SDC-92-229

SDC
SOLENOIDAL DETECTOR NOTES

OFFLINE COMPUTING SYSTEM USE SCENARIOS

SDC COMPUTING SECTION, COMPUTING MODEL AND OPERATIONAL CONCEPTS

February 9, 1992

Robert J. Ennis
SDC Computing Working Group
IBM Federal Systems

SDC-92-229

OFFLINE COMPUTING SYSTEM USE SCENARIOS

SDC COMPUTING SECTION, COMPUTING MODEL AND OPERATIONAL CONCEPTS

February 9, 1992

SDC Computing Working Group
IBM Federal Systems
Robert J. Ennis

ABSTRACT:

This SDC note presents a collection of scenarios that describe how physicists expect to use the SDC Offline Computing.

For additional information or questions please contact:
Bob Ennis (713) 282-8958, e-mail BOBENNIS @ HOUVMSCC.VNET.IBM.COM.

CONTENTS

1. PRODUCTION 1-1
1.1 Production Analysis 1-1
1.2 Express-Line Analysis 1-1

2. ANALYSIS 2-1
2.1 Data Filtering/Compression 2-1
2.2 Time-dependant Calibration 2-2
2.3 Event Scanning 2-2
2.4 Calculation of Total Reaction Cross Sections 2-3
2.5 Searches and Limits 2-5
2.6 Precision Measurement of a New State's Lifetime 2-5

3. SIMULATION 3-1
3.1 Generate Large, "Unbiased" Samples of Specific Events 3-1
3.2 Special Case Samples 3-1

1. PRODUCTION

1.1 PRODUCTION ANALYSIS

In the production stage of analysis, all of the data is taken from archival storage or directly across a network from Data Acquisition and passed through a suite of programs which perform reconstruction and filtering of the data. The output from this initial stage of analysis is used for all subsequent analysis of the data, with the original data being referred to only in very unusual circumstances or for a new complete pass of production analysis if that should become necessary. Reconstruction consists of the compute-intensive tasks of converting the raw or partially-processed data into physically meaningful quantities such as energy and momentum of individual particle tracks, calibrated energy depositions, clusters of tracks and/or energy, particle identification information, etc. The filtering part of production acts as a fourth-level trigger, using the finer view of the data after the reconstruction stage to reject events which have passed the triggering filters, but are now recognized as background. Production analysis is performed as mechanically and systematically as possible, keeping algorithms and parameters as constant as possible, so that measurements of biases and other systematic effects introduced at this stage can be applied to a large set of

The production analysis processing is preceded by extensive preparation aimed at making both the reconstruction and filtering aspects as reliable as possible, since all physics results from this data are directly dependent on the quality and reliability of this step. If production is done in line with data taking, without a period of storage, the preparation and tuning of production programs will have been done on initial samples of data on a detector that is understood well-enough and stable enough that frequent changes are not needed. A continuous process of updating calibration constants will be in place, with updated constants used in reconstruction to allow the reconstruction to follow slow changes in detector response. Preparation for production will include a process of agreement by the collaboration of what calculations should be made at this stage, what derived quantities should be output, and what filtering criteria should be applied, as well as extensive development and testing of the algorithms used for reconstruction and filtering.

1.2 EXPRESS-LINE ANALYSIS

Express-line analysis is the production phase of analysis for the small sample of events selected on-line as likely to contain high priority data that needs to be analyzed quickly. The selected data will be reconstructed as quickly as possible. In the early phases of data taking, express-line data may be reconstructed with early, incomplete versions of the production program, so that physics analysis can proceed expeditiously. Later, the significance of the express-line may be that it gets processed first, if production is not done immediately, or that it is maintained as a small sample that is likely to contain important physics and can be quickly used for later analysis stages to get an early look at the selected physics.

2. ANALYSIS

2.1 DATA FILTERING/COMPRESSION

Production takes the data passed by the level 3 triggers, reconstructs that data as quantities having physical meaning, filters the data to eliminate background, then classifies the surviving events into "production streams" based on specific selection criteria. These production streams are likely to number in the tens (not hundreds or thousands).

