

## OPTIMIZING DX12/DXR GPU WORKLOADS USING NSIGHT GRAPHICS: GPU TRACE AND THE PEAK-PERFORMANCE-PERCENTAGE (P3) METHOD

Louis Bavoil | Principal Engineer, Developer Technology

1 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**





# **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

#### DenoiseAO\_Before.nsight-gfxgpt 🗙

Color By: Stages 🔹 Ruler Relative: Capture 👻 🐻 Export... [] Reset Zoom Trace Compare...



Metric Graphs (for >= Turing GPUs)

+

 $\odot$ 

\*\*\*\*

## Nsight Graphics: GPU Trace

· · ·					Т		T	T.
iseAO_Before.nsight-gfxgpt 🗙				Summary	Metrics	Detailed Metrics	Capture In	formation
olor By: Stages 🔹 Ruler Relativ	ve: Capture 🔻 🕃 Export [] Reset Zoom Trace Compare			Start:	26.38 ms	Dura	ation: 5.99	ms
Time	26.3759ms 27ms 27.5ms 28ms 28.5m 28.7	635ms 9ms 29.5ms 30ms	30.5ms 31ms	End:	32.36 ms	Ran	ge: All V	isible
Frames		Frame 0 (39.82ms)						
▼ Q:0				Unit Through	put		•	Value
				SM Instr	uction Throu	ghput		70.95%
Synchronization				SM FMA	Pipe Throug	hput		69.98%
All Actions				SM Issue	e Active			60.59%
		Frame 531 (37.30ms)		L1TEX T	hroughput			56.57%
		RayTracedAO (6.05ms)		SM SFU	Pipe Throug	hput		52.02%
		DenoiseAO (5.99ms) Reconstruction (5.99ms)		SM ALU	Pipe Through	nput		25.13%
Markers	Sa BilateralFilter 1920x1080 (1.42ms) BilateralFilter 1920	0x1080 (1.55ms) BilateralFilter 19	920x1080 (1.36ms) B	L2 Throu	ghput			4.56%
				CROP Th	roughput			2.25%
				VRAM T	nrouahput			2.20%
				PROP Th	rouahput			1.56%
Unit Throughputs		SM Instruction Throughput	81.90%	RASTER	Throughput			0.29%
	Annual second	L1TEX Throughput	64.84%	PES+VP(	Throughpu	t		0.00%
F GPU Active		VRAM Throughput	1.21%	PD Throu	iahnut			0.00%
		CROP Throughput	1.15%		roughput			0.00%
SM Occupancy (TPC view)		PROP Throughput	0.74%		Toncor Din	a Throughput		0.00%
		RASTER Throughput     PD Throughput	0.16%		TTEOSOF FID			0.00%
		PES+VPC Throughput	0.00%	Warp Occupa	ancy	-1 .		Value
		ZROP Throughput	0.00%	Idle SM	Jnused War	o Slots		3.89%
				Active SI	M Unused Wa	arp Slots		20.25%
				Compute	Warps			0.00%
				Pixel Wa	rps			75.86%
				Vertex+1	Fess+Geom \	Narns		0 00%
	+ (							

٥

\_

 $\times$ 

## 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;

## 

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





# 

#### PSO Creation Stalls on Critical Path

#### Video-Memory Overcommitment

## THE P3 (PEAK-PERF%) METHOD





#### 9

#### DenoiseAO\_Before.nsight-gfxgpt imes



+



#### NVIDIA Nsight Graphics

– 0 ×

Ð

 Eile
 Connection
 Tools
 Window
 Help

 ↔
 Connect
 ↔
 Disconnect
 ×
 Terminate

SM Occupancy

#### DenoiseAO\_Before.nsight-gfxgpt $\times$

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	Summary	Metrics	Detailed Metrics	Capture Inf	ormation	
Frames						Frame	0 (39.82ms)						Start:	26.38 ms	Dura	ation: 5.99	ns	
▼ Q:0	-												End:	32.36 ms	Rang	ge: All Vis	sible	
													Unit Throug	hput			Value	
Synchronization													SM Instr	ruction Throu	ıghput		70.95%	3
													SM FMA	Pipe Throug	Jhput		69.98%	3
All Actions														e Active Throughput			56 57%	
						Frame 5	31 (37.30ms)						SM SFU	Pipe Throug	hput		52.02%	ā
						RayTrace	edAO (6.05ms)						SM ALU	Pipe Throug	hput		25.13%	3
						Denoise	AO (5.99ms)	<u></u>					L2 Thro	ughput			4.56%	4
Markers	Sa Bil	atoralEiltor 102	0v1080 (1.42mc)	Bilate	ralEiltor 1020v	(1080 (1.55ms)	liction (5.99ms)	) oralEiltor 1020v10	180 (1.36ms)	Bilator	Eilter 1020v108	0 (1 40ms)	CROP T	hroughput			2.25%	þ
	50 01		0,1000 (1.42115)	Dirace		(1.551115)	Dilate		00 (1.30ms)	Dilater	11 ILEE 1920X100	0 (1.45115)	VRAM T	hroughput			2.20%	0
													PROP TI	hroughput			1.56%	6
													RASTER	R Throughput			0.29%	6

