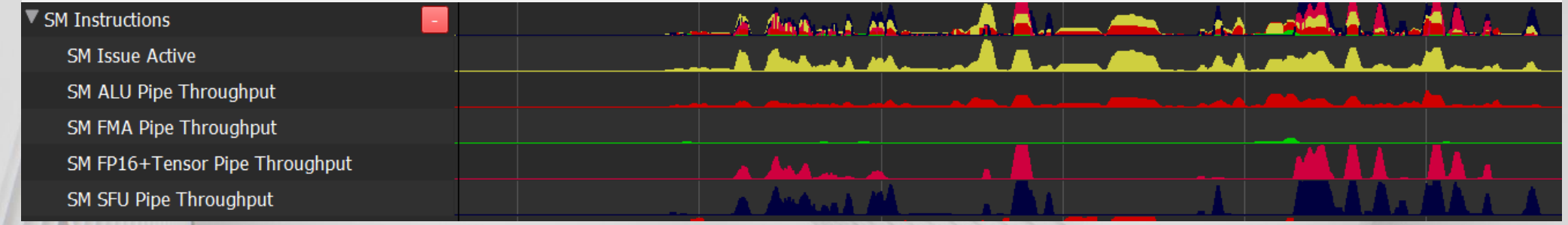
子单元吞吐量

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

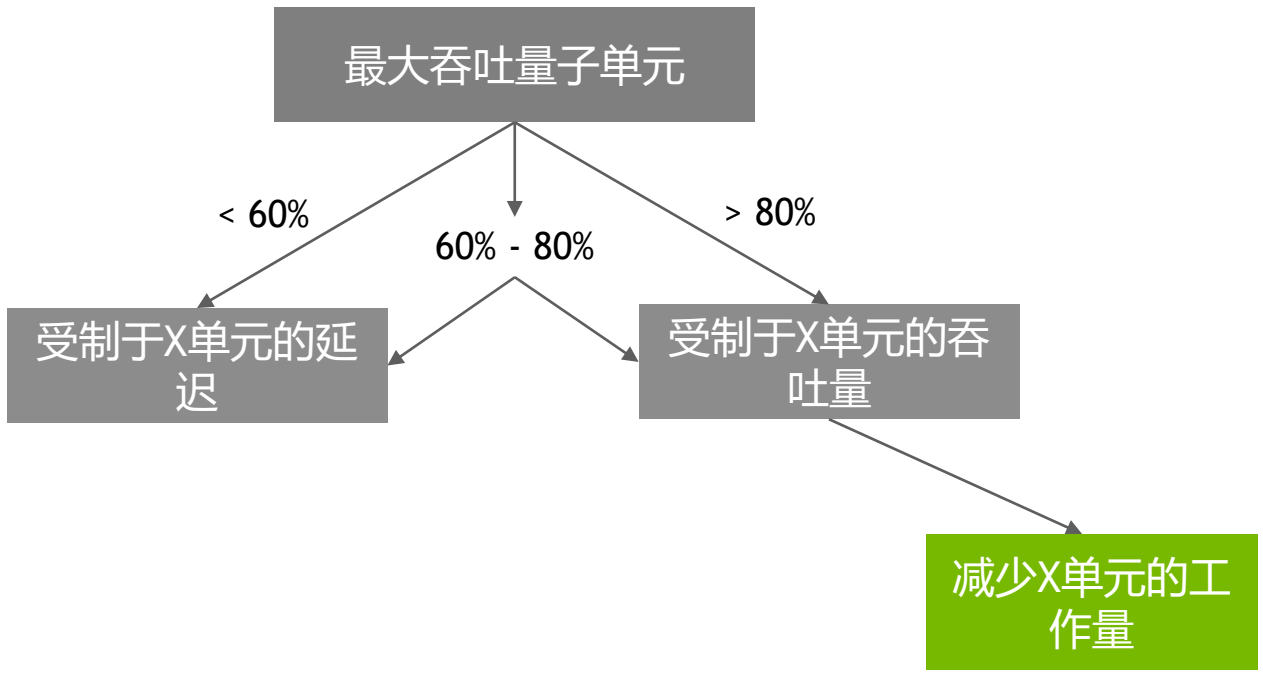
```
#if 0

        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

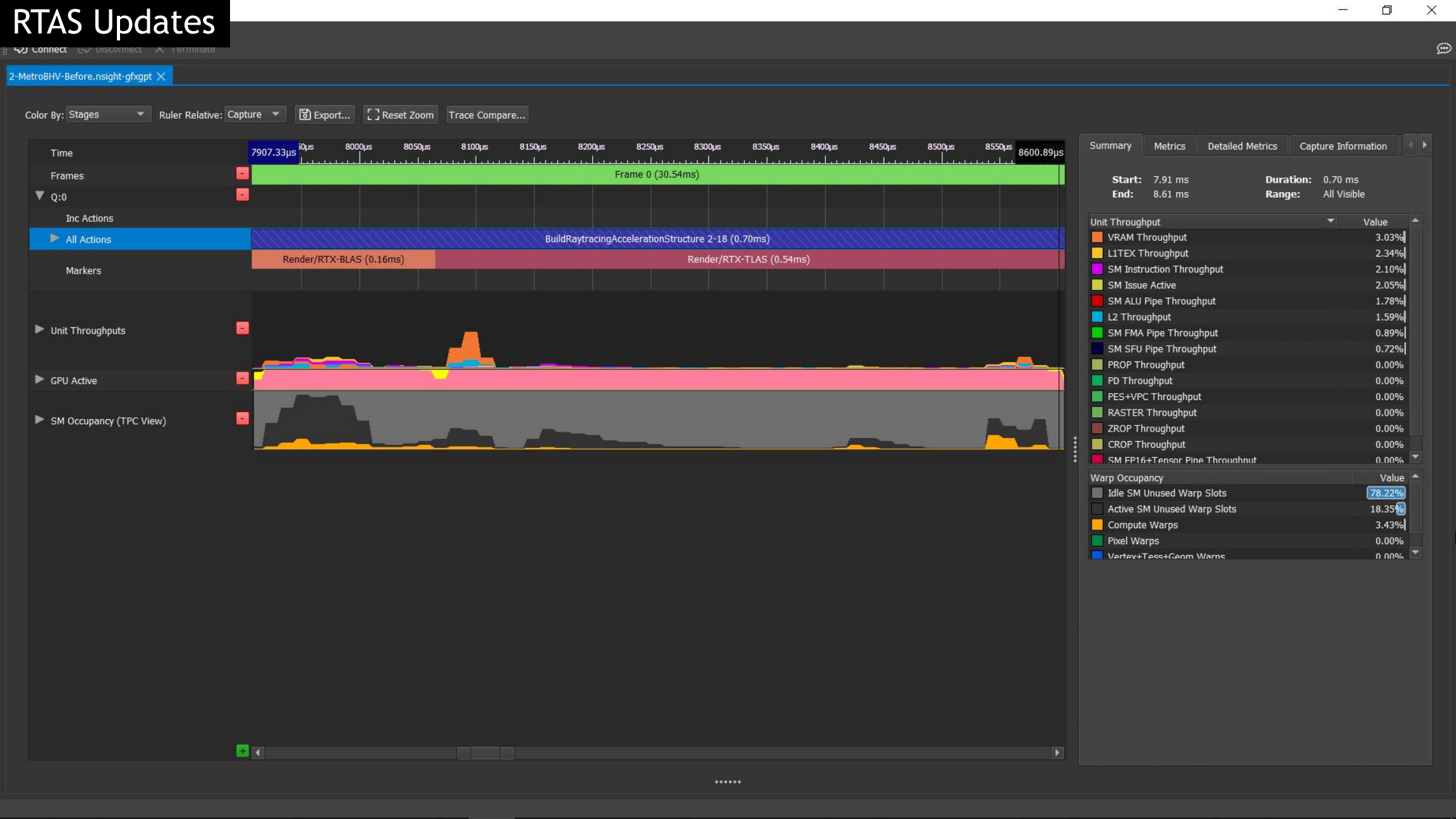
On RTX 2080 with SetStablePowerState(TRUE)

