PVDIS simulations requirements

R. Holmes*
(Dated: May 30, 2019)

I. INTRODUCTION

This note is a highly preliminary discussion of a number of issues regarding tasks and resources required to meet simulation needs for PVDIS.

II. DIS

To evaluate radiative corrections, kinematic corrections, tracking efficiency and accuracy, and so on, we will need a large data set of simulated DIS events, but exactly how large is not yet clear. GlueX plans to simulate 10 times their experimental data set[1] and I believe CLAS does the same. However, these experiments are very different from PVDIS.

In discussions with PVDIS collaboration members, a multiplier factor M of anything from 1 to 100 times the experimental data set has been suggested. Furthermore, the meaning of "experimental data set" is unclear. Approximately 20% of the data written to tape will be in the kinematic range of greatest physics interest, $x_{bj} > 0.55$ and $W > 2$ GeV. In truth the simulation requirements cannot be pinned down without detailed consideration of what the simulation data will be used for and how it will be used.

In the following, I will leave M unspecified and take "experimental data set" to refer to the subset of physics events written to tape which are in the kinematic range of interest.

The approved running time for PVDIS is 137 PAC days[2], half the collaboration's request[3] allocated as shown in Table I. From this we may estimate the total number of DIS events in the kinematic range of interest in the

TABLE I. PVDIS approved running time

LD2:		
60 days	11 GeV	50 μA production
30 days	6.6 GeV	50 μA production
9 days	4.4 GeV	50 μA radiative corrections
LH2:		
45 days	11 GeV	50 μA production
5 days	4.4 GeV	50 μA radiative corrections

experimental data set. The rate for EC+LGC coincidences is 230 kHz at 11 GeV with 50 μA on LD2.[4]. For this estimate we assume roughly 1/2 the LD2 rate on LH2, and roughly 1/2 the 11 GeV rate at 4.4 and 6.6 GeV. As stated above, kinematic cuts bring a factor of about 20%. Then we have

$$(0.2)(0.23\text{ MHz})(10^6)(3600)(24)\left(60+30\left(\frac{1}{2}\right)+9\left(\frac{1}{2}\right)+\left(\frac{1}{2}\right)\left(45+5\left(\frac{1}{2}\right)\right)\right)=4\times10^{11}\text{ triggers} \quad (1)$$

So we assume a need to simulate $4\times10^{11}M$ triggers. The number of simulated events needed to produce this number of triggers, accounting for efficiency and acceptance, may be estimated as about 5 times higher, or $2\times10^{12}M$ events.

(The above numbers are for the full PVDIS acceptance; a single sector would be a factor of 30 smaller. However, we probably will need ultimately to simulate all 30 sectors.)

With our current simulation software on a single core, the time required per DIS event is about 140 ms. This includes both the GEMC simulation stage and conversion of the evio output file to ROOT format; it does not include

* rsholmes@syr.edu

time required to generate the input DIS electrons, which should be relatively small, nor post-GEMC digitization, tracking, and analysis. Background simulation is considered separately. Then the DIS simulation processing time required is about $80M$ Mcore-hr.

For comparison, GlueX anticipates requiring 36 Mcore-hr per year starting in 2020 for their simulations.[1] This is a significant effort which has been tested with a run on 9 clusters on the Open Science Grid, achieving 1 Mcore-hr of simulations completed in 15 days.

As for storage, we have output files for 1 to 2 million DIS events which are about 40 GB per million events. However, these files include hits in virtual planes, and passthroughs (hits that deposit no energy) in real detectors. For production we likely would store files that include only energy depositing hits in real detectors, which are smaller by about a factor of 10. Additionally, for the long term we probably would keep not these output files but digitization output files. At this point I do not have an estimate of long term disk storage requirements.

III. GENERAL BACKGROUNDS

We need to analyze DIS events with backgrounds, which means merging beam on target events with DIS events. At 50 μA, the rate of electrons on target is 3.12×10^8 MHz. In a time window of 100 ns around any DIS trigger there are 31 million electrons on target. Most of these will not produce background of interest. For example, the fraction that will produce electrons or positrons in the EC is about 5×10^{-6}. There are several approaches to background merging one could imagine.

In GEMC, in principle, one could generate tens of millions of electrons on target to accompany each DIS electron, spread over a suitable time window, and simulate them all together. (The length of the time window needed would be dependent upon the yet to be determined time constants of the DAQ electronics.) A mechanism exists in GEMC to do something like this, though it was developed with CLAS in mind and the much higher luminosity of PVDIS likely would require modifications if it can be done at all. Indeed, CLAS does something slightly different: they simulate only the background electrons at a fraction, say 10%, of full luminosity and write to an output file the parameters (position, momentum, and particle ID) of all tracks that intersect their detectors. They then merge the tracks from 10 such events with each signal primary in a subsequent simulation. This allows them to simulate a smaller number of background electrons at a time, and permits them to recycle such simulated backgrounds as needed. The fact that CLAS finds it desirable to simulate backgrounds at only 10% of their luminosity suggests that this kind of approach would be extremely challenging if not impossible for PVDIS. Indeed, a test with "only" 30,000 electrons per event led to a fatal buffer size error in the evio to root conversion.

