The ATLAS Level-2 Trigger Pilot Project

R. Blair, J. Dawson, W. Haberichter, J. Schlereth, R. Bock, A. Bogaerts, M. Boosten, R. Dobinson, M. Dobson, N. Ellis, M. Elsing, F. Giacomini, E. Knezo, B. Martin, T. Shears, S. Tapprogge, P. Werner, J. R. Hansen, A. Wäänänen, K. Korcyl, J. Lokier, S. George, B. Green, J. Strong, P. Clarke, R. Cranfield, G. Crone, P. Sherwood, S. Wheeler, R. Hughes-Jones, S. Kolya, D. Mercer, C. Hinkelbein, K. Kornmesser, A. Kugel, R. Manner, M. Müller, M. Sessler, H. Simmler, H. Singpiel, M. Abolins, Y. Ermoline, B. Gonzalez Pineiro, R. Hauser, B. Pope, S. Sivoklokov, H. Boterenbrood, P. Jansweijer, G. Kieft, R. Scholte, R. Slopsema, J. Vermeulen, J. T. Baines, A. Belias, D. Botterill, R. Middleton, F. Wickens, S. Falciano, J. Bystricky, D. Calvet, O. Gachelin, M. Huet, P. Le Dû, I. Mandjavidze, L. Levinson, S. Gonzalez, W. Wiedenmann, and H. Zobernig

***Abstract*—The Level-2 Trigger Pilot Project of ATLAS, one of the two general purpose LHC experiments, is part of the on-going program to develop the ATLAS high-level triggers (HLT). The Level-2 Trigger will receive events at up to 100 kHz, which has to be reduced to a rate suitable for full event-building of the order of 1 kHz. To reduce the data collection bandwidth and processing power required for the challenging Level-2 task it is planned to use Region of Interest guidance (from Level-1) and sequential processing. The Pilot Project included the construction and use of testbeds of up to 48 processing nodes, development of optimized components and computer simulations of a full system. It has shown how the required performance can be achieved, using largely commodity components and operating systems, and validated an architecture for the Level-2 system. This paper describes the principal achievements and conclusions of this project.**

Manuscript received October 31, 2000.

R. Blair, J. Dawson, W. Haberichter, and J. Schlereth are with Argonne National Laboratory, Argonne, IL 60439-4812 USA.

R. Bock, A. Bogaerts, M. Boosten, R. Dobinson, M. Dobson, N. Ellis, M. Elsing, F. Giacomini, E. Knezo, B. Martin, T. Shears, S. Tapprogge, and P. Werner are with CERN, The European Laboratory for Particle Physics, CH-1211 Geneva 23, Switzerland.

J. R. Hansen and A. Wäänänen are with Niels Bohr Institute, University of Copenhagen, Copenhagen, Denmark.

K. Korcyl is with the H Niewodniczanski Institute of Nuclear Physics, Krakow, Poland.

J. Lokier is with the University of Liverpool, Liverpool, UK.

S. George, B. Green, and J. Strong are with RHBNC, University of London, London, UK.

P. Clarke, R. Cranfield, G. Crone, P. Sherwood, and S. Wheeler are with UCL, University of London, London, UK.

R. Hughes-Jones, S. Kolya, and D. Mercer are with the University of Manchester, Manchester M13 9PL, UK.

C. Hinkelbein, K. Kornmesser, A. Kugel, R. Manner, M. Müller, M. Sessler, H. Simmler, and H. Singpiel are with the University of Mannheim, Mannheim, D-68159, Germany.

M. Abolins, Y. Ermoline, B. Gonzalez Pineiro, R. Hauser, and B. Pope are with Michigan State University, East Lansing, MI 48824-1321 USA.

S. Sivoklokov is with Moscow State University, Moscow, Russia.

H. Boterenbrood, P. Jansweijer, G. Kieft, and J. Vermeulen are with NIKHEF, Amsterdam, Netherlands.

R. Scholte and R. Slopsema are with NIKHEF, Amsterdam, Netherlands. They are also with the University of Twente, Enschede, Netherlands.

J. T. Baines, A. Belias, D. Botterill, R. Middleton, and F. Wickens are with Rutherford Appleton Laboratory, Chilton, Oxon OX11 0QX, UK (e-mail: f.wickens@rl.ac.uk).

S. Falciano is with the University of Rome "La Sapienza," Rome, Italy.

J. Bystricky, D. Calvet, O. Gachelin, M. Huet, P. Le Dû, and I. Mandjavidze are with DAPNIA, CEA Saclay, 91191 Gif-sur-Yvette Cedex, France.

L. Levinson is with the Weizmann Institute of Science, Rehovoth, Israel.

S. Gonzalez, W. Wiedenmann, and H. Zobernig are with the University of Wisconsin, Madison, WI 53706, USA.

Publisher Item Identifier S 0018-9499(02)06119-1.

***Index Terms*—ATLAS, computer network performance, distributed computing, LHC, message passing, object oriented methods, parallel processing, protocols, technology assessment, triggering.**

I. Introduction

THE ATLAS Level-2 Trigger (LVL2) will receive events (each of typically 1–2 Mbytes) at up to 100 kHz. This rate has to be reduced to a rate suitable for full event-building of the order of 1 kHz. To reduce the data collection bandwidth and processing power required for the challenging LVL2 task it is planned to use Region of Interest (RoI) guidance from Level-1.