THE P3 METHOD

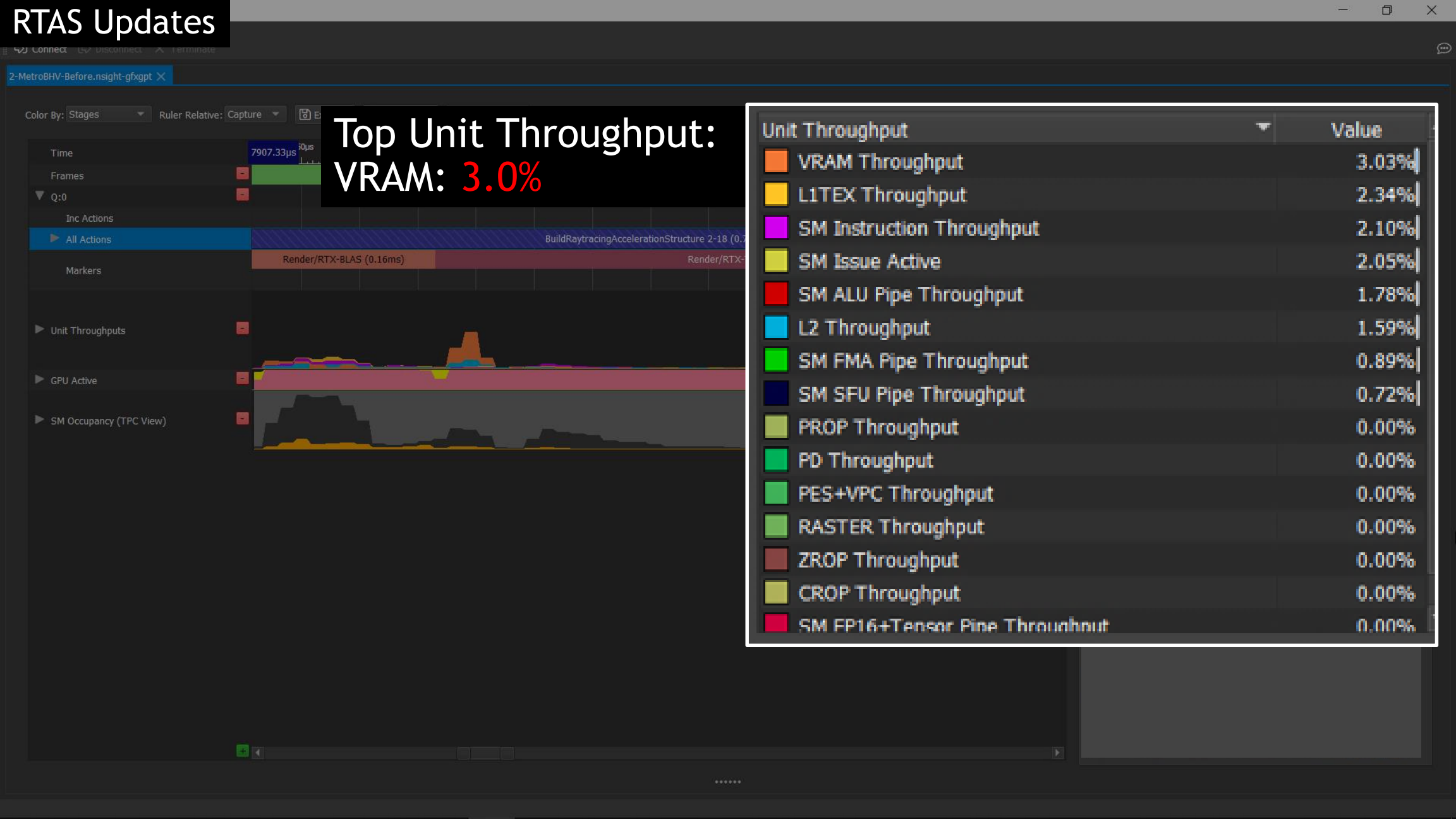


案例学习二
METRO EXODUS

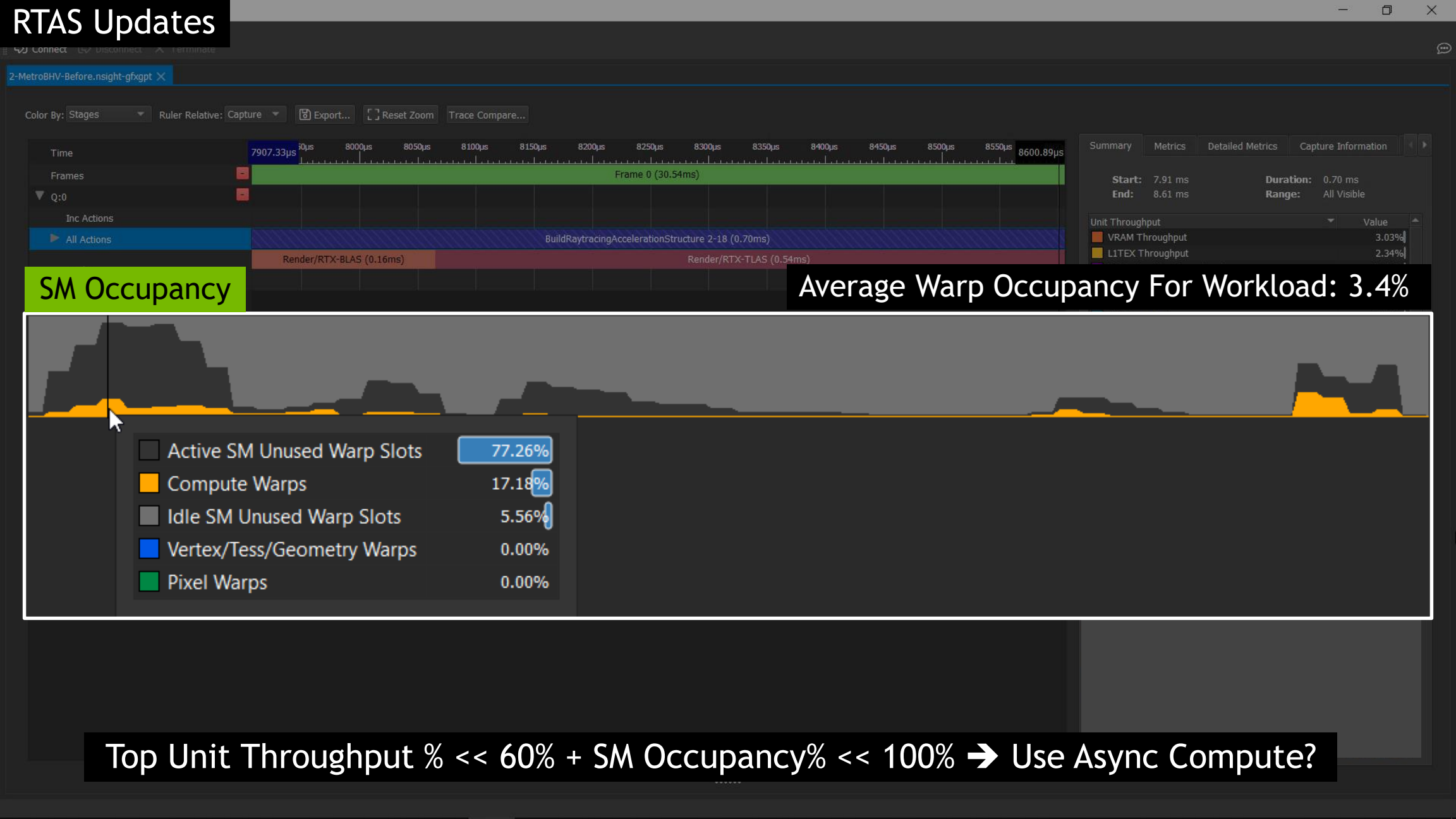
RTAS Updates



RTAS Updates



RTAS Updates

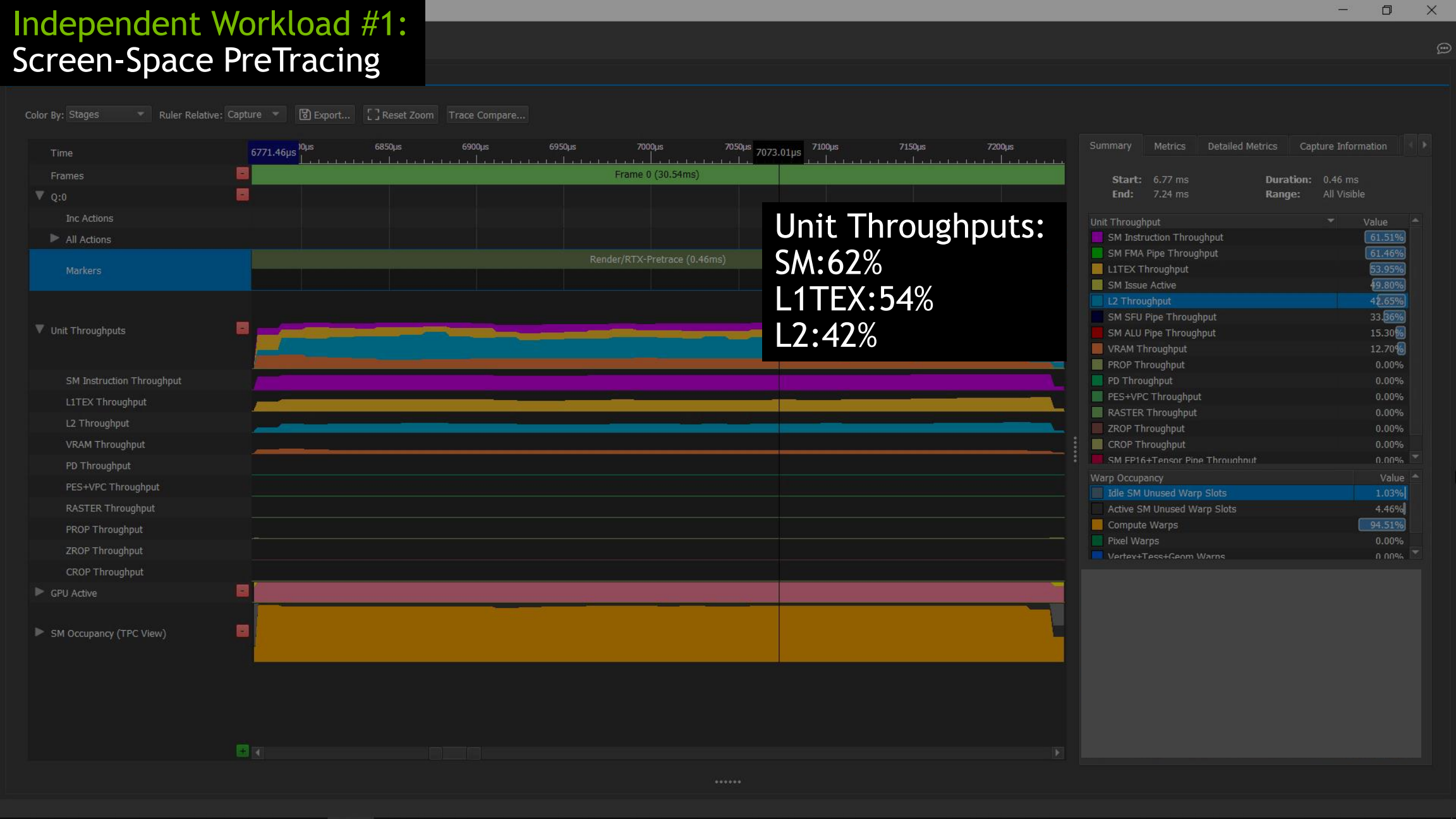


SM Occupancy

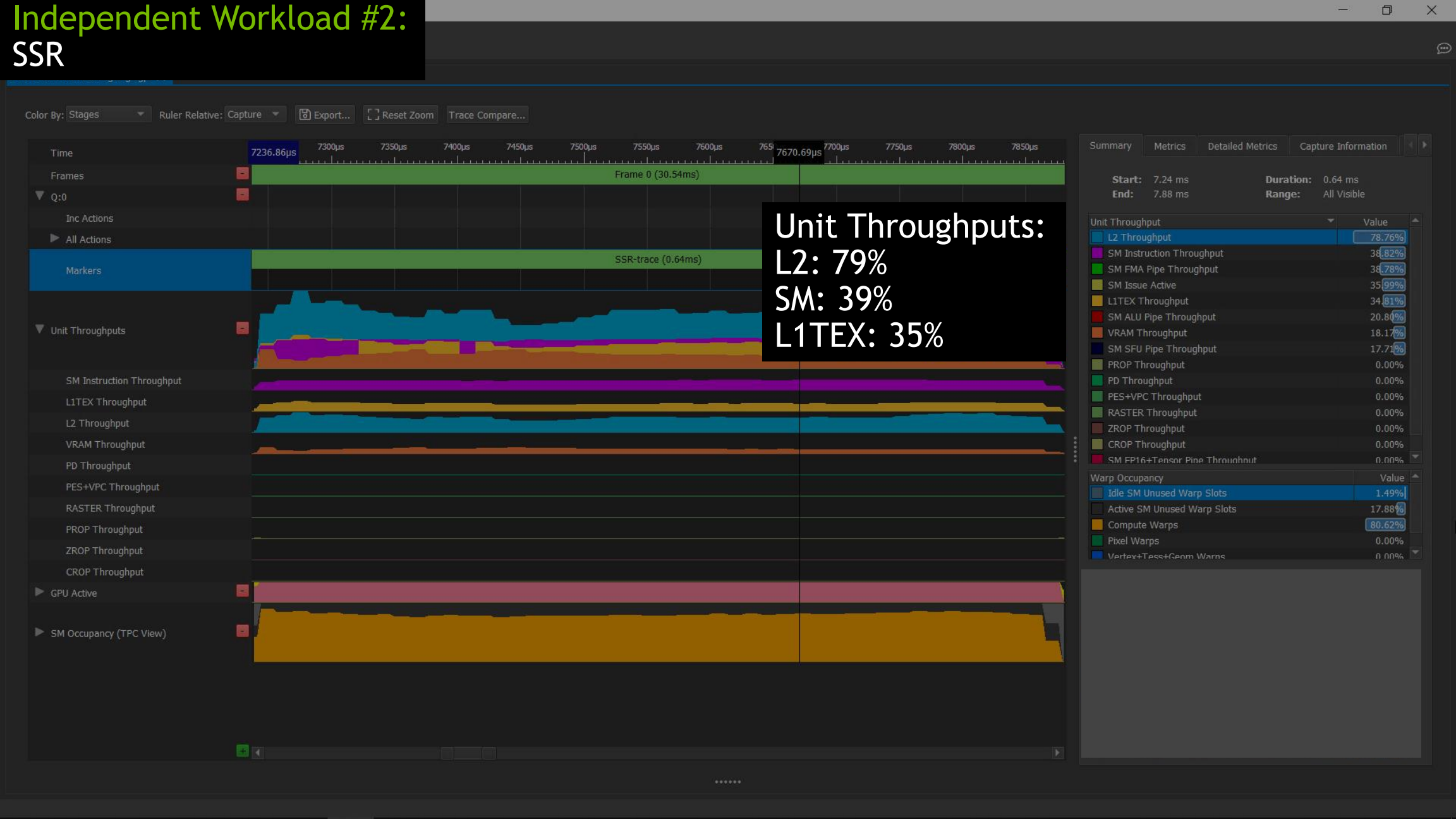
Average Warp Occupancy For Workload: 3.4%

