

The NERSC logo is located in the top right corner. It consists of the letters "NERSC" in a bold, white, sans-serif font, set against a dark blue background with a bright light flare effect behind the text.

NERSC

The main title text is centered on the slide. It reads "Introduction to CUDA" in a large white font, followed by "Session-4" in a slightly smaller white font, and "Profiling tools" in a smaller white font below it. The background is a photograph of a modern building with a glass facade reflecting the sunset sky.

Introduction to CUDA

Session-4

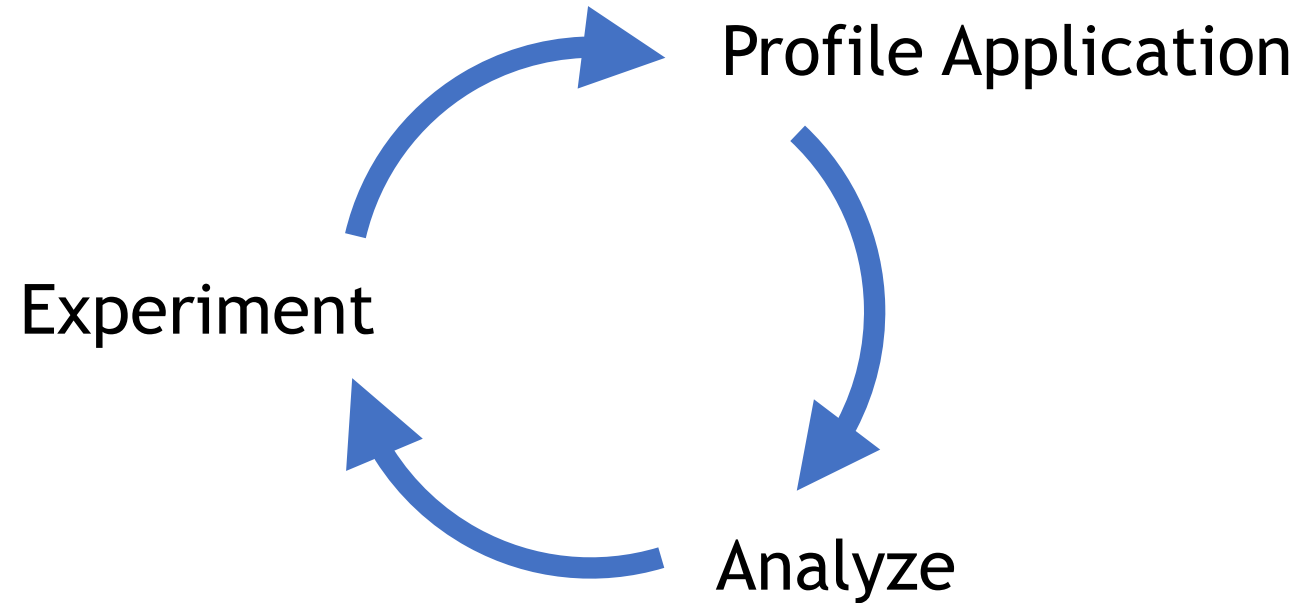
Profiling tools

The speaker's name is centered on the slide. It reads "Michael E. Rowan (NERSC)" in a white, sans-serif font. The background is the same photograph of the modern building at sunset.

Michael E. Rowan (NERSC)

Optimization workflow

Application optimization is an iterative process:



- **NSight Systems**
 - See application behavior in a cohesive timeline
- **NSight Compute**
 - Detailed analysis of individual CUDA kernels