By 1998 earlier ATLAS studies [1] had led to the conclusions that;

- — Affordable commercial networks would very likely be able to handle the LVL2 traffic in a single network—a total of a few Gbyte/s between of the order of 1000 ports.
- — Standard commercial processors (especially PCs) should be the appropriate choice for most of the LVL2 processing.
- — For some specific processing tasks (e.g., pre-processing and searching for tracks in a complete tracking detector) FPGA processors could offer an additional boost.
- — Sequential selection was favored, offering clear benefits (e.g., reduced network bandwidth and processor load) and easing the essential sequential steps needed for some triggers.

The Pilot Project, part of the on-going program to develop the ATLAS high-level triggers, was based on a hardware architecture shown in Fig. 1. The LVL2 processors include the option of FPGA coprocessors, which could be added if required. This architecture aims to use commodity items [processors, operating system (OS) and network hardware] wherever possible.

The RoI Builder combines the fragments of RoI information from the LVL1 system into one event record, which it passes to an RoI Processor (a general-purpose processor) within the Supervisor farm. The Supervisor assigns each event to one of the LVL2 processors. The LVL2 processor performs one or more steps of data collection and analysis from relevant Readout Buffers (ROBs). (Note in Fig. 1 the ROBs are shown

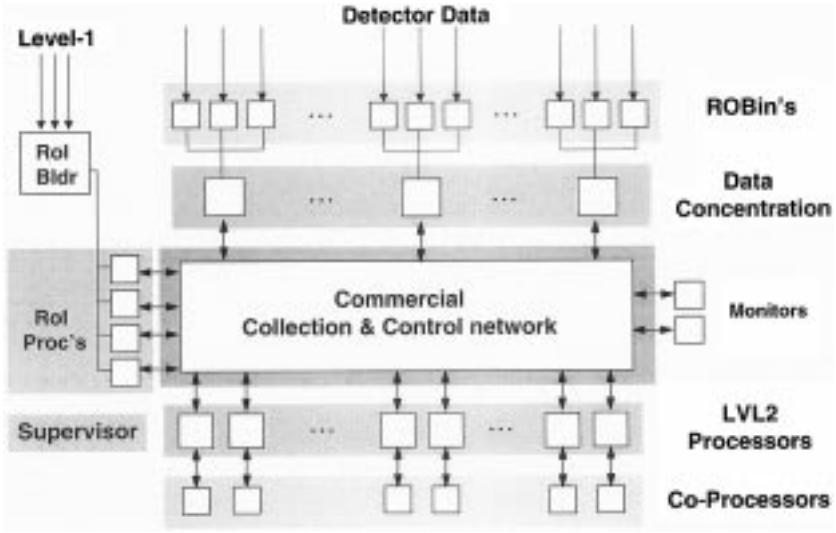

Fig. 1. The hardware architecture studied in the Pilot Project.

as separate input boards, ROBins, and a controller/output interface for a group of ROBins to provide a level of data concentration before the switch.) The controlling processor for an event can delegate part of the processing to another processor, for example the FPGA coprocessor for a Full Scan of the Transition Radiation Tracker (TRT) (an unguided search for track segments in the TRT). The trigger decision can be issued at any step. It is returned to the Supervisor which distributes it to the ROBs. Rejected events are discarded; accepted events are passed to the Event Filter (EF) for further analysis. Within this architecture the event selection process is sequential, all LVL2 processors can access all ROBs, and data collection from the ROBs is initiated by the processors using a request–response protocol.

The principal aims of the Pilot Project were to produce a validated LVL2 architecture and to investigate technologies likely to be required for its implementation. The Pilot Project built moderately large application test beds using the Reference Software; developed optimized components for the Supervisor and RoI Builder, the ROB Complex (a group of ROBins, together with its controller/output interface), networks and processors (especially FPGAs); and studied scaling to a full-scale system with computer simulations.

Indicative performance requirements for the LVL2 components are given in Table I.

TABLE I
INDICATIVE PARAMETERS AND PERFORMANCE REQUIREMENTS FOR LVL2 COMPONENTS

Parameter	Value
Total bandwidth in LVL2 network	~ 5 Gbyte/s
Max RoI request rate per ROB Complex	~ 14 kHz
Max output bandwidth to LVL2 per ROB Complex	~ 9 Mbyte/s
Max RoI Builder/Supervisor rate	100 kHz
Typical number of ROB Complexes per data request	~ 4
Data for TRT full scan	~ 200 kbyte from ~ 256 ROBs
Typical number of RoIs per event (primary RoIs contribute to the LVL1 trigger, secondary RoIs are below the LVL1 threshold)	1-2 primary ~3 secondary
Average number of sequential steps executed	~ 2

II. The Reference Software

The Reference Software aimed to provide a prototype implementation for the complete LVL2 process and a common software framework for the Pilot Project activities. These activities include evaluation of networking technologies, in particular ATM, Fast and Gigabit Ethernet, and SCI (Scalable Coherent Interface); evaluation of optimized components; development and evaluation of physics and run-time performance of LVL2 event-selection algorithms; measurements of critical parameters on multinode test beds to obtain indications of the system scalability; and validation of the software architecture.

