1. Introduction

"Computer modelling", i.e. system simulation, takes into account and provides insight in contention for resources and the resulting queuing. The latency (the time required to produce an accept or reject decision) distribution for the system modelled can be determined, as well as distributions of the filling degree of queues (i.e. of the required buffer sizes) and of the utilization of resources such as communication link bandwidth and processor capacity. The computer model can be checked against the paper model, as the average resource utilization should be the same.

2. Goals

The following goals have been formulated :

1. input to decision making,
2. aqcuisition of knowledge about the factors controlling system behaviour and resource requirements,
3. understanding relevant technologies with respect to behaviour and resource requirements if applied in the ATLAS environment,
4. provision of help in understanding results obtained with test set-ups, with as side effect "confidence building" with respect to modelling results.

It should be noted that simulation is necessary to acquire a good understanding of the factors controlling the behavior of the LVL2 trigger system. This is due to the large number of processors in combination with the networks and switches providing the communication facilities required and the use of RoIs for control of the dataflows, possibly in combination with sequential processing.

3. Method

The type of simulation used is called discrete event simulation. Events (not to be confused with events due to particle interactions, observed in an experiment and in this document referred to as "physics events") occur at certain simulation times. For each event the response of the simulated system is determined. The response can consist of the generation of new events at the same simulation time as the original event occurred or at future simulation times.

In order for the results of simulation to make sense many input parameters need to be set to realistic values. The operation of the system also depends on the type of events selected by the first level trigger and the number and types of RoIs associated with them. Two approaches are possible here. In the first the relevant information is extracted from simulated events and used as input for the simulation on an event-by-event basis. Alternatively an estimate can be made of the number of each type of events selected by the first level trigger, of the number of RoIs associated with each type and of the distribution of and correlation between the RoI positions. With this information the relevant properties of the events can be generated during running of the simulation program. This approach in principle is less accurate than the first, but provides a good first order estimate without requiring access to large samples of Monte-Carlo events.

For simulation of the ATLAS second level trigger the SIMDAQ program is developed. Its first implementation has been in MODSIM-II. This version [R1], apart from the code for the simulation of the SCI and ATM technologies, has been translated and extended in C++. The emphasis until now has been on the organization of the simulation program, on simulating "generic" models and on the correlation with paper models. Implementation of models of the various network technologies of interest has been partially done in C++.

The C++ program makes use of a platform independent graphical user interface. UNIX, Windows95 / WINDOWS NT and MacOS versions are available.

The model to be simulated is specified, at the level of processors and switches and their connections with a configuration file. The details at lower level can be controlled by parameters, that can be specified in the configuration file.

4. Present status

The current version of the program supports efficient simulation of the full model B architecture as used for the paper model and as outlined in figure 1. It has been found that results of the simulation program for processor and communication link utilization are in excellent agreement with the paper model results. Results with respect to the decision latency are discussed in section 5.

Since last summer the work force has expanded from a single person to a core group with persons from Argonne, Krakow, NIKHEF, Saclay and UCL.

Recently persons from Copenhagen and Manchester have started to work with Ptolemy [R2], a general purpose simulation tool that also supports discrete event simulation. This tool seems to be well suited for studying relatively small systems. This type of studies may provide valuable input for large scale simulations

5. Some results

A detailed study has been conducted on the model B second level trigger system (see figure 1), using the models and parameters documented in [R3]. However, for the switches and network technology no models are provided in this document. For the study mentioned the switches are crossbar switches with unlimited buffering on input and output links and with arbitration for access to the output buffers. This arbitration can be switched off for studying the effect of it. Aggregate switches can be built from these switches. The links transport data with a fixed speed of 10 Mbyte / s.

