References

[1] ATLAS Technical Proposal for a General-Purpose pp Experiment at the Large Hadron Collider at CERN, CERN/LHCC/94-43, LHCC/P2, 15 December 1994.

[2] CMS Technical Proposal, The Compact Muon Solenoid, CERN/LHCC/94-38, LHCC/P1, 15 December 1994.

[3] M. Costa et al., ATM-based Event-Building, ATLAS Internal Note DAQ-NO-024, 1st December 1994

ger rates.

The time to complete a full cycle is N x (T_c + T_t). Depending on T_t, and on the event fragment size distribution, several cycles might be necessary to complete the event-building of 1 event. Furthermore, several events might be concurrently in course of event-building at any one destination.

The choice of the time interval (T_t) during which a set of connections is active is made in order to satisfy some decisive requirements of the system (e.g. low latency, high data throughput, buffer size in the sources and destinations, etc.) while keeping a good bandwidth utilization efficiency.

Appendix 1. Barrel shifter traffic organization over a cross-bar switch.

A “barrel-shifter” organization can be used to organize the traffic with a cross bar switch, instead of polling for free connections. All the sources are correlated. At any one time, every source is connected with a destination. At regular time intervals, all the connections are changed, so that after N time slots, all possible connections have been established exactly once between the N sources and the N destinations (fig 4).

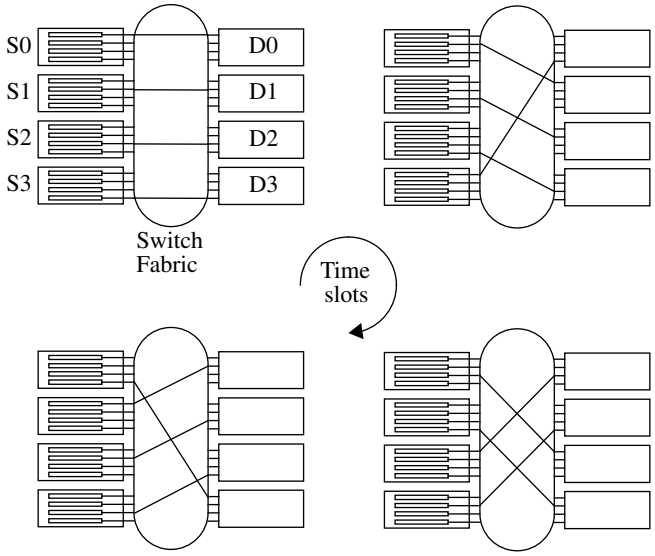

Figure 4. A 4 X 4 Barrel shifter

In the sources, a queue of event-fragments is maintained for every destination. Events k modulo N are stored in the kth queue. During 1 time slot, when the source is connected with the kth destination, as much data as possible from the kth queue is transmitted, irrespective of the event fragment boundaries.

The efficiency of bandwidth utilization is given by

$$\eta = T_t / (T_t + T_c)$$

where T_t is the effective transmission time during 1 time slot and T_c is the time to change a connection (time to establish a connection + time to break a connection).

If B_0 is the nominal bandwidth of 1 connection, the maximum usable bandwidth, B_{max}, is given by: $B_{max} = \eta \cdot B_0$. A high average bandwidth utilization is obtained by choosing $T_t >> T_c$, if possible.

During T_t, data is transferred at full bandwidth (B_0). In principle B_{max} is fully available for data transfer, but in practice one will set a safety margin to allow for variations in the trig-

5. Presentation of the results.

We are interested in the distributions of event-building latencies, buffer occupancies in the sources and destinations as function of various parameters like the trigger rate, the load on the switch, etc.

Those results can also be presented in the form of 'tail distributions' giving the probability distribution that a value of the parameter exceeds the value in abscissa. (e.g. probability that the latency is larger than a given latency, for all latency values). Tail distributions are helpful to estimate the quantity of a resources (e.g. memory) that would provide acceptable performance.

The percentage of nominal bandwidth effectively used for event-building is also an interesting parameter.

The collection of results must start only when the system has reached a steady state, although information on how the system reaches this steady state may be interesting in itself.

We are interested in the safety margin of the system: if it works well for some trigger rate, what are the trigger rates above at which it still operates well and what is the rate above which it becomes unstable or never stabilizes.