Development of the Reference Software started at the beginning of 1998, using an object-oriented (OO) approach with C++ as the main implementation language. The required versatility—multiple platforms (Linux, Windows NT), multiple networking technologies (ATM, Ethernet, SCI), multiple environments (desktop, online test beds, hybrid systems with coprocessors) and extensibility to allow for new concepts—has been achieved by organizing the software as a set of layered packages. The software included simple run control, error reporting and monitoring adequate to support test beds of up to ~100 nodes. Fig. 2 shows the software architecture with a request—response protocol to transfer data between functional components.

The implementation had to provide a Trigger Processor (combining steering and feature extraction), and emulations of ROB Complexes and the Supervisor. These objects could be implemented on the same processor, e.g., for development of algorithms on the desktop, or distributed over multiple nodes for online test beds. For some tests the prototypes of the ROB Complex and Supervisor would be used in place of the emulations.

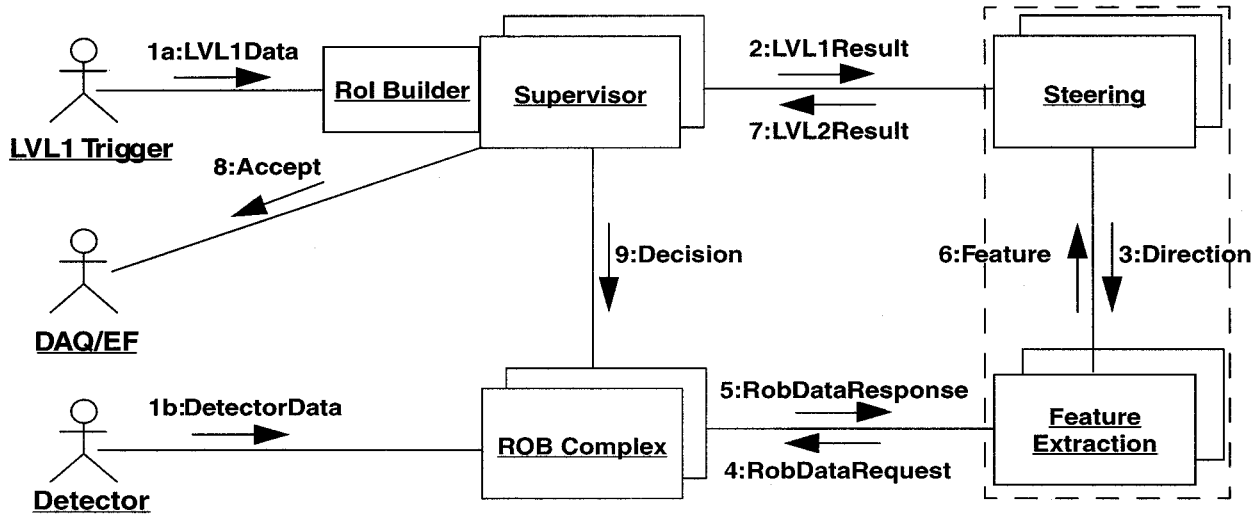

Fig. 2. The software architecture studied in the Pilot Project.

The full functionality is implemented as one application for a single node or split into three applications (Supervisor, Trigger Processor, ROB Complex) for the distributed version.

A farm of each of these is used to obtain a high trigger frequency, enough CPU power for the algorithms and enough bandwidth to supply the detector data to the Trigger Processors. Use of the (remote) proxy design pattern [2] provided transparent communications.

The standard thread scheduling provided by the native OS (Linux, Windows NT or Solaris) has been used in the ATM, SCI and MPI [3] test beds. The Ethernet test bed used an optimized package MESH [4] to provide fast thread switching and support the optimized Ethernet driver. The software is Symmctric Multiprocessor (SMP)-ready but tests have concentrated on single and dual CPU PCs.

Performance measurements with ATM and Ethernet used optimized drivers. The much heavier TCP and UDP protocols were used mainly for software development and testing. Some of the tests with SCI used MPI, with only a small performance loss.

The Reference Software has been run in many configurations, using various network technologies. As is described in the following sections it has been shown to scale to systems of up to ~100 nodes. The tests indicate that the required component performance can be obtained with commodity hardware (PCs) and OS software (such as PC/Linux). The request–response based architecture has been validated. Though not yet fully optimized the I/O performance obtained with the test beds gives good indications that the requirements of the LVL2 trigger can be met with this software architecture.

III. The Test Beds

The test beds were established to use the Reference Software: to check that individual components meet the required performance; to provide information on scaling up to moderate size systems; and to provide data for the full-system computer models. The ATLAS specific test beds varied in size from 25 to 50 nodes. In addition use was also made of a commercial cluster of 96 nodes at Paderborn University [5]. These systems correspond to a few per cent of the final ATLAS system. Ethernet (Fast and Gigabit), 155 Mb/s ATM and SCI technologies have been studied for the network. All the test beds were based on the hardware architecture shown in Fig. 1, although for most tests there were no coprocessors and emulations running on PCs were used for the Supervisor/RoI Builder and ROB Complexes. However, prototype components (i.e., Supervisor/RoI Builder, ROB Complex, FPGA processor) developed in the functional-component activities were integrated for some tests.

The Reference Software has been run on all test beds under Linux and Windows NT, and at Paderborn under Solaris. The ATM test bed was also run with the C-based ATM Test Bed Software developed for the Demonstrator Program under Windows NT, Linux and LynxOS. A series of measurements were devised for the test beds to obtain performance results of the various LVL2 components. The following sections consider each of the components in turn. Results from the test bed measurements are given together with the description of the functional-component activities.

