A second level data buffer with LHC performance

B.J. Green and J.A. Strong

Department of Physics, Royal Holloway, University of London

R. Cranfield and G. Crone

Department of Physics and Astronomy, University College, London

Abstract

For the ATLAS experiment at the Large Hadron Collider (LHC) buffers are required which will accept data at rates in excess of 100 MBytes/s and distribute selected data to the second and third level trigger systems. A design for a buffer using a digital signal processor as the memory manager is described and the performance of a prototype module is given. Tests of the system in a beam line will be used to develop and check discrete event simulation models of the buffers and a second level trigger system.

1. Introduction

For the ATLAS experiment [1] at the Large Hadron Collider (LHC) it is proposed to use a three stage trigger and data acquisition system where data will be moved to buffers off the detector after the first level trigger. The latency of the second level trigger determines the capacity of these buffers which must hold the data during the level two decision time. For a first level trigger rate of up to 100 kHz and a data volume to be moved off the detector of one to two MBytes for each trigger, the total data rate into the buffers is one to two hundred GBytes per second.

It is assumed that digitised data from the front-end electronics arrives at the second level (T2) buffers on links operating at around 1 Gbit/s so that up to two thousand buffers are required. For programmable processors to be considered at the second trigger level, latencies of several milliseconds must be contemplated. Approximately five hundred events can be stored in a buffer with 0.5 MBytes of memory corresponding to a mean second level latency of 5 ms.

To avoid problems associated with moving all the data to processors at the second level, it is proposed to use regions of interest (RoI) found at the first trigger level to guide the data processing at the second level [2] and thus reduce the data rate in the second level processing networks [3]. A preliminary investigation using a simplified monte carlo [4] indicates that the average number of regions of interest for each first level trigger is small (<5) and that the number of buffers containing information for the RoIs for that event is a small

Figure 1. Second level trigger local / global architecture showing key position of Buffers

fraction (~5%) of the total. Figure 1 shows a second level trigger local / global processing architecture and the key position of the buffers in this architecture.

A buffer containing information for second level processing for a particular event will receive a message from the first level trigger instructing it to send relevant data to the second level processor assigned to process that RoI for that event. Using the event identifier and RoI information in the message, the buffer must select the data and push it to the local processor. The second level decision for every event is sent to all buffers and the buffer must free the space in memory if the event is rejected (99%) or push the data to the event building system and then free the memory space if the event is accepted (1%). Table 1 shows the trigger and data rates expected for the inputs and outputs to the buffers.

	Event rate	Data rate	Comment
Inputs			
Data	100 kHz	100 MBytes/s	e.g. Fiber channel
L1 RoI	5 kHz	100 kBytes/s	Note that high transfer rates are needed to maintain low latency
L2 Decision	100 kHz	1 MBytes/s	
Outputs			
L2 system	5 kHz	5 MBytes/s	
Event builder and L3	1 kHz	1 MBytes/s	

Table 1. Trigger and data rates expected at the second level data (T2) buffers in ATLAS

2. Second level buffer hardware

The basic element of the second level data buffer is a multi-port memory. Data is written into the memory from sub-detectors and read out to the second and third level triggers systems. In this design we have separated the data paths for level 2 and level 3 off the board to avoid introducing additional latency into the second level trigger from level 3 data transfers. On board, some increase in latency results from the fact that the memory accesses are sequential. To limit the size of the derandomising buffers on the sub-detectors, we have assumed that a push architecture is used i.e. the T2 Buffer must accept data presented to it and not request it from the sub-detector. This requires a small buffer between the input and the memory and prompt access to the memory. From Table 1 it can be seen that the total data rate out of the buffer is an order of magnitude lower than the input rate and, therefore, we have opted for a design which uses a hardware engine to control writing data into the memory with data reads interleaved under software control.

Figure 2. T2 Buffer basic design

The basic T2 buffer design is shown in Figure 2. Because of the early stage of the experimental design, no choice has been made on data transfer protocols from the sub-detectors and the current design accepts a 32-bit parallel word with a data valid signal and has a programmable input protocol which can be adapted to match sub-detectors requirements. An input FIFO allows the input data to be synchronised to its own clock. The 2 kByte input FIFO on the prototype board can operate at up to 50 MHz giving a peak input rate of 200 MBytes/s.

Figure 3. T2 Buffer detailed design

The 512 kByte memory is page managed with a nominal page size of 1 kByte corresponding to a transfer period of 10 μ s at 100 MBytes/s. A new event starts a new page and an event can occupy more than one page. The page address is taken from the list held in the free page FIFO and the memory address is incremented until either the end of the page is reached when the next page is allocated or the end of the event is reached when the current address is written into the used page FIFO. The used page FIFO contains a list of used pages, a last page of event marker and the end of event address. A more detailed diagram of the circuit is shown in Figure 3.

A page managed memory was chosen as the best way to deal with the problems of zero suppressed data resulting in variable length data blocks and of variable second level trigger decision order from an asynchronous processing system. A circular buffer with an event start address list accommodates zero suppressed data but has a maximum latency equal to the time to fill the buffer. Setting a page size where the majority of events are contained in a single page gives a good compromise between efficient use of the memory and limited buffer manager actions.

