SBS Software and Simulation Outline and Requirements

Seamus Riordan
Argonne National Laboratory
sriordan@anl.gov

May 30, 2017

This document specifies the goals and requirements for the Super Bigbite simulation and analysis software frameworks. It summarizes present progress within the software development efforts and identifies key issues that have yet to be addressed within the collaboration.

1 Goals and Requirements

The Super Bigbite suite of experiments push to unprecedented luminosities and rates for Jefferson Lab and present a unique set of experimental challenges. One crucial aspect of these challenges is to accurately simulate and analyze these experiments. For the simulation, there are a number of critical aspects to be evaluated. These include

- Physics and background rates
- Experimental requirements and configuration optimization
- Realistic detector responses in production
- Shielding requirements and optimization
- Data sizes and DAQ requirements
- Pseudo-data for reconstruction algorithms such as tracking and calorimeter clustering
- Calibration Scripts
- Event Displays

In conjunction with these efforts, the reconstruction and analysis of the data, both online and offline, requires further development. The main requirements include

- Online and offline analysis available before the first experiment

- Robust reconstruction algorithms such as tracking and calorimeter clustering
- Particle identification
- Scaling of reconstruction algorithms as a function background rates
- Coherent event reconstruction between detectors in a single arm and multiple arms

While many of these goals have been satisfied or are ongoing, there remains a considerable amount of work to be done. A single simulation has been developed which describes the three main form factor experiments as well as the SIDIS experiment and includes many important details for physics, geometries, and detector responses. The analysis framework to be used is based upon the Hall A analyzer, which has provided a strong foundation for many unique experiments within Hall A. However, with all of this advanced progress, there remains a significant number of issues which must be addressed before running. These broadly include the completion of writing decoders for raw data, production of suitable pseudo-data, interfacing that data with the analysis framework, further development of specific algorithms for reconstruction within the anticipated detector configurations. Each of these has varying levels of complexity and interplay.

2 Organization and Development

To meet the aforementioned requirements, there are a significant number of development considerations to be evaluated. Foremost, a well organized project produced as simply and modularly as possible is important for a physics project of this scope. It is anticipated that use and stewardship for the simulation and analysis software will involve many different parties over the lifetime with varying experience levels, so it is critical that design is done with in a way which allows rapid familiarity. In addition, the capabilities of the software should be somewhat flexible to allow for growth with future needs. Unfortunately capability and complexity necessarily correlate, so to minimize this impact, careful design decisions must be made.

The areas of primary topics which require development are

- Simulation implementation
- Analysis framework
- Algorithms
- IO formats and standards
- Databasing

While an advanced simulation, `g4sbs`, and analysis framework, the Hall A analyzer, already exist, one major challenge is producing an interface between the two. We recognize this as an important requirement for the Super Bigbite experiments as they reach to unprecedented luminosity and rates. It is critical to the success of the experiments that reconstruction and analysis algorithms are tested well before the experiments are run.

One possible schematic of a simulation/analysis framework is shown in Fig. 1. While specific implementation details are not specified, it incorporates many of the important ideas to be considered. These are discussed in the following sections.

2.1 Simulation Implementation

The present Super Bigbite simulation, `g4sbs`[1], has been in development for several years at the time of the writing. It is based in Geant4 and compiled against ROOT to allow for ROOT tree output. In includes implementations which address many crucial aspects such as

- Elastic, DIS, parameterized resonance, single pion production, and Pythia event generators
- Deuterium and ^{3}He Fermi smearing for events based on realistic momentum distributions
- Capability of low energy backgrounds through Geant4 physics processes
- Detailed geometries for all kinematics and materials including detector, magnetic, and beamline elements
- Inclusion of detailed magnetic field maps simulated in Tosca where available
- Detailed detector responses including ionization in GEMs and optical photons in calorimeters and the gas Cherenkov detectors
- Output into ROOT trees
- Automatic build tests for new changes via Travis CI

Separate from this simulation is also the development by the INFN group of GEM responses down to the individual channel based on a simple diffusion model and tuned to real data along with the APV25 digitization. While this post-processing has been used to evaluate hit occupancies and tracking capabilities, it has not bee interfaced with the present `g4sbs` simulation.

A few minor aspects of event generation remain to be settled. The minimum bias Pythia generator requires mixing with low energy events. No event generators presently include pre-vertex external brehmstrahlung or multiple scattering. The code itself could be organized more efficiently. Aspects such as spin transport and spin-dependent processes are under investigation.