IV. The ROB Complex

Studies of the Readout Buffers (ROBs) combined paper design and system modeling with prototyping of hardware. Key parameters for the assessment of different LVL2 processing strategies and for different system scenarios were obtained from the paper models [6] and computer models [7], whilst the feasibility of different implementation approaches was demonstrated in performance measurements of several hardware prototypes.

A key focus of the work was to investigate grouping sets of buffers into a ROB Complex. This comprises a number of input buffers (ROBins) with a single controller and output interface. Extreme cases considered are the simple ROB with a single input buffer and the Active ROB [8] with many buffers and considerable local processing power. These options are documented in [9].

The prototype studies demonstrated that buffering from ATLAS compatible readout links at the projected ATLAS event rates can be achieved with a number of designs; the NIKHEF ROBin prototype [10] has directly achieved a 160 Mbyte/s input rate, whilst prototypes [11] built to an earlier 100 Mbyte/s specification have been successfully operated in test beds and are now being upgraded.

Measurements of ROB Complex emulators (running on PCs) in test bed configurations have demonstrated the operation of ROBout interfaces. The principal aim was to provide a data source, consistent with that expected from a ROB Complex, for testing other components in the test bed. The performance of this emulation was measured in different test beds. The performance is consistent with that used in paper models and gives confidence that the network connection from the ROB Complex to the LVL2 system is attainable over the network technologies studied.

In addition, a prototype of the Saclay ROBin [12] was integrated into the ATM test bed. A ROB Complex composed of 1, 2, or 3 ROBins was tested on VME, CompactPCI (LynxOS) and PC-Linux platforms. The ROB controller was a 400 MHz Linux-PC that served the requests from $\sim$10 processors via a 155 Mbit/s ATM link. For typical fragment sizes of 1–2 kbyte, the maximum measured service rate almost reached the bandwidth limit of the ROB Controller link (16 Mbyte/s in this implementation).

Further performance gains can be made if pre-processing (such as data selection or reformatting) is performed on the data within the ROB Complex. Examples of this have been demonstrated in [13] and [14].

It has been demonstrated that the ROB requirements can be satisfied with current technology: projected input rates can be handled; output to LVL2 and to the Event Builder (EB) is achievable at the necessary rates and bandwidth; some on-the-fly pre-processing is possible. The implementation studies showed a compact design for the ROBin is achievable, allowing the construction of a ROB Complex with several (3–6) ROBins. COTS hardware seems to be able to support the output requirements but perhaps not the input rates. The Pilot Project has also provided checklists for future ROB design work.

V. The RoI Builder and Supervisor

The basic functions of the RoI Builder (RoIB) and Supervisor are as follows: On each LVL1 accept, the RoI Builder receives RoI information fragments from the LVL1 system. These RoI fragments are formatted into a single record for each event. The RoIB then transfers the record to a selected RoI processor within the Supervisor. The RoI processor manages the event through LVL2. It allocates the event to a LVL2 processor; forward the RoI record to this target processor; receives the decision back; updates the statistics; packs the decisions and multicasts them to the ROBs.

A prototype RoIB was designed, produced and tested [15]. Since the RoIB must operate at event rates as high as 100 kHz without introducing dead-time, the prototype was implemented entirely in hardware, using FPGAs with a design emphasizing parallelism. This allowed an architecture where several RoIB channels operate in parallel. The LVL1 event ID, embedded in every RoI fragment, is used to identify the RoI fragments belonging to a given event and for the assignment to an RoIB channel. The assigned channel builds the RoI fragments of an event into a record, the other channels discard these fragments.

The RoIB was tested with an input card which supplies preloaded RoI fragments for up to 1024 events. These tests included use of the Supervisor/RoIB integrated into ATM and Ethernet test beds. The hardware was demonstrated to run at the required rates without errors or introducing deadtime.

Within the test beds, the Supervisor concepts were also tested using Supervisor emulators. Measurements included scaling with the number of Supervisor emulators and the dependence of the rate of a single emulator as a function of the number of RoIs per event. With a single RoI, a rate of $\sim$11 kHz per Supervisor emulator is reached. The results also show that the system rate scales with the number of Supervisors and a rate of 120 kHz was achieved with twelve Supervisor emulators (no RoIB) on the Paderborn cluster.

Thus a prototype RoIB has been built using FPGAs in a highly parallel architecture. The design uses custom hardware for the most demanding tasks, which reduces the demands on the other Supervisor components, so that they can be implemented using standard processors. The RoIB has been integrated with a Supervisor farm into ATM and Ethernet test beds. The RoIB plus a small Supervisor farm have been shown to satisfy the requirements for the LVL2 trigger, i.e., up to a rate of 100 kHz.

VI. Processor Requirements and Measurements

The first task of a LVL2 processor is to collect data from many sources and the second task is to process the data received. The emphasis in the test beds has been on the first task and on quantifying the resources needed for this. It was demonstrated that with all of the network technologies a single processor could collect data corresponding to a typical RoI (1–4 kbytes from across a few ROBs) at a few kHz. Whilst this data collection time is significant, the total farm sizes envisaged imply much lower rates per processor, leaving the larger part of the time for algorithm processing.