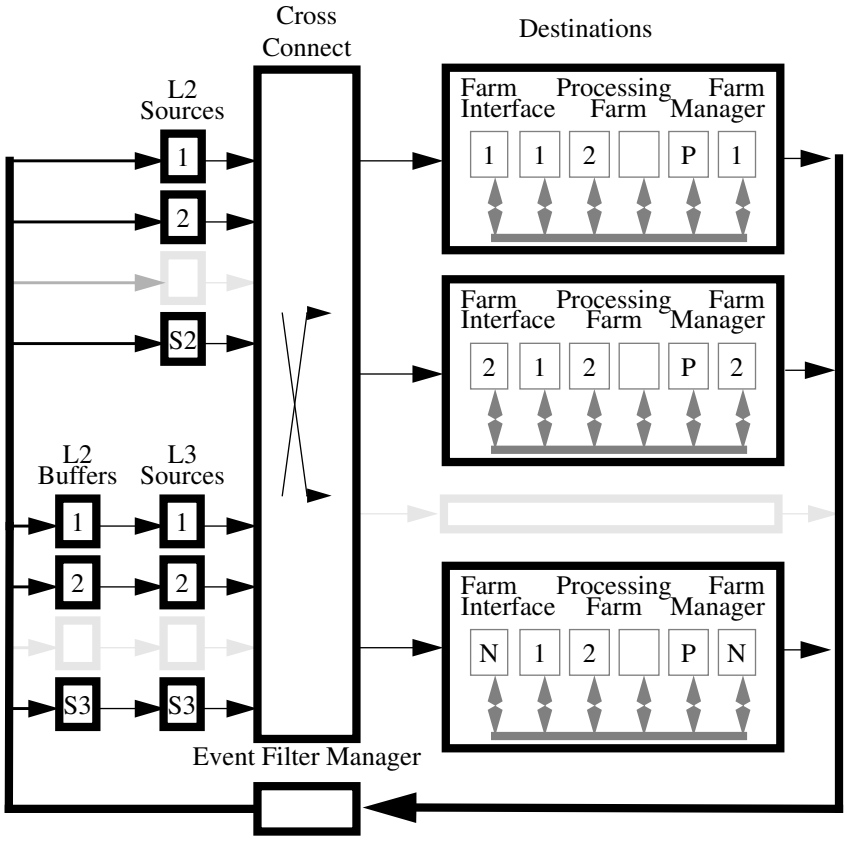

Figure 3. Combined Level 2 and Level 3 traffic

Number of sources:	S3 = 75% of N
Number of destinations:	N
Destination assignment:	The same destination that has processed Level 2 for this event.
Trigger rate:	10 kHz
Event fragment size:	Normal distribution (>0), average: 1000 bytes, sigma: 600 bytes

Results: Latency distribution for L2 and L3 event building.

Source buffers occupancies, destination buffers occupancies for the event-building part.

Level 2 only:

This is one part of the data acquisition system, where the traffic for Level 2 selection is considered. For every event, only a fraction of the sources have data to send (25%). The total size of the event has to be distributed (uniformly or by a normal distribution with mean value as given) over a selection of sources that varies randomly for every event. The total number of sources corresponds to a 'reasonable' value dictated by upstream read-out considerations.

A low event-building latency is an advantage, since the remaining of the data will be waiting for result from the Level 2 selection in the source buffer.

Trigger rate:	100 kHz
Event size:	7 KByte (to be distributed randomly over 25% of the sources)
Number of sources:	32
Number of destinations:	32
Result:	Event-building latency distribution

Combined Level 2 and Level 3 event-building:

This is a difficult case to simulate. It is given to suggest the level of complexity that is likely to be implemented in a real event-builder system.

Let N be the number of data sources. It is composed of 2 fixed sub-sets of respectively S2 and S3 sources (N = S2 + S3) (fig. 3). The S2 sources provide the data for Level 2 selection, all N destinations being available for Level 2 processing. The S3 remaining sources keep the rest of the data in the L2 buffers until a decision is taken. When an event is accepted by Level 2, the rest of the data for this event, residing in the L2 buffers, is transferred to the Level 3 sources and submitted for full event-building in the destination that had processed the Level 2 part. The data in L2 buffers belonging to rejected events are removed. Both traffics go on concurrently over the same switch.

	Total number of sources:	N = 1024 (may be modified, but values given below have to be scaled accordingly)
Level 2:		
	Number of sources:	S2 = 25% of N
	Number of destinations:	N
	Destination assignment:	Sequential.
	Trigger rate:	100 kHz
	Event fragment size:	Normal distribution (>0), average: 300 bytes, sigma: 200 bytes
	Processing time:	1 ms (constant)
Level 3:		

in the scenario'Level 2 only', the number of ports results from a detailed study of the read-out system [3] and should not be modified without good reasons.

Single level event-building:

We attempt here to select the events with one level only, namely Level 2 and Level 3 operations will be carried out in the same processor which will receive all the data from an event that has been selected by Level 1 (fig. 2).

