

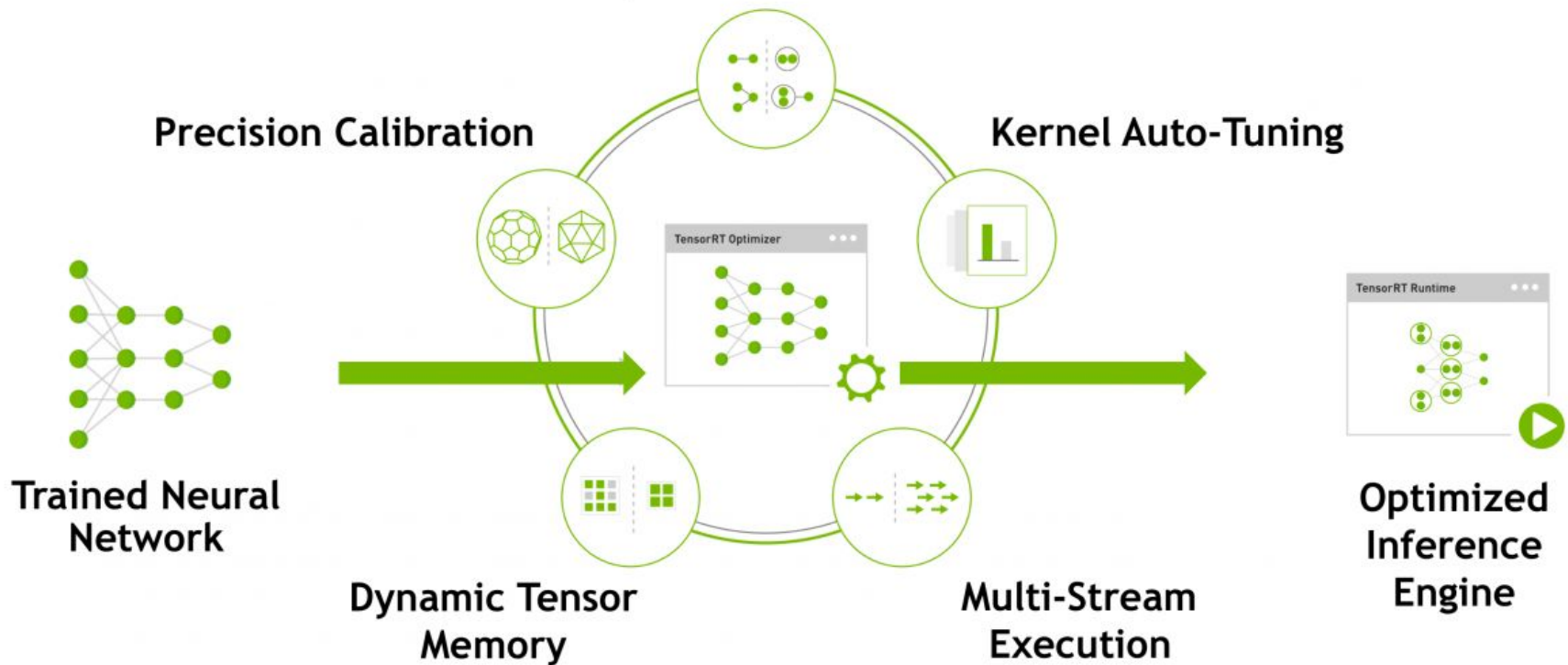
Fast inference in CMSSW with NVIDIA TensorRT

A. Di Pilato, A. Di Florio

Goal

- Integrate fast inference in CMSSW on GPU with **NVIDIA TensorRT**:
<https://developer.nvidia.com/tensorrt>
- NVIDIA TensorRT is a platform for high-performance deep learning inference. It includes a deep learning inference optimizer and runtime that delivers low latency and high-throughput for deep learning inference applications.
- Test the performance on the model developed for Particle ID and Energy Regression in HGAL within TICAL framework.

Layer & Tensor Fusion



Day 1 scrum

- Installed the latest release of NVIDIA TensorRT (TRT 6.0.1.5) within CMSSW environment
- Verification still ongoing
 - Apparently the tool for the conversion of .pb model into .uff model is not working
 - Testing is being made on a code that worked with TRT 4.0 outside CMSSW for doublets classification in the Tracker
- A simple code will be written to work with the new problem (ParticleID and EnergyRegression in HGAL)
- Need to adapt **PatternRecognitionByCA** inside TIGL framework to work with TensorRT instead of TensorFlow