## Occupancy% = ActiveWarpsPerCycle / NumWarpSlots \* 100

Ū		
7		
	Active SM Unused Warp Slot	s 69.58%
	Pixel Warps	27. <mark>33%</mark>
	Idle SM Unused Warp Slots	3.10%
	Vertex/Tess/Geometry Warps	0.00%
	Compute Warps	0.00%

+

#### NVIDIA Nsight Graphics

Markers

#### - 0 X

File Connection Loois	<u>W</u> indow <u>H</u> elp		
🛛 🖒 Connect 🔅 Discor			
Color By: Stages	<ul> <li>Ruler Relative: Ca</li> </ul>	pture 🔻 🐻 Expor	rt [] Reset Zoom
Time		26.3759ms	27.5ms 28ms 2
	l	3	

-

S Bilater

ms 27.5ms 28ms 28.5ms 29ms 29.5ms 30ms 30.5ms 31ms 31.5ms 32.3615ms	Summary Metrics Detailed Metrics Capture Information Annotations	
Frame 0 (39.82ms)	Start: 26.38 ms Duration: 5.99 ms	
	End: 32.36 ms Range: Selected	
	Metric Value	
	tpcwarp_launch_cycles_stalled_shader_gs.avg.pct_of_peak_sustained_elapsed	0.00
	tpcwarp_launch_cycles_stalled_shader_ps.avg.pct_of_peak_sustained_elapsed	92.89
	tpcwarp_launch_cycles_stalled_shader_tcs.avg.pct_of_peak_sustained_elapsed	0.00
	tpcwarp_launch_cycles_stalled_shader_tes.avg.pct_of_peak_sustained_elapsed	0.00
Narn Launch Stalle	tpcwarp_launch_cycles_stalled_shader_vs.avg.pct_of_peak_sustained_elapsed	0.00
valp Launch Stalls,	tpcwarp_launch_cycles_stalled_shader_vtg.avg.pct_of_peak_sustained_elapsed	0.00
	tpcwarp_launch_cycles_stalled_shader_ps_reason_ooo_warp_completion.avg.pct_of_peak_sustained_elapsed	
non warn completion 7/%	tpcwarp_launch_cycles_stalled_shader_ps_reason_register_allocation.avg.pct_of_peak_sustained_elapsed	0.00
$100_warp_completion. 24/0$	tpcwarp_launch_cycles_stalled_shader_cs_reason_barrier_allocation.avg.pct_of_peak_sustained_elapsed	0.00
	tpcwarp_launch_cycles_stalled_shader_cs_reason_cta_allocation.avg.pct_of_peak_sustained_elapsed	0.00
	tpcwarp_launch_cycles_stalled_shader_cs_reason_register_allocation.avg.pct_of_peak_sustained_elapsed	0.00
	tpcwarp_launch_cycles_stalled_shader_cs_reason_shmem_allocation.avg.pct_of_peak_sustained_elapsed	0.00
	tpcwarp_launch_cycles_stalled_shader_cs_reason_warp_allocation.avg.pct_of_peak_sustained_elapsed	0.17
	smspwarp_issue_stalled_barrier_per_warp_active.pct	0.00
	smspwarp_issue_stalled_dispatch_stall_per_warp_active.pct	0.43
	* smsp_warp_issue_stalled_drain_per_warp_active.pct	0.15
	smspwarp_issue_stalled_imc_miss_per_warp_active.pct	0.12
and the second	smspwarp_issue_stalled_lg_throttle_per_warp_active.pct	0.00
	smspwarp_issue_stalled_long_scoreboard_per_warp_active.pct	29.38
	smspwarp_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
	smspwarp_issue_stalled_selected_per_warp_active.pct	9.86
	smsp_warp_issue_stalled_short_scoreboard_per_warp_active.pct	4.62
	smspwarp_issue_stalled_sleeping_per_warp_active.pct	0.00
	smsp_warp_issue_stalled_tex_throttle_per_warp_active.pct	0.41
	amon warn iccus stalled wait nor warn active not	44 64 /

# Solution: Convert the PS to a CS

## THE P3 (PEAK-PERF%) METHOD



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

## **RTAS Updates**

.....

