The NA48[1] Data Acquisition System

F. Bal[2], P. Brodier-Yourstone[2], O. Boyle[3], S. Galagedera[3], A. Gianoli[2], A. Lacourt[2], S. Luitz[2], J.P. Matheys[2], N. Mckay[3], B. Panzer-Steindel[2], H. Parsons[3], K. Peach[3], B. Segal[2], G. Wirrer[2], M. Wittgen[4]
[2]CERN,CH-1211 Geneva 23,Switzerland
[3]Dept. of Physics, University of Edinburgh , JCMB, Mayfield Road, EH9 3JZ, UK
[4]Institut für Physik, Universität Mainz, D-55099 Mainz, Germany

Abstract

NA48 is a CP violation experiment that is starting its period of data taking at the CERN SPS accelerator. The expected average data rate of a few thousand events/*s* (with the possibility to go up to 10 kHz), a data volume up to 100 MB/*s* sustained over the beam burst, and the necessity to handle different sub-systems with very wide variation in the number of channels have lead to the design of a modular pipelined data acquisition system with a sampling rate of 25 *ns*.

I. Introduction

The NA48 Collaboration faces data-taking problems such as: up to 100 MB/*s* sustained data throughput and event rates up to 10 kHz during bursts[2]; the need to assemble events in real-time out of sub-events generated by over a dozen sources; stringent data processing requirements to use the online third-level event filtering; and data archival capacity in excess of 20 TB per year (this last request can increase if the data selection is not as efficient as expected).

To handle this situation a solution based on commercial and custom-designed components has been developed. It includes: custom-designed fiber-optic links to transfer sub-event data from the sub-detectors to the event builder; a custom-designed backplane used with additional arbitration logic to merge the sub-events into events; commercial HiPPI links and a HiPPI switch to send a "spill-worth" of assembled events to one of several workstations; custom-designed HiPPI to TURBOChannel interfaces capable of sustained data transfers in excess of 83 MB/*s* into workstation memory; FDDI links and concentrators to handle the data transfer from the experimental area to the computer center where a parallel processing Meiko CS-2 multi-node system is used to perform the third-level filtering, data archival and, later, the full data analysis.

II. Requirements

The aim of the NA48 experiment at CERN is to measure the direct CP violation parameter ε'/ε with a precision of 2×10^{-4}[1], by comparing the relative decay rates of long- and short-lived kaon beams into two neutral and charged pions. To achieve this the experiment measures all four decays concurrently using two simultaneous and nearly collinear beams.

In order to achieve this precision an intense kaon beam is required; this is expected to produce an average trigger rate of 3 kHz during bursts - taking into account calibration and monitoring triggers. To allow future expansion, the specification requires that the system run with negligible dead time up to 10 kHz. The maximum expected data volume can be calculated assuming a total of 10,000 channels at 10 bytes/channel and a 10% occupancy, leading to 10 kB/event. The maximum data bandwidth required is therefore: 10 kB/event × 10,000 events/*s* or 100 MB/*s* during bursts. Depending on the detector performance, out-of-burst calibration events may be needed, with a trigger rate of up to 1 kHz. Of course there will be significant online reduction of data recorded but, nevertheless, the experiment will eventually produce 20-60 TB per year.

III. Design philosophy

It is perhaps easier to understand the implementation if the underlying philosophy of the complete scheme is understood. The experiment has three basic constraints. Firstly, the possibility to have out-of-burst calibration events forces us to assemble the events in real-time. Secondly, at the detector level, the system is driven by a common 40 MHz (25 *ns*) clock; associated memories are therefore fast and consequently expensive static RAMs are used. Likewise, when the event is assembled from the components (sub-events) distributed over the various data sources (sub-detectors) the total volume of up to 100 MB (25 Mwords) each second also requires fast (40 *ns*) memories, which again must be SRAMs. But in the assembling stage, assuming that the data are reduced (zero-suppressed) before reaching the sub-event buffer, the volume of data from individual sub-detectors (or parts of sub-detectors) are modest; there is no need for big memory buffers and the data transfer between the sub-event buffers and the data merging engine may be addressed with conventional backplane technology.