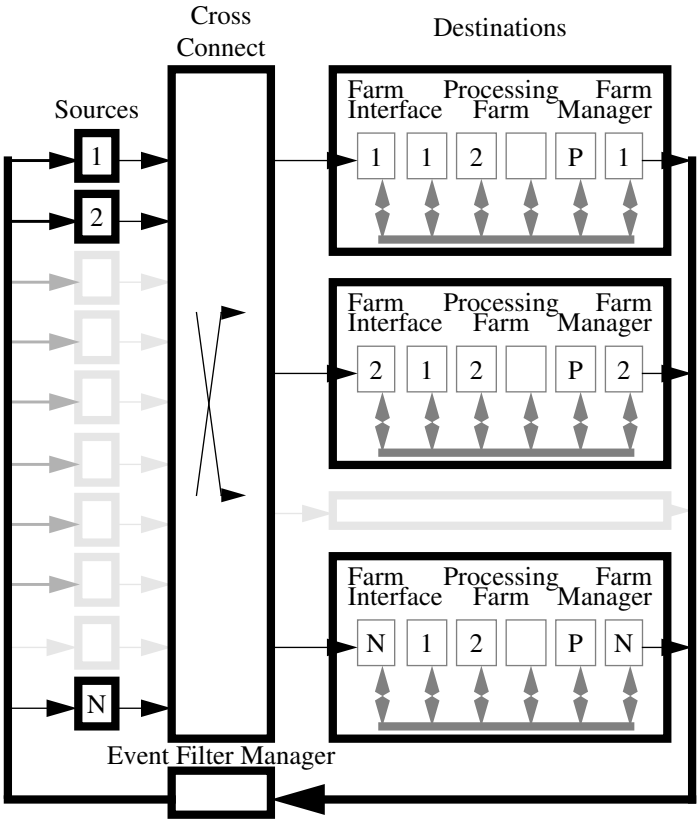

Figure 2. Single level event-builder

Total Event size:	1 MByte,
Event fragments:	normal distribution: average = 1 MByte/N sources, sigma = variable
Switch size:	128 x 128, 256 x 256... 1024 x 1024, different link speeds.
Results:	Average event-building latency as a function of trigger rate, how does it depend on sigma?
	What is the maximum trigger rate that one can reach (or the maximum load of the switch)?
	What would be the switch size necessary to sustain a rate of 100 KHz?

fication for a first approach in modelling is to assume that the identity of the sources is known by the destination, by some mechanism that needs not to be modelled.

Several assumptions are made in order to concentrate the study on the functioning of the switching network by eliminating all bottlenecks upstream or downstream of the event-building system which would influence the characteristics of the data stream across the switch. They are the following:

- The system that multiplexes the data and provides event-fragments in the sources has enough bandwidth and does not introduce overheads. When an event trigger occurs, all the data are available, without delay, in some shared memory in the source.
- The memory in the sources is assumed to be infinite (although the amount of memory actually used will be one of the useful results provided by the simulation).
- The memory in the destinations is assumed to be infinite (although the amount of memory actually used will be one of the useful results provided by the simulation). We distinguish between the memory where the events are in the phase of being assembled from the memory containing fully assembled events queueing for processing. If processing time does not play a role in the network traffic (examples 1 and 2), the latter buffer has not to be included in the model.
- The processing power is sufficient to analyse all the data arriving at any destination (queues should not grow for ever). However the processing time is finite and will constitute one of the important parameters of the models in some cases, but, like for the size of event fragment, very little is known about the shape of its distribution. In the models proposed here, either the processing time does not play a role (in the first 2 examples) or a constant value is assumed (3rd example).

3 types of event-building scenarios:

We can propose 3 types of event-building scenarios as follows:

- A simplified approach with a one level event-building, assuming that Level 2 and Level 3 operations are carried out in the same processor which receives all the data from an event.
- A Level 2 only event-builder where the data come from a subset of sources, different for every event.
- A decision scheme with 2 levels where the traffic for Level 2 and Level 3 are intermixed over the same switching network. The subset of sources delivering the data for Level 2 is fixed.

In all cases, values of rates and data fragment sizes are average values. In a first approach, uniform values (i.e. constant rate, same event fragment size in all sources) can be used, but it is necessary eventually to use distributions, namely exponential for inter trigger delay and normal distributions for event fragment sizes.

Depending on the available bandwidth, the number of sources can be adapted. However

The destination assignment logic does not need to be a sophisticated one. For simulation purposes, we can simply assign destinations sequentially. For the technology investigated so far (ATM), this scheme has proven to be practically as good as more complicated ones.

