STUDIES ON SWITCH-BASED EVENT BUILDING SYSTEMS IN RD13

C.P. Bee[a], S. Eshghi[b], R. Jones, S. Kolos[c], C. Magherini, C. Maidantchik[d], L. Mapelli[e], G. Mornacchi[f], M. Niculescu[i], A. Patel[g], D. Prigent, R. Spiwoks[h], I. Soloviev[c.]

CERN, Geneva, Switzerland

M. Caprini[i], P.Y. Duval, F. Etienne, D. Ferrato, A. Le Van Suu, Z. Qian,

Centre de Physique des Particules de Marseille, IN2P3, France

I. Gaponenko, Y. Merzliakov

Budker Institute of Nuclear Physics, Novosibirsk, Russia

G. Ambrosini[j], R. Ferrari, G. Fumagalli, G. Polesello

Dipartimento di Fisica dell'Universita' e Sezione INFN di Pavia, Italy

One of the goals of the RD13 project at CERN is to investigate the feasibility of parallel event building system for detectors at the LHC. Studies were performed by building a prototype based on the HiPPI standard and by modelling this prototype and extended architectures with MODSIM II.

The prototype used commercially available VME-HiPPI interfaces and a HiPPI switch together with a modular software. The setup was tested successfully as a parallel event building system in different configurations and with different data flow control schemes. The simulation program was used with realistic parameters from the prototype measurements to simulate large-scale event building systems. This includes simulations of a realistic setup of the ATLAS event building system. The influence of different parameters and scaling behavior were investigated. The influence of realistic event size distributions was checked with data from off-line simulations. Different control schemes for destination assignment and traffic shaping were investigated as well as a two-stage event building system.

1 Introduction

The future experiments at the LHC will need event building systems with an unprecedented bandwidth of 1 to 10 GB/s and which will be able to assemble event fragments from 100 to 1000 data sources at rates of 1 to 10 kHz [1]. Since bus based systems cannot be used, parallel event building based on high speed interconnects and switching elements will have to be envisaged [2].

a. Now at University of Zurich, Switzerland.

b. Now at University of Utrecht, The Netherlands.

c. On leave from the Petersburg Nuclear Physics Institute, St. Petersburg, Russia

d. Also with Federal University of Rio de Janeiro, Brazil.

e. Spokesperson.

f. Contact Person.

g. On leave from School of Computing and Information Systems, University of Sunderland, UK.

h. Also at the University of Dortmund, Germany.

i. On leave from the Institute of Atomic Physics, Bucharest, Romania.

j. Now at University of Bern, Switzerland.

The RD13 project [3] is studying the feasibility of such systems based on commercial communication switches based on HiPPI and FibreChannel. Two complementary approaches are followed: a small-scale prototype has been built and successfully tested. On the other hand simulations using parameters from the prototype measurements are used to investigate big systems as needed in real applications.

2 Event Building Prototype

A prototype based on the HiPPI standard [4] was built and used to gain realistic parameters for further simulations of big systems. This prototype can be regarded as a testbed where different hardware components and control schemes can be combined in order to investigate different technologies and architectures.

2.1 Hardware

The whole prototype is housed in one VME crate except the switch itself. VME-HiPPI interfaces based on the RIO module [5] act as either data sources or data destinations depending on the type of HiPPI interface because HiPPI is a simplex data transfer standard. The IOSC HiPPI switch [6] has 8 input and 8 output ports and an aggregate bandwidth of 800 MB/s. The arbitration of the source requests is done in a round-robin manner making the requests “camp on” as long as the destination is busy. A RAID processor [5] provides the functionality of data flow processes and their control and monitors the performance of the prototype.

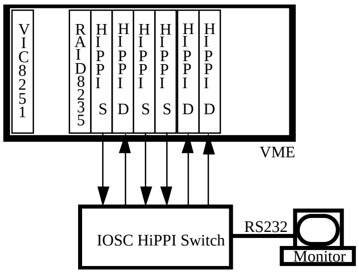

Figure 1. Event Building Prototype Setup

2.2 Software

The software developed [7] consists of two layers: the firmware and the data flow protocol. The lower layer is hardware specific and runs on the VME-HiPPI interfaces, sending and receiving event data by using the HiPPI protocol, one connection per event fragment. It communicates with the higher layer by using VME interrupts.