The architecture therefore consists of simple point-to-point links from sub-detectors to FIFO memories. There is no horizontal linkage between the transfers from different sources, since such linkage necessarily creates dead-time. Instead, each link is independently controlled by an XOFF protocol.

Each source continues to send data at the maximum rate,

[1]The NA48 Collaboration: Cagliari, Cambridge, CERN, Dubna, Edinburgh, Ferrara, Firenze, Mainz, Orsay, Perugia, Pisa, Saclay, Siegen, Torino, Vienna, Warsaw.

[2]The CERN SPS accelerator delivers protons to the fixed target program in bursts of 2.5 *s* every 14.4 *s*

0018-9499/98$10.00

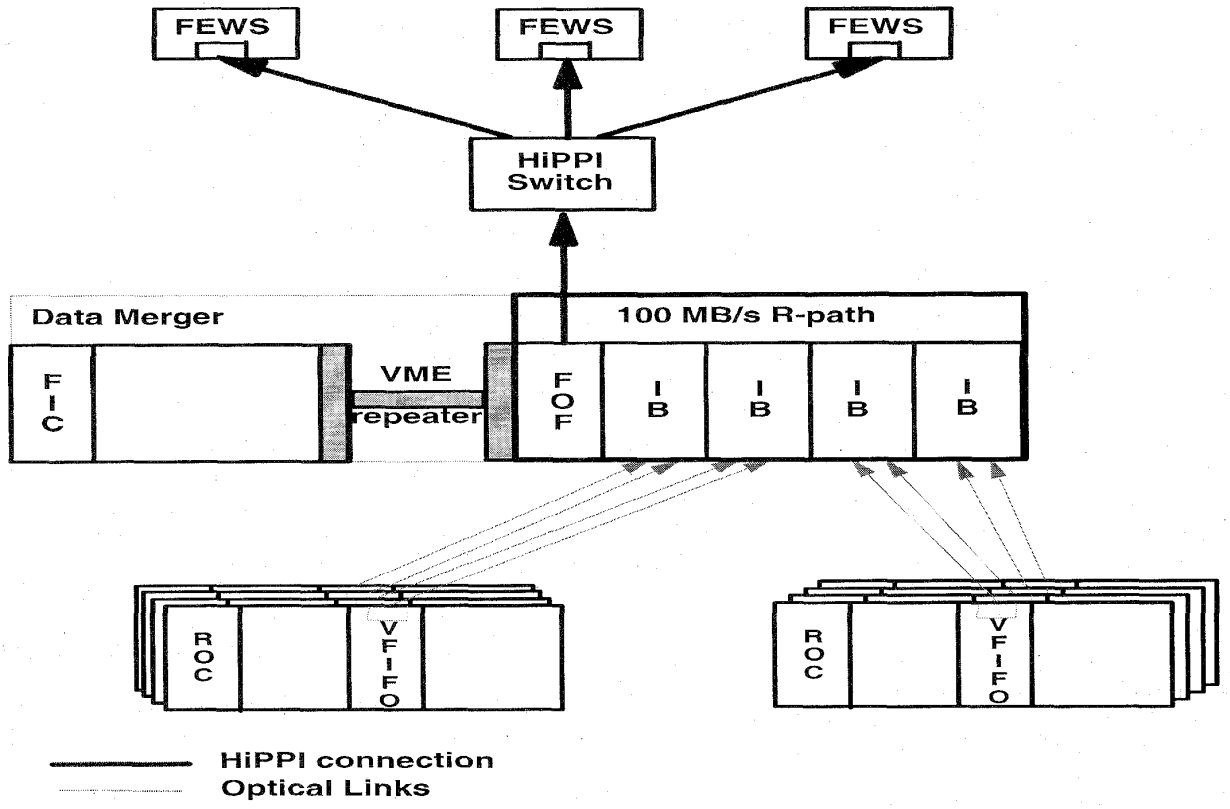

Fig. 1 : The NA48 data acquisition system architecture

until inhibited from the destination by a signal XOFF; it is the duty of the destination to guarantee to be able to accept all data in transit between the issuing of the XOFF signal and the source acting thereon.