Levels 2 and 3 require event-building over a switching network. The traffic patterns differ in both cases. For Level 2, it is characterized by a high event trigger rate (100 kHz) but low amounts of data per event. Only a small subset of the sources will be involved (typically 25%). This subset can either be a fixed one (CMS) or different for every event (ATLAS). For Level 3, most of the sources will contribute with all the data for the event selected by Level 2. If possible, the same switching network could be used for both operations, thus mixing Level 2 and Level 3 traffics.

The number of sources and the number of destinations will probably be nearly the same in order to use efficiently the bandwidth of the switch. The actual number of sources is not fixed. It depends predominantly on the capabilities of the switching technology. However, if one can choose between several nominal bandwidths per port, a trade-off will sometime be made between the cost of the switching fabric, including the interfaces, on one hand and the cost of multiplexing the data in the sources and/or demultiplexing them in the destinations, on the other hand.

A very interesting parameter of an event-builder architecture is the *event-building latency* which is defined as the time interval between a trigger signal, which starts the event-building operation for an event, assuming that all the event-fragments are available in the sources, and the arrival of the last fragment of that event in the allocated destination. It may be significantly larger than the sum of the latencies of the individual data packets (event fragments) through the network, due to queueing and/or contention in the sources and, possibly, inside of the switch.

4. Modelling

The previous section gave a general overview of the event-building concept. The choice of an architecture to implement an event-builder depends on the requirements of a particular experiment and on the capabilities of the switching technology. Very detailed simulations are required to decide whether a particular technology is suited. However a first approach with simplified assumptions is useful in order to identify the main features of possible architectures. The aim of this note is to provide simplified assumptions about various event-building requirements in order to carry out a preliminary exploration of the Fiber Channel technology as a possible switching network technology.

The different sources are assumed to provide event fragments with the same distribution of size, although in real systems the load will certainly not be uniformly distributed.

In many cases, only a reduced number of sources will have data to send. As an example, the Level 2 selection is done with a reduced set of the full event's data (e.g. 25%) that is provided by a small sub-set of the sources. One problem in this case is, at the destination, to determine when the event-building is terminated. The real systems will implement some protocol, depending on the characteristics of the data distribution and on the requirements of the system (e.g. short latencies). Possible protocols include the build-up of a data distribution mask, or the minimal notification by each source, or a time-out based detection in the destination. A simpli-

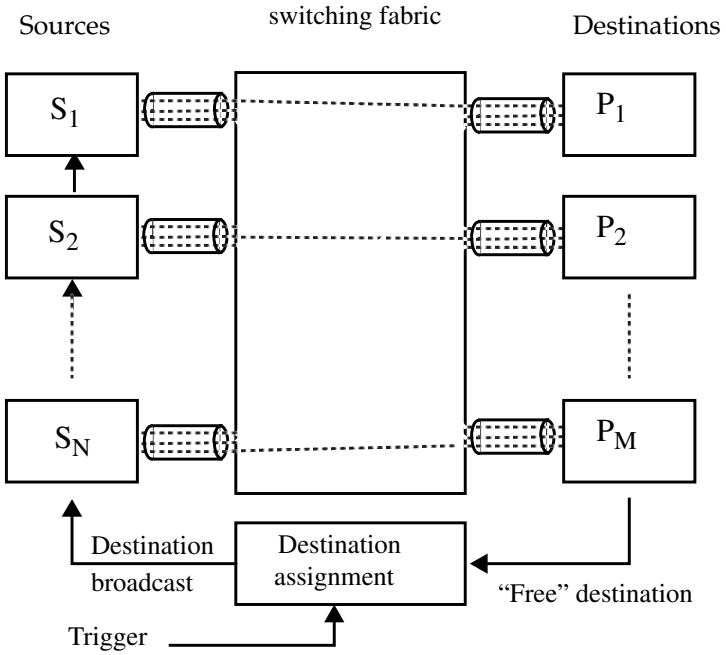

Figure 1. A generic event-builder

Equivalently, the inter trigger delay follows a negative exponential distribution.

The event-building traffic over a switching network has the very peculiar characteristic to require a concentration of the traffic, for a short time, from many (possibly all) sources towards one destination. Some switching technologies are not able to sustain this type of traffic without an external traffic organization. This is the case, for instance, when it is not possible to establish more than 1 simultaneous connections towards the same destination or if the mechanisms to absorb traffic burstiness in the switch are not sufficient. Nevertheless those characteristics are well predictable and stable, e.g. there are no such things as new users requesting connection in a random fashion. It is then relatively easy to implement simple traffic shaping schemes like a barrel shifter (See appendix 1).