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

Independent Workload #1:
Screen-Space PreTracing



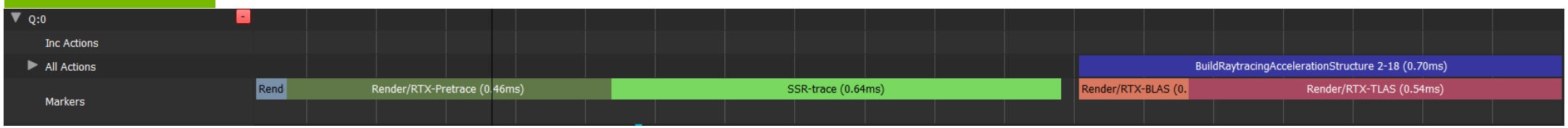
Independent Workload #2:
SSR



ASYNC COMPUTE DIFF

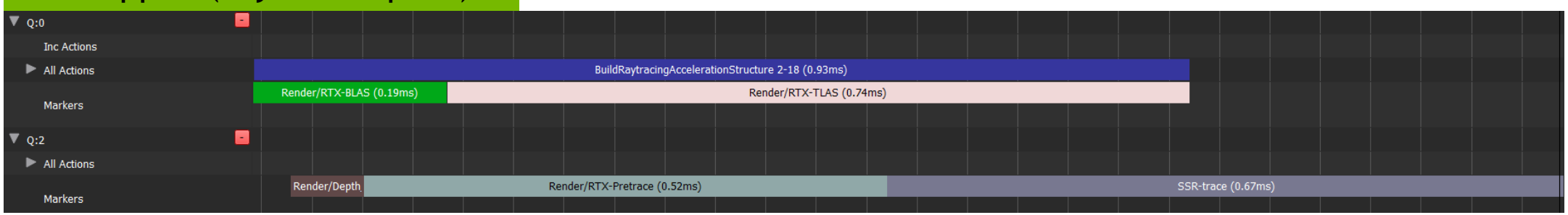
Serialized

1.83 ms



Overlapped (Async Compute)