The Trigger Supervisor is responsible for ensuring that the overall rate at which triggers are issued does not exceed the capacity of the system, so that the XOFF signal can never propagate back to the front-end pipelines; receipt of the XOFF signal at the front-end pipeline is to be treated as an error condition.

IV. System Implementation

A. The Trigger System

The readout electronics, even if the implementation differs from one sub-detector to another, is based on a pipeline structure, to fulfill the requirements of zero dead time and to withstand the expected rate. The pipeline is a digital one, 204.8 μs deep circular buffer, driven by the common 40 MHz clock. The trigger system also works for the most part in a pipelined mode: the neutral trigger system[2] continuously evaluates energy sums, number and time consistency of energy peaks and decay vertex position based on transverse projections of the electro-magnetic calorimeter, providing a response every 25 *ns*. An array of fast processors analyses candidate charged events as indicated by the fast trigger logic, and calculates the longitudinal decay vertex position and the particle pair invariant mass[3].

The final trigger decision is taken by a central pipelined device (Trigger Supervisor[4]) which synchronizes and combines the information and, in case of good events, sends within 200 μs to each sub-detector Read-Out Controller (ROC) a trigger word which contains the trigger type and a "timestamp" i.e. the time in 25 *ns* units from the start of burst.

B. The Event builder

When the readout electronics of each sub-detector receives a timestamp, it reads a predefined number of samples from the pipeline and sends them via an optical link to the Data Merger. The system is organized as a series of data-push links arranged around data-merging and data-distribution components. A block diagram of the architecture of the data flow system is shown in Figure 1.

The sub-detectors typically use VME or Fastbus components for their electronics. Upon arrival of a trigger,

each sub-detector Read-Out Controller (ROC) selects data pertaining to that time window and passes it in the form of a sub-event to the data flow system's first link, the Optical Link.

The Optical Links (OL) extend over 200 m, and use FiberChannel components and custom designed boards (called VFIFOs). Each link is capable of sustained data transfers up to 10 MB/*s*. This link was developed specifically for the experiment and is described in detail elsewhere[5].

The OL transfers the data into an Input Buffer (IB), which is a large format (9U×400 *mm*) VME board, that receives the sub-events and stores them in a 2 MB FIFO buffer where they await readout. This buffer is needed to cope with the different sub-detector data handling speeds and latencies. Each board contains two IB channels and there is one channel per OL.

Thus shortly after a trigger has been issued, the IBs contain a set of sub-events that together make up an event. The IBs signal that they have data to a controller, known as the FIFO Output Formatter (FOF), over a custom backplane known as the RPATH. This uses FUTUREBUS+ connectors and bus drivers and is designed to operate at 100 MB/*s*. The FOF issues a token to the first IB which takes control of the RPATH and sends its data to the FOF. The first IB then passes the token to the second IB in the chain, and so on. When all IBs have been read out, the token returns to the FOF. In this way the sub-events are concatenated into an event. The FOF adds a header and a trailer to each event and sends it, via a HiPPI link to the next stage, a 4×4 cross-bar HiPPI switch.

The set of modules comprising the IBs, the RPATH, and the FOF is the Data Merger, and is fully described in [6]. The Data Merger is housed in two 21-slot VME crates. Crate 1 is a 6U type and contains a FIC8234 single-board computer[7] which acts as the Data Merger Controller (DMC). This crate also contains a Bit-3 repeater output board[8] which allows the DMC to access the second crate. Crate 2 is fitted with a standard VME backplane in positions J1 and J2; the RPATH occupies the J3 position. Slot 1 is occupied by the Read Out Latch Enable Transmitter (ROLEX) that generates the clock signals which control the data transfer on the RPATH bus. Slot 2 is occupied by the slave board of the Bit-3 repeater. Slots 3-20 are available for IB boards and slot 21 is occupied by the FOF. The number of IBs read can be changed by software.