While the number of ROBs involved in data collection for an RoI is small, several other tasks require gathering data from all the ROBs of a given subdetector (e.g., all TRT ROBs in a complete search for tracks for the B-physics trigger), several subdetectors, or in the case of event building the whole detector. This data collection was investigated with the ATM test bed software. With 20 processors, each collecting data from 20 ROBs, a sustained global throughput of 260 Mbyte/s and 328 Mbyte/s is measured for ROB data fragments of 2 kbyte and 4 kbyte respectively. Data blocks of 80 kbyte equally spread across 20 ROBs are gathered at 4 kHz.

As already noted the use of sequential selection reduces the network and processor requirements and allows more complex algorithms to be run at lower rates. To validate the principle of multistep data transfers and processing a test was run on the cluster at the University of Paderborn [5] with the Reference Software including algorithms for three detectors: calorimeter

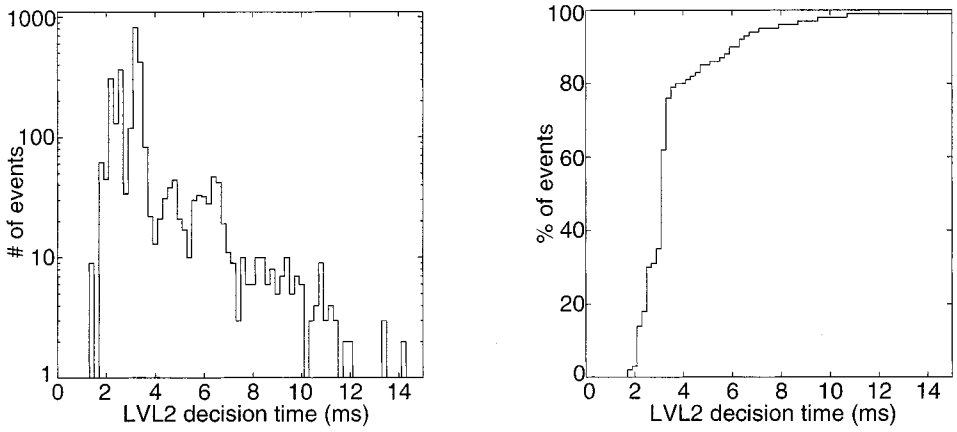

Fig. 3. The latency for 3 step sequential selection.

e.m. clustering, TRT tracking and precision tracking in the Semiconductor Tracker (SCT) and pixels. The Supervisor and RoB emulators were preloaded with data for ∼3000 jet events, preselected to contain at least one LVL1 e.m. RoI (about 15% of the events contained two RoIs). The menu consisted of three consecutive steps, the fraction of events accepted after each step was 0.19, 0.05 and 0.02.

The latency as measured in the Supervisor (the time interval between sending the LVL1Result and the reception of the LVL2Result—see Fig. 2) is shown in Fig. 3(a). It includes communication, data preparation and actual processing time. Communication delays contribute ∼500 μs. The RoI data size is ∼10–20 kbyte, contributing another ∼300 μs for each RoI/detector combination. The distribution reveals the sequential execution of the three algorithms: calorimeter e.m. clustering predominantly below 4 ms, subsequent TRT tracking at 4–7 ms and final SCT/pixel precision tracking extending beyond 7 ms. The average and median values are 3.7 ms and 3.1 ms, respectively. The effect of rejection at early stages in the sequential process is shown explicitly in Fig. 3(b): 50% of the events finish within 3.1 ms, 95% within 7.2 ms and 99% within 10.8 ms. (Note the processors are 450 MHz dual Pentium II.)

Thus processors in test beds using the Reference Software have demonstrated: data collection within an RoI from a single detector at an acceptable rate; a three-step sequential selection strategy with prototype algorithms running on a multinode test bed with the processor requesting simulated event data from ROB emulator nodes.

VII. Use of FPGAs as Coprocessor

Modeling [6] shows that the size of the LVL2 trigger farm required may be determined primarily by the need of the B-physics trigger to execute a track search in the full inner-detector volume. The full-scan algorithms allow considerable parallelism and are good candidates to run in FPGA-based processors.

The FPGA implementation studies during the Pilot Project focused on the ATLANTIS processor system [16]. This is a combined FPGA and CPU-based computing system housed in a CompactPCI crate. A standard Intel Pentium PC—a CompactPCI computer—which plugs into one of the ATLANTIS active backplane slots is used for external connections. The FPGAs are mounted on the ATLANTIS Computing Board (ACB). Communication between the ACB and the PC is via the PCI backplane. When connected to a test bed the ATLANTIS system appears as a normal PC with accelerator features.

The full-scan TRT algorithm [17] was implemented on ATLANTIS, with the most time-consuming parts performed in the FPGAs. FPGA execution is performed on-the-fly as the TRT hits arrive in the ACB and the execution time is largely determined by the PCI data transfer rate. Tests demonstrated a factor 6 improvement in the total execution time compared to a 300 MHz Pentium II [18].

In addition to standalone tests, the ATLANTIS system, running Windows NT, was successfully integrated into the ATM test bed [18]. The ATLANTIS system appeared to the Reference Software as a normal trigger processor. It was demonstrated that event data could be transferred from ROBs to the ATLANTIS system and that the algorithm quality was identical to the CPU-only implementation.

This work indicates how FPGA coprocessors could be included in standard processors in a transparent way, offering significant performance improvements for suitable compute-intensive algorithms and hence a reduction in the size of the processor farms required.