The production streams that result from this process represent an initial data reduction, but each such stream will typically still contain massive amounts of data. Additional filtering and/or compression need to be applied to the data from production streams to produce sets of data that are suitable for analysis. "Suitable for analysis" in the sense that:

- Data volume is manageable for the tools to be used for the analysis,
- The data set represents a concentration of events with characteristics of interest to the given analysis.

A lower limit to the size of each data set produced is established by physics criteria. The upper limit is established by the capabilities of the data handling and data management technology in use. The possible mitigation of the constraints imposed by the limitations of current data handling methods has motivated research into the use of RDBMS or object-oriented data structure and organization.

The data reduction is accomplished by:

- filtering (keeping only a certain % of all events)
- event compression (or reduction)

  Event compression is accomplished by...

  - Utilizing data compression schemes

    The compression must be lossless. Historically, data compression has seldom been used.

  - Discarding nonessential subsets of events.

For each event, the filtering/compression process procedes as follows:

- Check event integrity against database; look for any reason to eliminate (online problems, etc.)
- Make global event cuts (primary vertex, cosmic ray rejection, ...) using selection criteria data (eliminates events)
- Classify, based mostly on information generated by the reconstruction

- Reduce and sort; data is organized into "banks" or arrays which are units marked for inclusion into output streams, or compressed and marked for output, etc.

- Write into output streams

This additional filtering/compression is like-in-kind to the filtering and classification done during production, with the following differences:

- Production operates on the data passed by the level 3 triggers. This filtering/compression is typically done to data from the output streams produced by production (although nothing precludes going back to the raw level 3 output data as input to this process).

- Production is global in that it is done to all the data passed by level 3. This filtering/compression will be applied to subsets of the production output (to one or more production streams, or subsets of a stream).

- Production data selection and classification criteria are "controlled" (agreed-to, then set) by the Collaboration. The criteria used in this filtering/compression are generally decided upon by the individuals or groups planning specific analysis.

- Production does not arbitrarily discard data. Filtering/compression may, in order to reduce data volume to a manageable level (e.g., only use 10% of otherwise qualified events).

2.2 TIME-DEPENDANT CALIBRATION

Time-dependent calibration studies will need to be done on several different time scales, ranging from a single data run (in the case of temperature-related gain variation, for example) to the lifetime of the detector (aging of scintillator, for example). Access to a requested fraction of events over a possibly very large sample of events, and to centrally-maintained time-keyed databases of calibration and detector monitoring constants, would be required. In cases where the calibration data are being accumulated in real-time, there will need to be access to events from the data acquisition pool and detector monitoring data, as well.

2.3 EVENT SCANNING

Especially in the case of new physics searches, the physics user will need to take a reasonably small sample (<200) of events culled by hard selection criteria, and subject them to hand scanning. This will require processing of a large sample of reconstructed events provided by the regional server, selection and streaming of the events to be scanned (to local storage possibly), and access to common (platform independent) graphical single-event display software. The user will, for example, want to be able to zoom the event picture or view alternate 2-dimensional representations of the event, or zoom or fetch an enhanced display of individual reconstructed "objects" (e.g. jets) or raw detector channels.

Event displays will also be important for commissioning the detector and event reconstruction software in the early stages of operation. In this case, the physics user (or detector operator) will want to select events from the data acquisition or expressline pools.

2.4 CALCULATION OF TOTAL REACTION CROSS SECTIONS

Calculations of cross sections requires three independently measured quantities:

1. # of detected events, say N, satisfying a particular condition.

2. The overall efficiency E, representing the probability that an event of the appropriate type will be detected and recorded if it is produced.

3. The integrated flux or luminosity L, representing the total number of collisions that took place during the period of time that the data was collected.

Given these quantities, the cross section can be expressed as N/L, where:

- The number of detected events "N" is measured from the experimental data.