C. The FrontEnd Workstations

The HiPPI switch is set up at the beginning of each burst by the FOF to select one out of several Front End Workstations (FEWS) as output for that burst.

The FEWS are TURBOChannel based DEC 3000/900 workstations, each of them equipped with a 275 MHz ALPHA processor, 384 MB of main memory and 12 GB of fast disks. In addition each FEWS contains the HiPPI-to-TURBOChannel interface[9], developed in co-operation with Digital, two FDDI network interfaces and an additional dual port SCSI host adaptor. The disks are combined into striped logical volumes (RAID0) giving a throughput of about 8 MB/*s* for concurrent reading and writing of independent files. Each FDDI interface allows a sustained data transfer rate of about 10 MB/*s* using TCP/IP.

The HiPPI-to-TURBOChannel was developed for maximum throughput and has been shown to be capable of a sustained data rate of at least 83 MB/*s*. The interface consists of two boards: a high-speed (up to 100 MB/*s*) TURBOChannel option board, and a HiPPI destination board. A fast FIFO is used to receive the data from the link. This FIFO is emptied into the workstation memory using DMA transfers. For optimum performance the DMA transfers use a scatter-gather table to translate logical addresses into physical ones. This table (which allows transfers of up to 512 MB) is loaded before the data transfer by the driver software, according to the application's requirements. Data transfers are thus not hampered by any software intervention.

All FEWS have an ethernet connection to the local NA48 network for login and file sharing. The high speed networks are reserved for the transfer of data and monitoring information. To overcome the limitations of a single FDDI link, a serial HiPPI connection (100 MB/*s*) is used between the NA48 counting room and the computer center. Since for the FEWS no software for IP over HiPPI exists, 4 FDDI networks (one for each FEW) are multiplexed onto a serial HiPPI link to the computer center using a GigaRouter[10]. In the computer center a second GigaRouter demultiplexes the HiPPI link into 4 separate FDDI rings which are connected to dedicated interfaces on several nodes of the CERN Central Data Recording and Processing (CDRP) system, a Meiko CS-2 computer[11]. The other FDDI interface of each FEWS are connected to a common FDDI ring which is part of the CERN central data processing FDDI network and it is used for exchange of monitoring data between the FEWS and monitoring workstations as well as a fall-back link to the central data recording in case the HiPPI-FDDI network fails.

The data from each burst are stored in the workstation memory as they arrive during the burst, and then copied to a disk file whose size can be as big as 256 MB, depending on the number of triggers received during the burst and the number of sub-detectors read out. During this year's run NA48 is using four FEWS; the HiPPI switch can be easily upgraded to an 8×8 configuration if more FEWS are needed to cope with the data volume.

As soon as a burst is successfully transferred into the FEWS, the workstation starts writing it to a local disk, and sends a message to a process running on the Meiko CS-2. The CS2 computer is a distributed memory scalable parallel system. The CERN CS2 has 64 nodes, each with two 100 MHz HyperSPARC processors, 128 MB of memory and 1 GB disk each, and it is connected to a disk farm with a capacity of over 1 TB. Upon arrival of the message, the CS-2 starts a process which copies the file containing the burst to parallel filesystems. Normally the file is sent by the FEWS while it is copying it to disk, thus the request can be satisfied from the buffer cache of the FEWS and do not require real disk activity. After the successful arrival of the data on the CS-2, autonomous daemons start the transfer of the data to magnetic