VIII. Meeting the Network Requirements With Available Technologies

Networking technologies for the ATLAS HLT/DAQ system have to support large data collection networks connecting the ROBs to hundreds of destination processors. Depending on the detector readout and event selection strategy, the raw bandwidth requirement is estimated to be in the range of 4–6 Gbyte/s. The networks have to transport various types of traffic with different requirements in terms of bandwidth, message rate and latency. Protocol messages are characterized by a relatively small size (∼tens of bytes), a high rate (∼tens of kHz per node), and mainly flow from the destination processors toward the ROBs. Multicast capability is likely to be required (e.g., to distribute trigger decisions to the ROBs). Data traffic, characterized by the concentration of messages toward the processors from a number of ROBs, requires a high bandwidth. Data collection of RoIs re-

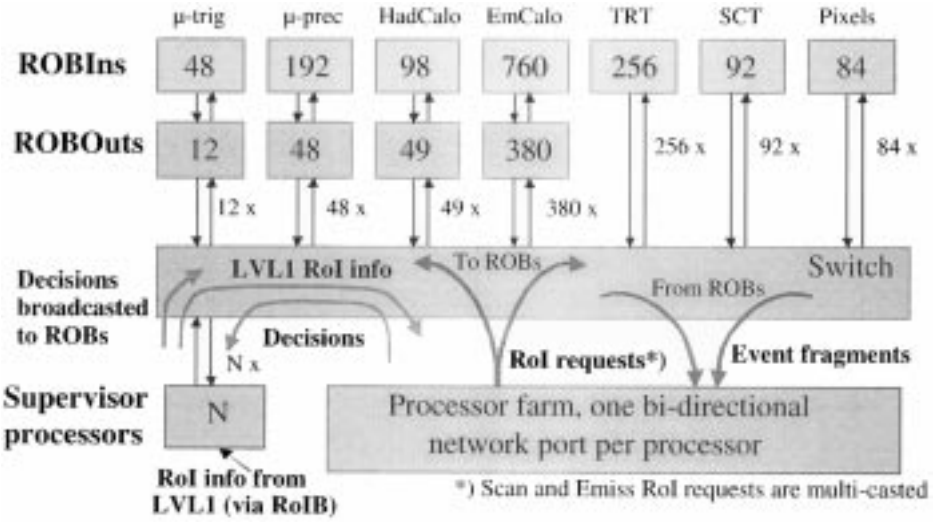

Fig. 4. The full-scale system model.

quires low communication overheads, whilst for the full scan or event building, care must be taken to resolve network congestion, minimize data loss and sustain the event rate.

Extensive tests have been made with ATM, Ethernet and SCI, beyond the scope of this short paper. The architectural, conceptual and technology studies done with ATM are described in [19] and [20]. Ethernet developments, evaluations and modeling studies are documented in [4], [21]–[24]. The SCI studies are documented in [25]–[27].

None of the technologies used the TCP/IP stack, ATM used a custom driver over ATM Adaption Layer 5, Ethernet used a custom driver with raw Ethernet frames and SCI used a simple message passing over the shared memory driver. ATM and Ethernet tests included multiswitch tests, and ATM also investigated mixing LVL2 and Event Building traffic (studying the congestion avoidance mechanisms of this technology).

In the test beds all of the technologies were able to demonstrate the level of component performance (ROBs, Supervisor and processors) required by the final trigger system. However, although SCI is becoming more widely adopted it is likely to remain in a niche market with small volumes and few sources. The most likely relevance to ATLAS is inside commercial clusters. Therefore no further studies of SCI are planned for ATLAS.

The other two technologies studied are commodity and therefore particularly interesting candidates for ATLAS. Link speeds are increasing from the 100 Mbit/s to the Gbit/s range—already a wide range of high-performance switches exists in these technologies. In both technologies the cost of a large network ($\sim$1000 ports), including switches and host adapters, appears to be within the ATLAS cost estimates.

IX. Extrapolation to a Full System

Extrapolation to the full system has been done using both paper models [6] and discrete event simulations. A discrete event model has been written in C++ [7]. In addition a separate model [28] has been developed based on Ptolemy. The simulations include sequential processing based on the LVL1 trigger menus, as used for the paper models. The behavior of the test beds has also been modeled, to obtain a deeper understanding of the test bed results and to calibrate and check the models.

The architecture of the model of a full-scale system is shown in Fig. 4. Events are generated internally in the program on the basis of a trigger menu. A realistic mapping of the detector into the ROBs is used to determine for each RoI position which ROBs should be sent data requests. Very good agreement has been found between results from the paper model and the C++ program (using average values, rather than distributions, for processing times and event-fragment sizes). The Ptolemy simulation work, which started in the last year, is expected to help with further verification of results and will also provide an environment for detailed network switch models.

A typical run of the C++ model has 1 million triggers. Results indicate that for the low-luminosity trigger (with farm processors at 80% load) requires less than 1000 fragments to be buffered in the ROBs, so a ROBin buffer memory size of 2 Mbyte would be sufficient. The peak for decision times is at a few milliseconds, with the B-physics triggers giving a second peak at around 45 ms.

A Ptolemy model of a large system with a multistage Ethernet switch (a central Gigabit Ethernet switch connected to a number of Fast Ethernet switches) has been implemented [28]. The model was calibrated using the Ethernet test bed measurements.