A Texas Instruments TMS320C40 digital signal processor is used as the buffer manager as its six communications ports can be used for all I/O transactions apart from the main data input. In the current design, memory reads are under the control of the manager which can access the memory at a maximum of one in four clock cycles. For the 40 MHz devices currently used this corresponds to a read rate of up to 40 MBytes/s and a memory write rate of at least 120 MBytes/s. The communications ports are capable of operating at up to 16MBytes/s within a total rate of 40 MBytes/s. In the prototype, other hardware limits the clock frequency to 30 MHz. The current design has all the hardware associated with the buffer on a board which can plug into a commercial board holding the buffer manager DSP. The commercial board, an LSI DBV44 [5], can hold up to four C40s sited on TIM40 modules [6]. One of the TIM modules is replaced by the buffer board which then connects to the global bus of the C40 on another TIM site.

3. Second level buffer software

The main functions of the buffer manager software are (see Figure 4):

- 1) page-fifo servicing and event indexing
- 2) region-of-interest record handling and transmission of RoI data
- 3) global second level trigger decision (GT2) handling and transmission of DAQ data

Event indexing and global decision handling must be performed at an average rate of 100 kHz; the average rate for the other activities will in general be much lower (see Table 1).

Since it is unlikely that anything would be gained by buffer manager functions interrupting

Figure 4. Buffer manager software context

one other, we have avoided the complexity and overhead of interrupts altogether, using instead a simple loop in which input sources and output queues are polled in turn and handled immediately if they require servicing.

The software is implemented on the C40 which accesses the T2 buffer hardware over the DSP's global bus and receives messages from the RoI and GT2 sources over DSP links. In the current prototype, Level 2 and DAQ data are also transferred over links, and error and status messages are sent over a link to a VME based processor via the VME link-interface-adaptor on the LSI DBV44 mother-board. Code has been written in ANSI C for portability and all the current code and variables fit into the 8 kByte on-chip RAM with only the index table held in external memory.

Event data is split into pages by the T2 hardware. It is the job of the buffer manager to keep track of these pages, passing free addresses to the hardware free-page fifo and reading used addresses from the hardware used-page fifo. The buffer manager reads the event id from the first page of an event and uses this to index a table in which it stores all the page addresses for the event; when an event is eventually released on reception of a global trigger decision the page addresses for that event are looked up and put back into the free-page fifo.

The current indexing algorithm simply uses the least significant part of the event id as an offset into the index table array. This assumes that event id's arrive approximately in sequence and that the distribution of global latencies does not have too long a tail. More complex algorithms have been devised in case these assumptions prove too severe.

Timing tests have been made with a software emulation of the buffer hardware in which a second DSP reads and writes pages via a link acting in place of the buffer hardware FIFOs. The measurements shown in Table 2 indicate an overall event rate of around 75% of the LHC goal of 100 kHz. DMA transfer of data should introduce minimal overhead since the DMA is between the global bus and a DSP link and the global bus itself is used minimally for access to the FIFOs. As these results are for modular code written in (optimised) C, without any attempt to hand-tailor in assembler, and as we are running on the lower speed (40 MHz) C40 chip, we are optimistic that we can achieve the projected LHC rate with current DSP technology.

Operation	Time or frequency range
Insert a 1-page event into index	88 clocks (30 instruction fetches)
Release a 1-page event	61 clocks (19 instruction fetches)
Loop overhead	108 clocks (55 instruction fetches)
total time for high frequency subtasks	257 clocks
Total event rate with 20 MHz instructions	77.5 kHz

Table 2. Timing measurements for software operations of the T2 Buffer manager

4. Beam Tests

Figure 5. Beam test lay-out

Two of the prototype buffers will be used in a test beam at CERN in 1994 as part of a minimal test of a local / global architecture. The link unit and feature extractors will be C40 DSPs on LSI DBV44 boards. The global network will consist of a small SCI ring and Alpha single board computers. The results obtained will allow comparison with modelling studies and the development of modelling techniques. Figure 5 shows the beam layout.

5. Conclusions

A prototype T2 buffer with performance close to that required at the LHC has been built and will be tested in 1994. The software currently runs at 75% of the LHC requirement but it is confidently expected that we will see substantial improvements when speed critical parts of the code, currently in C, are replaced by assembler code.

6. References

- [1]. ATLAS Letter of Intent. CERN/LHCC 92-4
- [2]. D. Crosetto et al.: A local/global architecture for level 2 calorimeter triggers. LHC Workshop, Aachen, 1990
- [3] J.A. Strong. Local processing for a farm-based second level trigger at LHC. Proc. Conf on Computing for High Energy Physics, San Francisco, 1994
- [4]. D. Johnson and J. Strong.: A study of a calorimeter based trigger system using single electron and two jet events. ATLAS CAL-NO-030 October 1993
- [5] Loughborough Sound Images Ltd. Leics. UK.
- [6] TIM40 spec., Texas Instruments, Bedford, UK