[1] `https://github.com/JeffersonLab/g4sbs`

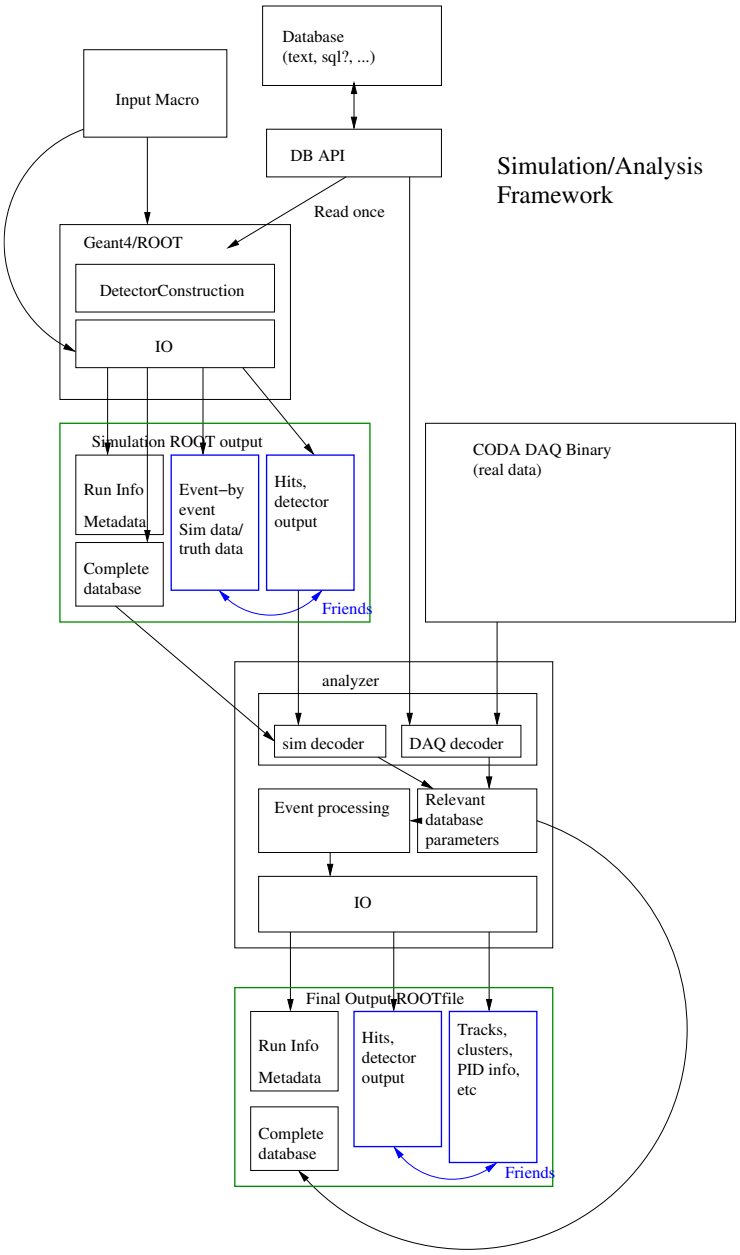

Figure 1: A possible schematic framework for a coherent simulation and analysis framework.

In principle, the output for the simulation is flexible so that it could be cast into a form for digitization post processing, we recognize the need to include in the output (and ideally within the same output file) parallel information about the running and geometry conditions which are necessary for the analysis. This requires a more advanced input and databasing system which has not been included. Presently, input is done through a text file based input system within the Geant4 macro system. This will need to be expanded to include more details such as individual GEM positions, wire orientations, and detector channel maps. A system within the Hall A analyzer exists and is discussed in following sections.

While many aspects of the detector responses are already handled within the existing simulation, details about the digitization and a standardized output into individual channel signals (resembling pseudo-data) are not. A organized system to translate the present Geant4 hit information into a pseudo-digitized format suitable for analysis within the present framework must be developed. Such a basis for at least the GEMs presently exists from the INFN GEM digitization package and could provide a strong starting point for this.

