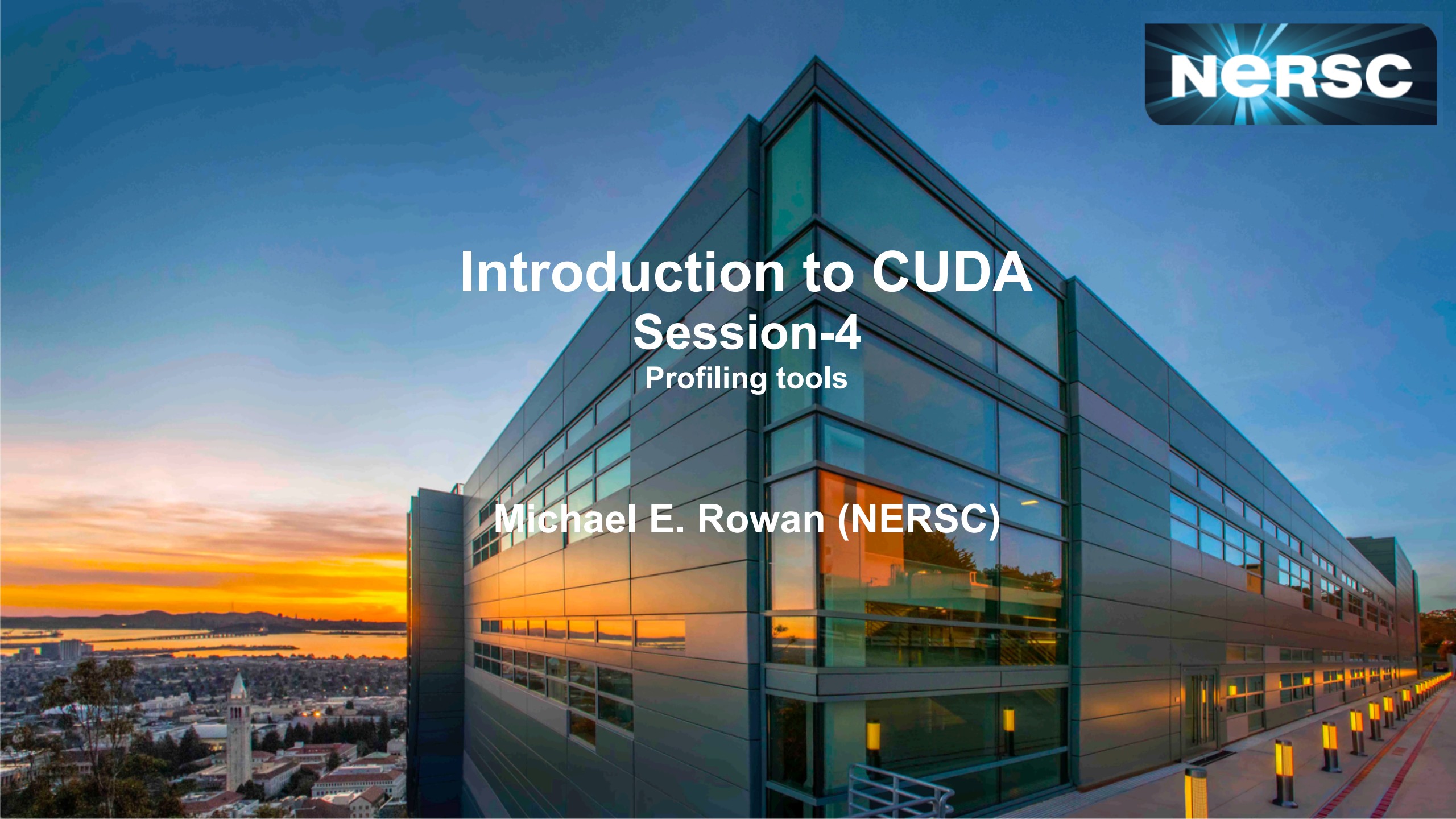
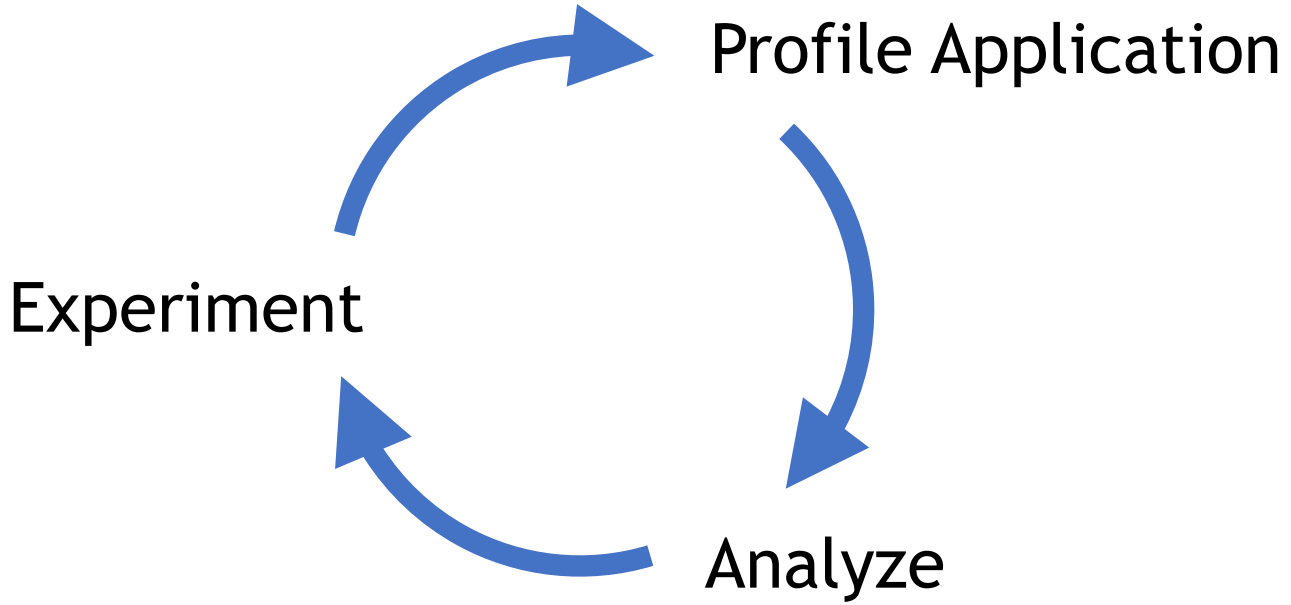


