PROCEEDINGS OF SCIENCE

Generating the full SM at linear colliders

Mikael Berggren[†,*]

DESY
Notkestrasse 85
D-22607 Hamburg
Germany

E-mail: mikael.berggren@desy.de

Future linear e+e- colliders aim for extremely high precision measurements. To achieve this, not only excellent detectors and well controlled machine conditions are needed, but also the best possible estimate of backgrounds. To avoid that lacking channels and too low statistics becomes a major source of systematic errors in data-MC comparisons, all SM channels with the potential to yield at least a few events under the full lifetime of the projects need to be generated, with statistics largely exceeding that of the real data. Also machine conditions need to be accurately taken into account. This includes beam-polarisation, interactions due to the photons inevitably present in the highly focused beams, and coherent interactions of whole bunches. This endeavour has already been partly achieved in preparing design documents for both the ILC and CLIC: Comprehensive samples of fully simulated and reconstructed events are available for use. In this contribution, we present how the generation of physics events at linear colliders is categorised and organised, and the tools used. Also covered is how different aspects of machine conditions, different sources of spurious interactions (such as beam-induced backgrounds) are treated and the tools involved for these aspects.

40th International Conference on High Energy physics - ICHEP2020
July 28 - August 6, 2020
Prague, Czech Republic (virtual meeting)

[*]Speaker
[†]On behalf of the generator group (LCGG) of the Linear Collider Collaboration (LCC)

PoS(ICHEP2020)903

1. Linear colliders

All proposed Linear colliders would collide polarised electrons with positrons. Centre-of-mass energies would be ranging from 250 GeV to 3 TeV. As the e^+e^- initial state implies electroweak production, the background rates will be quite low at such machines. This has consequences for the detector design and optimisation. The detectors can feature close to $\sim 4\pi$ coverage, and they do not need to be radiation hard, so that the tracking system in front of calorimeters can have a thickness as low as a few percent of a radiation-length. In addition, the low rates means that the detectors needn't be triggered, so that *all* produced events will be available to analysis. Furthermore, at an e^+e^- machine, point-like objects are brought into collision, meaning that the initial state is fully known.

Two options for linear colliders are currently under study, ILC and CLIC. The ILC [1] has a defined 20 year running scenario, yielding integrated luminosities of 2 and 4 ab^{-1} at E_{CMS} = 250 and 500 GeV, respectively, and would be up-gradable to 1 TeV. An $E_{CMS} = M_Z$ option is also foreseeable. At the ILC, the positron beam would also be polarised. To construct the ILC is currently under high-level political consideration in Japan. Likewise, CLIC [2] has presented a 20 year staged running scenario, yielding integrated luminosities of 5, 2.5 and 1 ab^{-1} at E_{CMS} = 3 TeV, 1.5 TeV, and 380 GeV, respectively. CLIC is one possible future CERN project.

Linear colliders aim for extremely high precision measurements, requiring excellent detector performance, and well controlled machine conditions. But it also requires the *best possible estimate of backgrounds*. A corollary to this is that MC statistics or lacking channels *must not* be a major source of systematic errors. Therefore, all SM channels yielding at least a few events under the full lifetime of the projects need to be generated, with *statistics largely exceeding that of the real data*. In addition, machine conditions need to be accurately taken into account. Furthermore, at a linear collider, *all* events are interesting, and they are often fully reconstructed. In a sense, physics analysis at a linear collider might have more in common with a B-factory than with the LHC.

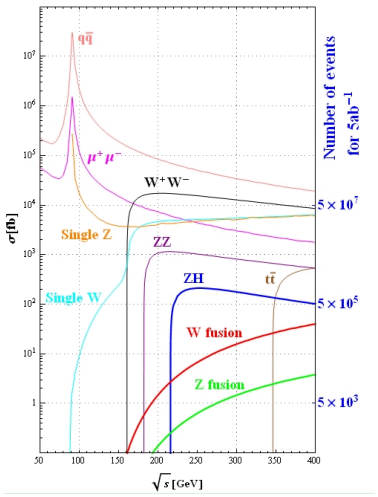
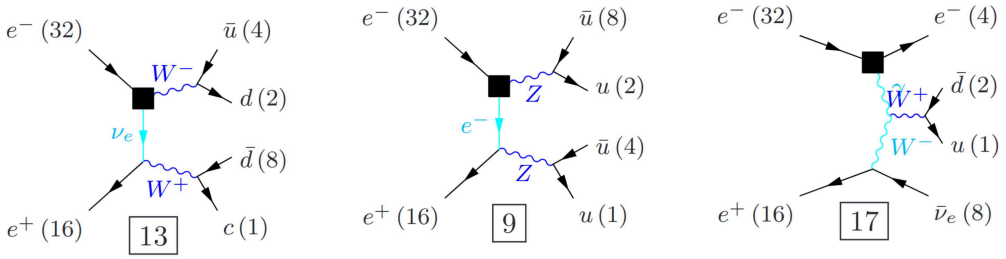


