Enabling the Next Generation of Computational Graphics with NVIDIA® Nsight™ Visual Studio Edition

## Jeff Kiel Director, Graphics Developer Tools



## **Computational Graphics Enabled**



- Problem: Complexity of Computation Graphics
- Solution: Kepler with Maximus<sup>™</sup> & CUDA<sup>®</sup> 5.0
- How To: NVIDIA® Nsight<sup>™</sup> Visual Studio Edition
- Questions?









#### Real-World Physics During Interactive Design

Mechanical Dynamics
 Fluid Dynamics
 Light Simulation



#### Real-World Materials During Interactive Design

- Physical Properties
- Photorealism
- ✓ Global Lighting







# Real-World Physics in Scene Design

Smoke
Fluid Dynamics
Light Simulation



#### High-Performance Parallel Processing

Color Correction
 Effects
 GPU Rendering



Smoke effects created on NVIDIA Maximus for MIB3 courtesy Sony Pictures

Real-time rendering on NVIDIA Maximus for GDC courtesy Dawnrunner Prod.



Traditional Workstation





## Solution: Kepler, Maximus & CUDA 5.0 ....tamed by Nsight



## Kepler Architecture

 SMX, Dynamic Parallelism, Bindless Textures, FXAA/TXAA, PCIe Gen3, HW H.264 encode, 4 displays

Maximus

- Quadro runs graphics (OpenGL/Direct3D)
- Tesla handles computation
- Seamlessly supported by over 22 professional applications
- CUDA 5.0
  - Dynamic Parallelism
  - RDMA for GPUDirect
  - GPU Library Object Linking





















# **NVIDIA® Nsight™ Visual Studio Edition**



## Visual Studio integrated development for GPU and CPU



# NVIDIA Nsight® Visual Studio Edition





#### **GPU Debugger**

GPU native Compute and Graphics debugging GPU breakpoints including complex conditionals GPU memory views and exception reporting Dynamic Shader Editing



#### Frame Debugger

Real-time inspection of Graphics APIs Investigate GPU pipeline state See contributing fragments with Pixel History Profile frames to find GPU bottlenecks



## **System Analysis**

View CPU & GPU events on a single timeline Examine workload dependencies/memory transfers CPU/OS, File I/O, Compute, Graphics API Trace Capture call stack and jump to source



## Free and Integrated into Visual Studio 2008 and 2010

## **NVIDIA Nsight 3.0 – CUDA Features**

- Maximus support debug CUDA and Graphics
- CUDA 5.0
- Advanced CUDA Profiling
  - Source correlated profiling
  - Dynamic kernels and memory transfer tracing
- View interop surfaces while debugging CUDA
- Debug and trace dynamic kernels
  - **CUDA library usage**







## **NVIDIA Nsight 3.0 – Graphics Features**

- Single GPU Debugging for Graphics
- Support for OpenGL
  - All new GUI
  - GLSL Shader Debugging
  - Frame Debugger
  - Frame Profiler
  - **Pixel History**











## **Demo Application: FluidsGL**



- Developed by Cliff Woolley & Nolan Goodnight
- Compute the fluid simulation on one GPU
  - Simple fluids solver for the Navier-Stokes equations for incompressible flow
  - Jos Stam's FFT-based stable fluids system
- Render the results in OpenGL on second GPU
  - Diffuse texture & normal map
  - Sample in UV space & screen space then blend
  - Gives look that there is a 2<sup>nd</sup> draw call in the background





## **Demo: Setup your CUDA project**







## **Demo: Launching...**



#### Start Nsight Monitor



#### Configure Parallel Nsight Project Settings

Launch	Launch	
Synchronization Shader Compilation	Launch Action	
	C Launch project	
	Iaunch <u>external program</u> :	E:\sw\devtools\Nexus\Sandbox\FluidsGL.Demo\b
	Launch Options	
	<u>C</u> onnection name:	localhost
	Command line <u>a</u> rguments:	
	Working directory:	E:\sw\devtools\Nexus\Sandbox\FluidsGL.Demo\F
		OK Cancel

#### Launch Your Application





## **Demo: Single GPU & Memory Checker**





## **Demo: CUDA Kernel Debugging**





## **Demo OpenGL Shader Debugging**





## **Demo: OpenGL Frame Debugger**





## **Demo: FluidsGL Analysis**





# 

#### 🔤 SampleViewer 9.4 fps

Scene 1: Column Fracture General: 1-8 or F10: Reset scene p: pause o: single step w,a,s,d,q,e: camera x render mode: optix F1: toggle help F2: num shadows F3: reflection -F4: change weapon ES: toggle camera mode F6: togole slow motion EZ: print profiling info Shift-mouse: drag objects Space: Shoe

