Commodity Multi-processor Systems in the ATLAS Level-2 Trigger

M. Abolins[4], R. Blair[1], R. Bock[2], A. Bogaerts[2], J. Dawson[1], Y. Ermoline[4], R. Hauser[4], A. Kugel[3], R. Lay[5], M. Muller[3], K.-H. Noffz[5], B. Pope[4], J. Schlereth[1], P. Werner[2]

[1] Argonne National Laboratory, Argonne, IL 60439-4812, USA
[2] CERN, The European Laboratory for Particle Physics, CH-1211 Geneva 23, Switzerland
[3] University of Mannheim, Mannheim, D-68159, Germany
[4] Michigan State University, East Lansing, MI 48824-1321, USA
[5] Silicon Software GmbH, Mannheim, D-68161, Germany

Abstract

Low cost SMP (Symmetric Multi-Processor) systems provide substantial CPU and I/O capacity. These features together with the ease of system integration make them an attractive and cost effective solution for a number of real-time applications in event selection. In ATLAS we consider them as intelligent input buffers ("active" ROB complex), as event flow supervisors or as powerful processing nodes.

Measurements of the performance of one off-the-shelf commercial 4-processor PC with two PCI buses, equipped with commercial FPGA based data source cards (microEnable) and running commercial software are presented and mapped on such applications together with a long-term programme of work.

The SMP systems may be considered as an important building block in future data acquisition systems.

DISCLAIMER

DISCLAIMER

Portions of this document may be illegible in electronic image products. Images are produced from the best available original document.

INTRODUCTION

This work is based on the idea of using commercial commodity components in the ATLAS high-level trigger and DAQ systems [1]. One of these components is the SMP - Symmetric Multi-Processor system. It provides substantial CPU power and allows access to the main memory and input/output interfaces from all processors through a very high-speed system bus or switch. The workload is balanced among the processors by the single operating system. Low-cost commercial SMP systems have become generally available since 1998 and presently are limited to 4- or maximum 8-processor systems and of the order of 20 PCI slots. We discuss a number of possible real-time applications in DataFlow and LVL2 selection; present results of our measurements and suggest long-term programme of work.

APPLICATION AREAS

In the ATLAS high-level trigger system we consider several areas where SMP systems can be used - intelligent input buffers ("active" ROB complex), an event flow supervisor or as powerful processing nodes.

A. Active ROB complex

The proposed application for such components is the detector-adapted computing station with multiple processors, all having access to all ROBs (Read-Out Buffers) of a full detector, using commercial components from the HPCN (High Performance Computing and Networking) market - several multi-processor boards with proprietary interconnect, all working under a shared-memory paradigm. The availability of components has reduced the initial goal to a multi-ROB station; we call it "active" ROB complex because the processors actively contribute to alleviate the critical traffic over the LVL2 selection network. Processors in the SMP based "active" ROB complex are assumed to share memory and access to a number of ROBs. The grouping of ROBs is adapted to individual detectors; the limit is set at 16 ROBs per "active" ROB complex. Open questions include the achievable aggregate bandwidth for multiple ROBs on multiple PCI buses, limits set by the internal memory bus of the SMP system and the overhead for the multi-thread implementation of the "active" ROB tasks.

B. Event flow supervisor

A detailed description of the LVL2 Supervisor can be found in the literature [2]. It consists of several processors (VME or PC based), connected to the LVL2 network. It is designed to be simply scalable by adding more processors. An additional unit, an ROI Builder [3] combines the different data streams from the LVL1 into a record for each event and distributes the data to processors within the Supervisor farm. The Supervisor processor manages the event through LVL2 - allocates a LVL2 processor, forwards the ROI record to it, receives the decision back, packs the decisions and multicasts them to the ROBs. The Supervisor also transmits the LVL2 decisions, which may be grouped to reduce the rate of messages, to the Event Filter. The SMP based Supervisor might prove advantageous for LVL2 processor allocation, for grouping of LVL2 decisions and for interaction with the Event Filter and Supervisor monitoring tasks. The main issues are input/output bandwidth and scalability.

C. Processing node

Grouping LVL2 processors into sub-farms is made easier with SMP systems. Workload balancing is automatically provided by the operating system and communication and synchronization tasks among the different sub-farm components are easily accomplished.

ONGOING ACTIVITIES

A. Active ROB complex

Two activities were proposed in September 99. First is modeling the active ROB complex to assess the effect on general LVL2 selection network traffic for different detectors. A report on this work is available [4]. Second is measuring the internal performance of a present-day commercial SMP. Our results [5] are obtained on a PC server based on a four-processor board (Intel SC450NX, 4 x 550 MHz Pentium Xeon III, 2 PCI buses of 32 bits/32 MHz, 6 free slots, 512 MB shared memory, 100 MHz memory bus). Standard commercial FPGA based microEnable boards [6] were used as ROB emulators.

