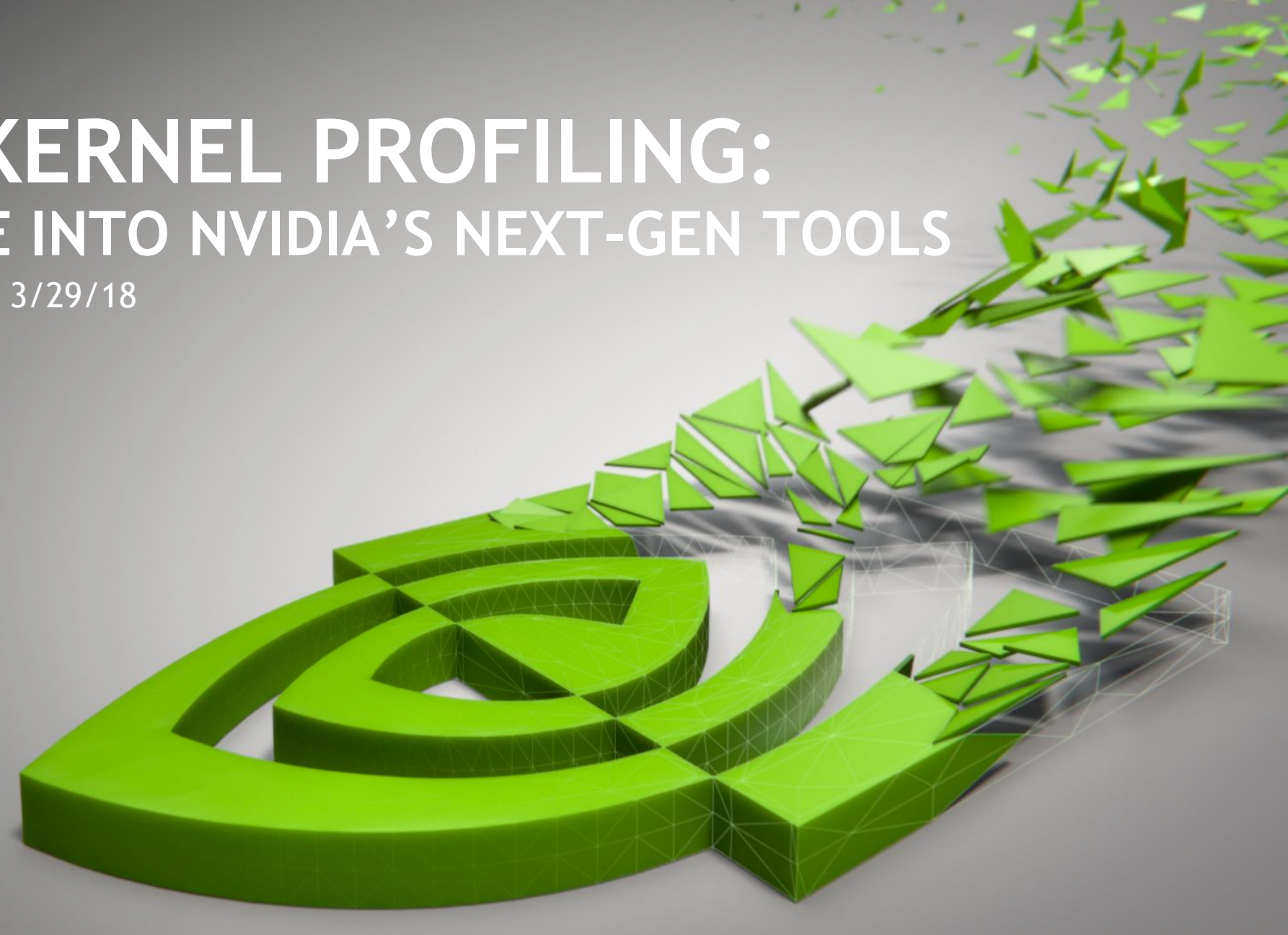
CUDA KERNEL PROFILING:
DEEP-DIVE INTO NVIDIA'S NEXT-GEN TOOLS

Magnus Strengert, 3/29/18

NVIDIA



UPDATES FOR CUDA 9.2

NVPROF

Many New Metrics:
- Tensor Core Metrics
- L2 Metrics
- Memory Instructions Per Load/Store