Nsight Systems

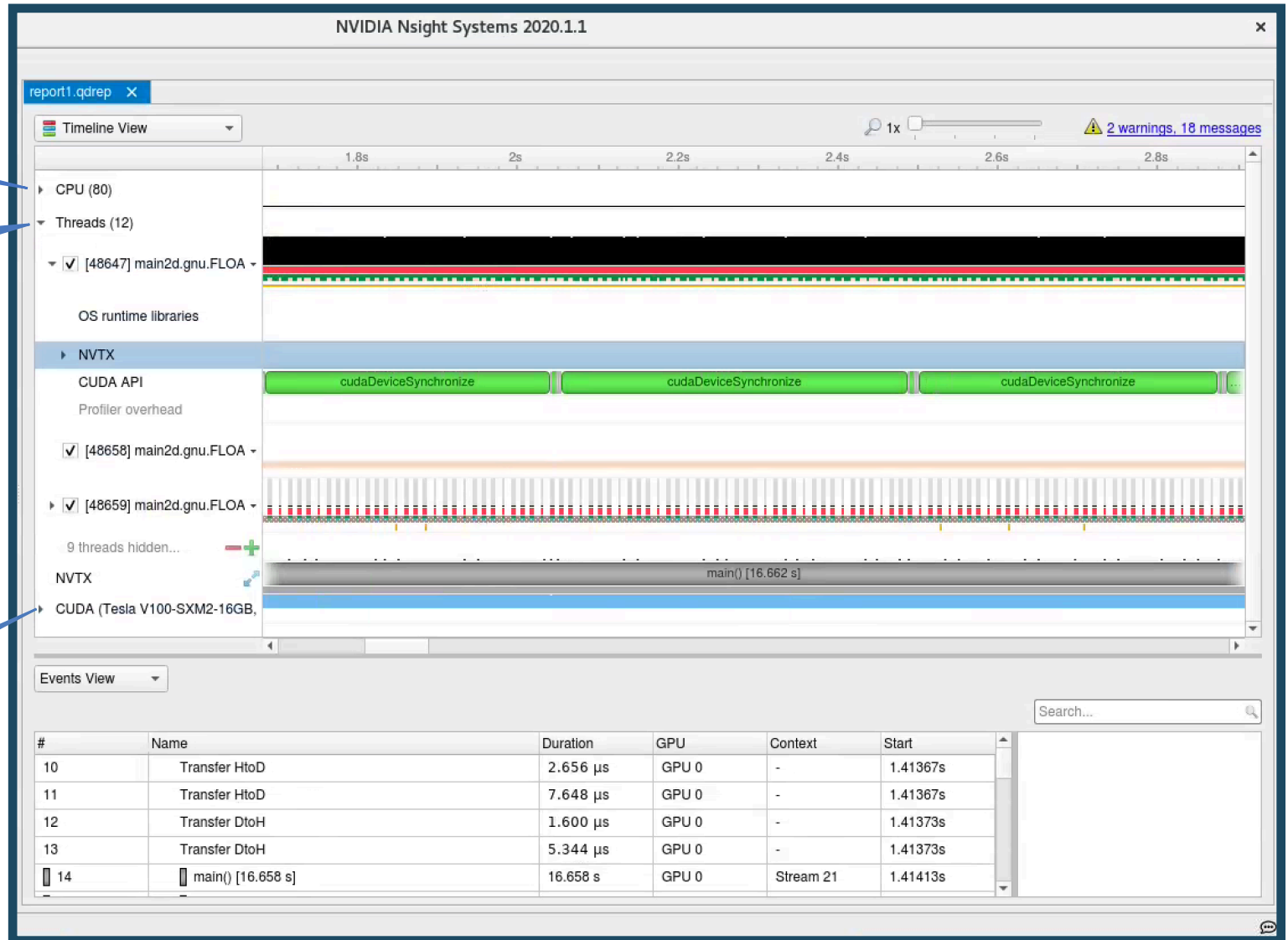
Cohesive picture of application behavior on a unified timeline

CPU workload

Thread activities:

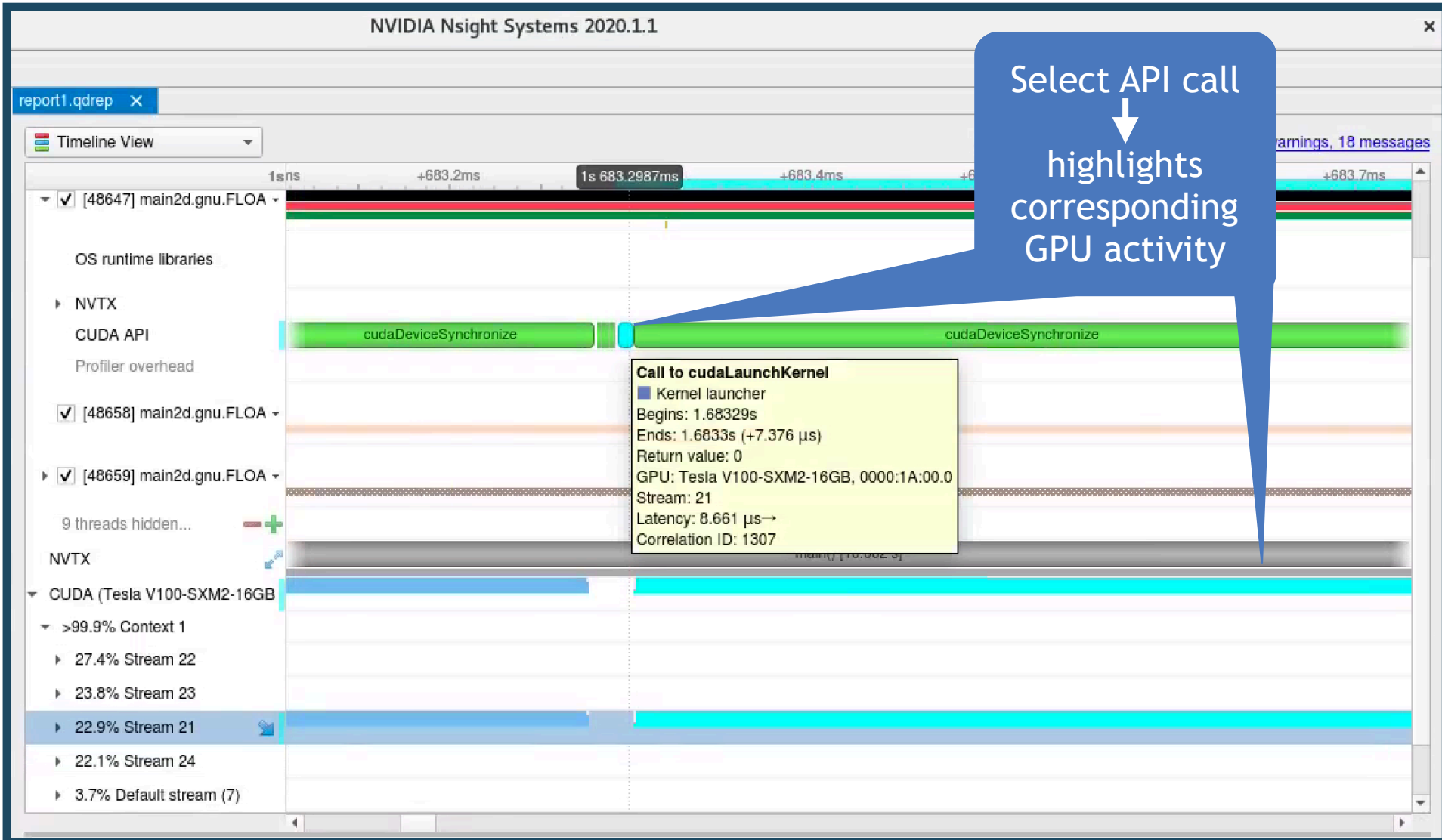
- OS runtime
- CUDA API
- NVTX (Nvidia Tools eXtension)
- Support for many others (cuBLAS, cuDNN, ...)

GPU workload



NSight Systems

Understand correlation between runtime API calls and GPU activity



NSight Compute

Detailed performance metrics for individual kernels

Baselines: track how your code changes affect kernel performance

Automatic detection of bottlenecks/potential issues

This kernel is neither compute nor memory bandwidth bound. What could be an issue?

GPU Speed Of Light

Metric	Value
Duration [msecond]	642.87
Elapsed Cycles [cycle]	700861185
SM Active Cycles [cycle]	700019975.96
SM Frequency [cycle/nsecond]	1.09
Memory Frequency [cycle/usecc]	728.84

GPU Utilization

SM [%]: ~5%
Memory [%]: ~10%

SOL SM Breakdown

Metric	Value
Mio Pq Read Cycles Active [%]	2.05
Issue Active [%]	0.30

SOL Memory Breakdown

Metric	Value
Xbar2Its Cycles Active	2.80
T Sectors [%]	2.75

Output Messages

ID	Origin	Source	Message
1	Host	NVIDIA Nsight Compute	SASS analysis failed. Some information might not be available (warning : Section '.text._Z11wp_ast_evalP7wp_no

(demonstration)

Summary

- Get an idea of overall application behavior with **NSight Systems**
 - Identify kernels for further analysis
 - Can use NVTX to correlate program logic to CPU and GPU activities
- Use **NSight Compute** for in-depth analysis of individual kernels
 - Track the effects of code changes using Baselines
 - Use the automated suggestions to identify bottlenecks
- Profile, Analyze, Experiment/test → repeat

Additional resources

- Official documentation (NVidia):
 - <https://docs.nvidia.com/nsight-systems/>
 - <https://docs.nvidia.com/nsight-compute/NsightCompute/index.html>
- Blog posts (NVidia):
 - <https://devblogs.nvidia.com/nsight-systems-exposes-gpu-optimization/>
 - <https://devblogs.nvidia.com/transitioning-nsight-systems-nvidia-visual-profiler-nvprof/>
- GTC 2018 NSight Systems talk:
 - <http://on-demand.gputechconf.com/gtc/2018/video/S8718/>
- Blue Waters tutorials:
 - NSight Systems
 - <https://www.youtube.com/watch?v=WA8C48FJi3c>
 - <https://bluwaters.ncsa.illinois.edu/liferay-content/document-library/content/NVIDIA%20Nsight%20Systems%20Overview%20by%20Sneha%20Kottapalli.pdf>
 - NSight Compute
 - <https://www.youtube.com/watch?v=nYSdsJE2zMs>