1.30 ms



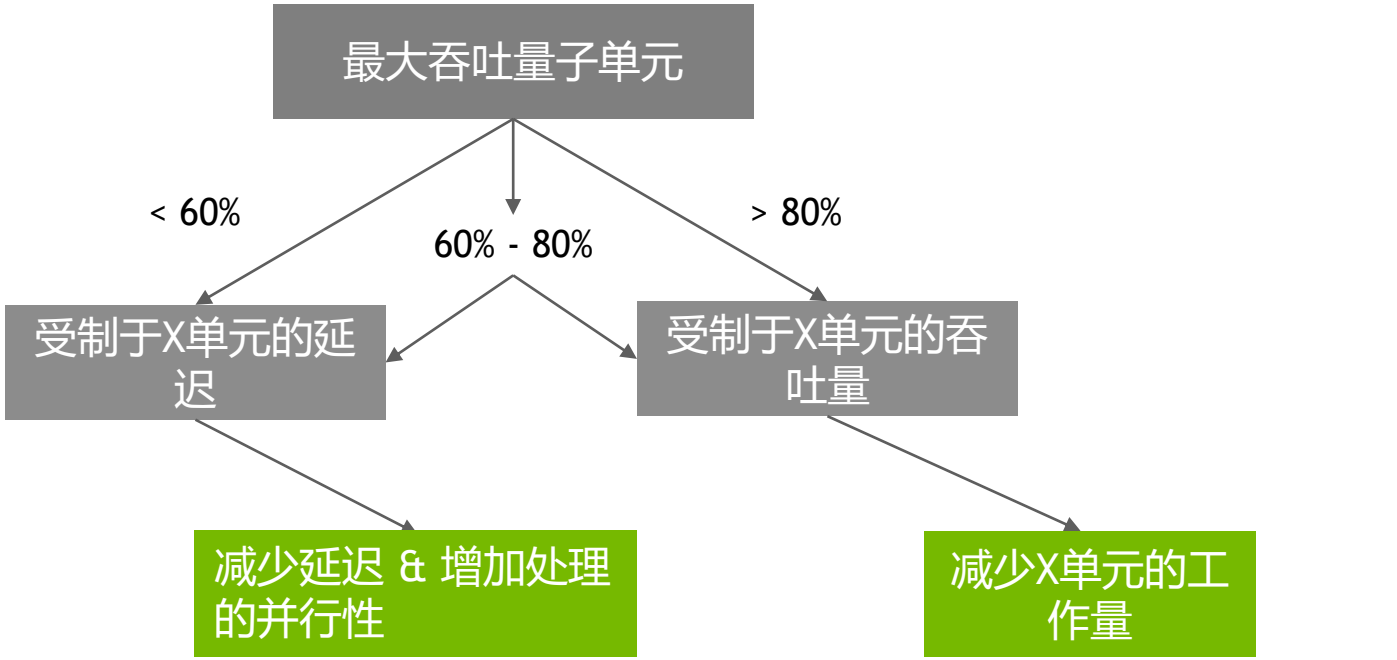
ASYNC-COMPUTE OVERLAP

(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



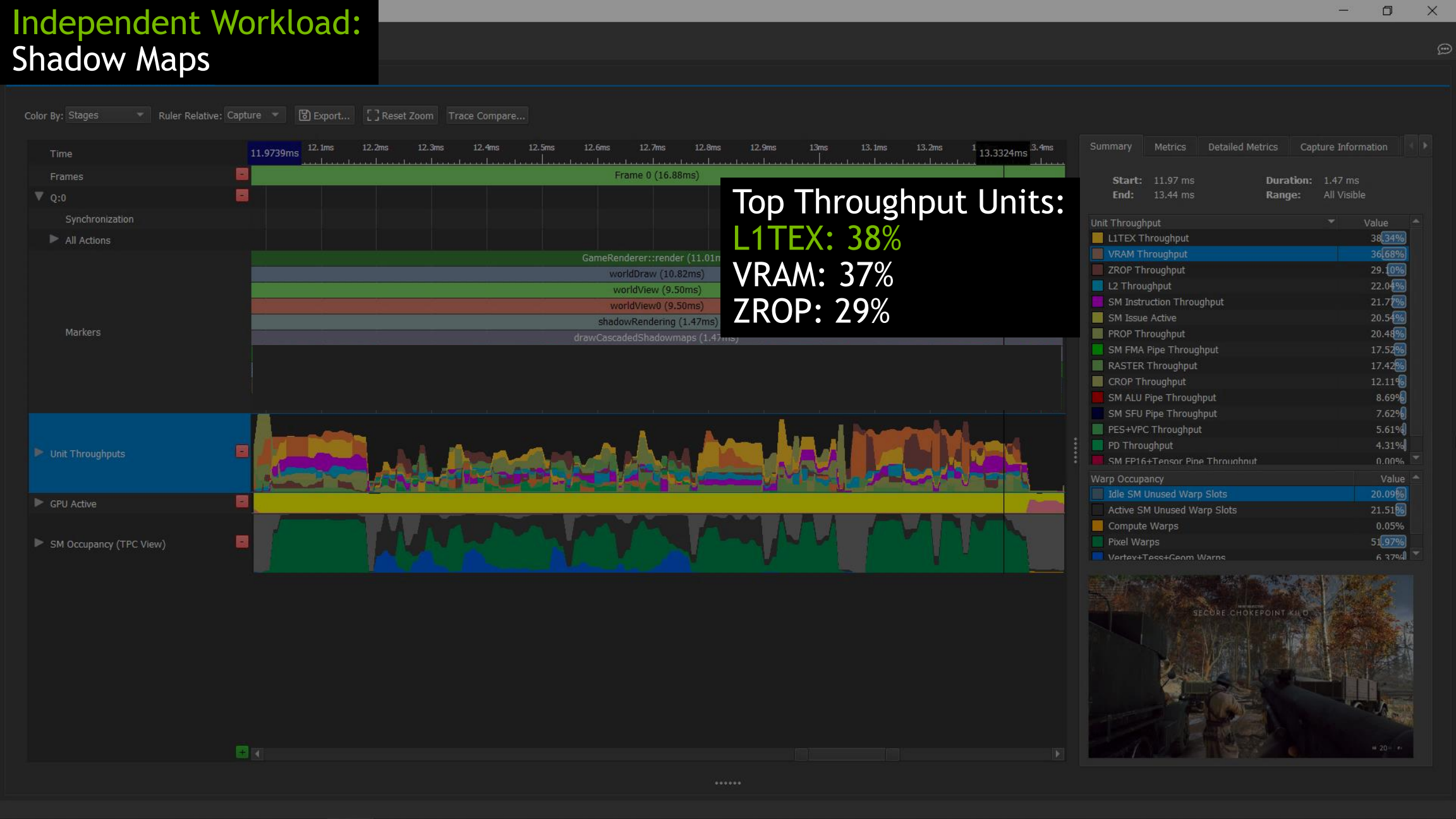
使用异步队列
增加缓存命中率
减少阻碍Warp执行的操作
减少执行指令的数量
…

假如是FMA: 减少 FP32 +/*

案例学习三
阴影贴图//异步计算

HBAO + SSR + Light Culling