For the "extended trigger menu without missing energy trigger items" a surprising result was found for the latency distribution, using the nominal first level trigger accept rate of 40.0 kHz. With this trigger menu the utilization of the processing resources of the hadron calorimeter ROBs would be more than 100 % when the parameter values of [R3] are used. Therefore the task switching time was reduced from 50 to 35 µs in order to obtain utilizations below 100 %. All other parameter values were set as specified in [R3], resulting in processor utilizations ranging from 28 to 42 %. For a bandwidth of 10 Mbyte / s the link utilizations are all below 30 % except for the hadron calorimeter ROB output link with an average utilization of 59 %. All processing times were fixed, as well as the length of the interval between consecutive first level trigger accepts. The switches were modelled as single crossbar switches of the type described earlier in this section. The internal bandwidth was set to 50 MByte / s per connection.

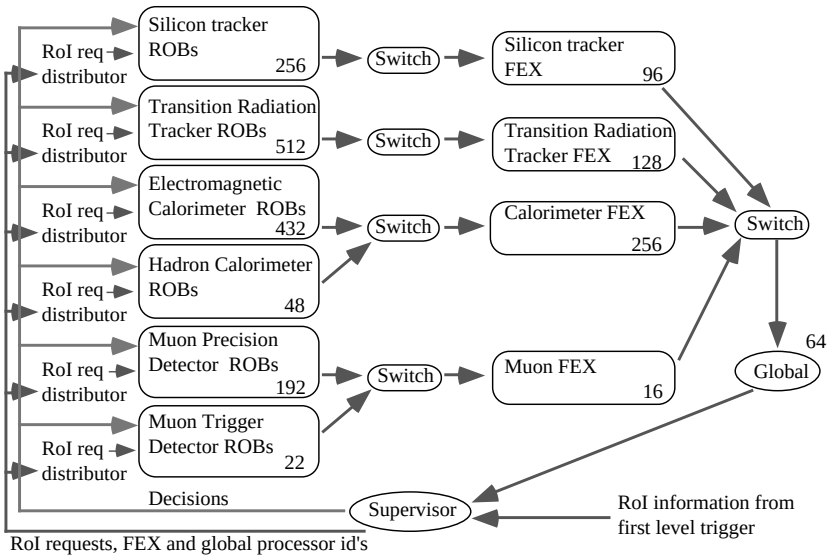

Figure 1: The full architecture B system, as defined for the paper model. The numbers indicate the number of Robs or number of processors, "FEX" stands for "feature extraction". The switches are either single switches or aggregate switches built from two layers of switches . The 100 Mbyte / s data links (one per ROB) transporting the raw data to the ROBs are not shown.

The latency distribution shows a number of peaks with equal distances between the peaks. Figure 2 shows latency distributions taken at different times. For this simulation 1 hr of running time corresponded to about 7.5 s of simulated time on a 200 Mhz Pentium Pro machine with Windows NT as operating system.

It was found that the peaks in figure 2 are due to the round robin allocation of the feature extraction processors. The distance between the peaks is equal to the time of one round robin cycle for the processors handling data from the calorimeter (2.7 ms for 256 processors, as the total RoI rate for the calorimeter is 94 kHz) and changes as expected when their number is changed. A smooth distribution is obtained when the processors with the smallest number of events queued, as determined by the supervisor from previous assignments and from the event id's contained in the decisions received from the global processors, are allocated. However, this does not lead to a decrease of the width of the distribution : the system modelled even becomes instable, i.e. the maximum latency grows without bounds.