#### 2-MetroBHV-Before.nsight-gfxgpt imes

Sy Connect Con Di

Color By: Stages 🔹 Ruler Relative: Capture 👻 🔀 Export... [] Reset Zoom Trace Compare...

•

5 (0.16ms)	BuildRaytracingAccel	me 0 (30.54ms) elerationStructure 2-18 Render/R <sup>1</sup>	0.70ms) X-TLAS (0.54ms)		Unit Throu Unit Throu VRAM LITED SM In SM Is SM AI L2 Th SM FI SM FI SM FI	rt: 7.91 ms : 8.61 ms ughput 1 Throughput K Throughput istruction Throughput istruction Throughput LU Pipe Throughput FU Pipe Throughput FU Pipe Throughput	Duratio Range:	n: 0.70 ms All Visible 3.03% 2.34% 2.10% 2.05% 1.78% 1.59% 0.89% 0.72%
5 (0.16ms)	BuildRaytracingAccel	lerationStructure 2-18 Render/R	0.70ms) X-TLAS (0.54ms)		Unit Throo VRAM LITED SM In SM Is SM Al L2 Th SM Fi	: 8.61 ms ughput 1 Throughput K Throughput Istruction Throughput Issue Active LU Pipe Throughput IVA Pipe Throughput FU Pipe Throughput	Range:	All Visible Value 3.03% 2.34% 2.10% 2.05% 1.78% 1.59% 0.89% 0.72%
s (0.16ms)	BuildRaytracingAccel	erationStructure 2-18 Render/R	0.70ms) X-TLAS (0.54ms)		Unit Throu VRAM LITEX SM In SM Is SM AI L2 Th SM FH SM SF	ughput I Throughput K Throughput Istruction Throughput Isue Active LU Pipe Throughput MA Pipe Throughput FU Pipe Throughput		▼ Value 3.03% 2.34% 2.10% 2.05% 1.78% 1.59% 0.89% 0.72%
5 (0.16ms)	BuildRaytracingAccel	lerationStructure 2-18 Render/R	0.70ms) X-TLAS (0.54ms)		VRAM LITEX SM In SM Is L2 Th SM FN SM SN	1 Throughput K Throughput Istruction Throughput Isue Active LU Pipe Throughput Iroughput MA Pipe Throughput FU Pipe Throughput		3.03% 2.34% 2.10% 2.05% 1.78% 1.59% 0.89% 0.72%
5 (0.16ms)		Render/R	X-TLAS (0.54ms)		LITEX SM In SM Is L2 Th SM FN SM SN	K Throughput Istruction Throughput Isue Active LU Pipe Throughput Iroughput MA Pipe Throughput FU Pipe Through <u>put</u>		2.34% 2.10% 2.05% 1.78% 1.59% 0.89% 0.22%
					SM In SM Is SM Al L2 Th SM FM SM SM	Istruction Throughput Isue Active LU Pipe Throughput Iroughput MA Pipe Throughput FU Pipe Throughput		2.10% 2.05% 1.78% 1.59% 0.89% 0.72%
					SM IS SM AL L2 Th SM FM SM SM	sue Active LU Pipe Throughput iroughput MA Pipe Throughput FU Pipe Through <u>put</u>		2.05% 1.78% 1.59% 0.89%
					SM AL L2 Th SM FN SM SF	LU Pipe Throughput roughput MA Pipe Throughput FU Pipe Through <u>put</u>		1.78% 1.59% 0.89% 0.72%
					L2 Th SM Fi	roughput MA Pipe Throughput FU Pipe Through <u>put</u>		1.59% 0.89% 0.72%
					SM FM	MA Pipe Throughput FU Pipe Through <u>put</u>		0.89% 0.72%
					SM SF	FU Pipe Through <u>put</u>		0.72%
					E PROP	Throughput		0.00%
					PD Th	nroughput		0.00%
					PES+	VPC Throughput		0.00%
					RAST	ER Throughput		0.00%
	-				ZROP	Throughput		0.00%
					CROP	Throughput		0.00%
					🕴 📕 SM FF	P16+Tensor Pine Thro	ouahnut	0.00%
					Warp Occ	cupancy		Value
					Idle S	M Unused Warp Slots	<b>1</b>	78.22%
					Active	e SM Unused Warp Slo	ots	18.35
					Comp	oute Warps		3.43%
					Pixel \	Warps		0.00%
					Verter	x+Tess+Geom Warns		0.00%
						Comp Pixel Verte	<ul> <li>Rest Ext Introdytput</li> <li>ZROP Throughput</li> <li>CROP Throughput</li> <li>SM EP16+Tensor Pine Throward Comparison of the SM Unused Warp Slots</li> <li>Active SM Unused Warp Slots</li> <li>Active SM Unused Warp Slots</li> <li>Compute Warps</li> <li>Pixel Warps</li> <li>Vertex+Tess+Genm Warns</li> </ul>	<ul> <li>KASTEX Throughput</li> <li>ZROP Throughput</li> <li>CROP Throughput</li> <li>SM FP16+Tensor Pine Throunhnut</li> <li>Warp Occupancy</li> <li>Tidle SM Unused Warp Slots</li> <li>Active SM Unused Warp Slots</li> <li>Compute Warps</li> <li>Pixel Warps</li> <li>Vartav+Tess+Genm Warns</li> </ul>