The higher level layer is hardware independent and modular. It can be regarded as a stripped-down version of the RD13 DFP [7] and functions as a mini-DAQ system. It contains data flow processes for the data sources (*Src* process) and data destinations (*Dst* process) and provides the event building functionality of event assembly and destination assignment. It can easily be extended to different hardware components and to more processes. The independent module of event assembly can be run either on the HiPPI/D module or the RAID processor.

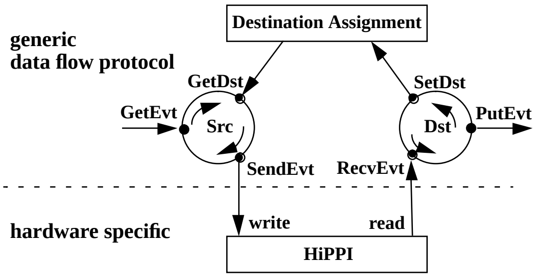

Figure 2. Software Layout of the Event Building Prototype

2.3 Performance

Simple data transfers from one source to one destination were used to measure the minimum latency to be 49 µs. This is mainly due to the firmware and the HiPPI protocol, the switch itself contributes with less than 1 µs. The interrupt handling could be measured to be minimally 32 µs, so that the latency between data source and data destination process is 81 µs in total. The maximum frequency for sending events is 30.3 kHz, and for receiving events 23.8 kHz. The link speed is 41.5 MB/s.

Several data sources and data destinations were combined to build a parallel event building system using a simple *PUSH* algorithm for the destination assignment which was done in a round-robin manner. The maximum throughput reveals a scalability with the number of data destinations for sizes above 10 kByte. This scalability, however, is limited for small event fragment sizes only by the single processor. Good agreement can further be seen between the measurements and the simulations carried out with the simulation program which uses the parameters from the one-to-one measurements.

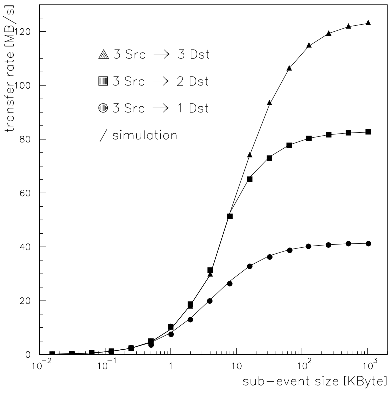

Figure 3. Performance of the Event Building Prototype

Strong variations in the event size disturb the parallelism and decrease the throughput. Individual exponential event fragment size distributions can be regarded as worst case scenarios which reduce the efficiency to about 76% compared to fixed event fragment sizes for a 2×2 setup.

Different schemes for the destination assignment were tested. Random assignment shows bad performance compared to round-robin (34% less throughput for a 2×2 setup). Alternatively to the *PUSH* scheme two schemes were tested which send the event fragments only after receiving a signal from the destination. VME interrupts were used for synchronization and the two schemes differ if they wait for the HiPPI/S module to have sent the previous event fragment (*SYNC*) or not, queuing the event fragments on the HiPPI/S module (*PULL*). No essential differences were found in the three schemes and the little differences can be explained by the single control process which has to deal with a different number of VME interrupts.

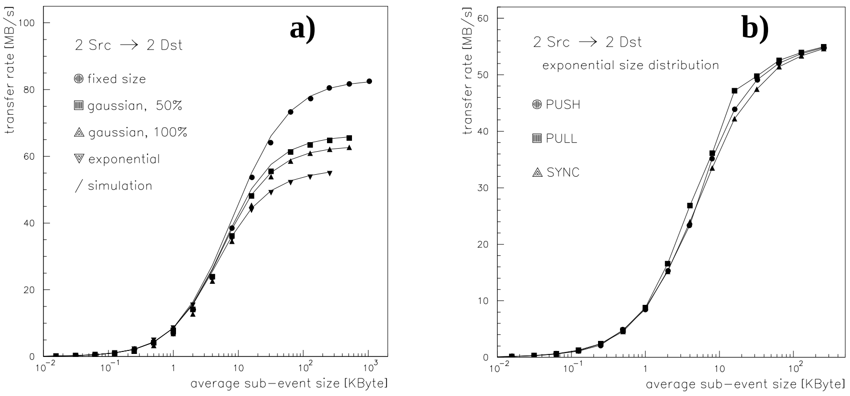

Figure 4. a) Variation of Event Fragment Sizes;
b) Destination Assignment Schemes

3 Modelling of Event Building Systems

Complementary to building prototypes, discrete-event simulations can be used to model big systems with many active elements. MODSIM II [8] was used to simulate generic parallel event building systems. The program is verified on the prototype measurements and extrapolated to model ATLAS event building.

