Demonstrator for a High Performance Second Level Trigger

H.Bertelsen, M.Dam, J.D.Hansen, J.R.Hansen *and* B.Rensch

Niels Bohr Institute, DK-2100 Copenhagen, Denmark

We developed and evaluated a scenario for the implementation of the Second Level Trigger in the HERA-B experiment at DESY. The architecture is characterized by a global switching network, based on the IEEE 1355 technology, connecting the SHARC based Second Level Buffer to a large farm of commodity processors. Data flow and resource allocation is controlled by a single EventController. Events are processed in parallel leading to an automatic load balancing at the processor level. Our results show that the system can cope with trigger rates in excess of the maximum expected ones.

Key words: IEEE 1355, Data Acquisition, Triggering;

1 Introduction

HERA-B[1] is an experiment designed primarily to search for CP violation in decays of B mesons into the "gold plated" decay mode $B \rightarrow J/\Psi K^0_S$. The B mesons are produced in interactions of 820 GeV protons in the HERA proton beam with an internal wire target in the beam halo. A four level trigger scheme (Fig. 1) has to reduce the primary interaction rate by almost seven orders of magnitude to a level suitable for recording on permanent storage. The Second Level Trigger(SLT) operates only on Regions of Interest(RoI), a small fraction of the total data that is stored in the Second Level Buffer(SLB) during SLT processing. The SLB is implemented with 134 DSP boards each carrying six SHARC processors[2] and is able to hold 384 events. The RoIs are determined by the FLT and are updated (refined) after each SLT processing step. The SLT has to be able to digest at least 50 kHz and should be easily upgradeable to cope with the maximum possible input rate of 97 kHz.

Preprint submitted to Elsevier Preprint *28 April 1997*

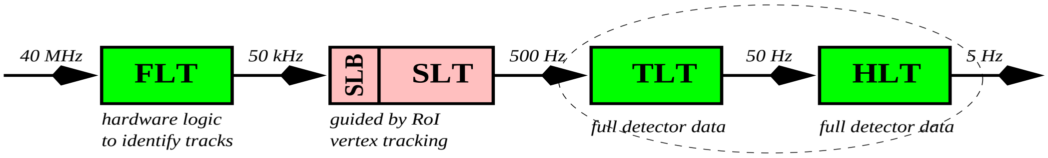

Fig. 1. The HERA-B four level trigger-scheme.

2 SLT Architecture and Functionality

The architectural view of the SLT is shown in Fig. 2. A global switching network links the Second Level Processor(SLP) farm to the SLB and a single Event Controller(EvC). A single SLP processor executes all triggering algorithms and makes the final decision for a particular event. At an FLT accept, the detector data is moved from the detector front-end to the SLB, where it is stored during SLT processing in a globally managed buffer. The FLT decision record, which contains a list of identified RoIs, is sent to the EvC. The EvC assigns the event to an SLP by passing the RoI list to this unit. After having identified – for each RoI – the SLB boards which could contain hit-data, data requests are passed from the SLP to these boards. The SLB responds by sending the hit-data inside RoIs back to the SLP for processing. The SLT processing is continued through multiple steps (*see* Fig. 3) until a decision is reached. For each step additional data is pulled from the SLB. The SLT decision is communicated to the EvC, which adds the SLP to its list of free processors. In case of a SLT reject, the SLB is informed to release the event buffer. In case of a SLT accept, the SLB is instructed to push the full (sparsified) event data to the Third Level Trigger farm(TLT). EvC commands to the SLB system are broadcast via a dedicated and implicitly acknowledging

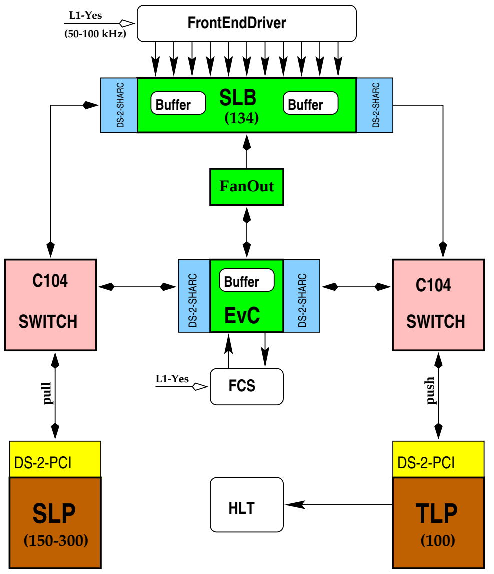

Fig. 2. Architectural view of the SLT

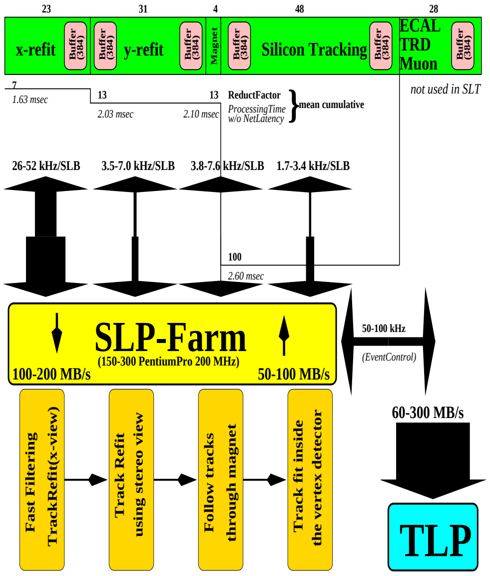

Fig. 3. Modelling asumptions

FanOut. After assuring that all data is safely out of the SLB, the corresponding event buffer will be released. Via a throttle mechanism to the FLT, the EvC is able to control that the SLB is not overrun by data.