## **RTAS Updates**

# olor By: Stages Ruler Relative: Capture Option By: Stages Ruler Ru

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%



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

## Independent Workload #1: Screen-Space PreTracing

- 0 X

œ

olor By: Stages 🔹 👻 Ruler Relative: Capture 👻 🐻 Export... [ ] Reset Zoom 🛛 Trace Compar

Time	6771.46µs 10µs 6850µs 6900µs 6950µs 7000µs	7050µs 7073.01µs 7100µs 7150µs 7200µs	Summary Metrics Detailed Metrics	Capture Information
Frames	Frame 0 (30.54ms)		Obachi 6.77 ma	
			End: 7.24 ms Range:	All Visible
Inc Actions		I lot Theory house	Unit Throughout	
All Actions		Unit Inrougnputs:	SM Instruction Throughput	61.51%
- All Activity		™ SM·67%	SM FMA Pipe Throughput	61.46%
Markers			SM Issue Active	53.95%) 49.80%
		L1TEX:54%	L2 Throughput	42.65%
V Unit Throughouts			SM SFU Pipe Throughput	33.86%
• one moughputs			VRAM Throughput	15.30%
			PROP Throughput	0.00%
SM Instruction Throughput			PD Throughput	0.00%
L1TEX Throughput			PES+VPC Throughput	0.00%
L2 Throughput			RASTER Throughput	0.00%
VRAM Throughput			CROP Throughput	0.00%
PD Throughout			SM FP16+Tensor Pine Throughout	0.00%
PEC IVPC Throughout			Warp Occupancy	Value 🐣
FESTVPC Infoughput				1.03%
RASTER Throughput			Active SM Unused Warp Slots	4.46%
PROP Throughput			Pixel Warps	0.00%
ZROP Throughput			Vertex+Tess+Geom Warns	n nn%
CROP Throughput				
F GPU Active				
SM Occupancy (TPC View)				

#### Independent Workload #2: SSR

9

Color By: Stages 🔹 👻 Ruler Relative: Capture 👻 🐻 Export... [ ] Reset Zoom 🛛 Trace Compa



# ASYNC COMPUTE DIFF

#### **Serialized** 1.83 ms 7ms 7.06575ms 's 7.2ms 7.3ms 7.4ms 7.5ms 7.6ms 7.7ms 7.8ms 7.9ms 8.1ms 8.2ms 8.3ms 8.4ms 8.5ms 6.9ms 8ms 6.77262ms Time Frame 0 (30.54ms) Frames -V Q:0 Inc Actions All Actions BuildRaytracingAccelerationStructure 2-18 (0.70ms) Rend Render/RTX-Pretrace (0.46ms) SSR-trace (0.64ms) Render/RTX-BLAS (0. Render/RTX-TLAS (0.54ms) Markers



# **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

52) Connect Es2 Disconnect X Terminate

 $\odot$ 

#### BF5\_AsyncCompute-Before.nsight-gfxgpt $\times$

Color By: Stages 🔹 Ruler Relative: Capture 👻 🐻 Export... [] Reset Zoom Trace Compare...



## Independent Workload: Shadow Maps

Color By: Stages Ruler Relative: Capture 12.8ms 12.9ms 12.4ms Frame 0 (16.88ms) Top Throughput Units: -L1TEX: 38% GameRenderer::render (11.01n **VRAM: 37%** worldView (9.50ms) **ZROP: 29%** shadowRendering (1.47ms



Start: End:	11.97 ms 13.44 ms	Duration: Range:	1.47 r All Vis	
Unit Through				Value 🏾 🛎
LITEX TH	roughput			38,34%
VRAM TH	roughput			36.68%
ZROP Th	roughput			29.10%
L2 Throu	ghput			22.04%
SM Instru	action Throughput			21.77%
SM Issue	Active			20.54%
PROP Th	roughput			20.48%
SM FMA	Pipe Throughput			17.52%
RASTER	Throughput			17.42
CROP Th	roughput			12.11%
SM ALU F	Pipe Throughput			8.69%
SM SFU	Pipe Throughput			7.62%
PES+VPC	: Throughput			5.61%
PD Throu	ighput			4.31%
SM FP16	+Tensor Pine Throughn			0.00%
Warp Occupa				
				20.09%
Active SM	1 Unused Warp Slots			21.51%
Compute	Warps			0.05%
Pixel Wa				51.97%
Vertex+T	ess+Geom Warns			6 37%