Independent Workload:
Shadow Maps



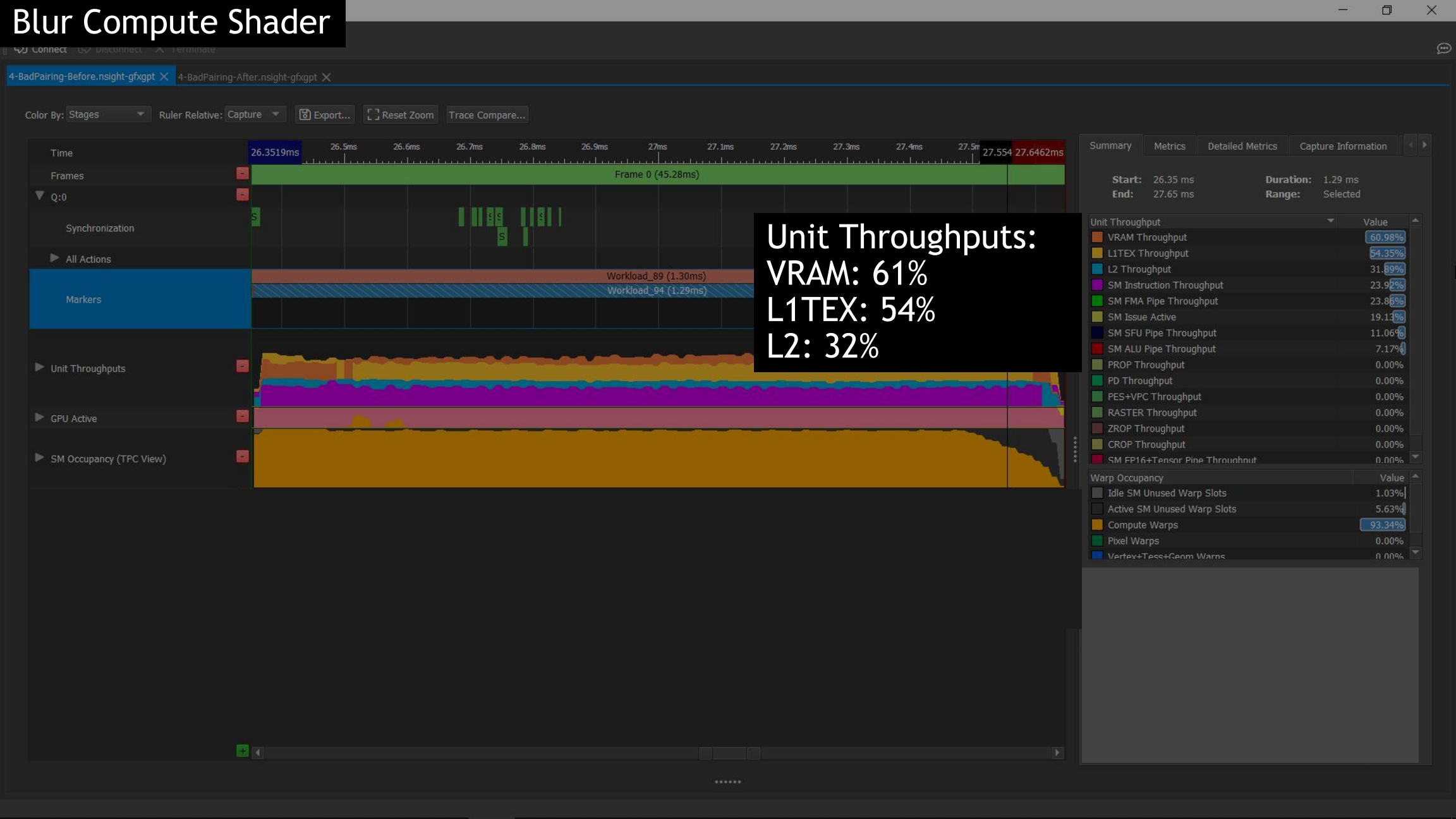
ASYNC-COMPUTE OVERLAP
(SHADOW MAPS) // (HBAO+SSR+LIGHT-CULL)

	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

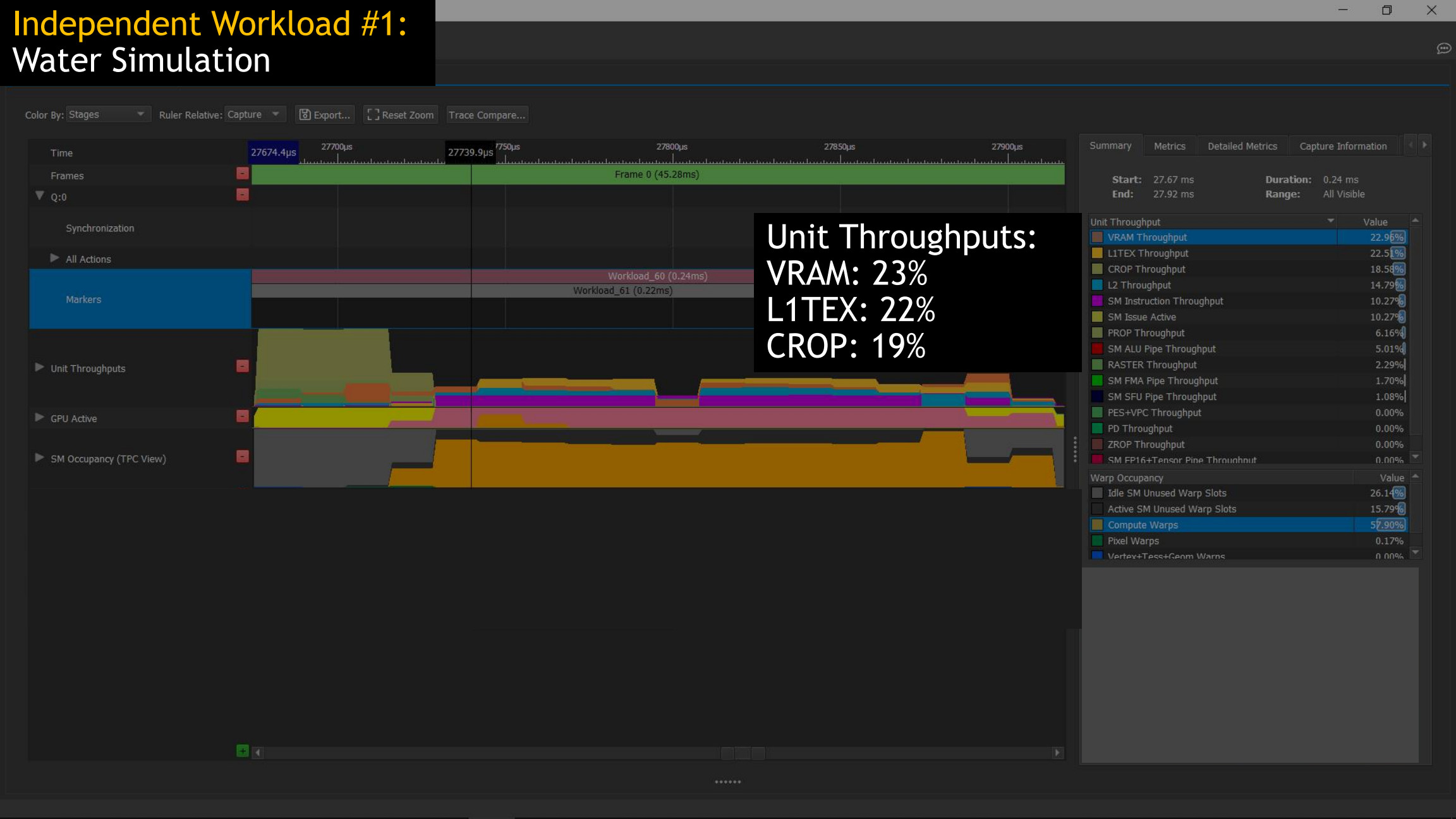
On RTX 2080 + SetStablePowerState(TRUE)

案例学习四
不宜使用异步计算的情况

