Introduction into NVIDIA® Nsight™ Systems

Dr. Momme Allalen | LRZ | 16.07.2020

PRACE Training

LRZ as part of the Gauss Centre for Supercomputing (GCS) and IT4Innovations belong to the 14 **PRACE Training Centres** that started in 2012-2017-2020:

- Barcelona Supercomputing Center (Spain)
- CINECA Consorzio Interuniversitario (Italy)
- CSC – IT Center for Science Ltd (Finland)
- EPCC at the University of Edinburgh (UK)
- Gauss Centre for Supercomputing (Germany)
- Maison de la Simulation (France)
- GRNET – Greek Research and Technology Network (Greece)
- ICHEC – Irish Centre for High-End Computing (Ireland)
- IT4I – National Supercomputing Center VSB Technical University of Ostrava (Czech Republic)
- SURFsara (The Netherlands)
- TU Wien – VSC Research Center (Austria)
- University ANTWERPEN – VSC & CÉCI (Belgium)
- University of Ljubljana – HPC Center Slovenia (Slovenia)
- Swedish National Infrastructure for Computing (SNIC) (Sweden)

Mission: Serve as **European hubs and key drivers of advanced high-quality training** for researchers working in the computational sciences.

http://www.training.prace-ri.eu/

DEEP LEARNING INSTITUTE

DLI Mission: Help the world to solve the most challenging problems using AI and deep learning

We help developers, data scientists and engineers to get started in architecting, optimizing, and deploying neural networks to solve real-world problems in diverse industries such as autonomous vehicles, healthcare, robotics, media & entertainment and game development.

CUDA® PROFILING TOOLS

nvvp: NVIDIA visual profiler
nvprof: tool to understand and optimize the performance of your CUDA, OpenACC or OpenMP applications,

Application level opportunities

Overall application performance

Overlap CPU and GPU work, identify the bottlenecks (CPU or GPU)

Overall GPU utilization and efficiency

-Overlap compute and memory copies

-Utilize compute and copy engines effectively.

Nsight Systems
Nsight Compute

Kernel level opportunities

- Use memory bandwidth efficiently
- Use compute resources efficiently
- Hide instruction and memory latency

There are more features, example for Dependency Analysis

Command: nvprof --dependency-analysis --cpu-thread-tracing on ./executable_cuda

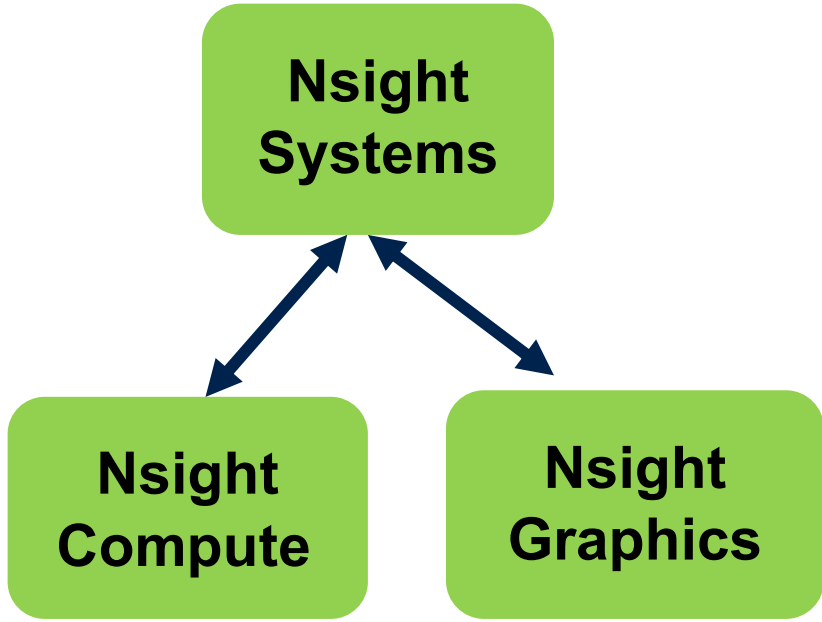
NSIGHT PRODUCT FAMILY

Standalone Performance Tools:

