MAGIS Simulations Integration Discussion Notes

Dylan Temples -- September 14, 2022

Action Items.

- [JM] Organize a meeting for NU/SLAC/Oxford to interface regarding phase extraction techniques, with a goal of comparing and benchmarking the algorithms.
- [DT] Create and distribute a survey about thoughts and opinions on simulation/analysis language choice.
- [DT] Sketch out fast and slow simulation/analysis chain (10,000 ft view).

Kickoff Questions.

- What is the most useful way to connect these simulation tools?
    - Don't need to be re-done in same language as long as standardized files can be passed between them.
- How can others best interface with the current simulations?
- Can "physics" sources (DM, GW) be added to Julia simulation?
    - Separately generated signals can easily be added to Julia simulation
- Should we break out "analysis" methods (phase extraction, etc) into separate tools?

Language Choice Considerations.

- Options discussed: Julia, Python, ROOT.
- A primary consideration for choice of simulation is execution speed and CPU usage. Python is not compiled which means processing time can be a big issue for heavy duty simulations. This is something Julia handles well.
- While it's not necessary that each "leg" of the simulation pipeline must be written in the same language, if we use multiple languages that means core software has to do more steering.
- Data analysis and simulation code do not need to be written in the same language
- At this point, most algorithms have been implemented in every language, which makes it hard to weigh pros/cons on functionality alone. It comes down to execution speed and resource usage.
- Fermilab experiments mostly use ROOT. Some neutrino experiments (e.g. NOvA) use ART which is an in-house software with links to ROOT and is supported by FNAL computing division.
- Fermilab supports ROOT, so if tools don't exist in ROOT we could potentially interface with the FNAL support teams to have those features implemented.
    - Also makes it easier for new people to contribute to MAGIS sim/ana efforts if there's not a big leap in language.
    - ROOT is focused on histograms, not necessarily time-series data. Time series statistical methods may not be robustly implemented in ROOT.

- Perhaps the best way to make the choice of language for MAGIS is to determine the most useful way of structuring MAGIS data, then determine the best language to support that (rather than going the other way).
- Simulation chain should likely be a single language, while the analysis tools can be different tools and languages that we can benchmark and test against one another.

Simulation vs Analysis Pipelines.

- In any "pipeline" where you're worried about execution time and/or CPU resources, one needs an official end-to-end simulator with all ingredients wrapped up in the same language.
    - However, analysis can be much more flexible "post-production" but generating large datasets needs to be standardized (platform(s)/language)
    - How flexible can we be in analysis tools? For instance consider implementation of systematic uncertainties of a particular technique in analysis. We would want everyone to apply them in a uniform way.
- Quality control and certification of analysis tools will be important.
- If everyone is working on the same codebase and using the same tools, bugs and quirks are found faster.
- How do we want to do signal extraction? Building a fully-differentiable simulation pipeline is useful to understand physics signals and extract them from data.
- ROOT is particularly useful for statistical inference (setting limits), but most if not all of this has been ported to Python, for example.
- We will likely need to use irregular/non-uniform FFTs in our analysis of real data due to realistic operation of the detector. This is implemented in Python, but not Mathematica for example. This may dictate part of the choice in analysis language.
- If we were to use ROOT, that makes it harder to decouple the analysis and simulation tools since ROOT data formats are very specific (e.g. nTuples). Also, ROOT is focused on "events" but that's not how MAGIS operates fundamentally.
- What is the separation between simulation and analysis?
    - One interpretation is things used on raw data are "analysis tools" and their output is databased.
- We may want two simulation pipelines: fast sim chain (aggregated time-domain RQs) vs slow sim chain (images)
    - Sim integration has hazy endpoint for what's the output of sims: images or RQs (e.g. phase)
    - Work on understanding systematics etc in images, analysis tools will require raw images as inputs. Once we have large datasets for runs we'll want aggregated RQs to do time-series -- simulations should do both, hence the need for two chains.
    - there may be sims that don't do the image generation
- We should develop and maintain a recommended standard phase extraction tool for collaboration

Data and File Format.

- Will need comprehensive metadata attached to the simulated data.
- Want to select a single format that keeps md+data together (json, maybe h5) and is usable in multiple languages.
    - Data size not a dominant factor for metadata/header - doesn't necessitate binary over formatted text files (e.g. JSON)
    - Physics signal simulations generate data written to a file w/ standard format, which can then be read by anyone's program for simulating detector effects.