3.1 Simulation Program

The simulation program [9] is a subset of simplified DSL objects [7]. It implements an event generator, the data source and destination processes and a simple switch model which parameterizes the transfer speed as a linear function of the event fragment size. This model can simulate different input event fragment size distributions, different control schemes and is fully configurable in the number of data sources and destinations and their parameters. This is appropriate for generic studies of parallel event building systems and was in particular used to simulate HiPPI and FibreChannel, class 1 based systems. A setup of 100×100 sources and destination with an average event fragment size of 10 kByte per source and an input frequency of 1 kHz was used as a reference system.

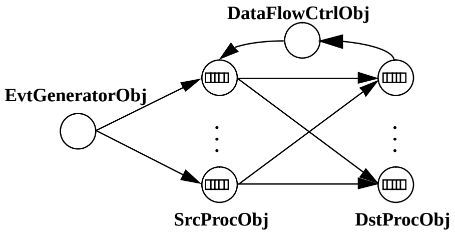

Figure 5. Simulation Program

3.2 Simple Model

Generic studies based on the parameters from the prototype show that with a round-robin destination assignment, the event building system runs in a “barrel shifter” mode. The efficiency ε is defined by

$$T^{max} = \varepsilon \cdot T^{max}_{ideal} = \varepsilon \cdot N_{Dst} \cdot speed$$

where T^{max} is the maximum throughput, N_{Dst} the number of destinations and *speed* the link speed. The efficiency can be factorized into a contribution from the overhead and a contribution from the event size variation:

$$\varepsilon = \varepsilon_{overhead} \cdot \varepsilon_{size} \quad \text{where} \quad \varepsilon_{overhead} = \frac{size}{size + overhead \cdot speed}$$

Typical values for the event size variations are 80% for a Gaussian with σ_{rel} = 50% and 60% for exponential size distributions. The correlation of event fragment sizes has a similar effect. The maximum throughput, latency and buffer occupancy show a scaling behavior with the number of sources and destinations.

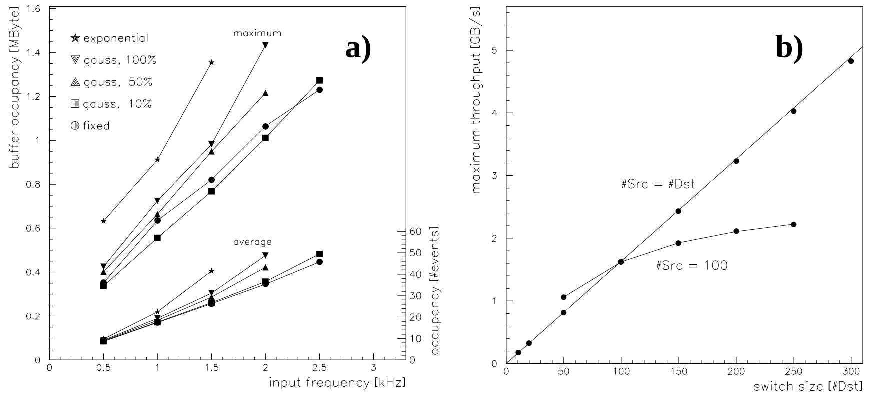

Figure 6. Simulation Results: a) Buffer Occupancy; b) Scaling

3.3 ATLAS Event Building

Realistic event fragment size distributions were obtained from ATLAS off-line simulations [10] using the SCT detector and jet events with 18 minimum bias events, passing a level 1 trigger for isolated electrons. The channels of the SCT detector were mapped on 16 event building data sources arranged in an array in $\eta\times\phi$. To simulate a realistic detector, 4 times this sample and 4 times a sample with Gaussian event size distributions from calorimeters were used. The resulting latency of 44 ms and maximum buffer occupancy of 900 kByte are in agreement with the simple model. The efficiency is ε =58%, with ε_{size} = 80%.

4 Data Flow Control Schemes

Data flow control schemes have to cover the contention resolution in the switch, the destination assignment and the synchronization between the data flow processes and the interconnecting network. The contention resolution has no visible influence on the performance [11] and is not discussed here.

