LAPP-EXP-96.14
NOVEMBER 1996

THE L3 SECOND LEVEL TRIGGER IMPLEMENTED FOR LEP-II WITH THE ST T9000 TRANSPUTER AND THE ST C104 ASYNCHRONOUS PACKET SWITCH FROM SGS-THOMSON.

J.J.BLAISING, F.CHOLLET-LE FLOUR, J.C.CRUZ, G.DAGUIN,
A.DEGRE, A.MASSEROT, G.PERROT
LAPP / IN2P3, BP110 74941 Annecy-le-vieux Cedex, France and
Université de Savoie
Email: Degre@lapp.in2p3.fr

B.MARTIN and M.ZHU
CERN, 1211 Geneva 23, Switzerland

X.CAI
L.N.S. M.I.T. Cambridge MA 02139, USA

ABSTRACT

A networked second level trigger has been installed in the L3 experiment at CERN. Made of 29 ST T9000 interconnected via 2 ST C104, it is embedded in the data acquisition since July 1995. The hardware and software implementation is briefly described. Event building performances have been measured in the real data taking environment under different hardware and software configurations. Parametrization in terms of maximum speed and overhead per data block quantifies the measurements and demonstrates that the network works as expected in the different configurations. No deadlock or slowdown versus rate has been observed. A maximum speed of 6 Mbyte/s per link, and a minimal overhead of 1μs per data block have been measured.

1. Introduction

The L 3 data acquisition system[1] is structured in three levels. The hardwired first level trigger makes a decision between two LEP bunch crossings (22 μs). The second level trigger, based on fast programmable processors, memorizes and builds-up the trigger data block, processes the data with a dedicated algorithm according to the level-1 trigger pattern and makes a decision (accept or reject) in a few milliseconds. The central event builder collects the detector data together with the trigger data and send it to the third level trigger. Based on a workstation network level-3 makes its decision in a few hundred milliseconds.

The initial second level trigger[2], implemented around XOP processors and dedicated fast dual ports memories with simultaneous access, ensured the LEP data taking between 1989 and 1994 with very good performances and reliability. But its hardware and software maintenance required experts who were no longer available. Driven by a microcoded instruction set, it had to be coded in assembler and didn't provide the desired flexibility to face the LEP phase II program. Consequently we decided to upgrade the system for LEP II with the following guidelines:

Presented at conference DAQ96 - OSAKA - 13/15 November 1996

1) The minimization of the hardware and software maintenance, by using commercial components whenever possible. 2) The implementation of a built-in test facility providing the playback of real events under standard data taking conditions. 3) Farm processors have to run applications coded in high level language and algorithms coded in Fortran. 4) Scalability (implementation of additional input ports or farm extension) has to be easy and controlled by software parameters. 5) The second level trigger total timing should not exceed 10 milliseconds per event. 6) Priority has been given to latency, as throughput requirement is rather low and a fast decision is required by some detector.

In this application the most sensitive function is event building. It has to assemble 76 blocks of data having a size from 4 to 400 bytes. Some blocks have a variable length from one event to the next one. The mean trigger total block length is 5 Kbytes, i.e 66 bytes per block. For a 2 milliseconds target time, of what the allocated event building time should not exceed 26 μs per block. A classical implementation with time shared bus access controlled by a monoprocessor cannot easily provide the requested performance essentially because of the software overhead required to initialize the data block readout. This target time is supposed to be more realistic with a networked event builder. That was the basic motivation of our interest for networked systems.

In 1992 the most promizing solution was the new transputer technology just announced: a network of ST T9000 transputers[3] interconnected by the asynchronous packet switch[4] ST C104. The ST T9000 transputer is a RISC processor with 4 serial DS link and a virtual channel processor implemented in silicone. This SGS-Thomson technology provides a global and integrated solution to the L3 second level trigger. There is no interface to develop, no traffic control software to implement and consequently no unnecessary lost of performances. A C-toolset developed by the manufacturer provides a C language compilation chain and the network configuration tools. Transputer communication is implemented in one line of C language code. The network communication traffic based on the DS link engine and its token-level flow control is managed by hardware.

2. System implementation.

The implementation has been described in details in reference 5. An input memory, made of 48 input parallel ports implemented on 12 Fastbus boards, named TMB, collects the data delivered by all trigger digitizers at each LEP crossing. Data are transmitted to the input ports FIFOs through 50 m long cables at a maximum speed of 60 ns per 16 bit word in ECLine standard. Its 1.6 Gbytes/s bandwidth allows the trigger data collection in less than 10 μs. The first level trigger decision made before the next crossing is sent to the second level. On a negative decision all FIFOs are cleared and ready to collect the next crossing. On a positive decision, a high priority process implemented on each TMB ST T9000 transputer starts the data transfer from the FIFOs to a local multi events RAM buffer via the transputer's data bus. This transfer, completed in less than 150 μs, is performed during the L3 data acquisition deadtime. It doesn't introduce any additional deadtime.