Introduction to CUDA
Session-4
Profiling tools

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

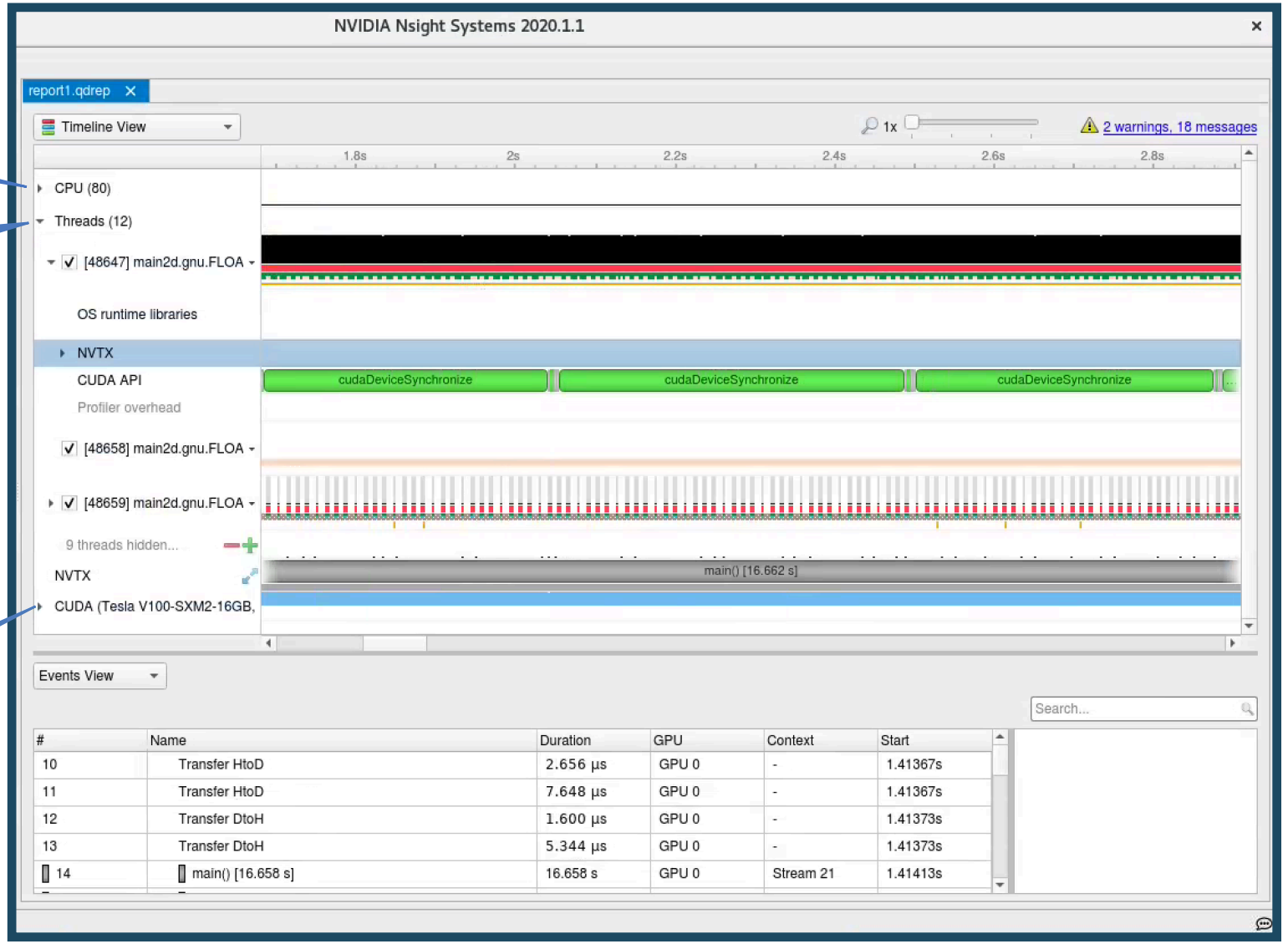
NSight Systems

Cohesive picture of application behavior on a unified timeline