Blur Compute Shader

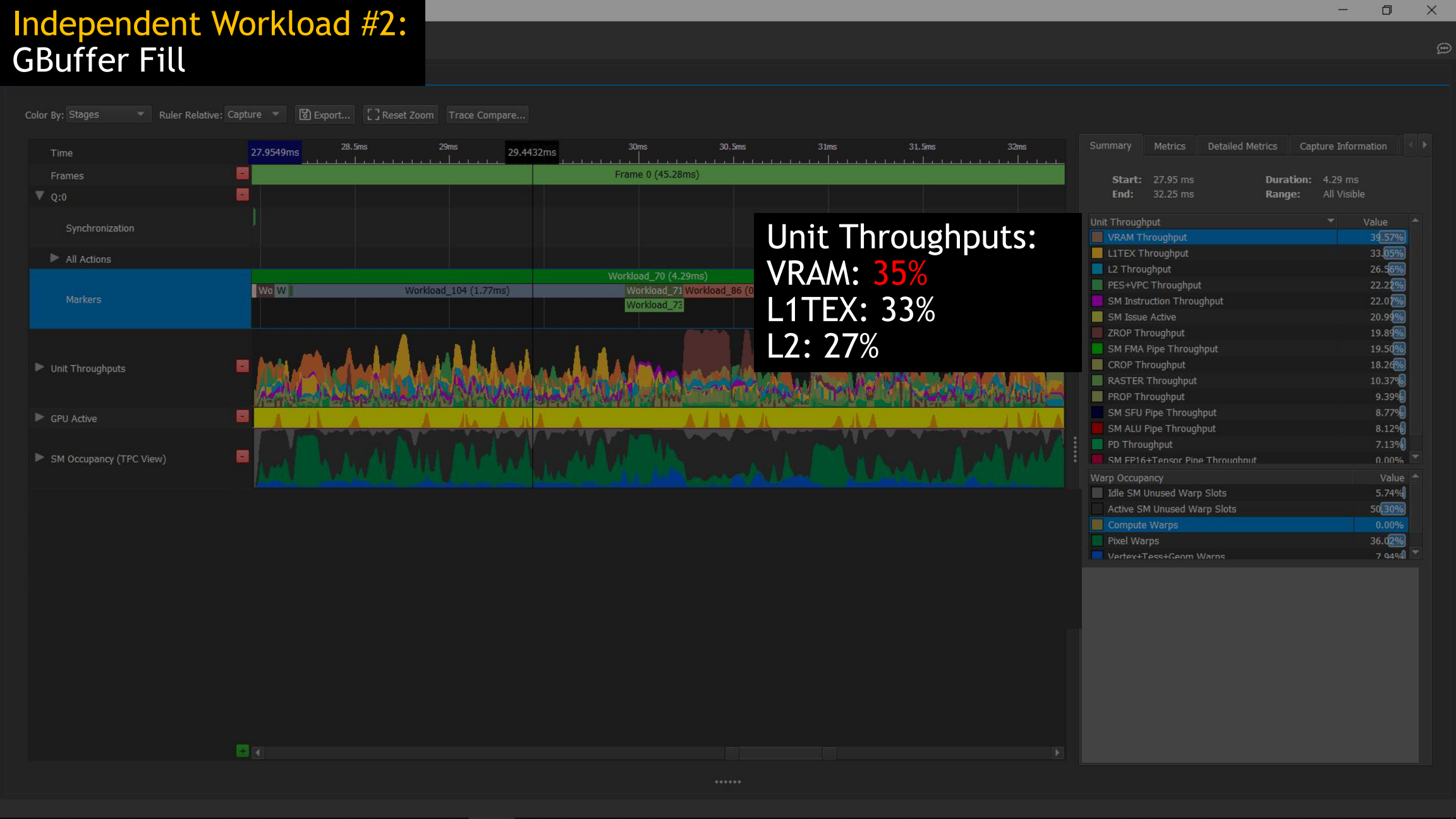


Independent Workload #1:
Water Simulation



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

Independent Workload #2:
GBuffer Fill

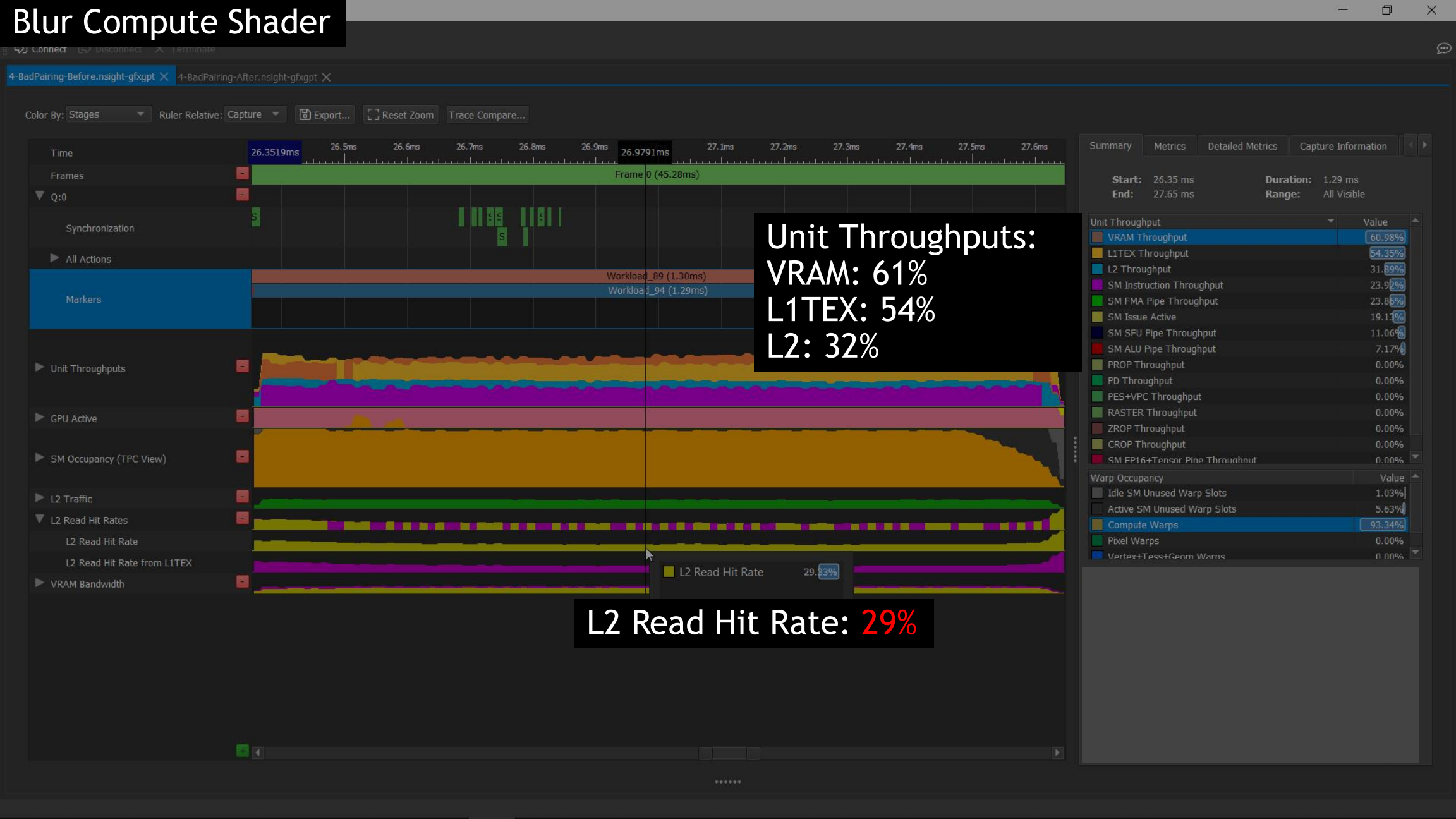


BAD ASYNC-COMPUTE PAIRING
(BLUR CS) // (GBUFFER + WATER SIM)

