Design of the Online PC Farm for the High Level Trigger of the NA62 Experiment at CERN

J. Kunze, R. Fantechi, G. Lamanna, M. Sozzi, R. Wanke

***Abstract*—We present a highly efficient data processing framework optimized for software based triggers of fixed-target HEP experiments with continuous data collection during burst time alternating with longer out-of-burst periods.**

I. Introduction

STARTING in 2014, the fixed-target experiment NA62 at the CERN SPS will examine about 10^{13} decays of the charged K meson to precisely measure the very rare decay $K^+ \to \pi^+\nu\bar{\nu}$.

During machine cycles of about 17 s more than 10 TByte/s of raw data have to be processed at an average event rate of 10 MHz. While the first, hardware-based trigger level (L0) reduces the rate by a factor of 10, the remaining data reduction by a factor of more than 100 will be performed by an online PC farm. For this PC farm, a new concept was developed and the corresponding framework was implemented and tested.

II. The NA62 experiment

NA62 is an about 275 m long fixed-target-experiment at the CERN SPS. It is currently under construction and is designed to precisely measure the branching ratio of the very rare decay $K^+ \to \pi^+\nu\bar{\nu}$ [2]. The Standard Model of particle physics predicts the branching ratio of this decay to be of the order of 10^{-10}. The NA62 experiment is expected to collect about 100 such events in 2 years with a kaon decay rate of about 10 MHz.

During a burst a positive hadron beam with a momentum of (75 ± 1 GeV/c) is produced by a primary proton beam of 400 GeV/c coming from the SPS accelerator. The hadron beam is composed of about 6% kaons and directed into the decay region of the experiment (see Fig. 1) [1].

To detect the signal events and to veto other K^+ decays, 11 sub-detectors are positioned along the beam direction (see Table I).

Kaon decay products and beam halo background induce a rate of about 10 MHz in the sub-detectors. For each event the data of every sub-detector is read out, leading to a total data rate of about 2.3 TB/s (see Table I).

Jonas Kunze and Rainer Wanke are with the Institute of Physics, University of Mainz, Germany (e-mail: Jonas.Kunze@uni-mainz.de, Rainer.Wanke@uni-mainz.de).

Gianluca Lamanna is with CERN, Switzerland (e-mail: Gianluca.Lamanna@cern.ch).

Riccardo Fantechi and Marco Sozzi are with the department of Physics, University of Pisa, Italy (e-mail: Riccardo.Fantechi@cern.ch, Marco.Sozzi@cern.ch).

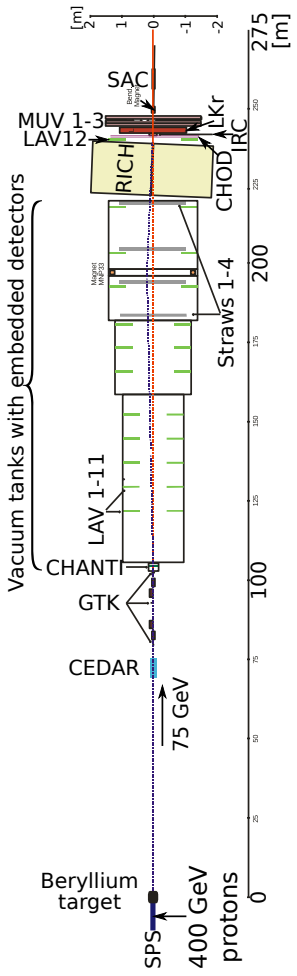

Fig. 1. Overview of the NA62 experiment [1].

III. The trigger system

The maximum data rate with which events can be written to tape is limited by 100 MB/s for the NA62 experiment. To reduce the incoming rate from the detector (2.3 TB/s) to this limit, a three level trigger system was developed:

- L0 FPGA-based readout and trigger (1 ms max. latency)
- L1 Sub-detector level trigger (1 s max. latency)
- L2 Event level trigger (12 s max. latency)

To cope with the high data rates and rigorous real time demands the first level trigger L0 is implemented in FPGA-based hardware. The higher level triggers are software architectures running on commodity processors in large PC farms.

At each positive L0 trigger decision (about 1 MHz) all sub-detectors but the Liquid Krypton calorimeter (LKr) are read out by the online PC farm software. This represents only

TABLE I
THE SUB-DETECTORS, THEIR EVENT SIZES AND THE RESULTING DATA RATES AT 10 MHZ EVENT RATE.

Sub-detector	Event size [B]	Data rate [GBps]
CEDAR	216	2.16
GTK	2250	22.50
CHANTI	192	1.92
LAV	160	1.60
STRAW	768	7.68
RICH	160	1.60
CHOD	$\ll 1000$	$\ll 10$
MUV	768	7.68
IRC & SAC	576	5.76
LKR	**222 k**	**2220**
Sum	**≈227 kB**	**≈2.3 TB/s**

about 2% of the data produced by the whole detector and hence induces a data rate of about 5 GB/s (Table I). At L1 the trigger decision is taken based on a partial event reconstruction and in case of a positive L1 trigger decision (∼100 kHz) the remaining data from the LKr detector is read out within the time period of a whole beam cycle. As the beam cycle will be more than two times longer than the proton burst (see section IV-C), the LKr data is read out with a data rate of about 7 GBps. Together with the LKr data the event building takes place and reconstruction based on the whole detector data is performed (L2). The maximum positive trigger rate by L2 is limited to about 440 Hz to meet the maximum output data rate of 100 MB/s.