Display PCIe Topology

Collect Trace and Profile in same pass (`--trace`)

CUPTI

New Activity Kind: PCIE

New Attribute: Profiling Scope (Device-Level, Context-Level)

Exposes New Metrics

VISUAL PROFILER

Summary View for Memory Hierarchy

Improved Handling of Segments for UVM Data on the Timeline

NVIDIA NSIGHT COMPUTE

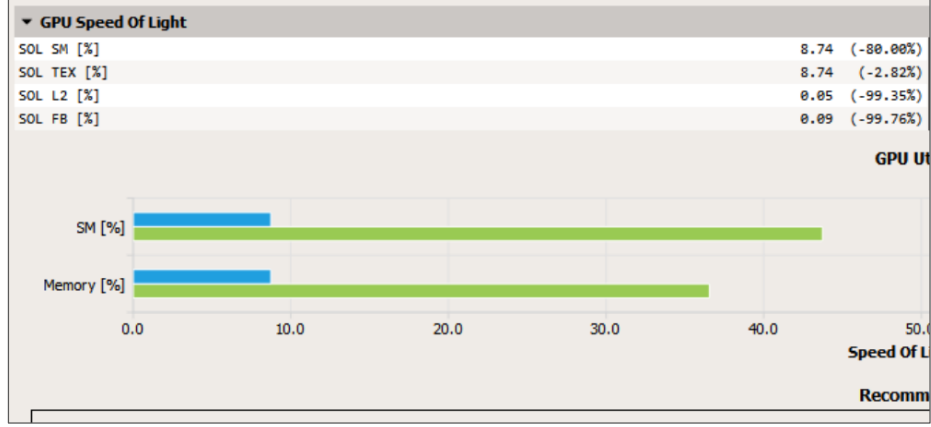
Next-Gen Kernel Profiling Tool

Key Features:

- Improved Workflow (Diff'ing Results)
- Fast Data Collection
- Command Line, Standalone, VS Integration
- Fully Customizable (Programmable UI/Rules)

OS: Windows, Linux

GPUs: Pascal, Volta



inst_executed [inst]	16,528.00; 16,528.00; …	13,476.00; 13,476.00; …
l1tex__sol_pct [%]	14.33	n/a
launch__block_size	128.00	128.00
launch__function_pcs	47,611,587,968.00	12,273,728.00
launch__grid_size	4,132.00	3,369.00
launch__occupancy_limit_blocks [block]	32.00	32.00
launch__occupancy_limit_registers [register]	21.00	21.00
launch__occupancy_limit_shared_mem [bytes]	384.00	384.00
launch__occupancy_limit_warps [warps]	16.00	16.00
launch__occupancy_per_block_size	3,638.00	3,638.00
launch__occupancy_per_register_count	5,792.00	5,792.00
launch__occupancy_per_shared_mem_size	2,260.00	2,260.00
launch__registers_per_thread [register/thread]	17.00	17.00
launch__shared_mem_config_size [bytes]	49,152.00	49,152.00
launch__shared_mem_per_block_dynamic [bytes/block]	0.00	0.00
launch__shared_mem_per_block_static [bytes/block]	20.00	20.00
launch__thread_count [thread]	528,896.00	431,232.00
launch__waves_per_multiprocessor	3.23	42.11
ltc__sol_pct [%]	6.93	7.18
memory_access_size_type [bytes]	2.00; 32.00; 32.00; 32…	2.00; 32.00; 32.00; 32…