<sup>1</sup> 13.3324ms <sup>.3.4ms</sup>



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

#### BEFORE AFTER RATIO 2.15 ms 1.14x Gain **GPU Elapsed Time** 2.45 ms 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

# Case Study #3: Bad Async-Compute Pairing

## Blur Compute Shader

Sconnect Lev disconnect A terminate

#### -BadPairing-Before.nsight-gfxgpt 🗙 4-BadPairing-After.nsight-gfxgpt 🕽

– 0 ×

54.35% 31.89% 23.92% 23.86% 19.13% 11.06% 7.17%

93.34%

 $\odot$ 

	26.3519ms 26.5ms	26.6ms 26.7ms 26.8ms	26.9ms 27ms 27.1ms	27.2ms 27.3ms 27.4ms	27.5m 27.554 27.6462ms		Metrics Detai	led Metrics
			Frame 0 (45.28ms)			Ctarts	26.25 mc	Dure
						End:	27.65 ms	Rang
						10.1		
nchronization				Unit Through	nute	Unit Through	out	
				Unit mougi	puls.		rougnput	
Actions							rougnput	
			Workload_89 (1.30ms)	<b>VRAM: 61</b> %		EZ Throu	Jupur Inpur	
arkers			Workload_94 (1.29ms)			SM EMA	Dine Throughput	
				L11EX: 54%		SM Issue	Active	
						SM SEU	ine Throughput	
				17.37%		SM ALU F	ipe Throughput	
broughoute		Construction of the second				PROP Thr	oughput	
nioughputs						PD Throu	ghput	
						PES+VPC	Throughput	
ctive						RASTER -	Throughput	
						ZROP Thr	oughput	
	_					CROP Th	roughput	
ccupancy (TPC View)						SM FP16-	FTensor Pine Throu	iahnut
						Warp Occupa		
						Idle SM L	nused Warp Slots	
						Active SN	I Unused Warp Slot	
						Compute	Warps	
						Pixel War	ps	
						Vertex+T	ess+Geom Warns	

#### Independent Workload #1: Water Simulation

œ

Ruler Relative: Capture 👻 Color By: Stages 27739.9µs 27850µs 27700µs Frame 0 (45.28ms) Duration: 0.24 ms Range: Unit Throughputs: 22.51% L1TEX Throughput **VRAM: 23**% CROP Throughput 18.58% 14.79% L2 Throughput L1TEX: 22% 10.27% SM Issue Active 10.27% PROP Throughput 6.16% **CROP: 19%** 5.01% RASTER Throughput 2.29% SM FMA Pipe Throughput 1.70% 1.08% PES+VPC Throughput PD Throughput 26.14% Idle SM Unused Warp Slots 15.79% 0.00%

## Independent Workload #2: GBuffer Fill

Color By: Stages 💎 Ruler Relative: Capture 🔻 🐻 Export... [] Reset Zoom Trace Compare...



D

X

# **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	<b>54.9</b> %	57.5%	1.05x
L2 Read Hit Rate	52.3%	44.5%	0.85x

On RTX 2080 with SetStablePowerState(TRUE)

## Blur Compute Shader

W Connect Lev Disconnect X Ferminate

#### I-BadPairing-Before.nsight-gfxgpt 🗙 4-BadPairing-After.nsight-gfxgpt 🕽

- 0 X

.....

Color By: Stages	Ruler Relative: Ca		Export	[] Reset Zoom	Trace Compare					
Time		26.3519ms	26.5ms	26.6ms	26.7ms	26.8ms	26.9ms	26.9791ms	27. 1ms	27.2ms
	I		an.					Frame 0 (45.2	.8ms)	
	ļ									
Curcheronization					2 2					