Figure 1: Production cross-sections for various e^+e^- processes. From [5].

The endeavour to achieve these requirements on event generation has been organised as a common effort between the ILD and SiD detector concept groups at ILC [3] and CLICdp group at CLIC [4]. The work is done within the generator group, the LCGG, of the Linear Collider Collaboration (LCC).

2. Generating the full SM

To generate the full SM, there are many details to consider, beyond the pure physics generation. One must determine what is colliding, since not only electrons and positrons are present in the beams, but also photons. The incoming particles are not strictly mono-energetic, so the beam-spectrum must be known and specified. The degree of polarisation of the beams must also be assigned,



(a) WW_hadronic **(b)** ZZ_hadronic **(c)** SingleW_semileptonic

Figure 2: Examples of four-fermions diagrams with different flavour groupings.

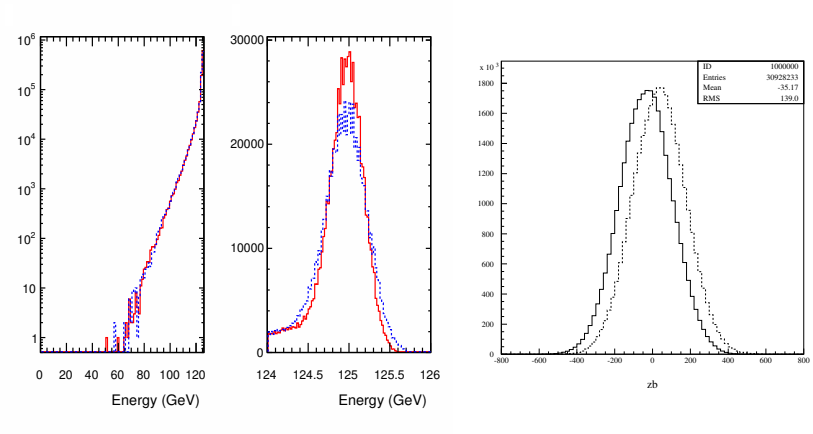
and the distribution in space of the interaction point. One must also consider what else happens during the beam-crossings, namely beam-strahlung in the very strong fields at the interaction point and parasitic $\gamma\gamma$ interactions in the same bunch crossing as the physics events. These sources of additional particles must be generated to be able to correctly treat them in the subsequent detector simulation. In addition to the specification of the beam-properties and determination of spurious interactions, the physics channels themselves needs to be considered in detail, when determining how to proceed with the generation. At a linear collider, all events are interesting, and feature a huge spread in cross-sections, as illustrated in fig 1. At generation time, one cannot apriori know if a given physics study needs to consider a tiny cross-section process as an important background, or possibly a tiny fraction of a huge cross-section one. The general characteristics of different sources of background will be different in different cases. Therefore, processes should be grouped at generation time in sufficiently well thought-through and well documented way as to serve as many physics analyses as possible, in the most convenient way.

The process classification starts at defining the initial state, either e^+e^-, $e^{+(-)}\gamma$ or $\gamma\gamma$. For electrons and positrons, the polarisation is specified, and for γ:s whether they are real or virtual ones. The final state is then classified in several levels. Firstly, the number of final fermions (1 to 8) is defined. Then a flavour-grouping is performed, determining if the final state can arise from intermediate W or Z bosons, or if it is ambiguous. Finally, fully leptonic, fully hadronic, and semi-leptonic final states are separated. In the "Z-leptonic" case , final states with neutrinos were also separated out. Figure 2 illustrates a few examples of four-fermion flavour groupings.

