Proposal Guidelines:
Database and Python Analysis Package for AWAKE

Google Summer of Code 2019

Overview:

The GSoC Project for AWAKE addresses two topics:

- Creating a database and search functionality for the AWAKE dataset.
- Creating tools for analyzing AWAKE data.

The search functions and analysis tools must be written in Python and executable from a Jupyter notebook.

Database:

The AWAKE dataset consists of over 600,000 HDF5 files totaling 13 TB of data. The files are organized as a directory tree consisting of groups and datasets. The files do not all contain the same groups and datasets. A sample of 10 files is located here.

The first task for the project is to create a searchable database for the AWAKE data. The database should provide two types of search functionality:

1. String searches to identify available datasets. For example, searching for 'BCTF' would return '/AwakeEventData/TT41.BCTF.412340' and all the datasets therein.
2. Boolean searches to identify subsets of the data matching some conditions. For example, return all events matching the condition: /AwakeEventData/TT41.BCTF.412340/Acquisition/totalCurrentPreferred > 1

The database can be SQL, NoSQL, or simply a large table containing the relevant data and metadata. The only requirement is that it is based on open source libraries and that a Python API for search is created or already provided.

The database only needs to be created once. No new data will be added after the database is created. The data is located on CERN servers and accessible through the CERN SWAN service. In general, only a few people (less than 12) will be requesting data at a given time.

Analysis Tools:

Once the data is retrieved, it will be analyzed in the Jupyter notebook. Data analysis proceeds a long two lines:

1. Extracting data from images.
2. Correlating and plotting data across events.

Image analysis is primarily based on filtering images and fitting shapes to the image. For example, you could use SciPy curve_fit to compare data with Gaussian curves. Your goal is to create a highly visual Jupyter notebook which will serve as an example for future users. The exact details of what will be visualized do not need to be addressed in the project proposal.

Additionally, you are encouraged to develop functions that search for correlations in the data and make predictions about success/failure in the experiment. For your proposal, you do not need to address in detail how you will approach this topic. Simply suggest some algorithms that are used for finding correlations and point to some examples.

Timeline:

The coding period is from late May to late August. Expect to spend the first 3-4 weeks creating the database and the final 8-9 weeks working on analysis tools, creating example notebooks, and creating functions that search for correlations in the data. In addition, you will also spend 10%-20% of your time learning about the AWAKE experiment and the physics of plasma wakefield acceleration. No prior knowledge is required.

Pre-proposal:

Please send your pre-proposals in PDF format with file name
"<first name>_<last name>_AWAKE_GSoC.pdf" to:

spencer.j.gessner@cern.ch

I will not provide detailed feedback on individual proposals, but I will do my best to answer specific questions as they arise.