In Fig. 3 we present the set of assumptions made for the current design study. The reduction factors and algorithm processing times (excluding the network latencies) were derived from a small sample of simulated data within the sequencial selection strategy. Combined with the architectural choice the resulting traffic pattern can now be predicted. Most importantly, the proposed global switch has to link at least 450 nodes to each other and has to support at least 300 MB/s aggregate throughput at low latencies.

3 Technological Choice

We have chosen the by now well proven IEEE 1355 technology[6] for the implementation of the global switching network. A 512 terminal Clos[5] (or indirect multistage, Fig. 4) adaptive routing network built from STC104 asynchronous packet switches[7], that link 32 bidirectional DS link ports via a 32×32 non-blocking crossbar switch, connects the 134 SLB nodes to about 150 (300 for the high rate option) processors, leaving plenty of free connections to add more processors or to multiply the bandwidth to critical nodes.

To adapt the SLB system to the switch, we built a SHARC-to-DS interface (*see* Fig. 7) which handles the SHARC-link protocol by an FPGA that is connected to an INMOS T9000 Transputer[3] via two FIFOs, one for each direction, and allows for a bidirectional message transfer rate of up to 45 kHz.

A commercially available DS-to-PCI interface[8] has been chosen to hook the SLP processors to the switch. The T9000 Transputer on the interface communicates with the EvC via the global switch, pulls the relevant data from the

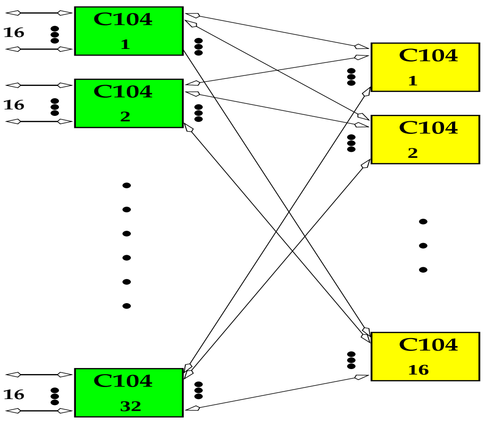

Fig. 4. 512-way folded Clos network

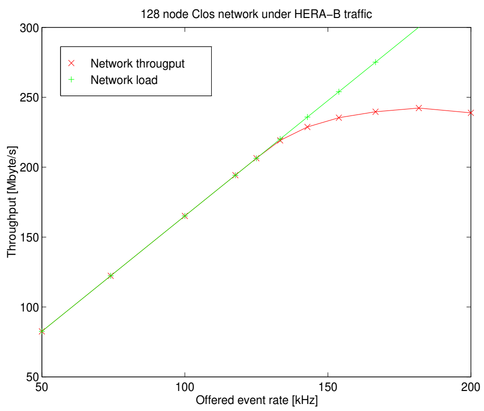

Fig. 5. MACRAMÉ results

SLB system and uses the attached processor, a 200 MHz PentiumPro running LINUX, as a coprocessor to take a decision on the event. All real time tasks are performed by the T9000 to keep packet transmission overheads at a minimum. In addition, the T9000 can assemble a second event while the previous one is still processing in the Pentium to significantly reduce the average network latencies.

We emphasize that all traffic in to and out from the switch is handled by T9000 Transputers, an important fact since the T9000 imposes a higher level protocol (restricting the payload of a packet to a maximum of 32 bytes), supports the concept of *virtual channels* and includes efficient multiplexing hardware, allowing a number of virtual channels to use the same physical link transparently. The mapping of the channels onto physical links is carried out by configuration tools[4].

4 Results

To demonstrate that the proposed configuration is capable to stand the expected event rates, we factorized the system into independent components as indicated in Fig. 2, namely the EvC, the global switch and the interfaces (DS-to-PCI and DS-to-SHARC).