- The overall efficiency "E" is calculated from a combination of simulation, special calibration data, and normal data samples.

- The integrated luminosity "L" is measured run-by-run as part of the experiment monitoring.

The evaluation of these quantities involves the following steps (Note: The steps will not necessarily be performed in precisely this order. The process is somewhat iterative; certain steps may be redone on the basis of information gained in other steps.):

1. Select the data samples (runs or running periods) which will be used for this analysis. This selection involves understanding which data samples are of sufficiently high quality to be used. Samples might be deemed to be of insufficeint quality if pieces of the detector were not working or if unacceptable biases were introduced by triggering or other conditions.

2. For the selected data samples, accumulate and sum the run-by-run luminosity (including correction for experimental dead time). This requires access to the database containing all experiment rates and conditions over the entire time of the experiment. This determines L.

3. Using a selected sample of events from a small fraction of the running time, selection cuts must be determined that select the "best" sample of events (for counting N). This will require repeated (10 - 100 times) access to data samples of a few thousand events, where the few thousand events have been selected from a much larger sample (but not the entire run).

4. These same selection cuts must be studied and applied to data samples from other parts of the run to insure their sensitivity.

   Problems require re-doing step 3.

5. Once the final selection criteria have been established, the cuts must be applied to the entire data sample. This requires accessing all the data (all events) once to apply the selection criteria.

6. From the final full sample of events, "N" is determined through a separation of "signal" and background events. This may involve fitting procedures, counting of background events and subtraction, or more sophisticated methods. In addition, this may require multiple passes over selected data samples. This determines N.

7. To calculate "E".

   The overall detector efficiency has 2 components:

   - a "static" component, the geometric acceptance, representing how likely a produced event would be detected by a "perfect" detector

   - a dynamic component, the efficiency induced by detector performance, trigger selecting, etc.

The first component is calculated from simulation and subsequent reconstruction of large numbers of events of the appropiate type. Note that it may be critical to feed back experimentally measured parameters into the simulations when acceptances and detector resolutions are intertwined. Complicated fitting procedures may be needed to determine acceptances, even to the extent of calculating individual weights for each event (requiring step 6 to be redone and N to be calculated in a more complex manner). Note that the analysis cuts determined in step 3 must also be applied to the simulated data.

The second component requires measurement of detector and trigger performance over the entire running period. Some of the contributions of this efficiency are made from standard event data (e.g. efficiency of the tracking chamber), while others must be determined from special calibration or monitoring data (such as trigger efficiencies). These calculations will require access to significant amounts of data from the entire run to create a database of time dependent detector efficienciesthat can be used in many such analyses. If there is too much variance in these efficiencies, the efficiency corrections will need to be applied run-by-run, requiring recalculation of L in step 2.

2.5 SEARCHES AND LIMITS

- Carry out Monte Carlo study of signal and background to learn about optimization of selection/cuts to maximize signal to background ratio in search parameters (Example: SUPERSYMMETRY (SUSY) search in MISSING TRANSVERSE ENERGY).

- Define a broad selection/cut set to isolate data sample for more detailed study. This allows some subsequent tuning of tighter cuts without the need to return to the overall data collection.

- Check correspondence of Monte Carlo background and data sample (broad) for same cuts.

- Apply cuts etc. from initial Monte Carlo study to the data.

- Iterate and refine cuts to suppress detector characteristics/performance (e.g. special treatment of dead regions).

- Study "final" distribution of parameter of interest.

  - If an excess of events found over background expectation - PUBLISH DISCOVERY and study signal nature.

  - If no excess is found, use the results of Monte Carlo study (a function of mass) to set lower limit on SUSY particle mass, using a statistical argument based on observation of events above background in the research region of E/T. PUBLISH LIMIT.