The event size, the event building latency and the processing time being all variable, the source and destinations must include buffers large enough to store several events. The modelling of the system will help to determine a safe value for their dimension.

Every source includes a buffer to hold the data that are either queueing for network access or waiting for a Level 2 decision. The data structure in this buffer depends on the switch technology and on the event-builder architecture. As examples, it can be organized as a simple FIFO, or, to avoid head of line blocking, as separate queues, one per destination.

Each destination can include a single processor or a farm of processors, depending on the computing power required to execute the selection algorithm and on the effective bandwidth delivered by the switching network. Destinations also include a buffer sufficient to hold several events: 1 or more events being assembled, the event being processed and, possibly, several events completely assembled and waiting for processing.

The selection aims at enriching the sample in events that present an interest for the particular research done by the experiment. It cannot be 100% efficient due to the selection criteria that must remain very simple in order to be applied in a very short time. In order to maximize the efficiency, the selection is usually split in several levels, with a complexity increasing at each level. The scheme usually considered in the experiments for which we are developing data acquisition systems includes 3 levels of selection briefly described below. The rates correspond to the 2 planned LHC experiments, ATLAS [1] and CMS [2]. With the exception of the initial rate of 40 MHz which is fixed. the rates at the next levels are average values of Poisson distributed event triggers:

Level 1: performed by hardware, relies on local properties of the data (energy deposition, clusters of data). Input rate: 40 Mhz (constant)

Level 2: relies on larger size local features and combines interesting features from several regions. Input rate: 100 kHz (average) (actual value will be lower, but the system must be designed to sustain this rate)

Level 3: Has access to the totality of data of the event. Input rate: 1-10 kHz. (average)

Each selection level eliminates a large fraction of events that are discarded from the buffers. Levels 2 and 3 are performed by processors and necessitate that a part (level 2) or all (level 3) of the data are transferred from local buffers into a processor. A first data multiplexing is performed, after the level 1 decision, by the local data acquisition electronics, in order to reduce considerably the number of sources from which data will be collected for analysis.

3. Event Building

Event building is the operation of collecting a part or the whole of the data of 1 event into 1 processor. This operation is repeated for every event, selecting another destination each time. The general principle is shown in fig. 1. For every event trigger, some destination assignment logic generates and broadcasts the information used by the sources to select the appropriate connection over which they send the partial information on the event (an *event fragment*) that they store. At the destination, the event fragments are linked together and, once completely assembled, the event is processed.

The size of event fragments is variable both between the sources and, for a given source, between the different events. Information on the distributions of the event fragment sizes is only partially available. It depends on the physics and on the architecture of the detector read-out system. Provisionally we make some simple assumptions, like normal distributions, although we can expect that the real distributions will not be so simple, and one should keep in mind that large differences in event fragment sizes will occur, due to local concentrations of energy depositions. The inter-trigger delay is also variable, since the selection of events by the previous level(s) is a stochastic process. One can assume that the triggers follow a Poisson distribution with the average provided by the currently estimated rates which are well agreed.

RD-31 Note 95-02

January 1995

Problem statement for Fiber Channel event-builder modelling.

J.-P. Dufey

(dufey@sunvlsi.cern.ch)

1. Introduction

Fiber Channel is one of the switching technologies that is considered for the implementation of high rates data acquisition systems in High Energy (or Particle) Physics (HEP). The very specific characteristics of the data traffic requires extensive modelling studies in order to evaluate the feasibility of Fiber Channel based "event-builders" and to find which system architectures are best suited for various classes of data traffics.

This note is intended for specialists of Fiber Channel Technology, not familiar with HEP, who are interested to evaluate how the technology could be applied in this specific field. It gives an elementary description of the problem and proposes several simplified architectures of data acquisition systems suitable for a first modelling research.

2. Data selection and data acquisition in High Energy Physics experiments

In High Energy Physics (HEP) experiments, we study collisions of elementary particles which produce a certain number of secondary particles (charged and neutral, short or long lived) some of which are detected by a set of large size detectors of various types. The result of a collision is a collection of digitized values stored in many local buffers scattered over the detectors. In the HEP jargon, this set of data resulting from 1 collision is called an *event*.

The number of collisions is huge (millions per second) and the amount of data per event can also be large (megabytes). Obviously it is not possible to store all this data in permanent storage for later analysis. On the other hand, not all collisions lead to interesting phenomena. Therefore the experiments use various methods to reduce the amount of data, in order to keep a much reduced set of events that they are able to store on permanent media.