The Event Controller(EvC) that supervises both the SLP and TLP farm has been implemented on a single SHARC-board in a completely data driven way(Fig. 6). The EvC has to detect the presence of a new event, get an SLP to process the event, await the decision, instruct the SLB to push the data into a selected TLP, await the acknowledge from the TLP and free the occupied buffer locally and in the SLB. Any congestion or fault will result in a throttle request to the FLT. A monitor on the EvC watches the flow of control to locate faults and to report and suggest possible actions to Run Control. Each

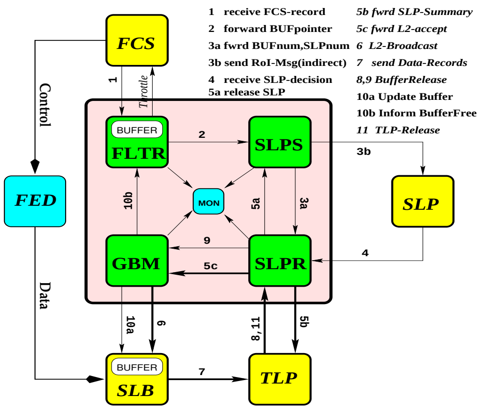

Fig. 6. Data Driven EventController

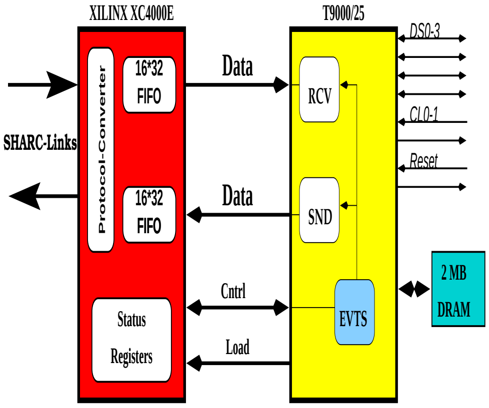

Fig. 7. SHARC-to-DS interface

event in the SLT is uniquely identified by the buffer number it occupies until it has left the domain of the SLT. Measurements in a compactified environment (i.e. no switch and minimal processing times in SLP and TLP) showed that the EvC can handle trigger rates in excess of 130 kHz.

The behaviour of the proposed switching network has been studied in detail for different traffic patterns within the MACRAMÉ[9] project. The scaling properties have been used to extrapolate to a 512-node switch. We made the conservative assumption of a random traffic pattern and concluded that the required low network latencies can be obtained if the average load per DS link is kept below 4.5 MB/s, resulting in a lightly loaded network ($\approx$10%). The most demanding and critical parts in our design (*x-refit* and *EventControl* in Fig. 3) have been fully emulated on an existing 128-node network using the expected HERA-B traffic pattern. The result in Fig. 5 shows that the network saturates only at 125 kHz. The measured throughput and latencies are in full accordance with the expectations.

A paper model of the chosen architecture has been developed. It uses average values for data volumes and algorithm execution times, as well as measured and estimated processing and data transfer overheads. The model is used to evaluate system resource requirements and trigger decision latencies and to match the capabilities of the various subsystems to each other. We found that we need four DS links on two interfaces(Fig. 7) to connect the EvC to the switch and that the *x-refit* data have to be distributed over twice as many SLB nodes as foreseen(Fig. 3). On average, only 15% of the total execution time is spent in the network. At 50 kHz input rate 152 SLP processors are sufficient and on average 172 buffers are occupied (of which 8 are blocked by the buffer management and another 12 by TLP event building), leaving 212 for short term upward fluctuations. The corresponding numbers for the maximum input rate of 97 kHz (where no upward fluctuations are possible) are 304 processors and 362 occupied buffers.

5 Summary and Conclusion

We have developed a scheme to implement the Second Level Trigger for the HERA-B experiment. The architecture is based on a global switching network and a large farm of commodity processors, supervised by a single centralized event controller. We have chosen the by now well proven IEEE 1355 technology to implement the switching network and to interface it to the Second Level Buffers and to the processor farm. We have shown that this simple and scalable system can handle the highest possible trigger rates.

References

[1] HERA-B Design Report, DESY–PRC 95/01, January 1995.

[2] ADSP-21060, `http://www.analog.com/products/sheets/ADSP21060.html`

[3] *The T9000 Transputer Hardware Reference Manual*, INMOS Ltd., INMOS document number 72 TRN 23801.

[4] *T9000 ANSI C Toolset User Guide*, INMOS Ltd., INMOS document number 72 TRN TDS 40801.

[5] C. Clos, *A Study of Non-blocking Switching Networks*, Bell Systems Technical Journal 32, 1953.

[6] IEEE 1355 (ISO/IEC 14575), *Standard for Heterogenous Inter-Connect(HIC), Low Cost Low Latency Scalable Serial Interconnect for Parallel System Construction*, IEEE Inc., 1995.

[7] *STC104 Asynchronous Packet Switch, Data Sheet*, April 1995. SGS THOMSON Microelectronics.
Networks, Routers and Transputers: Function, Performance and Application. IOS Press 1993, ISBN 90-5199-129-0, ISSN 0925-4986.
R.W. Dobinson et al., *Triggering and Event Building Results using the C104 Packet Routing Chip*, CERN–ECP 95/20, November 1995.

[8] SN9204/PCI(PCI-to-DS Interface) and TA9108(TransAlpha) `http://www.parsys.com/`

[9] R.W. Dobinson et al., *Realization of a 1000-node High-speed Packet Switching Network*, CERN–ECP/95–16.
R. Heeley et al., *Results from the Macramé 1024 Node Switching Network*, These Proceedings.