2.1 Main generator: Whizard

`Whizard` [6] is the generator of choice for e^+e^-. It features a full tree-level matrix-element evaluation. It treats polarised beams, and contains a full treatment of helicity densities. The code traces the colour flow in full, and passes this from the hard interaction generation to the code used to develop the subsequent parton-shower. Using it's `Circe2` component, `Whizard` can handle arbitrary beam-spectra, and, using `Tauola` [7], decays of polarised τ:s are correctly treated. Finally, `Whizard` can generate $2 \to 8$ processes. The subsequent parton-shower and hadronisation is done by other codes, typically `Pythia6.4` [8]. LCGG has tuned hadronisation using input from OPAL at LEPII [9].

The process-definition given in the `Whizard` steering file (known as the *sindarin* script,



(a) **(b)**

Figure 3: Beam-characteristics: (a) e^- (dashed) and e^+ (solid) beam-spectra. (b) Position of interaction point for $e^+\gamma$ (solid) and $e^-\gamma$ (dash) events, in μm.

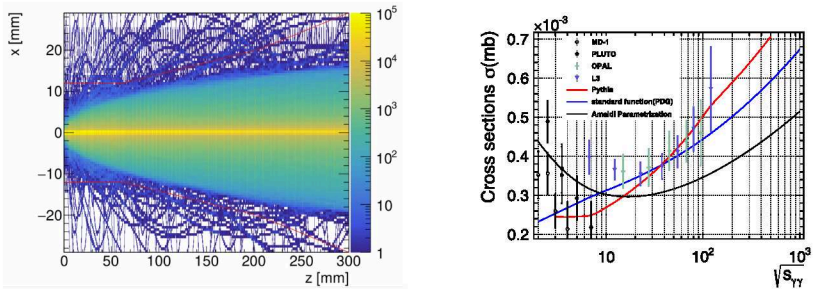
which is a component of the `Whizard` package) is also the driver for the scripts that organises the production along the lines explained above. Sindarin contains powerful grouping and aliasing capabilities, which are exploited to assure that no processes are over-looked.

2.2 Generating beam properties

The electron and positron beam-spectra are determined both by the incoming beam-spread, and also by beam-beam interactions, which are due to the need to very strongly focus the beams at linear colliders. In addition to electrons and positrons, the beams also contain photons, and one must determine how many they are, and whether they are virtual or real.

The incoming beam-spread is due to the workings of the damping-rings and, for the electron beam, the ondulator (used for producing the polarised photons needed to create the polarised positrons). For these properties, external input from the machine-scientists are utilised, while the interaction region is simulated using `GuineaPig` [10]. This program supplies the beam-spectrum for electrons and positrons individually, the amount and spectrum of real photons, and the spacial distribution of the interaction point. Figure 3 shows the results of `GuineaPig` for ILC-250.

In addition, there are spurious interactions of two types: The *pair-background*, which arises from pair-creation of photons in the beam by the strong fields. The `GuineaPig` code can also be used to generate the full activity during a beam-crossing (a "BX"). The second type of spurious interactions are *low-$p_\perp$ hadrons*, i.e. $\gamma^{(*)}\gamma^{(*)}$ interaction with small $M_{\gamma\gamma}$ and multiplicity, amounting to $O(1)$ event in each BX. These are either generated by `Pythia`, or a custom generator developed by LCGG relying on fitting to available data. For both these types of spurious interactions a pool of events were pre-generated, from which events were picked at random and overlaid on the main event during detector simulation. In the case of the pair-background, the pool of events where obtained by simulating $\sim 10^5$ BXes, and using the fast detector simulation SGV [11] to select those tracks actually reaching any element of the tracking system (~10/BX). A large fraction of the rest of the pairs would hit the very-forward calorimeter system (the BeamCal), and were used to build a map of background on the system, used in the full detector simulation to simulate the BeamCal.



(a) Pair-background in one BX. **(b)** Cross-section of $\gamma\gamma \to hadrons$.

Figure 4: Sources of spurious particles. In (a), the red line is the beam-pipe.

3. Setup, integration, event generation, and documentation

The generation is preformed in a number of steps, with verification done between them. The initial few steps are done interactively. First, the `Whizard` process definition is parsed to build a directory-tree structure, with one unique directory per process. This is to avoid possible race-conditions later, in the actual generation step. The process-specific code is generated and compiled, then the tree is traversed to do a "pre-integration" of all channels, to flag zero cross-section ones.