To reduce the necessary amount of memory within the readout electronics, the processing time of L0 (L1[1]) is limited to 1 ms (1 s) per event while the processing time of L2 is only limited by the time between two proton bursts to reduce the number of necessary PCs.

IV. THE NA62 ONLINE PC FARM

The NA62 PC farm concept physically and logically combines the trigger levels 1 (L1, sub-detector level) and 2 (L2, event level and event building) into one single PC farm. This means that each PC can read out and process the data of every sub-detector and process L1 and L2 algorithms simultaneously. This way several advantages can be achieved compared to the more common scheme of separated L1 and L2 farms.

A. L1 event building

Instead of spreading the data of one event over several PCs (sub-component wise), the digitized raw data of a series of L0-accepted events are sent to one single farm node, which performs all L1 computations simultaneously as well as in case of a positive L1 trigger decision the event building and the L2 trigger decision.

Having the data of all sub-components (except LKr) available at one single PC a premature event building can be done (L1 event building). This allows to implement highly efficient trigger algorithms, since each event can be processed sequentially. As soon as one part of the trigger algorithm declines the event (e.g. one sub-detector produces a veto signal) the processing can be stopped for all sub-detectors. In case of distributed event processing this would not be feasible.

Additionally, due to the L1 event building, no L1 trigger decisions have to be distributed through the farm. Instead all PCs are running self-sufficiently making the farm much more scalable.

Even though it is currently planned to implement only L1 algorithms based on the data of single sub-detectors the L1 event building allows to combine several sub-components within one algorithm at an early stage of the trigger chain.

[1]The L1 processing time is limited to reduce the time until the LKr data is read out which decreases the necessary memory of the LKr readout electronics.

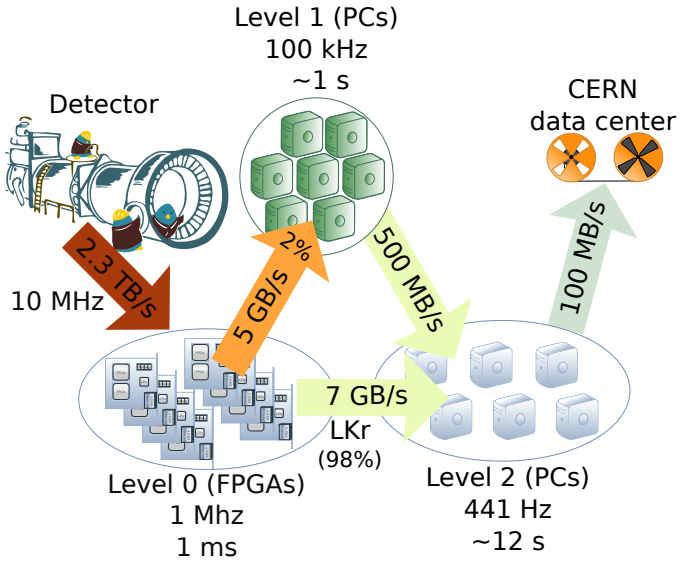

Fig. 2. The logical dataflow from the electronics over the three trigger levels to the tapes.

B. High scalability and maintainability

As the combined L1/L2 PC farm is completely homogeneous with respect to the software, all PCs are optimally used and the system is easily scalable by simple addition of more computing power. Therefore almost no assumptions on the performance of the L1 algorithms and PCs need to be made.

Due to the combination of both trigger levels only one software framework has to be implemented. Also the installation of the PCs is facilitated as only one farm has to be maintained.

C. High efficiency

Within the NA62 experiment data is only collected during limited burst periods ($T_b \approx 5$ s). Between two bursts the out-of-burst period will last about $T_{oob} \approx 12$ s[2].

Due to the negligible latency at L0 (1 ms), the L1 trigger will only read out data during the burst. As the L1 computation is limited to one second a separate L1 PC farm would only be processing data for $T_b + 1 \approx 6$ s and stay idle during the

[2]These numbers will vary during the data taking period. The given values are the expected averages.

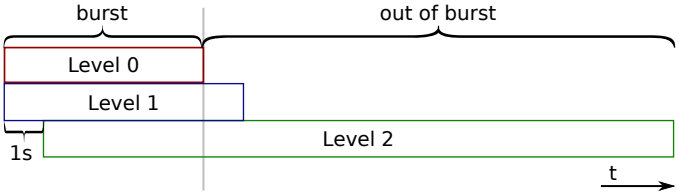

Fig. 3. Processing times of the three different trigger Levels with separate L1/L2 PC farms.