CPU workload

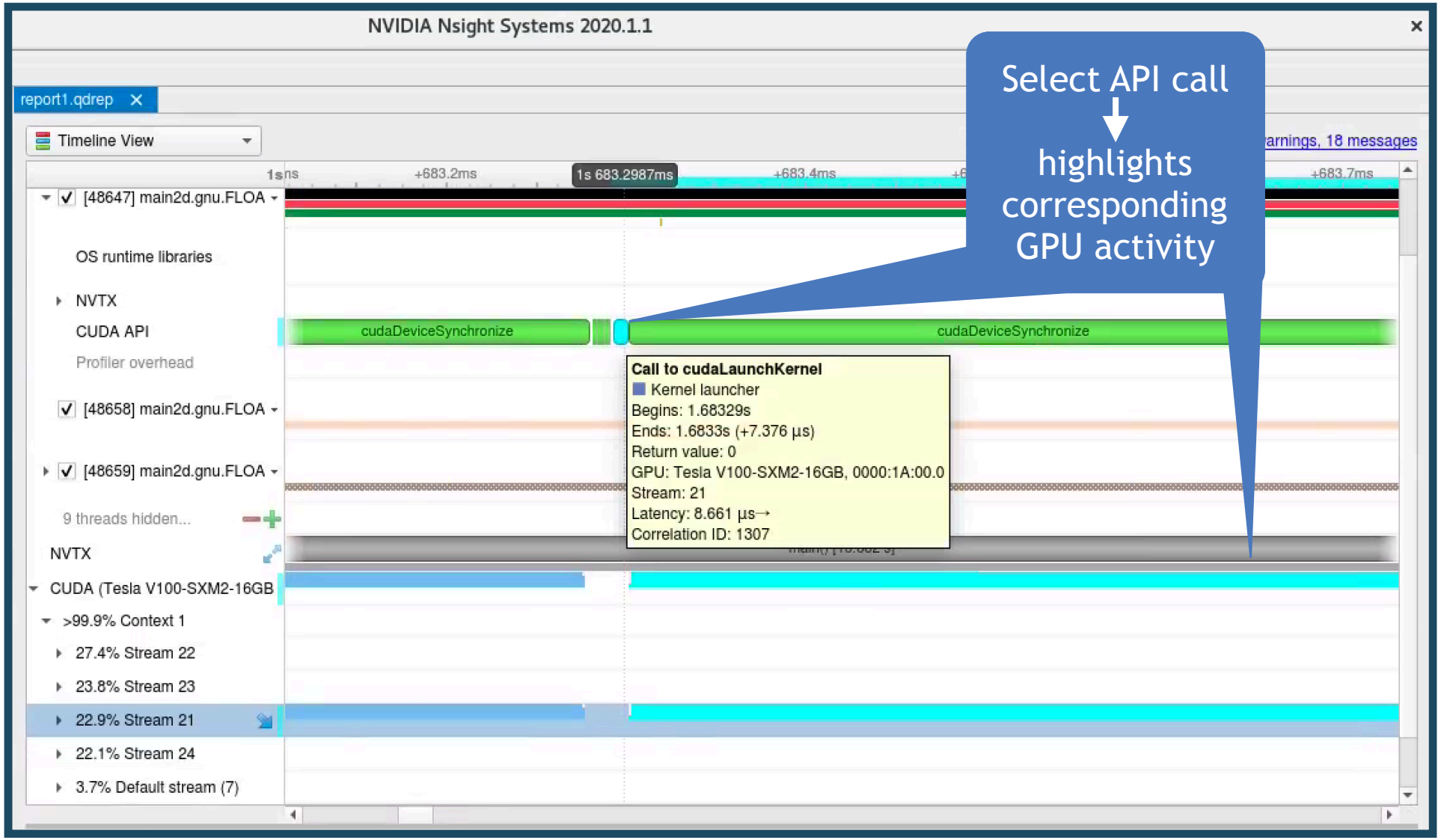
Thread activities:
- OS runtime
- CUDA API
- NVTX (NVidia Tools eXtension)
- Support for many others (cuBLAS, cuDNN, …)