Synchronization	Unit Throughputs:	Unit Throughput VRAM Throughput L1TEX Throughput	✓ Value ▲
All Actions Markers	Workload_89 (1.30ms) Workload_94 (1.29ms) L1TEX: 54%	L2 Throughput SM Instruction Throughput SM FMA Pipe Throughput SM Issue Active	31.89% 23.92% 23.86% 19.13%
Unit Throughputs	L2: 32%	SM SFU Pipe Throughput SM ALU Pipe Throughput PROP Throughput PD Throughput PES+VPC Throughput	11.06% 7.17% 0.00% 0.00%
<ul> <li>GPU Active</li> <li>SM Occupancy (TPC View)</li> </ul>		ZROP Throughput CROP Throughput SM FP16+Tensor Pine Throughput	0.00% 0.00% 0.00% 0.00%
L2 Traffic		Warp Occupancy Idle SM Unused Warp Slots Active SM Unused Warp Slots	Value 1.03%
L2 Read Hit Rates L2 Read Hit Rate L2 Read Hit Rate from L1TEX	L2 Read Hit Rate 29.	Compute Warps Pixel Warps Vertex+Tess+Genm Warns	( <u>93.34%)</u> 0.00% ∩ ∩∩% ▼
	L2 Read Hit Rate: 29%		

\*\*\*\*\*

#### Independent Workload #1: Water Simulation

œ

Color By: Stages 🔹 Ruler Relative: Capture 👻 🐻 Export... [] Reset Zoom Trace Compar


### Independent Workload #2: GBuffer Fill

Color By: Stages 🔹 Ruler Relative: Capture 🤟 🐻 Export... [ ] Reset Zoom Trace Compare.



### THE P3 (PEAK-PERF%) METHOD



# Case Study #4: VRAM-Limited Denoiser CS



File Connection Tools Window Help

♦) Connect (▷) Disconnect × Terminate

#### .....

#### 5-Denoising\_CS-Before\_2.nsight-gfxgpt 🗙





<u>F</u>ile <u>C</u>onnection <u>T</u>ools <u>W</u>indow <u>H</u>elp

◆) Connect ↔ Disconnect × Terminate

6-Denoising\_CS-Before\_2.nsight-gfxgpt

Color By: Stages 🔹 Ruler Relative: Capture 👻 🐻 Export... [] Reset Zoom Trace Compare...



o ×

\_

◆) Connect ↔ Disconnect × Termina B Export... [] Reset Zoom Trace Compare... Color By: Stages Ruler Relative: Capture 22ms 22.5ms Metrics Detailed Metrics Capture Information 21.5ms 23ms 23.4052ms 21.0545ms -Frame 0 (27.66ms) Start: 21.05 ms Duration: 2.37 ms -Range: All Visible End: VRAM Inc Actions Unit Throughputs: VRAM Bandwidth 28.05% VRAM Read: 54% VRAM Read Throughput 54.19% Dispatch[PSO\_0\_0\_0\_0\_584F3225] (2.37ms) Markers VRAM Throughput VRAM Write Throughput 1.90% VRAM Write: 2% GPU Active V L2 Read Hit Rates L2 Read Hit Rate L2 Read Hit Rate from L1TEX VRAM Bandwidth VRAM Read Throughput VRAM Write Throughput

<u>File Connection Tools W</u>indow <u>H</u>elp

Connect ⊕ Disconnect × Terminat





- 0 X

.....

<u>File Connection Tools Window Help</u>

 $\Rightarrow$  Connect  $\Rightarrow$  Disconnect imes Terminat

6-Denoising\_CS-Before\_2.nsight-gfxgpt imes

Color By: Stages 🔹 Ruler Relative: Capture 💌 🐻 Export... [] Reset Zoom Trace Compare...

Time	21.0545ms	21.5ms	22ms	22.5ms	23ms	23.4163ms	Summary	Metrics	Detailed Metrics	Capture Information	Annotations
Frames			Frame 0	(27.66ms)			Start	21.05 ms		Duration: 2.37 n	ne
▼ Q:0	8						End:	23.43 ms		Range: All Vis	ible
Synchronization							Metric				Value
Inc Actions							l1tex_t_sec	tor_hit_rate.	oct		55.52
							smspwarp	p_issue_stalle	d_barrier_per_warp	_active.pct	0.00
All Actions							smspwarp	p_issue_stalle	d_dispatch_stall_per	_warp_active.pct	0.15
Markers			Dispatch[PSO_0_0_0	_0_584F3225] (2.37ms)			smspwarp	p_issue_stalle	d_drain_per_warp_a	active.pct	0.26
							smspwarp	p_issue_stalle	d_imc_miss_per_wa	rp_active.pct	0.12
				– Warn Issu	e Stalled	on	smspwarp	p_issue_stalle	d_lg_throttle_per_w	arp_active.pct	0.00
Unit Throughputs				Marp 155a			smspwarp	p_issue_stalle	d_long_scoreboard_	per_warp_active.pct	77.27
			and the second se	l ang Scar	cohoard	770/	smspwarp	p_issue_stalle	d_math_pipe_throttl	e_per_warp_active.pct	1.31
				LUNG SCU	epuaru.	/ /0	smspwarp	p_issue_stalle	d_membar_per_war	p_active.pct	0.00
GPU Active							smspwarp	p_issue_stalle	d_mio_throttle_per_	warp_active.pct	4.80
							smspwarp	p_issue_stalle	d_misc_per_warp_a	ctive.pct	0.10
							smspwarp	p_issue_stalle	d_no_instruction_pe	r_warp_active.pct	1.72
SM Occupancy (TPC View)					V latana		+ ~ •			:tive.pct	2.95
				<b>FIGURIE</b>	іх іагенс	v aue	LO:			pct	4.76
										rp_active.pct	1.67
L2 Read Hit Rates		IIIII IIIIIIIII		💶 I 7 Read F	lit Rate	37%				pct	0.00
VRAM Bandwidth	•				IIC Matter					ive.pct	80.0
				11 Hit Da	$t_{0}, 55\%$						4.39
				LIIIL NA							