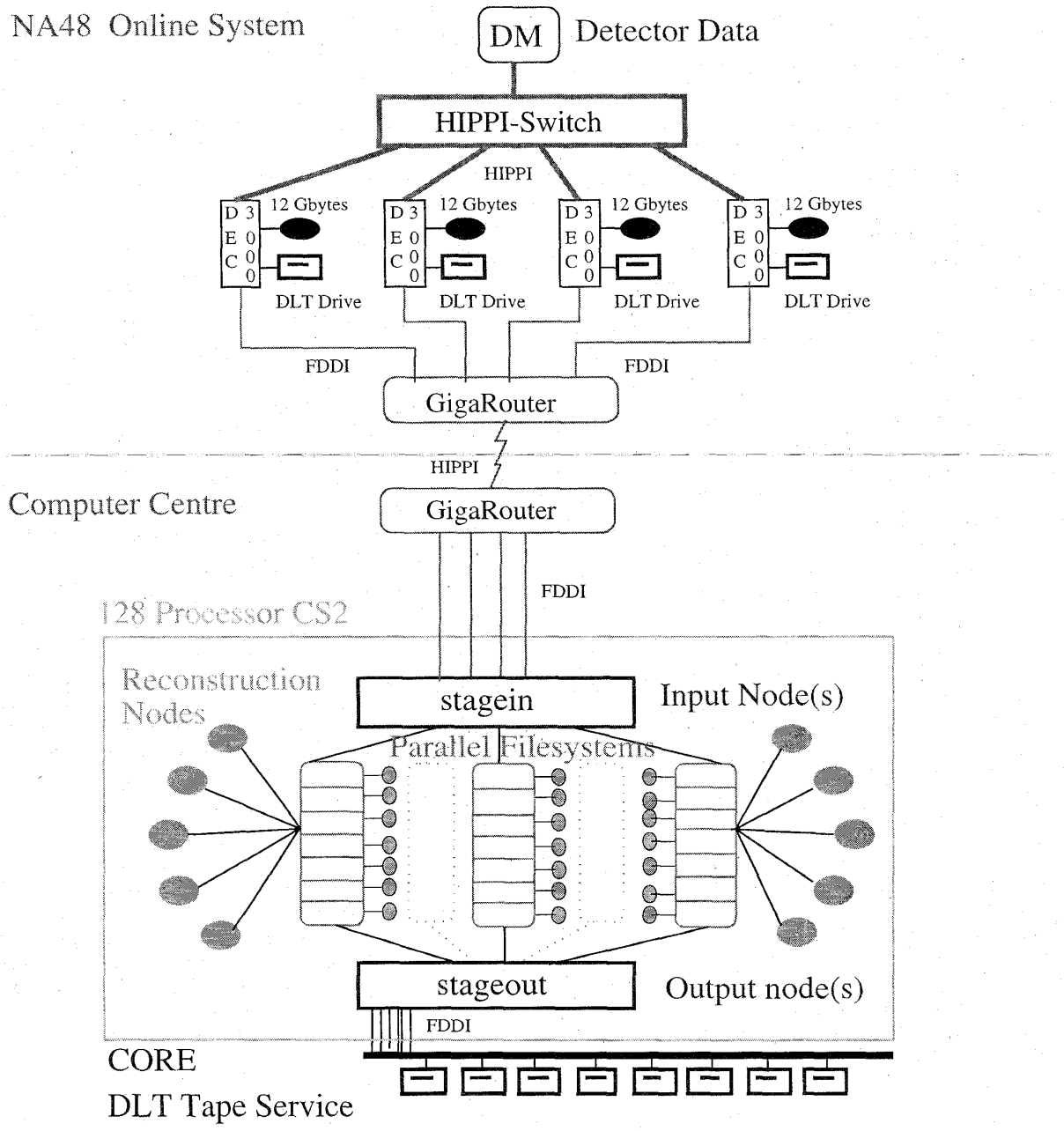

Fig. 2 : The Central Data Recording and Processing system.

tape. Dependent on the transfer time several transfers from different FEWS might be active simultaneously. The maximum aggregate measured via one FDDI link was between 8 and 10 MB/*s*. For the copying task as many links as the number of Fews can be used (up to a maximum of 8).