GPU workload



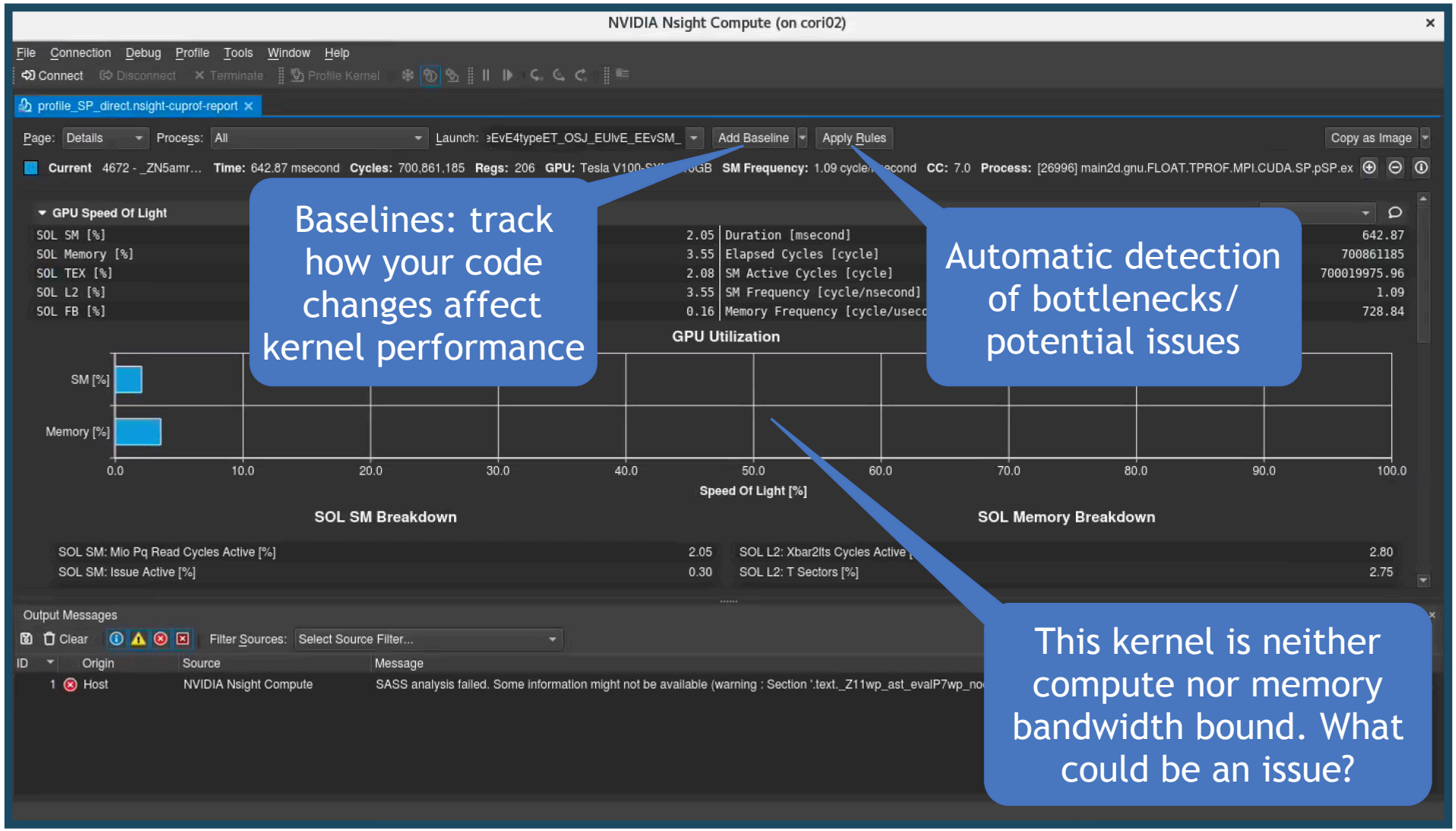
NSight Systems

Understand correlation between runtime API calls and GPU activity



NSight Compute

Detailed performance metrics for individual kernels



NSight Systems, NSight Compute demonstration

(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://bluewaters.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