All programs were developed on standard PCs. The commercial Micro-Enable drivers were adapted to the multi-processor / multi-bus environment. A high-level interface was written to handle requests for ROB data (by event number), and to provide a polling mechanism. An application program was written in C++ to perform the basic measurements. Multi-threaded programs allow the system to distribute the tasks to different processors.

The measurements (Figure 1) show that the double PCI bus can be put to use. The rate increase for the second PCI bus (i.e., going from two to four ROBs) is 88% for large packets, 55% for the preferred packet size of 1 KB.

The input/output rate also depends on the memory bus loading: available PCI bandwidth drops by about 20%, when loading the memory bus with a read/write flow of 130+130 MB, i.e. it goes from the 160 to 130 MB/s for 1 KB ROB fragments.

Measurements of internal communication at application level have also shown that the substantial CPU capacity in the SMP system can be largely made available to user programs in situations approaching those of Atlas LVL2 traffic. More work is required to show the user and system level effects of multithreading.

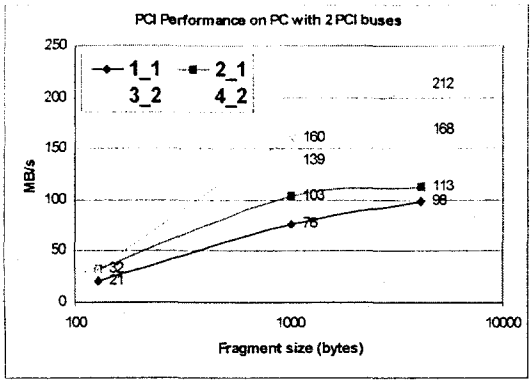

Figure 1. The bandwidth as a function of the data size for 1 to 4 ROBins.

B. Event flow supervisor

Investigations are under way to determine the feasibility and possible benefits of using an SMP system as the Supervisor in place of the current complement of PCs.

The main concern is the input/output bandwidth of the Supervisor. In the final implementation of the LVL2 ROI Builder [7], the input buffers on 7 links from the LVL1 will accommodate up to 63 32-bit words per event fragment. This leads to a maximum aggregate input bandwidth to the LVL2 Supervisor of about 180 MB/s at 100 kHz event rate and 1.8 Kbytes event size.

Measurements on the four-processor boards show that a bandwidth of 40 MB/s per input port could be achieved for packet sizes of 1 Kbytes. Therefore 5-6 links from the ROI Builder to the Supervisor may be necessary to carry the event data for the maximum input load.

A similar output bandwidth from the Supervisor to the LVL2 selection network will be necessary for communication with the LVL2 processors and ROB systems. Further measurements need to be done in order to estimate the necessary number of output links.

It is quite obvious that the SMP system we are using for the preliminary measurements will not be suitable for the SMP Supervisor implementation. Another SMP architecture needs to be investigated, such as the Compaq 8-way multiprocessor [8]. It provides eight Pentium Xeon III processors on two 100 MHz 64-bit memory buses, a crossbar switch with a peak throughput of 4 GB/s, a 100 MHz 64-bit input/output bus and 11 PCI slots.

REFERENCES

[1] "ATLAS High-Level Triggers, DAQ and DCS", CERN/LHCC/2000-17, 31 March 2000.

[2] R. Blair et al., "The ATLAS Level-2 Trigger Supervisor", in Proc. Second Workshop on Electronics for LHC Experiments, Balatonfured, Hungary, 23-27 September 1996, pp. 191-193.

[3] R. Blair et al., "A Prototype ROI Builder for the Second Level Trigger of ATLAS Implemented in FPGA's", LEB'99, Snowmass, September 20-24 1999, ATL-DAQ-99-016, 7 Dec 1999.

[4] R. Bock et al., "The active ROB complex", ATL-DAQ-2000-022, 31 March 2000.

[5] R. Bock et al., "The use of low-cost SMPs in the Atlas level-2 trigger", ATL-DAQ-2000-010, 13 March 2000.

[6] MicroEnable is manufactured by Silicon Software GmbH, Mannheim, D-68161, Germany.

[7] M. Abolins et al., "Specification of the LVL1 / LVL2 trigger interface, Version 1.0", ATL-DAQ-99-015, 7 October 1999.

[8] "Compaq ProLiant 8500 Server Technology", Technology Brief, 0091-0899-A, Compaq Computer Corporation, August 1999.