Ns- Systems – System-wide application algorithm tuning

Ns- Compute – Debug/&Profile specific CUDA kernels

Ns- Graphics – Analyze/&Optimize specific graphics workloads



IDE Plugins

Nsight Eclipse Edition/Visual Studio – editor, debugger, some perf analysis

Nvprof will be replaced with **nsys –profile=true**

Docs/product: **https://developer.nvidia.com/nsight-systems**

NSIGHT SYSTEMS

System-wide application algorithm tuning
- Multi-process tree support

Locate optimization opportunities
- Visualize millions of events on a very fast GUI timeline
- Or gaps of unused CPU and GPU time

Balance your workload across multiple CPUs and GPUs
- CPU algorithms, utilization, and thread state
- GPU streams, kernels, memory transfers, etc

Multi-platform: Linux & Windows, x86-64, Tegra, Power, MacOSX (host only)

GPUs: Volta, Turing

Docs/product: **https://developer.nvidia.com/nsight-systems**

NSIGHT COMPUTE

CUDA Kernel profiler

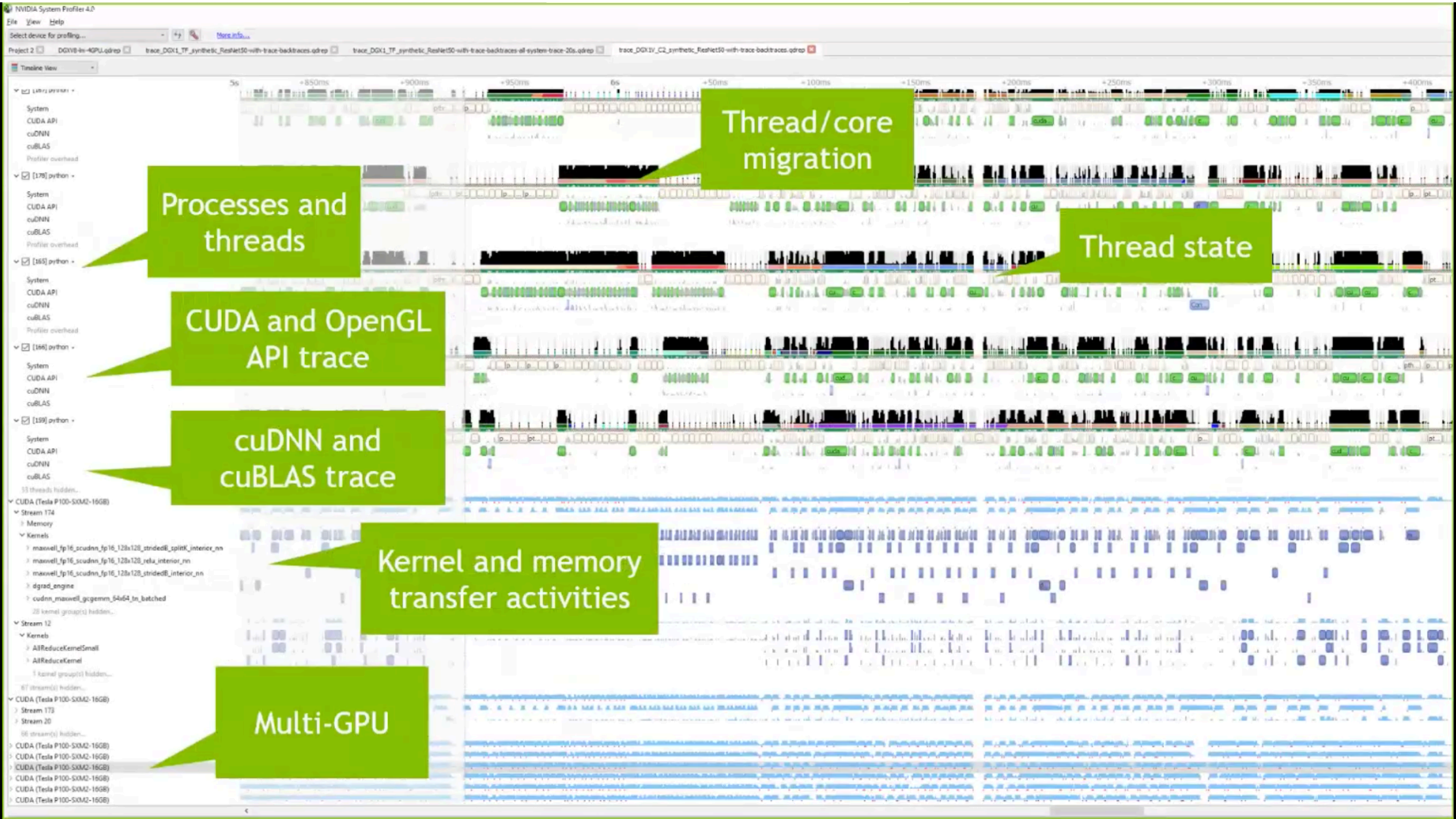
Targeted metric sections for various performance aspects (Debug/&Profile)