The third-level filtering is performed on this machine, using a number of CPUs: the number of CPUs used for this task can vary depending on the data load and the type of analysis wanted (up to 64 CPUs are reserved for the experiment). The filtering also splits the data in different streams depending on the event type. Finally an automatic process writes the files on tape, relying on parallel tape streams to provide the necessary capacity. The CDRP has a fully scalable architecture, ensuring that one can cope with the required data volume, processing and

archival. Figure 2 shows the data flow from the Data Merger to the CDRP.

V. Conclusions and Performance

To summarize, the data flow is hardware-based and data-driven up to the FEWS farm. The experiment uses custom-designed components for the OL, the Data Merger and the HiPPI to TURBOChannel interface. From the OL to the FOF the data rate is controlled by use of an $\overline{\text{XOFF}}$ signal. At the level of the FEWS and beyond, commercial hardware is used. We have the possibility to upgrade and expand this area of the system as new technology becomes available.

During this year's run the system is used to collect sub-events from 15 OLs, using 4 Fews, and 4 FDDI links and up to 12 tape streams. The average event size has turned out to be 14 to 16 kB, bigger than the value used to design the system. For this reason the system has never reached a trigger rate of 10 kHz with all links active. Moreover the HiPPI-to-TURBOChannel interface has shown its limit at around 90 MB/*s* sustained transfer rate. For these reasons the system has run with an average trigger rate of 5-6 kHz during bursts, equivalent to 170-190 MB/burst. The detector and its electronics have proved to be stable and there was no need for out-of-burst calibration events. More than 100,000 bursts have been recorded, representing a dataset of more than 20 TB.

VI. Acknowledgements

We would like to thank our colleagues of the NA48 Collaboration for their efforts and their help in the construction of the system and during the data taking period.

The Mainz group was supported in part by the German Federal Minister for Research and Technology (BMBF) under contract 7MZ18P(4)-TP2. The Edinburgh group thanks the UK Particle Physics and Astronomy Research Council for financial support.

VII. References

[1] G. D. Barr et al., "Proposal for a precision measurement of ϵ'/ϵ in CP violating $K^0 \to 2\pi$ decays", CERN/SPSC/90-22/P253.

[2] G.D. Barr et al., "A fully software-programmable pipelined trigger-processing module", *IEEE Trans. Nucl. Sci.*, vol. 43, pp. 1689-1694, 1996.
C. Avanzini et al., "The PeakSum Processing System for the NA48 esperiment", *IEEE Trans. Nucl. Sci.*, vol. 43, pp. 1694, 1996.
B. Gorini, "A 40 MHz Pipelined trigger for $K^0 \to 2\pi$ Decays for the CERN NA48 Experiment", this conference.

[3] M. Mur et al., "The charged trigger system of the NA48 CERN-SPS experiment", this conference.

[4] F. Bertolino et al., "The NA48 Trigger Supervisor Design", *IEEE Trans. Nucl. Sci.*, vol. 41, pp. 274-279, 1994.

[5] P. Brodier-Yourstone et al., "A 10 Mbytes/s fibre optic link", *IEEE Trans. Nucl. Sci.* vol. 42, pp 870, 1995.

[6] O. Boyle et al., Conference Record, IEEE Nuclear Science Symposium and Medical Imaging Conference, Norfolk, Virginia USA, 1994, vol. 2, pp. 991-993.

[7] Creative Electronic Systems S.A., 70 r.te du Pont-Butin, P.O. Box 107, CH-1213 Petit-Lancy 1, Geneva, Switzerland.

[8] VME bus repeater mod. 418 made by SBS Bit 3 Operations, 1284 corporate Center Drive, St. Paul, MN 55121-1245, USA.

[9] W. Bozzoli et al., "Data Transfer and Distribution at 70 MB/s", Conference Record, IEEE RT93 Conference, Vancouver, Canada, 1993, pp. 105-107.

[10] NetStar Inc., now part of Ascend Communications Inc., One Ascend Plaza, 1701 Harbour Bay Parkway, Alameda, CA 94502, USA.

[11] Meiko Ltd., University Gate, Park Row, Bristol BS1 5UB, UK.