Models of a full-scale system indicate that: the high-p_T LVL2 triggers may be handled by a similar number (100–200) of 1000 MIPS processors at both high and low luminosity; the maximum data volume for LVL2 through the network is 20–40 Gbit/s for low luminosity (including the inner detector full scan for B-physics) and 10–25 Gbit/s for high luminosity; the TRT full scan increases the processing power and network bandwidth required in the LVL2 processors significantly. Current estimates indicate that $\sim$450 processors, each of 1000 MIPS and with an FPGA coprocessor, would be needed. Without the coprocessors the number increases to $\sim$770.

X. Conclusion

In addition to the conclusions drawn in each section it can be concluded that: The main aims of the Pilot Project have been reached; The options chosen at the end of the Demonstrator Program have been validated; The software architecture has been validated.

Choices of components and candidate technologies for LVL2 now have a large overlap with choices for DAQ/EF—Data-acquisition and Event Filter—(PCs for processors, ATM and Fast/Gigabit Ethernet for networking).

References

[1] The ATLAS Collaboration, "ATLAS DAQ, EF, LVL2 and DCS Technical Progress Report,", CERN-LHCC-98-16.

[2] E. Gamma, R. Helm, R. Jonson, and J. Vlissides, *Design Patterns*: Addison-Wesley, 1995.

[3] The Message Passing Interface (MPI) standard. [Online]. Available: http://www-unix.mcs.anl.gov/mpi/

[4] M. Boosten, R. W. Dobinson, and P. D. V. van der Stok, "Fine-grain parallel processing on commodity platforms," in *Architectures, Languages and Techniques*, B. M. Cook, Ed. Amsterdam, The Netherlands: IOS Press, pp. 263–276.

[5] A. Bogaerts, F. Giacomini, S. Gonzalez, and S. Tapprogge. (2000, Mar.) Running the ATLAS second level trigger software on a large commercial cluster. CERN, ATLAS internal note, ATL-DAQ-2000-024. [Online]. Available: http://weblib.cern.ch/Home/Internal_Notes/ATLAS_Notes/DAQ_and_Trigger/

[6] J. Bystricky and J. Vermeulen. (2000, Mar.) Paper modeling of the ATLAS LVL2 trigger system. CERN, ATLAS internal note, ATL-DAQ-2000-030. [Online]. Available: http://weblib.cern.ch/Home/Internal_Notes/ATLAS_Notes/DAQ_and_Trigger/

[7] J. Vermeulen. (2000, Mar.) Computer modeling of the ATLAS LVL2 trigger. CERN, ATLAS internal note, ATL-COM-DAQ-2000-031. [Online]. Available: http://weblib.cern.ch/Home/Internal_Notes/ATLAS_Notes/DAQ_and_Trigger/

[8] M. Abolins, R. Blair, R. Bock, A. Bogaerts, J. Dawson, and Y. Ermoline *et al.*, "Commodity multiprocessor systems in the ATLAS Level-2 Trigger," in *Conference Record IEEE NSS*, 2000, (Paper 92), pp. 12–204.

[9] R. Cranfield and J. Vermeulen. (2000, Mar.) Options for the ROB complex. CERN, ATLAS internal note, ATL-DAQ-2000-027. [Online]. Available: http://weblib.cern.ch/Home/Internal_Notes/ATLAS_Notes/DAQ_and_Trigger/

[10] H. Broterenbrood, P. Jansweijer, G. Kieft, R. Scholte, R. Slopsema, and J. Vermeulen. (2000, Mar.) A SHARC based ROB complex. CERN, ATLAS internal note, ATL-DAQ-2000-021. [Online]. Available: http://weblib.cern.ch/Home/Internal_Notes/ATLAS_Notes/DAQ_and_Trigger/

[11] G. Boorman, P. Clarke, R. Cranfield, G. Crone, B. Green, and J. Strong. (2000, Mar.) The UK ROBin, a prototype ATLAS read-out buffer input module. CERN, ATLAS internal note, ATL-DAQ-2000-013. [Online]. Available: http://weblib.cern.ch/Home/Internal_Notes/ATLAS_Notes/DAQ_and_Trigger/

[12] D. Calvet, O. Gachelin, M. Huet, P. Le Du, I. Mandjavidze, and M. Mur, "A read-out buffer prototype for ATLAS high level triggers," IEEE Trans. Nucl. Sci., vol. 48, no. 4, Aug. 2001, to be published.

[13] O. Brosch, P. Dillinger, K. Kornmesser, A. Kugel, R. Manner, and M. Sessler *et al.*, "MicroEnable—A reconfigurable FPGA coprocessor," in *Proc. 4th Workshop on Electronics for LHC Experiments*, Rome, Italy, 1998, CERN/LHCC/98-36, pp. 402–406.

[14] D. Calvet, O. Gachelin, M. Huet, and I. Mandjavidze. (1999, Nov.) A scheme of read-out organization for the ATLAS high-level triggers and DAQ based on ROB complexes. CERN, ATLAS internal note, ATL-DAQ-2000-014. [Online]. Available: http://weblib.cern.ch/Home/Internal_Notes/ATLAS_Notes/DAQ_and_Trigger/

[15] R. Blair, J. Dawson, W. Haberichter, J. Schlereth, M. Abolins, and Y. Ermoline. (1999, Dec.) A prototype RoI builder for the second level trigger of ATLAS implemented in FPGAs. CERN, ATLAS internal note, ATL-DAQ-99-016. [Online]. Available: http://weblib.cern.ch/Home/Internal_Notes/ATLAS_Notes/DAQ_and_Trigger/

[16] K. Kornmesser, T. Kuberka, A. Kugel, R. Manner, S. Ruhl, and M. Sessler *et al.*, "ATLANTIS—A hybrid approach combining the power of FPGA and RISC processors based on CompactPCI," in ACM/SIGDA Seventh International Symposium on Field Programmable Gate Arrays, Feb. 1999, [Online]. Available: http://www.acm.org/pubs/citations/proceedings/fpga/296 399/p245-kornmesser/.

