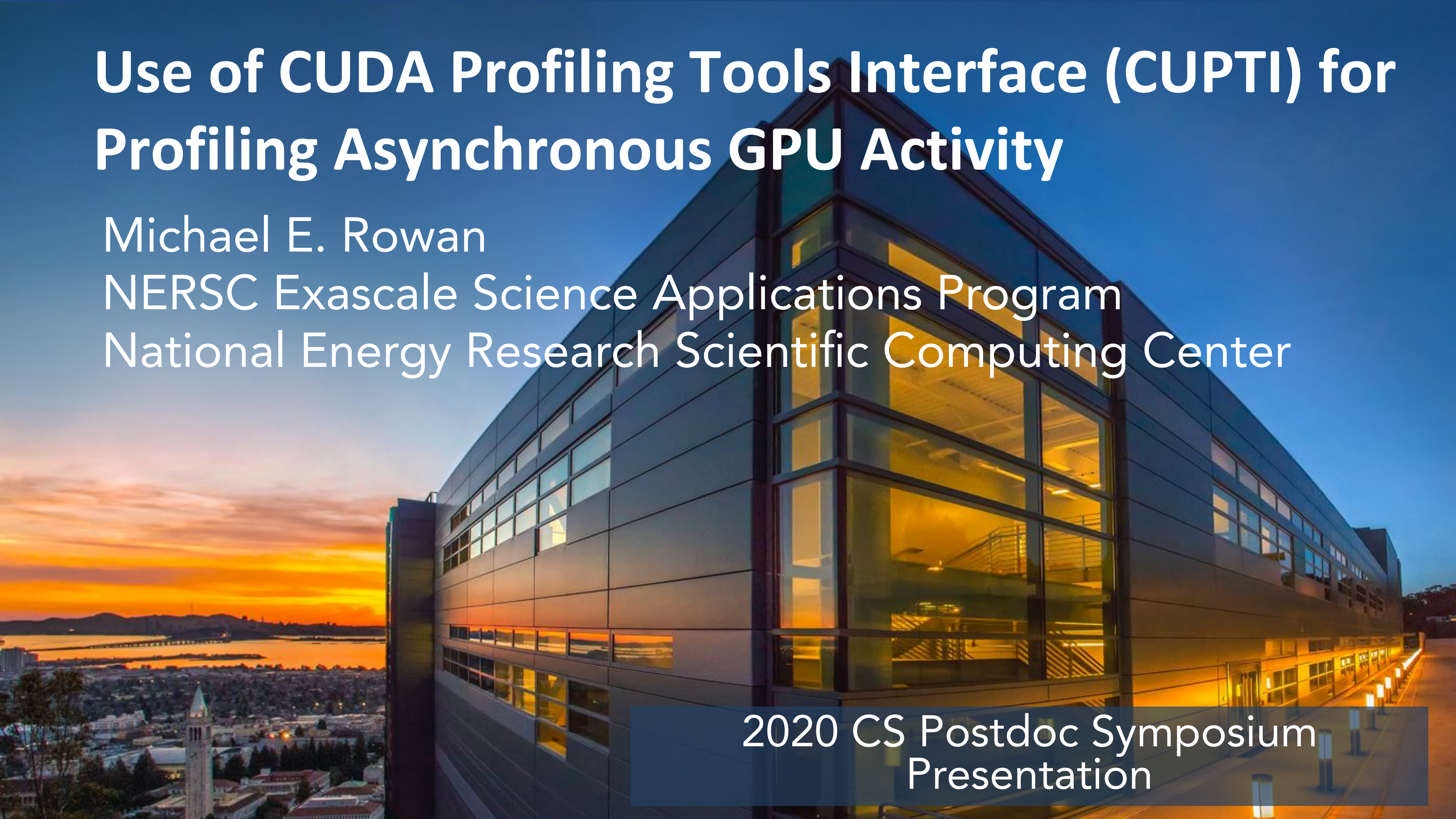
Use of CUDA Profiling Tools Interface (CUPTI) for Profiling Asynchronous GPU Activity

Michael E. Rowan
NERSC Exascale Science Applications Program
National Energy Research Scientific Computing Center

2020 CS Postdoc Symposium Presentation



Real-time measurement of kernel execution time is needed for correct load balancing in GPU-accelerated codes

- On-the-fly measurement not possible with standard profiling tools (NVProf, Nsight)
- Developed a method (CUPTI Callback timing) for real-time measurement of kernel time
- Impact:
  - Provides accurate kernel timing **on-the-fly**
  - Enables correct **load balancing** in WarpX