Finally, the preparation of pseudo-data requires suitable methods for "event mixing". As the generators consist of a mix of totally unbiased, minimally biased, and cross section-weighting, producing a "complete" event with appropriate correlated multiplicities and response distributions in all detectors, different generators will need to be combined. This is a non-trivial problem to do efficiently. A post-processing suite which is able to combine different output generation coherently should be developed to take general output from the existing generators and combine them together using only the beam current as an input. A method to generate all these runs simultaneously in an efficient way using existing computing facilities would be desirable. It is also desirable to be able to flexibly modify the data for special conditions such as channel failure or miscalibration.

2.2 Analysis Framework

The existing Hall A analysis framework [2] has been chosen as the basis for the analysis of these experiments. Similar experiments with unique detector configurations have already shown this framework to be effective, including experiments using the Bigbite spectrometer. There are a few aspects which need to be completed within the existing analyzer to complete it for the SBS experiments. Retrospective information needs to be included specific to our work in the output.

A repository has been formed[3] to accomodate a library, a central repository for scripts and examples, and provide tracking for features and bugs. The library is loaded into the analysis framework to provide all the necessary classes.

Most importantly, with the new sets of electronics, for example the GEM readouts, decoding of the raw data has not yet been included, though decoders

[2] `https://github.com/JeffersonLab/analyzer`
[3] `https://github.com/JeffersonLab/SBS-Offline`

presently exist. As has been done in the past, a library specific to the SBS experiments must be developed which includes representations of all the detector arms. For coherency, the selection of the different experimental configurations should be done easily to reuse code and flexibly as there are small variations using much of the same equipment.

The present databases rely upon sets of text files organized by timestamps and a variable-number association with a basic hierarchy system. While this databasing system has capabilities to handle more complex structures with arrays of numbers, it may be insufficient for situations involving pseudo-data (which is not necessarily organized into time stamps) or with large numbers of changes. Possible solutions involving a generic databasing system (such as SQL) and writing a flexible API should be explored. It is noted that large installation experiments such as SoLID face similar issues.

A standard system to process pseudo-data from the simulation should be implemented. This requires not only a specialized decoder to process the "raw" data, but also parallel facilities to keep track of run conditions and "truth" information. Information on run conditions should be obtained from data contained within the simulation output to maintain consistency in analysis conditions.

2.3 Algorithms

Algorithm development is a critical aspect to the analysis. Within the analyzer, analysis algorithms are implemented in a modular way within the detector classes in the analysis framework or in a post-processing class which is able to access and combine data from multiple detectors and spectrometers. Presently, much of the development of analysis algorithms has been done regarding GEM tracking, which is the most challenging. There have been two main efforts in tracking. One using a "tree search" algorithm which has been used for other Hall A experiments and offers a robust and workable solution in such high luminosities [1]. The other by the INFN group uses a neural network approach [2], and would need to be ported into the existing framework.

There remains work to be done on almost all aspects for analysis algorithms. These include

- Optimal clustering in electromagnetic and hadronic calorimeters
- Combining hit positions in all detectors within the tracking algorithms
- PID within the GRINCH and SIDIS RICH
- PID using data coherently from multiple detectors
- Clustering within the GEMs at high rate
- Effective GEM zero suppression
- GEM amplitude association
- Complete kinematic event reconstruction between arms

- Optimal timing reconstruction in FADC-based detectors

2.4 Input and output formats and standards

Input and output standards in a few aspects are yet to be defined. Presently, the raw data from the data acquisition is in the lab-wide standard CODA format and will continue to be used for these experiments. In addition the analyzer uses a ROOT tree output for analyzed data, which should continue to be used. This leaves a few aspects which are not yet included to be considered.

From the simulation, modifications should be made to the output such that it is in a well documented and organized format in a ROOT tree which can either be directly used or read in conveniently in post-processing (e.g. for digitization). Complete data about the run conditions and auxiliary information required for post-processing and pseudo-analysis by the analyzer needs to also be included. The latter should be in a form such that it is accessed transparently as run conditions from the database by the analysis code. This specification needs to be produced.

The format for the raw pseudo-data to be decoded by the analyzer must be specified. It should also have the simulation run conditions as well as the "truth" data contained with it to minimize the possibility for inconsistencies between the simulation and analyzing configuration conditions.

2.5 Databasing

Discussed in the analysis framework, databasing is an important aspect for data analysis. Included with the previous discussion is also the need for external, general databasing of parameters. These include run and time based data such as target polarization values, EPICs data, high voltage, and other slow controls data. In the past this has frequently been stored within an SQL database. Considerations should be made ahead of running on how best to organize and incorporate this data into the analysis to avoid ad hoc approaches.