[17] J. Baines, R. Bock, C. Hinkelbein, A. Kugel, R. Maenner, and M. Muller *et al.*. (1999, Mar.) Global pattern recognition in the TRT for B-physics in the ATLAS trigger. CERN, ATLAS internal note, ATL-DAQ-99-012. [Online]. Available: http://weblib.cern.ch/Home/Internal_Notes/ATLAS_Notes/DAQ_and_Trigger/

[18] C. Hinkelbein, A. Kugel, R. Manner, M. Muller, M. Sessler, H. Simmler, and H. Singpiel. (2000, Mar.) LVL2 full TRT scan feature extraction algorithm for B-physics performed on the hybrid FPGA/CPU processor system ATLANTIS: Measurement results. CERN, ATLAS internal note, ATL-DAQ-2000-012. [Online]. Available: http://weblib.cern.ch/Home/Internal_Notes/ATLAS_Notes/DAQ_and_Trigger/

[19] J. Bystricky, D. Calvet, O. Gachelin, M. Huet, P. Le Du, and I. Mandjavidze. (2000, Mar.) An integrated system for the ATLAS high level triggers; Concept, general conclusions on architecture studies, final results of prototyping with ATM. CERN, ATLAS internal note, ATL-DAQ-2000-011. [Online]. Available: http://weblib.cern.ch/Home/Internal_Notes/ATLAS_Notes/DAQ_and_Trigger/

[20] J. Bystricky, D. Calvet, M. Huet, P. Le Du, and I. Mandjavidze, "Studies of ATM for ATLAS high level triggers," *IEEE Trans. Nucl. Sci.*, vol. 48, no. 4, pp. 1318–1322, Aug. 2001.

[21] M. Boosten. Fine-grain parallel processing on a commodity platform: A solution for ATLAS. Draft thesis to be submitted to Eindhoven University of Technology. [Online]. Available: http://home.cern.ch/mdobson/mesh/M_Boosten_thesis_141 299.ps.gz

[22] R. Hughes-Jones and F. Saka. (2000, Mar.) Investigation of the performance of 100 Mbit/s and gigabit ethernet components using raw ethernet frames. CERN, ATLAS internal note, ATL-DAQ-2000-032. [Online]. Available: http://weblib.cern.ch/Home/Internal_Notes/ATLAS_Notes/DAQ_and_Trigger/

[23] K. Korcyl, F. Saka, and R. Dobinson. (2000, Mar.) Modeling ethernet switches for the ATLAS LVL2 trigger. CERN, ATLAS internal note, ATL-COM-DAQ-2000-021. [Online]. Available: http://weblib.cern.ch/Home/Internal_Notes/ATLAS_Notes/DAQ_and_Trigger/

[24] R. Dobinson, E. Knezo, M. J. LeVine, J. Lokier, B. Martin, and C. Meirosu *et al.*, "Testing and modeling ethernet switches and networks for use in ATLAS high-level triggers," *IEEE Trans. Nucl. Sci.*, vol. 48, no. 3, pp. 607–612, Aug. 2001.

[25] F. Giacomini, A. Belias, A. Bogaerts, D. Botterill, R. Hauser, R. Middleton, P. Werner, and F. Wickens. (2000, Mar.) Evaluation of commercial SCI components and low-level SCI software for the ATLAS second level trigger. CERN, ATLAS internal note, ATL-DAQ-2000-029. [Online]. Available: http://weblib.cern.ch/Home/Internal_Notes/ATLAS_Notes/DAQ_and_Trigger/

[26] A. Belias, A. Bogaerts, D. Botterill, F. Giacomini, R. Hauser, R. Middleton, P. Werner, and F. Wickens. (1999, Oct.) ATLAS LVL2 trigger SCI demonstrator evaluation report. CERN, ATLAS internal note, ATL-DAQ-2000-041. [Online]. Available: http://weblib.cern.ch/Home/Internal_Notes/ATLAS_Notes/DAQ_and_Trigger/

[27] F. Giacomini, A. Belias, A. Bogaerts, D. Botterill, R. Middleton, F. Wickens, and P. Werner. (2000, Mar.) Implementation of the message passing software layer over SCI for the ATLAS second level trigger test beds. CERN, ATLAS internal note, ATL-DAQ-2000-028. [Online]. Available: http://weblib.cern.ch/Home/Internal_Notes/ATLAS_Notes/DAQ_and_Trigger/

[28] M. Dobson, K. Korcyl, R. Hughes-Jones, P. Clarke, F. Crone, and S. Wheeler. (2000, Mar.) Ptolemy simulation of the ATLAS LVL2 trigger. CERN, ATLAS internal note, ATL-DAQ-2000-039. [Online]. Available: http://weblib.cern.ch/Home/Internal_Notes/ATLAS_Notes/DAQ_and_Trigger/