Very high freq GPU perf counter, customizable data collection and presentation (tables, charts ..,)

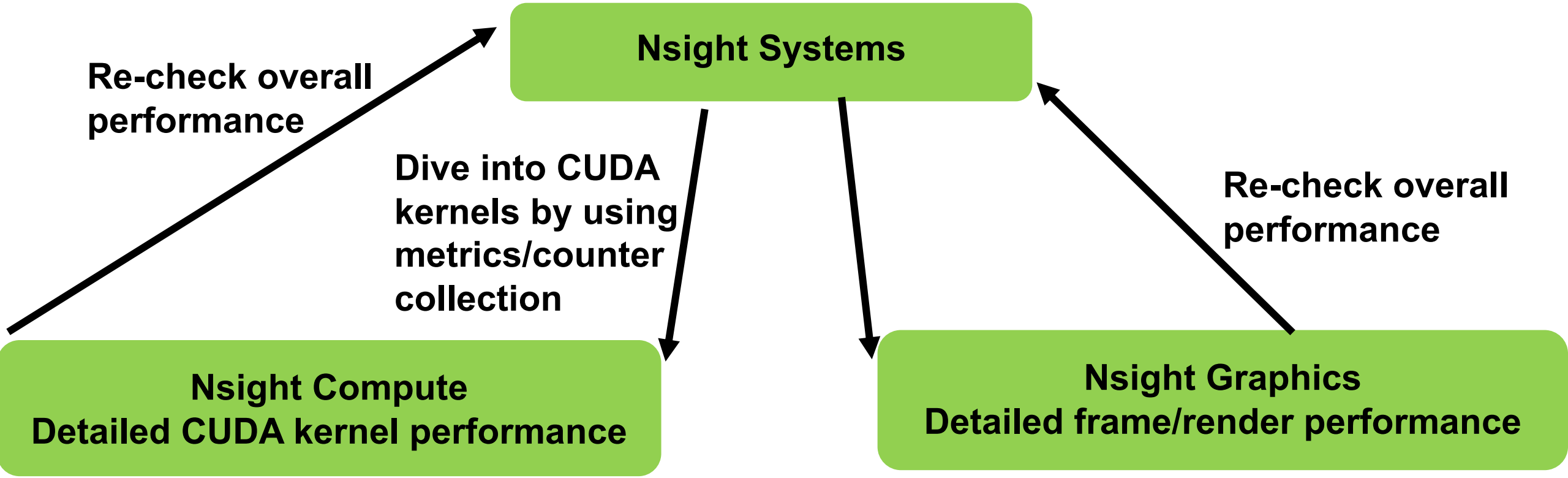
Python-based rules for guided analysis (or postprocessing)

GPUs: Volta, Turing, Amper

Docs/product: **https://developer.nvidia.com/nsight-systems**


Thread/core migration
Processes and threads
Thread state
CUDA and OpenGL API trace
cuDNN and cuBLAS trace
Kernel and memory transfer activities
Multi-GPU

NSIGHT PRODUCT FAMILY



Nsight Systems - Analyze application algorithm system-wide
Nsight Compute - Debug/optimize CUDA kernel
Nsight Graphics - Debug/optimize graphics workloads

Demo using Nsight

DEEP
LEARNING
INSTITUTE


THANK YOU

Instructor: Dr. Momme Allalen

www.nvidia.com/dli