	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)

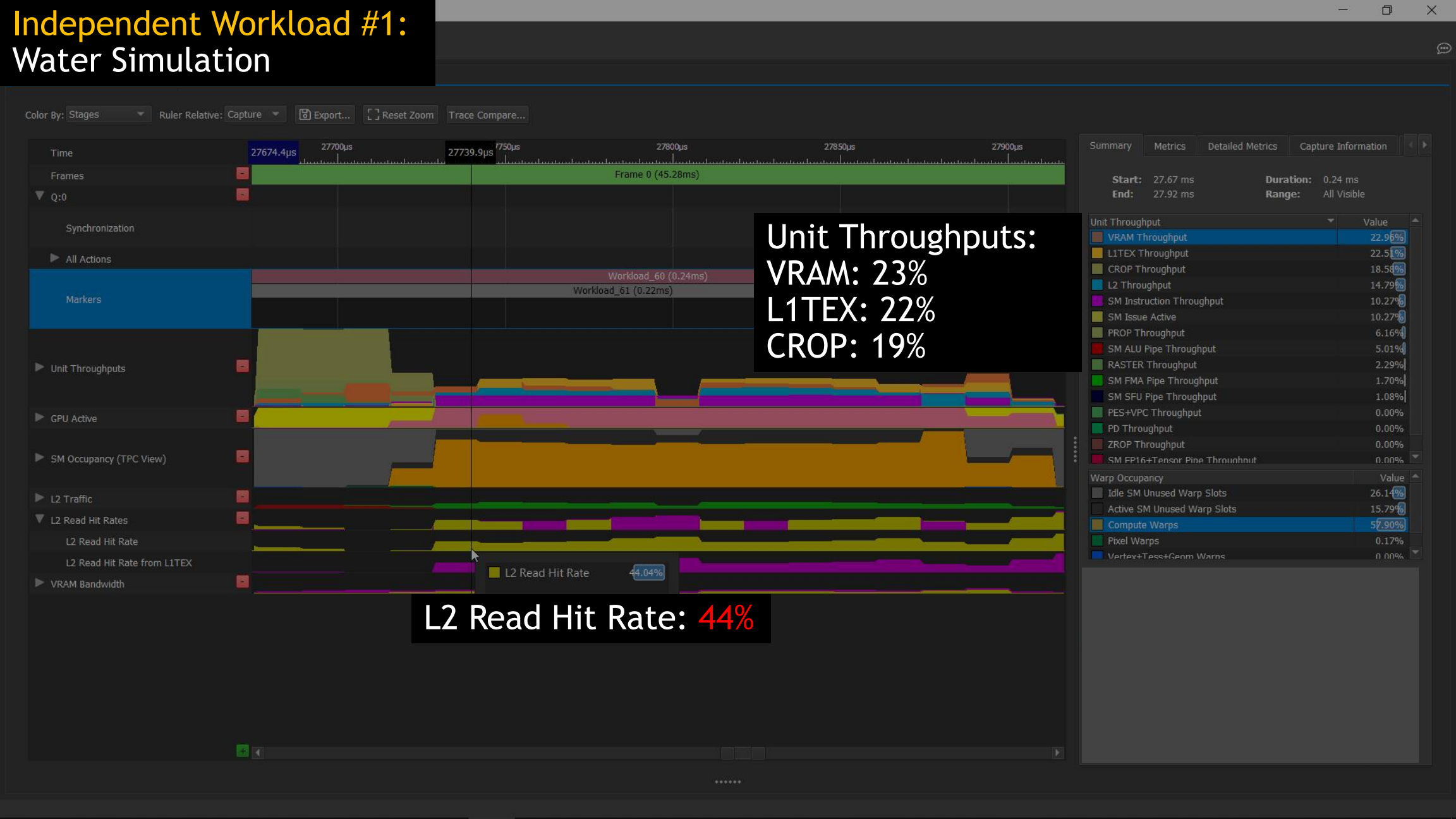
Blur Compute Shader



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

L2 Read Hit Rate: 29%

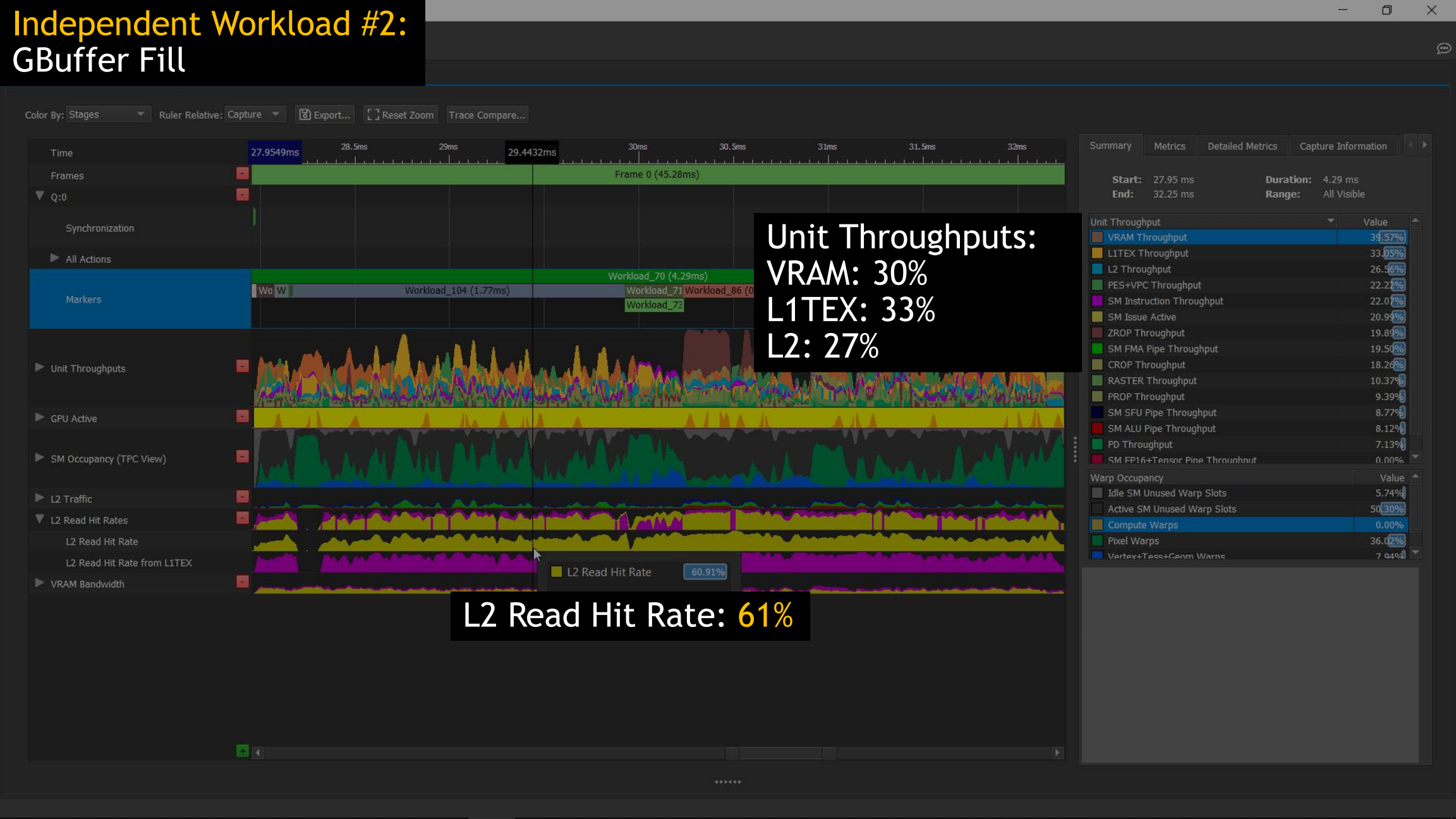
Independent Workload #1:
Water Simulation