Event building, algorithms processing and transfer to the output memory are sequentially performed by a single ST T9000 transputer of the processing farm. To give priority to latency, the farm is implemented in pull mode: when idle, a processor sends a request to the event server. Processors are served according to their chronological request.

The event building is performed through a two ST C104 network. It collects and orders consecutively, according to a predefined output format, the 76 blocks of trigger data distributed over the 48 input ports memorized in the 12 multi events buffer RAM. Note that because some input ports collect data sent by several subdetectors, data blocks collected by ports have to be reordered. This ordering is included in the event building. Consequently the number of blocks exceeds the number of ports. By saving a further reordering, this option saves a non negligible CPU time.

Some blocks being variable in length from one event to the next one, the event building proceeds in two steps. All wordcounts are transferred first. The event builder prepares the output format and reserves the appropriate space. In such a way, the relative block's reception order doesn't affect the output format. All wordcounts and data blocks are sent in parallel without traffic shaping.

3. Results

The programmable built-in test facility allows an event injection in real data taking conditions at any required rate up to saturation. The basic measurement set is the throughput rate measured for different combinations of the number of bytes per block (2 to 4096) and of the number of blocks (12 to 84), all blocks having the same size. A global 3-parameter fit on the measurement set provides the maximum transfer speed, the overhead per block, and the software overhead per event.

Throughput has been measured in the following « standard » conditions: Only one element in the farm (destination) is active. It is connected to the C104 switch via two DS links which are driven to saturation by multiple virtual links. Data are stored in a non cachable 100 ns access time 32 bits SRAM and code is connected to a 150ns access time cachable 32 bits DRAM. The event builder collects blocks of fixed length. Under these conditions event building is performed with a 12.2 Mbytes/s bandwidth, i.e 6.1 Mbytes/s bandwidth per DS link. This last number has to be compared with the 7 Mbytes/s reported in internal memory transfer[6]. For small size blocks, the transfer speed cannot reach the maximum bandwidth and the bandwidth loss is inversely proportional to the block size. This effect increases with the number of blocks. The full bandwidth is reached for blocks larger than 1 Kbytes. We measure a 1 μs overhead per block, and a 35 μs software overhead per event.

To check the performance sensitivity to the hardware and software implementation, we have repeated the same measurements after modification of one parameter at a time respectively: the number of links connecting the farm processors to the switch, the data memory allocation, and the high level event building protocol (fixed or variable length blocks).

The same measurement with a 4 DS links interconnection (instead of 2) between the switch and the processing unit gives a maximum throughput of 16.3 Mbytes/s instead of the expected 24 Mbytes/s. Overhead per block and per event are not modified. It is a clear confirmation[6] that the 20 Mhz ST T9000 cannot drive 4 links at full speed.

Another measurement in standard conditions but with code and data both connected to the same 16 Kbytes cache DRAM has been done. Cache effect is clearly seen. For event length smaller than 6 Kbytes, speed increases up to 15 Mbytes/s, and suddenly decreases to an asymptotic speed between 8 and 9 Mbytes/s for larger length. This value is induced by the 150 ns memory access time. Overhead per block and per event are identical to standard conditions.

Finally we have quantified the importance of the high level event building protocol. Fixed length blocks event building has been performed with a one step event building protocol because the event builder knows « a priori » the block length. Under standard conditions defined previously, the full bandwidth is reached with 1 Kbytes data blocks. Next, variable length blocks event building has been performed with a two steps event building protocol: the wordcount transmission preceding the data block transfer. This protocol introduces an important software overhead required to prepare the output format. The full bandwidth speed is not yet reached with 4 Kbytes data blocks. The protocol implemented in L3 includes some additional data integrity checks which degrade the performance a bit more. Bandwidth and overhead per event are identical in the three measurements. The overhead per block is 1.1 μs for fixed length protocol, it increases to13 μs for variable length protocol and to 16 μs in the L3 protocol.

data memory	nb DS link	event building protocol	maxim. speed in Mbytes / s	overhead/block in μs
SRAM	2	fixed length	12.21 ± 0.03	1.12 ± 0.14
SRAM	4	fixed length	16.35 ± 0.17	1.20 ± 0.60
SRAM	2	variable length	12.24 ± 0.04	12.97 ± 0.24
SRAM	2	L3 (data integrity)	12.32 ± 0.11	15.89 ± 0.57
DRAM	2	fixed (small blocks)	15.13 ± 0.52	1.10 ± 0.24
DRAM	2	fixed (large blocks)	8.5 ± 0.70	-

TABLE 1

Table 1 summarizes the maximum speed and the overhead per block measured in different experimental conditions. An additional constant 35 μs software overhead per event has been observed. It appears clearly that bandwidth value relies only on the network and memory implementation. There is no correlation between bandwidth and the high level event building protocol. Similarly, the overhead per data block relies only on the high level protocol, not at all on the network implementation. This is exactly what we expected.