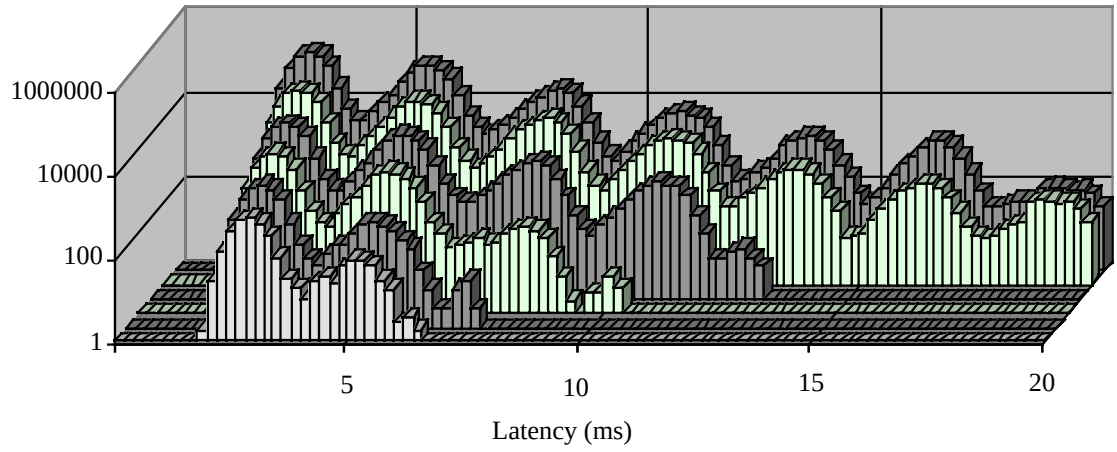

Figure 2: Distributions for the decision time (latency) of the architecture B second-level trigger, as obtained after about 0.1, 0.3, 1.0, 3.0, 10.0 and 30.0 s of simulated time for the system of figure 1, when single switches are used. The numbers along the vertical axis indicate the number of events. The relative large increase of the tail at longer simulated times shows that the system is nearly or just unstable (i.e. longer and longer latencies may occur, until buffer overflow reduces the number of events to be processed).

The width of the latency distribution can be explained by queuing in the switch connecting the calorimeter ROBs to the feature extraction processors. This is evident from the distribution of the time needed for transport of all data fragments of a single RoI across the switch : the distribution has approximately the same shape and width as the latency distribution. The queuing in the switch occurs predominantly for fragments sent by the hadron calorimeter ROBs : the distribution of the time needed for transfer of a single fragment across the switch is relatively narrow for the electromagnetic calorimeter ROBs, while for the hadron calorimeter ROBs it has again approximately the same shape and width as the latency distribution. This is due to the high RoI request rate for these ROBs (on average 5.9 kHz, for the electromagnetic calorimeter ROBs this rate is 2.5 kHz, for all other ROBs it is not higher than 1 kHz) in combination with the arbitration for access to an output port inside the switch. These two factors lead to queuing of the event fragments in the input ports of the switch. The time interval between the arrival of a RoI request at a hadron calorimeter ROB and the availability of the event fragment at the output of the FIFO queue in the switch input port receiving data from that ROB can be longer than the average time interval between two successive assignments of the same feature extraction processor (i.e. for round robin assignment the length of one round robin cycle). This leads to additional contention in the switch. In the case of round robin assignment the amount of contention changes periodically, which most likely causes the peaks in the latency distribution.

Results were also generated for a system with aggregate switches, each consisting of an input and an output layer of smaller switches. An almost smooth distribution for the latency was obtained, with a tail not extending beyond 20 ms, by choosing a suitable configuration of the switches (for the calorimeter : 54 switches with 8 inputs connecting to the electromagnetic calorimeter, 12 switches with 4 inputs connecting to the hadron calorimeter and 16 switches with 16 outputs connecting to the feaure extraction processors) resulting in a reduction of the probability of head of line blocking.

6. Outlook

The current workplan consists of completion of implementation and study of models used also for paper modelling, including a study of the impact of the different trigger strategies (parallel/sequential) on the behaviour of the system modelled. A next step is"generic" simulation (with relatively simple models) of laboratory setups and identification of main factors determining behaviour. This is to be followed by detailed simulation of technologies calibrated using results of laboratory measurements and re-using code / models developed earlier for ATM, SCI and DS link technologies.

References

[R1] S.Hunt et al, “SIMDAQ - A System for Modelling DAQ/Trigger Systems", IEEE Trans. on Nucl. Sci. 43, no.1, pp. 69 - 73

[R2] Ptolemy, http://ptolemy.eecs.berkeley.edu/

[R3] S. George et al., “Input Parameters for Modelling the ATLAS Second Level Trigger”, ATLAS Internal Note, DAQ-No-070, June 1997.