### → TEX-Latency Limited

– 0 ×

**\_** 











Thread groups launched sequentially have far-appart pixel coordinates

- $\rightarrow$  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

### THE P3 (PEAK-PERF%) METHOD



### Case Study #5: Lighting CS using Shared Memory

<u>File Connection Tools Window H</u>elp

-LightingCompose-Before.nsight-gfxgpt 🗙





<u>F</u>ile <u>C</u>onnection <u>T</u>ools <u>W</u>indow <u>H</u>elp

Connect ( Disconnect × Terminat

8-LightingCompose-Before.nsight-gfxgpt  $\times$ 

Color By: Stages 🔹 Ruler Relative: Capture 👻 🐻 Export... [] Reset Zoom Trace Compare.

Time	26 26.4996ms 27ms 27.5ms 28ms 28ms 28.5ms 29ms 29.5537ms	Summary Metrics Detailed Metrics Capture Information Annotations		
	Frame 0 (54.74ms)	Start: 26.50 ms Duration: 3.07 ms		
▼ Q:0		End: 29.57 ms Range: All Visible		
Synchronization		Metric Value toc warp launch cycles stalled shader os reason barrier allocation avo not of neak sustained elansed	0.00	
		tpcwarp_launch_cycles_stalled_shader_cs_reason_sta_allocation.avg.pct_of_peak_sustained_elapsed	0.00	
All Actions	ward launch stalls on	tpcwarp_launch_cycles_stalled_shader_cs_reason_register_allocation.avg.pct_of_peak_sustained_elapsed	1.19	
	shmem allocation: 93%	tpcwarp_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_drain_per_warp_active.pct	0.12	
Unit Throughputs		smspwarp_issue_stalled_imc_miss_per_warp_active.pct	3.80	
	Warn issue stalls on	smspwarp_issue_stalled_lg_throttle_per_warp_active.pct	0.00	
		smspwarp_issue_stalled_long_scoreboard_per_warp_active.pct	41.75	
GPU Active	long scoreboard: 12%	smspwarp_issue_stalled_math_pipe_throttle_per_warp_active.pct	4.99	
	$10119_3COTCDOATO, 42/0$	smspwarp_issue_stalled_membar_per_warp_active.pct	0.00	
SM Occupancy (TPC View)		smsp_warp_issue_stalled_mio_throttle_per_warp_active.pct		
		smsp_warp_issue_stalled_misc_per_warp_active.pct	0.10	
		smsp	6.60	
SM Warp Can't Launch		smspwarp_issue_stalled_selected_per_warp_active.pct	11.79	
		smspwarp_issue_stalled_short_scoreboard_per_warp_active.pct	4.43	
		smspwarp_issue_stalled_sleeping_per_warp_active.pct	0.00	
		smspwarp_issue_stalled_tex_throttle_per_warp_active.pct	0.00	
		shispwarp_issue_sumed_warper_warp_active.pcc		

## 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	<b>9</b> 4%	0%	+INF
Warp issue stalls on long_scoreboard	41.7%	36.7%	1.14x

### THE P3 (PEAK-PERF%) METHOD



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

### Battlefield V DXR Ultra

SP/

SECURE CHOKEPOINT KILO



•

Original.nsight-gfxgpt imes

 Eile
 Connection
 Tools
 Window
 Help

 (42)
 Connect
 (42)
 Disconnect
 X
 Terminate

.....