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

L2 Read Hit Rate: 44%

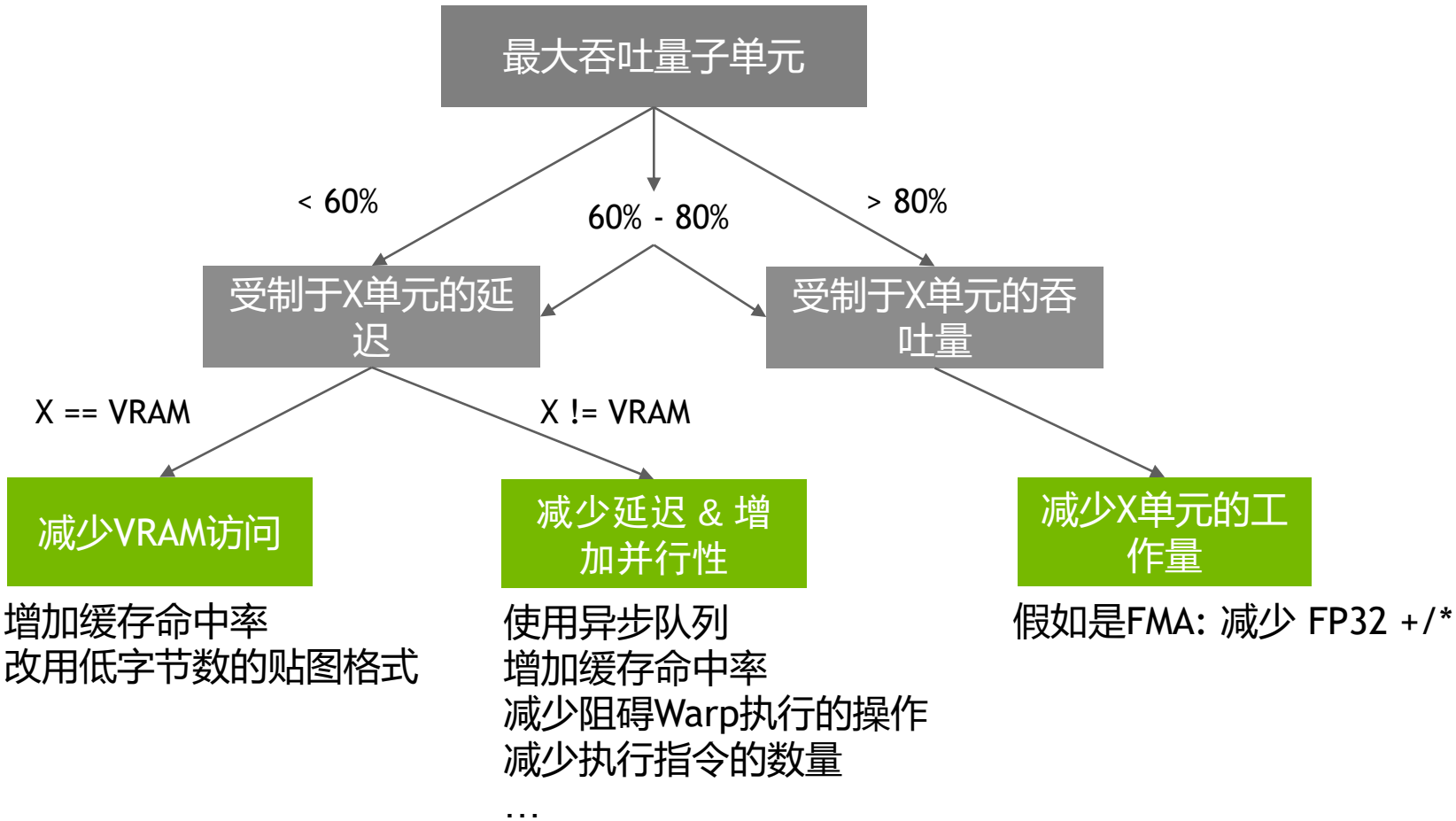
Independent Workload #2:
GBuffer Fill



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

L2 Read Hit Rate: 61%

THE P3 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