Source	Live Registers	Sampling Data (All)	Sampling Data (No Issue)
@!PT SHFL.IDX PT, RZ, RZ, RZ, RZ;	0	223	0
MOV R1, c[0x0][0x28];	1	13	44
S2R R0, SR_CTAID.X;	2	143	75
S2R R2, SR_TID.X;	3	0	38
IMAD R0, R0, c[0x0][0x0], R2;	3	599	94
ISETP.GE.AND P0, PT, R0, c[0x0][0x170]	2	125	26
@P0 EXIT;	2	259	86
MOV R2, R0;	3	386	29
@!PT SHFL.IDX PT, RZ, RZ, RZ, RZ;	2	0	0
MOV R4, 0x4;	3	0	0
IMAD.WIDE R4, R2, R4, c[0x0][0x160];	4	0	0
LDG.E.SYS R3, [R4];	3	0	0
BSSY B0, 0xb00976780;	3	0	0
SHF.R.S32.HI R0, RZ, 0x1f, R2;	4	0	0

NVIDIA Nsight Compute

File Connection Debug Profile Rules Tools Window Help

Connect Disconnect Terminate Scheduling: Step Warp

Untitled 3 *

Page: Details Launch: 0 - 120293 - jit_queue Set Baseline Apply Rules Copy as Image

Current: 120293 - jit_queue (132224, 4, 1) **Time:** 17,408 **Cycles:** 22,452 **Regs:** 17 **GPU:** TITAN V **SM Frequency:** 1,289.77 **CC:** 7.0
Baseline: 120294 - jit_queue (107808, 4, 1) **Time:** 81,088 **Cycles:** 101,586 **Regs:** 17 **GPU:** Quadro P1000 **SM Frequency:** 1,252.79 **CC:** 6.1

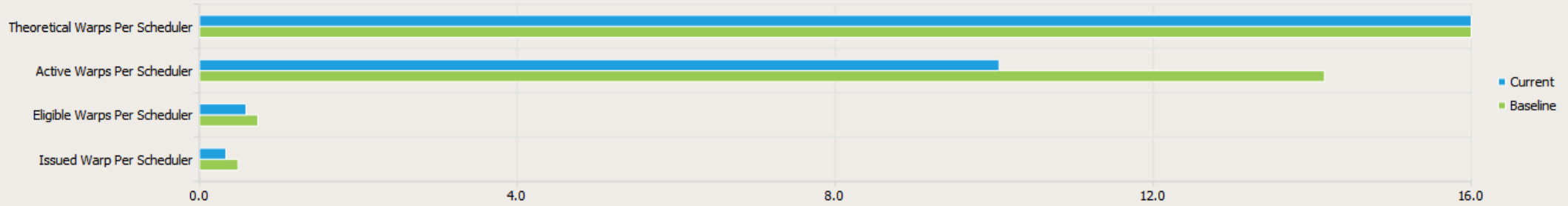
GPU Speed Of Light

SOL SM [%]	17.10	(-60.89%)	Duration [usec]	17.41	(-78.53%)
SOL TEX [%]	14.33	(+59.22%)	Elapsed Cycles [cycle]	22,452.33	(-77.90%)
SOL L2 [%]	6.93	(-3.48%)	SM Frequency [Ghz]	1.29	(+2.95%)
SOL FB [%]	29.20	(-20.12%)	Memory Frequency [Mhz]	919.12	(-81.85%)

Scheduler Statistics

Active Warps Per Scheduler [warps/cycle]	10.06	(-28.91%)	Instructions Per Active Issue Slot [inst/issue]	1.00	(-13.45%)
Eligible Warps Per Scheduler [warps/cycle]	0.59	(-20.19%)	No Eligible [%]	66.46	(+32.35%)
Issued Warp Per Scheduler [issue/cycle]	0.34	(-31.42%)	One or More Eligible [%]	33.53	(-31.42%)

Warps Per Scheduler



Recommendations

Issue Slot Utilization [Warning] Every scheduler is capable of issuing one instruction per cycle, but for this kernel each scheduler only issues an instruction every 3.0 cycles. This might leave hardware resources underutilized and may lead to less optimal performance. Out of the maximum of 16 warps per scheduler, this kernel allocates an average of 10.06 active warps per scheduler, but only an average of 0.59 warps were eligible per cycle. Eligible warps are the subset of active warps that are ready to issue their next instruction. Every cycle with no eligible warp results in no instruction being issued and the issue slot remains unused. To increase the number of eligible warps either increase the number of active warps or reduce the time the active warps are stalled.