The full integration of all (non-zero cross-section) channels is then submitted to the local batch-farm, with the goal that the calculated cross-section should be accurate to 0.1 %. For all channels with ≤ 5 fermions, this was found to be an over-night job. The last step before the full generation was a "pilot generation' of all channels, with 1000 events/channel. This was useful to evaluate the CPU time and storage needed for the full generation. A number of channels were identified for which either the precision or efficiency of the generation was problematic. These were channels with low cross-sections and complicated final states, and it was decided to defer their generation until after the full simulation of the bulk of the samples was completed.

The full generation was also done on a batch-farm. The production of event files - in LCIO [12] format - was supervised by a daemon, which was responsible to upload each completed file to the grid as soon as possible. Upon successful upload, the local file was deleted, avoiding disk-space issues on the batch-cluster. When *all* jobs of a channel were completed, a summary metadata file was created and uploaded to the grid, together with input and log-file tarballs. The existence of the metadata file triggered the full simulation system under `DIRAC` [13] to start processing the channel.

A number of precautions were taken to balance the size of the output files, both in physical size, and the number of events they contained: a demand from the down-stream detector simulation was that all files of any given process should contain the same number of events, and that single files should not (much) exceed 500 MB in size. As channels were split both in several files, but also in some cases in several jobs, care should be taken with event- and file-numbering: no gaps in the event-numbering sequence, nor in the file sequence numbers.

At the time of writing (December 2020), 104 of the total of 212 channels with ≤ 5 fermions have been generated (478 jobs), corresponding to 2.7 billion events - 5.4 TB in 15788 files. The total CPU time was 7233 hours, which was obtained in 10 days on the batch farm. The remaining deferred channels are expected to contain 0.5 billion events.

The full documentation of each channel was created by generation job, driven by the contents

of the process- definition Sindarin script and common conditions. This information can be found in the header of each event (Process-id, beam-polarisation, and cross-section), and, in full, in the generator meta-data files. Apart from being uploaded to the grid, this information is also available in browasble format on the Web. In addition to the condensed metadata, all steering-files, log-files, pdf:s showing the diagrams contributing, and the integration phase-space grids, are available on the Web, and in tar files, which are stored on the grid in a parallel directory to the generated files. This collection of information is sufficient to re-do the generation, if e.g. more events are requested in some channel, or if it is found that detailed debugging would be needed.

4. Conclusions

We have explained that the precision-goal of future e^+e^- colliders are such that it is *not permissible* that MC statistics or coverage could constitute an *major systematic uncertainty*. In this spirit, we showed how the generation of the full SM can be achieved. It consists of bringing a large number of different codes together, namely: `Whizard`, `Pythia`, and `Tauola` for physics generation, `GuineaPig`, and `Circe2` for beam-properties, and `SGV+GuineaPig` and the Peskin/Barklow generator for spurious interactions. In addition, input from machine-physics and data from LEPII was used. This full data is organised and documented in a physics-oriented fashion, for the benefit of the end-user. The system pivots around *one data-source*, the Whizard process definition file.

References

[1] C. Adolphsen, M. Barone, B. Barish, *et al.*, [arXiv:1306.6328 [physics.acc-ph]].

[2] M. Aicheler, P. Burrows, *et al.*, CERN-2012-007

[3] T. Behnke *et al.*, arXiv:1306.6329 [physics.ins-det].

[4] N. Alipour Tehrani *et al.* [CLICdp], CLICdp-Note-2017-001.

[5] X. Mo, G. Li, M. Q. Ruan and X. C. Lou, Chin. Phys. C **40** (2016) no.3, 033001 [arXiv:1505.01008 [hep-ex]].

[6] W. Kilian, T. Ohl and J. Reuter, Eur. Phys. J. C **71** (2011) 1742 [arXiv:0708.4233 [hep-ph]].

[7] S. Jadach, J. H. Kuhn and Z. Was, Comput. Phys. Commun. **64** (1990), 275-299

[8] T. Sjostrand, S. Mrenna and P. Z. Skands, JHEP **05** (2006), 026 [arXiv:hep-ph/0603175 [hep-ph]].

[9] A. Boehrer, Phys. Rept. **291** (1997), 107-217

[10] D. Schulte, CERN-PS-99-014-LP.

[11] M. Berggren, [arXiv:1203.0217 [physics.ins-det]].

[12] S. Aplin, J. Engels, F. Gaede, N. A. Graf, T. Johnson and J. McCormick, doi:10.1109/NSSMIC.2012.6551478.

[13] A. Tsaregorodtsev, *et al.*. J. Phys. Conf. Ser. **119** (2008), 062048