Time	12.4854ms 12.6ms 12.7ms 12.8ms 12.9ms 13ms 13.1m	3 40 mm 40 mm 40 Amm 40 Fmm	Summary Metrics Detailed Metrics Capture	Information Annotations
Frames	Frame 0 (20.39ms)	Unit Throughputs:	Strutt 12.40 mc	Duration: 112 mc
<b>V</b> 0:0			End: 13.61 ms	Range: Selected
Synchronization				
The Actions			Unit i nroughput	23.78%
Inc Actions		VRAM: IZ%	VRAM Throughput	12.16%
All Actions	DispatchRays 2 (1.13r		SM Instruction Throughput	7.63%
		SM: 8%	SM Issue Active	7.63%
Markors			L1TEX Throughput	7.42%
			SM FMA Pipe Throughput	5.81%
			SM ALU Pipe Throughput	5.33%) 2.42%
			PROP Throughput	0.00%
			PD Throughput	0.00%
Unit Throughputs			PES+VPC Throughput	0.00%
			RASTER Throughput	0.00%
			2ROP Throughput	0.00%
F GFU Active				
			Idle SM Unused Warp Slots	5.80%
SM Occupancy (TPC View)			Active SM Unused Warp Slots	57.11%
			Compute Warps	37.09%

### NVIDIA Nsight Graphics Elle Connection Tools Window Help

S Connect (S Disconnect X Termination			
Color By: Stages Ruler Relativ	ve: Capture * 👔 Export ; JReset Zoom Trace Compare		
	12.55mc 12.65mc 12.7		
	12.482.3M5 12.492.3M5 12.731.3M5 12.731.3M5 12.731.3M5 12.731.3M5 12.732.3M5 12.7323.3M5 12.7332.3M5 12.7323.3M5 12.7323.3M5 12.7323.3M5 12.7332.3M5 12.7323.3M5 12.7323.3M5 12.73323.3M5 12.7323.3M5 12.73323.3M5 12.73323.3M5 12.73323.3M5 12.73323.3M5 12.73323.3M5 12.73323.3M5 12.73323.3M5 12.73323.3M5 12.73323.3M5 12.7333333333333333333333333333333333333		
	Frame 0 (20.39ms)	Start: 12.48 ms Duration: 1.13 ms	
		End: 13.61 ms Range: Selecte	
Inc Actions		l1tex_t_sector_hit_rate.pct	
All Actions	Active Threads per Inst Executed% 79% << 50%		
	Active mileuds per mise Excedeed/0. 27% × 50%	smspwarp_issue_stalled_barrier_per_warp_active.pct	
	Indicates packing issue and (or high divorgence	smsp_warp_issue_stalled_dispatch_stall_per_warp_active.pct	0.01
	Indicates backing issue and/or high divergence	smsp_warp_issue_stalled_drain_per_warp_active.pct	
		smspwarp_issue_stalled_init_miss_per_warp_active.pct	0.09
	Peduce number of hit shaders	smspwarp_issue_stalled_ing_unotals_per_warp_active.pct	10.34
		smspwarp_issue_stalled_math_pipe_throttle_per_warp_active.pct	
		smspwarp_issue_stalled_membar_per_warp_active.pct	
Markers		smspwarp_issue_stalled_mio_throttle_per_warp_active.pct	
		smspwarp_issue_stalled_misc_per_warp_active.pct	
		smspwarp_issue_stalled_no_instruction_per_warp_active.pct	
		smsp_warp_issue_stalled_not_selected_per_warp_active.pct	0.23
		smspwarp_issue_stalled_selected_per_warp_active.pct	2.70
		smspwarp_issue_stalled_short_scoreboard_per_warp_active.pct	10.68
		smsp_warp_issue_stalled_steeping_per_warp_active.pct	0.60
		smspvarp_issue_stalled_wait_per_warp_active.pct	
		• sm_inst_executed.sum	
Unit Throughputs			
GPU Active			

smsp\_\_thread\_inst\_executed\_pred\_on\_per\_inst\_executed.pct: The percentage of active not predicated off threads per instruction executed



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  $\times$ 

## 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

SECURE CHOKEPOINT KILD

GeForce RTX 2080 + SetStablePowerState

iiii 20 /10, j#3

۲

Battlefield V DXR +OPAQUE flag for ALL geometries

DispatchRays: 0.58 ms (1.95x)

SECURE CHOKEPOINT KILO

GeForce RTX 2080 + SetStablePowerState

IIII 20 AV, J 3

٠

## 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

### THE P3 (PEAK-PERF%) METHOD



## **NVIDIA TOOLS FOR GPU PROFILING**

### Two Nsight: Graphics modules

	GPU TRACE METRICS	RANGE PROFILER
Graphs over time	Y	Ν
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 <u>a blog post</u> up-to-date as tools and GPUs evolve.

- Async Compute rule of thumb:
  - Do not overlap 2 VRAM-latency-limited workloads



## **QUESTIONS?**

Louis Bavoil | <u>lbavoil@nvidia.com</u>

www.nvidia.com/GDC