4.1 Destination Assignment

Different destination assignment schemes like in the prototype (i.e. *PUSH*, *PULL* and *SYNC*) were investigated. Random assignment in the PUSH scheme is much less efficient ($\varepsilon \approx$ 30%) than round-robin. The other schemes are very similar in terms of efficiency ($\varepsilon \approx$ 45..55%), latencies and buffer occupancies. An alternative to these schemes is to use a special processor connecting to the data flow processes which maintains tables for the status of the data flow processes and for the events already assigned. The processing time of this data flow manager has to be smaller than 6 µs after which the performance degrades.

4.2 Traffic Shaping

Another field of investigation is control schemes at the level of individual event fragments. Algorithms known as “traffic shaping” for packet-oriented networks [12] could be implemented in this domain of connection-oriented networks in the following way: an event fragment will be skipped temporarily if its destination is busy and the next will be tentatively sent instead. This scheme is based on the flow control in the HiPPI protocol. It improves the efficiency by about 15% and reduces the latency and occupancy by about 10% for exponential size distributions when allowing an event to be skipped at maximum of 10 times. However, the time to check a destination has to be taken into account and for fixed size events this leads to reduced performance.

4.3 Two-Stage Event Building System

The cost and availability of large switches is still an open question. The use of a single large switch presents integration and reliability problems. None of these problems are fatal but they could be overcome by building a network of smaller switches. A simulation of a two-stage system of 10 switches of 10×10 ports on each stage, using a *PUSH* scheme with round-robin destination assignment reveals that the efficiencies are similar while the latency is reduced by about 30%. The buffer occupancy requires a maximum of about

400 kByte on the first and about 700 kByte on the second stage for each node.

5 Summary

The prototype has shown that parallel event building is possible using a commercially available technology. A particular technology was chosen and integrated with generic software which can be extended for future implementations using other technologies (like FibreChannel or ATM). Measurements with the prototype have revealed realistic parameters for different overheads in hardware and software. With values of about 40 MB/s and about 100 μs the simulations show good agreement with the measurements.

Simulations of big systems with realistic event size distributions from off-line simulations revealed an efficiency of 58% with reasonable latency and buffer occupancy. The event building systems presented are scalable with the number of data sources and destinations. The destination assignment has little influence on the performance and if a data flow manager is used, its processing time must be small. Traffic shaping can improve the performance. A network of smaller switches has similar performance to a big switch but a smaller latency and might be easier to build.

Acknowledgments

We would like to thank E. v.d Bij and R. McLaren from CERN/ECP for their advice with the HiPPI hardware and software and for loan of some of the modules. We are also indebted to W. Greiman from LBL for the original version of the simulation program.

References

[1] L. Mapelli, The DAQ and Trigger System of the ATLAS Experiment at the LHC, these proceedings.

[2] E. Barsotti et al., Effects of various Event Building Techniques on DAQ System Architectures, FERMILAB-CONF 90/61, 1990.

[3] L. Mapelli et al. (RD13 Collaboration), A Scalable Data Taking System at a Testbeam for LHC, CERN/LHCC 95-47, LCRB Status Report/RD13, 1995.

[4] HiPPI Standard, ANSI X3T9.3/91-005.

[5] Creative Electronics Systems, Petit-Lancy, Switzerland.

[6] Input Output Systems Corporation, Mountain View, California.

[7] The RD13 Technical Notes and other documentation can be found on the WWW under http://rd13doc.cern.ch/welcome.html.

[8] CACI Products Corporation, La Jolla, California.

[9] W. Greiman, Design and Simulation of FibreChannel based Event Builders, Proc. of Int. Conf. on DAQ and Event Building, FNAL, Batavia, Illinois, 1994.

[10] R. Hawkings, The Level 2 Silicon Track Trigger (T2SI) Implementation in ATRIG, ATLAS internal note, 1995 (in prep.).

[11] R. Spiwoks, Evaluation and Simulation of Event Building Techniques for a Detector at the LHC, PhD Thesis, University of Dortmund, 1995 (in prep.).

[12] I. Mandjavidze, Review of ATM, FibreChannel and Conical Network Simulations, Proc. of Int. Conf. on DAQ and Event Building, FNAL, Batavia, Illinois, 1994.