Simulation and Ray Tracing run on GPU, display frame buffer in OpenGL







) Timeline																						
Target Process: SampleViewer.exe [1364] Clear																						
w Filters																						
Time L: 3.95254250s R: 8.99687110s	4	4	4.5	5		5.5		6		6.5		7			C	DI	11/	<b>hr</b> k		de f	or	
Processes															G	ГU	<u> </u>	JIN	IUa	<b>U</b> 3 I	UI	
<ul> <li>SampleViewer.exe [1364]</li> </ul>		_																				
Thread 81.9% [2632]									~ ~ ~	T					mu	ltin		n	nnu	ito r	and	
CUDA															IIIU	πρ			npu		anu	
Context 0																						
Context 1 [0]															0	rar	shic	re c	nn.	tovt	<u>e</u>	
Runtime API	Y														9	Tap	אחת	22		τολι	S	
Driver API	Y																					
Memory	Y			- U I 🗍 II			11 111			11 11 1			I I <b>II</b>									
Compute	Y				1111-111	1 111 1	II III I	1  111	11 11 11	11 1 1	1 11	11 1 11	I III					11 1				
Counters																						
Context 2 [0]																						
Runtime API	Y											· •				I I			I III I			
Driver API	Y																					
Memory	Y							1 1	10 10 10													
Compute	Y																					
Render Context 0x0																						
Render Context 0x10000																						
	CPU	1 3 6	9 18	19 20 2	1 22 23	24 25 2	26 27 28	29	30 31	32 33	34 35	36	37 3	8 39	40	41 4	2 43	44	45 4	46		
<ul> <li>Device Context 0xFE0110D9</li> </ul>	3PU GPU 0 - G 🚺	3 6	9	3 19 20	21 22 23	24 25	26 27 28	29	30 31	32 33	34 35	36	37	38 39	40	41 4	12 43	44	45	46		
Transfers	Y	11 111 111	111 111 111		1 1		1 1	1	1 1 1	1 1	1	1 1	1	1	1 1	1	1	1	1 1	1		
Draw Calls [GPU 0]	Y	111 111 111			i i	i i [	T. T.	1	I [ [			[]]	ľ	[	[	· [	[	[	[ [			
System											1											
CPU Usage																						
CPU Process Allocation																						
GPU Usages	1		AAA AAA AAA A		4. 44																	
- 15	1048848																					
± vSyncs	1048851																					
<ul> <li>GPU Driver Stats</li> </ul>																						
Adapter - 0xFFFFFA800E2F5040	Active Process	i sin i si s	n n de la constante de la constante en la const		واوددي					الكالكال					(1)							
GPU Queue	Active Context		<b>, , , , , , , , , , , , , , , , , , , </b>		وروديين																	
<ul> <li>csrss.exe [540] - DXGK Device 0xFFFFF8A</li> </ul>	Active Process												1		11 1							
<ul> <li>csrss.exe [540] - DXGK Device 0xFFFFF8A</li> </ul>	Active Process																					
<ul> <li>devenv.exe [384] - DXGK Device 0xFFFFF</li> </ul>	Active Process																					
<ul> <li>firefox.exe [3064] - DXGK Device 0xFFFFF</li> </ul>	Active Process																	1				
<ul> <li>SampleViewer.exe [1364] - DXGK Device 0</li> </ul>	Active Process			111 0 01	01 01	1 11 1	0.011	11.1111	1 1 1		0.0011	0.01	11	0	0111 1			11 1		11		
<ul> <li>SampleViewer.exe [1364] - DXGK Device 0</li> </ul>	Active Process	1110 010 010	THE HE HE	1 11			1 1		1 1 1	1 1		1 1	1	1	1		1		L I	1		
SampleViewer.exe [1364] - DXGK De	Active Process								LLL				- T	E		Γ Γ		E				
wmplayer.exe [4952] - DXGK Device 0xEE	Active Process					n dina in		11111					11	11				al (1				