3 Data Sizes and Tape Storage

Raw data rates, which are driven by the reduced GEM occupancies and the raw data storage for tape is lised in Table 3. Additional tape will be required for the replayed data and is anticipated to be approximately the size of the raw data in the form of ROOT files. Estimates of the replay time per pass is given in Table 3. These estimates assume a per-event analysis time of 100 ms which we consider very conservative and 500 available farm cores. This is based on the estimated 500 ms event analysis time for the Hall B and D detectors. Sheer scale of our detectors and the fact that tracking is done in field free regions greatly reduce the computational requirements. The Jefferson Lab farm is expected to have 10000 cores available by 2019 and Hall A will have 1000 dedicated to them. 2-3 passes per analysis is anticipated.

	Prod. [days]	Data Rate [MB/s]	Data PB
GMn	25	130	0.3
GEn	50	130	0.6
GEp	45	500	1.9
SIDIS	64	50	0.3
Total	184		3.1

Table 1: Estimated analysis raw data rates and accumulated size for the form factor and SIDIS experiments.

	Events $[10^9]$	Analysis Time per pass [core weeks]
GMn	11	1800
GEn	22	3600
GEp	19	3200
SIDIS	28	4600

Table 2: Estimated analysis time for the form factor and SIDIS experiments assumming 100 ms reconstruction time.

4 Timeline and Milestones

Milestones have been developed to have the software tested and ready before the first experiment, likely GMn, would run in spring 2019 according to the Fall 2016 anticipated Hall A schedule. The timing of these has been agreed to by the collaboration to be modest and acheivable.

- Nov 2016 - Software Review
- Jan 2017 - Start Digitized Simulation Output
- Apr 2017 - Decoders for all DAQ modules written
- Jul 2017 - Each detector system in analyzer, experiment configurations, basic reconstruction algorithms
  - Can fully analyze raw data at this point
- Dec 2017 - Simulation Interfaced to analysis, Have detector event displays, calibration scripts
- Jan 2018 - Start simulated analysis for detector reconstruction
- Jun 2018 - Begin simulated experimental analysis for core form factor experiments

General Purpose Software	
analyzer Development	Hansen
Front End Decoders	Camsonne
Event Reassembly	JLab DAQ Group
SBS Specific	
Repository Maintenance	Riordan
MPD Decoding	SBU, JLab, UVA, INFN
GEM Tracking	INFN, JLab
HCal Analysis	Franklin
ECal Analysis	Puckett
Coord. Det	CNU (Monaghan, Brash)
GRINCH	Averett
BigBite Legacy	Riordan

Table 3: Table of subsystems and assigned responsible parties.

- Jan 2019 - Ready for beam for form factor, start simulated experimental analysis for SIDIS and TDIS
- Spring 2019 likely earliest start of neutron experiments
- Spring 2020 likely earliest start for GEp

5 Workforce Organization and Responsibilities

Workforce and responsibilities have been divided into three main groups. General purpose software which is common to experiments beyond SBS and is maintained by laboratory staff, SBS specific subsystems that are core to a detector rather than a specific experiment, and experiment-specific software. The last group includes the full analysis chain and event reconstruction. Responsibilities are listed in Tables 5 and 5. Each group has agreed to produce the required software in accordance with the requirements listed above including calibration and event displays. For the experimental groups, a contact person for each experiment has agreed to be the point of contact for the software and will possibly delegate internally.

References

[1] SBS Collaboration, Super Bigbite Conceptual Design Report `https://userweb.jlab.org/~mahbub/HallA/SBS/SBS-CDR_New.pdf`

[2] C. Fanelli, Private communication, manuscript in progress

Experiment Analysis Specific		
GMn	Quinn	Bigbite, HCal
GEn	Riordan	Bigbite, HCal, ^{3}He target
GEp	Cisbani	ECal, Coord. det, SBS w/ trackers
SIDIS	Puckett	Bigbite, SBS w/ trackers and RICH
TDIS	Dutta	SBS e^- w/ trackers and RICH, LAC, RTPC

Table 4: Table of all approved SBS experiments and assigned responsible parties for individual experiment analysis. TDIS is conditionally approved at this time.