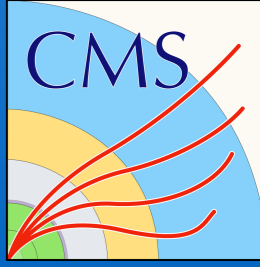TensorFlow 2.0 with GPU Support
in CMSSW

Marcel Rieger

7th Patatrack Hackathon

30.9.2019

- **Project objective**
  - Provide **all** the tools to perform fast inference of TF (2.0) models on GPUs in CMSSW

- **Current status in CMSSW**
  - CPU-only
  - TF 1.6 (!)
    - ▷ TF 2.0 way faster, even compared CPU vs. CPU
    - ▷ New models built with new operations, likely to be not supported at all in 1.6

- **Tasks to accomplish**
  1. Build TF 2.0 and all requirements in cmsdist
     - ▷ Done: updated Bazel, Protobuf, CUDA, added CUPTI & cuDNN,
     - ▷ ToDo: fix CUBLAS, **TF 2.0**
  2. Adapt customizations in CMSSW (mostly interference with different threading approach)
  3. Add a "resource scheduler", assigning resources (CPU/GPU*/GPU RAM) to CMSSW modules
  4. Handle GPU memory growth on demand and simultaneously prevent too large batch sizes