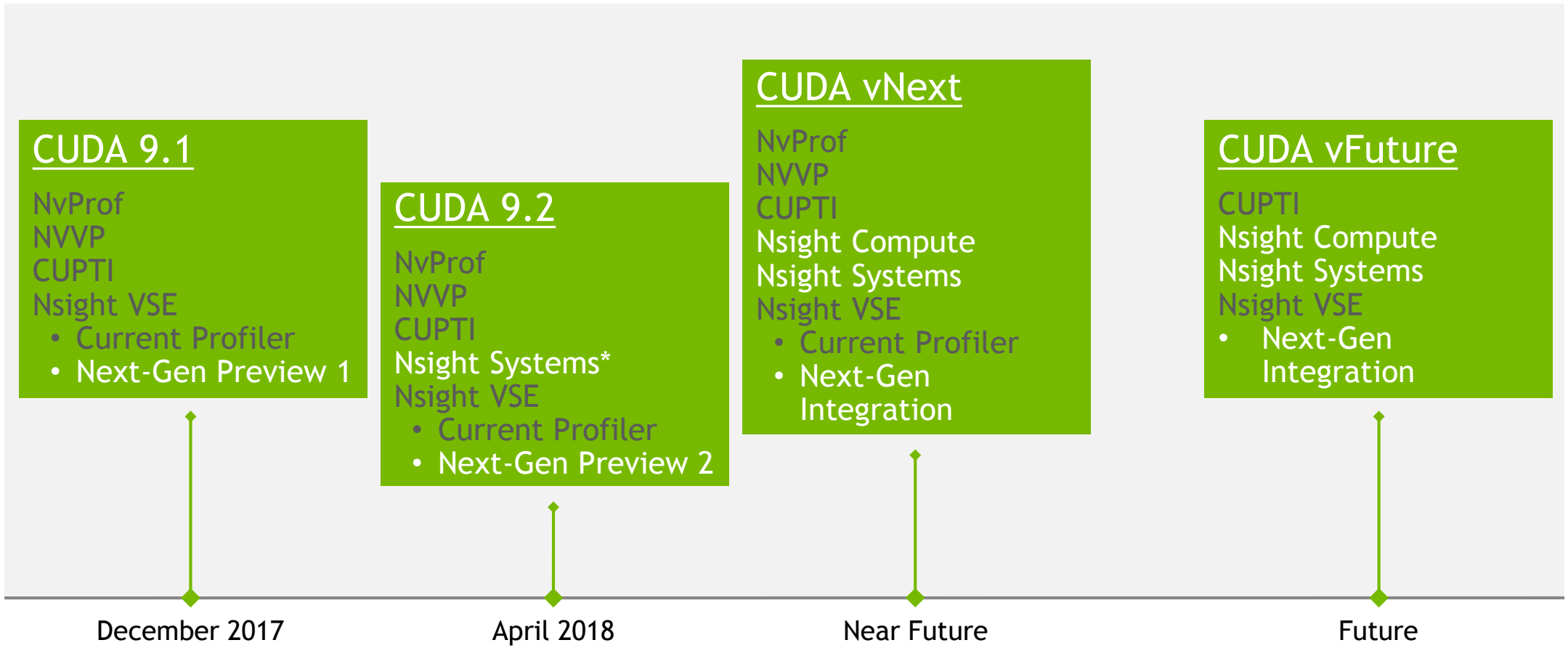
Warp State Statistics

Warp Cycles Per Executed Instruction [cycle/inst]	30.72	(+24.65%)	Threads Per Warp [thread/inst]	28.89	(+0.49%)
			Predicated Off Threads Per Warp [thread/inst]	25.74	(-1.85%)

GPU State: Unknown

DEMO: NVIDIA NSIGHT COMPUTE

ROADMAP



** Supported on Linux x86_64 only*

THANK YOU

Any Questions?

Upcoming Tools Talks/Events:

- Connect with the Experts: Jetson & DevTools
  2pm, LL Pod B
- Connect with the Experts : DevTools & CUDA
  3pm, LL Pod B
- Live Nsight Demos
  Tools Pod at Nvidia Booth
  on Showfloor

NVIDIA