## **Source Correlated Profiling**



essell	ation120801apture_000.nvreport ×												Corr	espor	ndir	ng A	<b>\S</b> N	/I fo	r		
6	CUDA Source View 🔹 🕅												se	lected	lso	urc	e li	ne			
💌 ti	esselateNURBSPatches_Kernel	Grid Dim	n: {100, 1, 1}								Block Dim: {	512, 1, 1}					•				
ile: r	nurbstessellation4.cu   View: Source and SASS		▼ Kerr	nel: tes	seiateNL	JRBSPatc	hes_Ken	nel	•		High to	Low: A V Low to								)	
	4			1		1	1	· I				1						ŧ	1	1	1
.ine S	jource	Instruct Execute Exec	read tructic cutec	d Thre ict Instr ite Pred	Brancł Taken	Brancł Not Taken	Brai Bra Dive Bao	Mer		Line	Source			Instru Exect	Instruct Execute	Thread Instruct Execute (False)	Threa Instru Brai Predic Tak Off	nci Not Taken	Bra Br Div Ba	Memory Ty	be A Tj
49				,						1125	/*2320*/	MOV R4, R2;		160	0 14000	0	0.0				
50	<pre>if ( pfKnots[i+p]!=pfKnots[i] )</pre>	10000 12	22000 1400	00 11.5				Gen		1126	/*2328*/	CAL 0x2198; #	Target=0x00029	160	0 14000	0	0.0				
51	{									1127	/*2330*/	LD R3, [R22];		160	0 14000	0	0.0			Generic, Glo	bal Lo
52	C1 = ComputeCoefficient( pfKnots, nInt	16000 14	10000	0 0.0						1128	/*2338*/	LD R0, [R24];		160	0 14000	0	0.0			Generic, Glo	bal Lo
53	C2 = ComputeCoefficient( pfKnots, nInt	1600 1	L4000	0 0.0						1129	/*2340*/	NOP;		V							
54										1130	/*2348*/	FFMA R4, R3,	-R4, R26;	160	0 14000	0	0.0				
55	fResult += ( C1 - pfKnots[i]*C2 )/( pf	8000 7	70000	0 0.0				Gen		1131	/*2350*/	FADD R5, R0,	-R3;	160	0 14000	0	0.0				
56	}									1132	/*2358*/	CAL 0x26f0; #	Target=0x000298	f0 160	0 14000	0	0.0				
57	<pre>if ( pfKnots[i+p+1]!=pfKnots[i+1] )</pre>	10000 12	22000 1680	00 13.8				Gen		1133	/*2360*/	F2F.S R26, R4;	;	160	0 14000	0	0.0				
58	{									1134	/*2368*/	LD R0, [R22+0:	x4];	200	0 24400	0	0.0			Generic, Glo	bal Lo
59	C1 = ComputeCoefficient( pfKnots, nInt	26000 21	18400							1135	/*2370*/	LD R3, [R24+0:	x4];	200	0 24400	0	0.0			Generic, Glo	bal L
60	C2 = ComputeCoefficient( pfKnots, nInt	2000		0.0	-			-		1136	/*2378*/	IADD R25, R25	, 0x1;	200	0 24400	0	0.0			,	-
								- <b>F</b>		- ا											•
tes	sselateNUDBCRatebox Konsel (CUDA Laurab)	2	gr	oup by th SASS Lin	at columr ne # 🍸	Instruc	tions Ex	ecuted S	Th	read Ir						rated	Off 🏹 🎜	Active Mas	k		Pr

111

Hot spots at the kernel source level

- -

Exp

CUDA Divergent Branch

SASS Line # 🏹	Instructions Executed $\nabla$	Thread I
<u>1127</u>	1600	
<u>1128</u> 🍋	1600	
<u>1130</u> 🖑	1600	
<u>1131</u> 🖑	1600	
<u>1132</u> 🖑	1600	
<u>1133</u> 🖑	1600	
Sugar	2000	

See which threads in the warp hit the current line





## **Supported Configurations: Desktop**



CUDA & Graphics Analysis CUDA & Graphics Profiling CUDA & Graphics Debugging CUDA Memory Checker **Graphics Frame Debugger Graphics Pixel History** 







2 GPUs Tesla®+GPU







**Remote PC** 



## **Supported Configuration: Laptop**



CUDA Analysis
 CUDA Profiling
 CUDA Debugging
 CUDA Memory Checker



**Optimus™ Laptop** 





# Recap: Nsight Tames Computational Graphics

- Maximus & CUDA 5.0 support
- Simultaneous CUDA & Graphics Debugging
- Single GPU Debugging for Graphics and CUDA
- OpenGL Frame Debugger and Profiler
- Tracing CPU & GPU workloads and memory transfers on all APIs
- Source correlated performance information





## Thank you...questions?

- Join our Registered Developer Program
- Grab latest version from

http://developer.nvidia.com/nsight





# GTC 2013 | March 18-21 | San Jose, CA

The Smartest People. The Best Ideas. The Biggest Opportunities.

## Showcase Your Work Among the Elite of GPU Computing!

#### **Call for Submissions**

- Call for Sessions: August 2012
- Call for Posters: October 2012

As a speaker you'll receive many benefits that include:

- A complimentary Full Conference pass
- Discounted passes for your colleagues
- Access to exclusive speaker-only on-site amenities
- Audio capture and hosting of your session
- A unique opportunity to promote your name, your expertise, and your company

# es

#### Learn more at www.gputechconf.com