2.6 PRECISION MEASUREMENT OF A NEW STATE'S LIFETIME

This is based on my experience with measurement of charm particle lifetimes. Let us assume the state is produced rarely and a 2% measurement of the lifetime is desired. The basic technique will be to measure the vertex position of the state's decay products relative to the main event vertex or collision point. The vertex position is determined by intersecting all the tracks decay products ion three dimensions to find a best common origin. Simultaneous measurement of the vertex position and the momentum of the decay products serves to measure the lifetime of a particular event and a fit of the distribution of these lifetimes to the form e**-t/tau is used to determine tau the mean lifetime.

Because the state is presumably rare, one will possibly use a combination of cuts on the reconstructed invariant mass of state; possibly cuts on the identification of decay particles, and undoubtedly a requirement that the state does not verticise at the collision point. These cuts will be used both to reduce the data to a manageable level and to obtain a suitable signal to noise to allow for a background subtraction. A good rule of thumb is that the particle detector used in intersecting tracks should have a (transverse) position resolution of about c*tau/5. This will allow one to require a 3.5 standard deviation vertex separation (reducing the background by a factor of 2,000-4,000) while only losing 1/2 the signal. The job of looking for detached verticies in events with a large number of tracks is likely to be the most compute intensive part of the task. If the state decays into special, easily identifiable particles such as muons or electrons - only these combinations need be intersected and the savings in computer time will be enormous.

To measure tau to 2% requires approximately 2500 signal events. It is important to skim off >= 10 times this number of events in order to study the lifetime distribution of the background. It is quite possible that one will choose a set of looser cuts initially so that one can optimize the vertex cuts to make the most systematics-free measurement. An isolated sample of about 1 million candidate events might be expected. These would be stripped from the full experimental sample and hopefully stored on disk for repeated analysis.

Once suitable cuts are found, candidates with masses consistent with the signal mass will be put into a file along with their observed proper life. Candidates with mass adjacent to the signal region will have their lifetimes placed into a background file. Histograms will be made of the lifetime distribution for signal and background events. The signal region time distribution will be fit to the form:

$$a*f(t)*e**(-t/tau) + b*B(t),$$

where f(t) gives corrections to a pure exponential decay owing to absorption of particles, inefficiencies in finding verticies at short lifetimes, resolution, and possible changes in geometrical acceptance. B(t) is the lifetime distribution measured for the background events.

The correction function f(t) is determined from Monte Carlo simulation of the new state decaying in the detector. Physical processes and detector response must be well modeled by the simulation program and the simulated events must be put through the same computer codes as actual data. One should endeavor to make f(t) as constant as possible by proper selection of the analysis cuts to minimize systematic errors. If about ten times the final data set is simulated, the statistical fluctuations in the Monte Carlo will only be about 5% of the statistical error in the data.

3. SIMULATION

3.1 GENERATE LARGE, "UNBIASED" SAMPLES OF SPECIFIC EVENTS

The user...

1. Initiates a session with the SDC Computing System
2. Examines and modifies a simulation data file, which specifies the type of event to be generated
3. Starts the simulation
4. Examines the intial generated event and checks that the event is OK
5. Optionally, repeat steps 2-4 for several more events
6. If satisfied with the examined event(s), restarts the simulation
7. Terminates session (logs off)

The system...

1. When the simulation is finished, automatically stores the data produced in a manner such that they are:

   - accesible to the user the next time he logs on
   - accesible to analysis procedures.

3.2 SPECIAL CASE SAMPLES

The user...

1. Initiates a session with the SDC Computing System.
2. Obtains simulation source code.
3. Modifies simulation source code to enable it to generate special-case event.
4. Compiles modified code and builds an executable.
5. Tests modified code
6. Generate single or multiple events and look for inconsistencies
7. Repeats steps 3-6 until results are satisfactory.
8. Starts simulation run

   The simulation program may write the event data produced to a file or present it to an analysis program for immediate analysis.

9. Starts analysis program

   The analysis program will...

   - Read simulated data
   - Display events graphically
   - Examine an event track by track
   - Generate histograms