CLAS does this because they do detector digitization in GEMC. Our approach, however, generally is to generate energy-depositing hits in GEMC and then digitize these hits in a subsequent step. The GEMs and EC are done in this way (as are SPD, MRPC, and HGC). The LGC is semi-digitized in GEMC, that is, what is written to the output is the number of photoelectrons produced for each event, but not the ADC output. For all our detectors, hits (or photoelectron counts) from DIS events and background events can in principle be generated separately and added together in the digitization stage. This indeed is the approach taken in the existing GEM digitization code.

Average simulation time for a single electron on target is about 1.4 ms. With $4 \times 10^{11} M$ DIS triggers to simulate, it would hardly be feasible to generate tens of millions times as many beam on target events. Instead we will have to recycle background events from a smaller pool. To minimize correlations it would make sense to generate hit information for individual beam on target events, or relatively small numbers of beam on target events together, and then choose at random enough such outputs from the pool to constitute full background luminosity to add to each DIS event. How many such events would be needed for the pool has not yet been determined.

IV. HADRONIC BACKGROUNDS

For studies focusing on hadronic backgrounds we have developed the bggen generator, based on the Hall D code. It uses photoproduction data from the MAID and SAID databases to handle hadron generation from photons up to 3 GeV, and a modified version of PYTHIA for higher energies.[5] An equivalent photon approximation then is used to get electroproduction.

Note that PYTHIA was originally designed for Tevatron/LHC energies, and the modifications for Hall D include an empirical fudge factor:

"Above 3 GeV the PYTHIA generator... was used, slightly adapted for low energies. PYTHIA was designed and tuned by the authors for much higher energies. Special efforts were taken by the HERMES collaboration to adapt it to the HERA electron energy of $\sim$ 30 GeV. We slightly adapted the version from HERMES to the energies as low as 3 GeV, and compared the PYTHIA results with some experimental data. At 9 GeV PYTHIA underestimates the

total photoproduction cross section, providing $\sim 80\mu$b instead of $\sim 120\mu$b. However, the partial cross sections from PYTHIA, scaled up by a factor $120/80 = 1.5$ are in a reasonable agreement with the data..."

A drawback of bggen is that the processes selected for photons under 3 GeV do not include strange particle production. Strange particles are produced by the PYTHIA code for higher energy photons, but the checks of the modified code mentioned above do not appear to have included strange channels.

Another problem bggen originally had was that the vertex distribution was incorrect, being made uniform along the length of the target. Recently changes have been made to more correctly generate the vertex distribution, with differing z dependence for electro- and photoproduction. The effect of these changes needs to be evaluated.

To help evaluate the performance of bggen, it would be useful to compare with the hadrons produced from beam on target events in GEANT. We have recently made a modification to GEMC which allows one to request output of only those events in which a hadron hit occurs. In combination with a sensitive version of the target this could be used to generate a hadron track input file for comparison with bggen. Roughly one in one ten thousand electron on target events produces a hadron, with about one in a million having a kaon. A census of hadrons produced from 252 million events is shown in Table II.

TABLE II. Census of hadrons produced from 252 million events with 11 GeV electrons on target. "All" is count of all hadrons in the target. "1st gen" is count of only those hadrons which do not have a hadron as their immediate ancestor.

Particle	**All**	**1st gen**
π^0	23546	22504
π^+	11875	11350
π^-	12016	11429
η	664	433
η'	354	354
ρ	108	0
ω	9	0
K_S	96	96
K_L	100	100
K^+	111	109
K^-	55	54
p	22829	22731
$\bar{p}$	7	7
n	22734	22651
$\bar{n}$	8	7
Λ	23	16
Σ^+	3	3
Σ^0	7	7
Σ^-	5	5
$\bar{\Sigma}^+$	1	1
nuclei	10240	6678

[1] R. Jones, "GlueX Experience with the Open Science Grid" (16 Apr 2019) (https://solid.jlab.org/DocDB/0001/000163/001/Solid-OSG-4-2019.pdf)
[2] "Report of the 37th Program Advisory Committee (PAC37) Meeting" (11 Jan 2011) (https://www.jlab.org/exp_prog/PACpage/PAC37/PAC37_Report.pdf)
[3] PVDIS proposal (15 Dec 2008)
[4] Y. Zhao, "PVDIS trigger rate" (https://hallaweb.jlab.org/12GeV/SoLID/download/sim/talk/PVDIS_trigger_rate.pdf)
[5] GlueX Collaboration, "Hall D Trigger Simulation" (30 Apr 2008) (https://hallaweb.jlab.org/12GeV/SoLID/download/sim/talk/trigger-review-2008.pdf)