Event building has been measured in a farm structure with a 12 sources and up to 4 destinations network without any traffic shaping. Each source is connected to a C104 switch by one DS link and each destination (processing element) by two links, according to the standard conditions previously defined. One event being built by each destination. With 2 and 3 destinations we measure the expected throughput respectively 24 and 36 Mbytes/s. With 4 destinations a total throughput of 43 Mbytes/s instead of the 48 expected is measured. This loss is induced by contention in the C104 switch. The total input bandwidth being 72 Mbytes/s, no contention is observed up to 50% of the maximum bandwidth traffic (3 events built in parallel). A 10% traffic slowdown induced by contention is observed at 67% occupancy of the maximum bandwidth (4 events built in parallel). Performance can be improved by introducing some traffic shaping or by increasing the number of DS links.

4. Conclusions

A networked second level trigger made of 29 ST T9000 transputers interconnected by 2 ST C104 has been imbedded in the L3 data acquisition system. It ensures the L3 data taking since July 1995. This system is very stable: less than a crash per week. It behaves as expected and no deadlock or slowdown has been observed.

Event building performance has been measured. Network and memory configuration affects the maximum transfer speed. With our system we have measured a 6.1 Mbytes/s bandwidth per DS link which compares well with the 7 Mbytes/s maximum value between two T9000 measured by using the internal memory[6]. The high level event building protocol introduces an overhead of 1.1 μs/block for fixed length blocks, and of 13 μs/block for variable length blocks. An additional constant software overhead of 35 μs per event has been observed. Negligible for blocks longer than 1 Kbytes with a fixed length protocol, these overheads remain important up to 4 Kbytes with a variable length protocol.

In a 12 sources (1 DS link) to 4 destinations (2 DS links) network, without traffic shaping, no contention is observed up to 36 Mbytes/s which is 50% of the maximum input bandwidth traffic, and a 10% traffic reduction is observed at 48 Mbytes/s for 4 events built in parallel (67% of the maximum input bandwidth traffic).

5. Acknowledgements

We very gratefully acknowledge the IN2P3 / CNRS national French agency for financial support of the system. We also thank our L3 , LAPP and CERN colleagues for their constant interest for this development.

6. References

1. T. Angelov, B. Bertucci, S. Falciano, M. Fukushima, D. Linnhofer, B. Martin and G. Medici, *Nucl. Instr. and Meth.* **A 306** (1991) 536.

2. Y. Bertsch, J.J. Blaising, H. Bonnefon, F. Chollet-Leflour, A. Degré, G. Dromby, J. Lecoq, R. Morand, M. Moynot, G. Perrot and X. Riccadonna, *Nucl. Instr. and Meth.* **A 340** (1994) 309-321. and

   S.P. Beingessner, J.J. Blaising, F. Chollet-Leflour, A. Degré, C.Dromby, G. Forconi, C. Goy, J. Lecoq, R. Morand, M. Moynot, G. Perrot, S. Rosier-Lees, *Nucl. Instr. and Meth.* **A 340** (1994) 322-327.

3. *The T9000 Transputer Hardware Reference Manual.* Inmos Ltd., Inmos document number 72 TRN 238 01. and
   The transputer Data Book, 2nd ed., SGS-THOMSON microelectronics, 1989.

4. *STC104. Asynchronous Packet Switch*, Preliminary Data Sheet, June 1994, SGS-THOMSON microelectronics.

5. Alain Masserot, *Mise en oeuvre et intégration dans l'expérience L3 d'un déclenchement de deuxième niveau avec assemblage de l'événement, développé autour d'un réseau de routeurs dynamiques C104 et de transputers T9000.* Thèse, Université de Savoie, Laboratoire d'Annecy-le-Vieux de Physique des Particules - September 1995.

   6. Stuart Fisher, *Low Level Benchmarking of the T9000 Transputer*, M.Sc dissertation, University of Liverpool, February 1995.

   Roger Heeley, *Real Time HEP Applications using T9000 Transputers, Links and Switches,* PhD Thesis, University of Liverpool, October 1996.