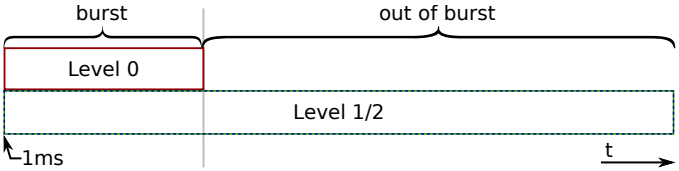

Fig. 4. Processing times of the three different trigger Levels with a combined L1/L2 PC farm.

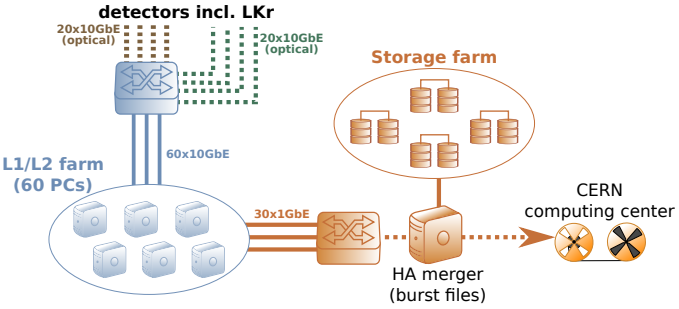

Fig. 5. The final topology of the NA62 PC farm.

remaining time of a burst cycle, namely $T_{oob} - 1 \approx 11$ s (see Fig. 3). This makes the concept with separate L1/L2 farms very inefficient at fixed-target experiments.

By combining the L1 and L2 PC farms all PCs can be optimally used during the whole burst cycle (see Fig. 4). At the NA62 experiment a separate L1 farm would have only been used about 55% of the time while only the L2 farm would have been used efficiently. Estimations show that by joining the L1 and L2 PC farms to one single homogeneous farm from 10% up to 25% of the whole computing power can be saved here.

D. The topology

In the final concept the L1/L2 farm PCs are connected via 10 Gb Ethernet through one 96 port router to the L0 read out electronics. These electronic boards send bunches of digitized raw data of a series of L0-accepted events to one single farm node (see section IV-A). Each bunch is sent to a different PC on a Round Robin basis. Events being accepted by L1 and L2 are sent to the high available merger PC which sorts the events by time and creates files with all accepted events of one proton burst. These files are sent to a local buffering disk pool and to a tape library at the CERN computing center (see figure 5).

E. Drawbacks

To meet the time limit of the L1 computation the software framework within the combined PC farm needs to ensure that the L2 processing does not overload the farm during the burst. This has been implemented via a simple load balancing algorithm which prioritises the L1 computation. Tests have shown that the average L1 computation time has become independent of the number of events being processed by L2 due to this load balancing.

As described in section IV-D all electronic boards (except the LKr readout system) send the data of a bunch of events to the same PC within a very short time. This could potentially overload the main router if it cannot buffer the short data peak going to one port (PC). At NA62 it is planned to send bunches of about 10 Events (5 kB each) to one PC. Hence, the buffer size of the main router must be at the order of 50 kB per port which is small compared to the typicall buffer sizes of at least the order of 100 kB per Port at modern Routers. Furthermore, problems related to that unsteady data transmission can be solved by implementing short delays of different lengths within the electronic boads.

V. The Software Framework

To receive the raw data from the sub-detectors, to perform the event building, to initiate the trigger algorithms, and to send the accepted events to the data storage a C++ based framework was implemented for the NA62 PC farm which is running on Linux. Tests have shown that using the standard Linux kernel sockets a relative packet loss of less than 10^{-5} at full data rate and high load using the UDP/IP protocol was not achievable. Therefore the network communications within the framework are based on a special socket called *pf_ring DNA* implemented by the company ntop [3]. Using the *pf_ring DNA* network driver the data packets are transmitted directly to the userland memory via direct memory access (DMA). This way the full data rate of 10 Gb/s can be received without any packet loss and almost no CPU load at any packet size.

As *pf_ring DNA* does not yet implement any layer of the UDP/IP protocol stack, Ethernet, IP, UDP and ARP had to be implemented by the authors. The resulting framework was tested on a Dell PowerEdge R710 with two Intel X5675 processors (12 cores and 24 threads total) and 24 GB memory with a clock frequency of 1333 MHz. The data reception (10 Gb/s), integrity checks, event building, and the transmission (about 0.1 Gb/s) altogether took less than 25% of the whole processing power. Thus, more than 75% of the processing power of each PC within the NA62 online PC farm can be used to compute L1 and L2 trigger algorithms.

References

[1] F. Hahn et al., *Technical Design Document* NA62 Collaboration, 2010, http://na62.web.cern.ch/na62/Documents/TD_Full_doc_v10.pdf
[2] G. Anelli et al., *Proposal to measure $K^+ \to \pi\nu\bar{\nu}$ rare decay at the CERN SPS* NA62 Collaboration, CERN-SPSC-2005-013
[3] http://www.ntop.org/products/pf_ring/dna/