The NA48[1] Event-Building PC Farm

M. Wittgen[2], A. Peters[2], P. Marouelli[2] S. Luitz[3‡],F. Bal[3], O. Boyle[3], A. Gianoli[3†], A. Lacourt[3], B. Panzer[3], O. Vossnack[3].

[2]Institut für Physik, Universität Mainz, D-55099 Mainz, Germany.

[3]CERN, CH-1211 Geneva 23, Switzerland.

[‡]Present address: SLAC, Stanford, CA 94309, USA.

[†]Present address: Sezione dell'INFN di Ferrara, I-44100 Ferrara, Italy.

Abstract

The NA48 experiment at the CERN SPS aims to measure the parameter $Re(\varepsilon'/\varepsilon)$ of direct CP violation in the neutral kaon system with an accuracy of 2×10^{-4}. Based on the requirements of:

- high event rates (up to 10 kHz) with negligible dead time
- support for a variety of detectors with very wide variation in the number of readout channels
- data rates of up to 150 MByte/s sustained over the beam burst
- level-3 filtering and remote data logging in the CERN computer center

the collaboration has designed and built a modular pipelined data flow system with 40 MHz sampling rate. The architecture combines custom-designed components with commercially available hardware for cost effectiveness and flexibility.

To increase the available data bandwidth and to add filtering and monitoring capabilities, the original custom-built event builder hardware has been replaced by a farm of 24 Intel PentiumII based PCs running the Linux operating system during the shutdown between the 1997 and 1998 data taking periods. During the data taking period 1998 the system has been successfully operated taking ca. 70 Terabyte of data.

I. Introduction

The NA48 Collaboration has built a detector to measure the direct CP violating parameter ε'/ε in neutral kaon decays. The experiment faces data-taking challenges such as:

- high sustained data throughput from the detector during beam bursts
- support of various sub-detectors with a very wide variation in the number of channels and amount of data
- the need to assemble all events in near-real time to ensure data consistency

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

Before the 1998 running period an online PC farm has been deployed to upgrade the capacity of the data flow system and to improve filter and monitoring capabilities. This farm completely replaces the hardware event builder ("Data Merger") and the Front-End-Workstations system previously used by the experiment[1].

The system includes custom designed components for the links between sub-detector electronics and the farm PCs as well as standard commercially available hardware for the PC themselves, their interconnecting network and the high speed link that provides the data transfer from the experimental area to the laboratory computer center where further selection and data archival are performed.

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}, by comparing the relative decay rates of long- and short-lived kaon beams into two neutral and charged pions, respectively. To minimize the systematic error the experiment records all four decays concurrently using two simultaneous and nearly collinear beams[2].

In order to collect enough statistics in a running time of three years an intense kaon beam is required which generates an average readout trigger rate of approx. 7 kHz during bursts including monitoring and calibration triggers. To leave room for future expansion, the design allows up to 10 kHz readout rate with negligible dead time. An average event size of 15 kByte thus requires a maximum data bandwidth of 150 MByte/s. But even 7 kHz trigger rate, 15 kByte event size and 120 days/year running time with a 50 percent overall efficiency of the accelerator and detector operation result in 75 Terabyte/year of raw data, posing a challenge to online and offline reconstruction as well as data handling and archival.

III. Design Overview

The SPS accelerator at CERN delivers protons to the NA48 targets in spills of 2.5 s length every 14.4 s. Additional calibration and guard time periods expand the effective spill length seen by the data acquisition system to 5.0 s, which leaves a 9.4 s gap between two bursts. This duty cycle allows a natural split between data taking and event building: during the burst all sub-detectors send their event fragment streams into the event builder that accumulate them into buffers. The peak

data rate from all sub-detectors during the burst can reach 150 MByte/s. The inter-burst gap is then used for event building.

From the beginning, the NA48 Data Acquisition system has been designed as a modular, scalable data flow architecture based on custom-designed and commercial components. It is organized as a series of data-push links connected to an event builder which merges event fragments from the individual sub-detectors into complete events. Each source continues to send data at maximum rate 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.

An important feature of the experiment is that all channels are sampled continuously synchronized by a global clock. Typically, the sub-detectors use custom designed VME or Fastbus components for this purpose. The trigger circuitry detects desired events and issues a trigger which results in a particular time window being selected. Each sub-detectors' readout controller reads the data belonging to that time window and passes it in form of an event fragment to the event builder.

At the level of the PC farm and beyond, only commercial hardware is used. This minimizes design and maintenance effort, improves reliability and allows for easy upgrading as new technology becomes available. The PC farm itself scales within the limits of the switch backplane bandwidth.

It should be noted that the data flow system does not modify data with the exception of simple reformatting and that the pathways for control and data are completely separate. Since core data flow only handles fast data transport, traditional data acquisition tasks like detector control, configuration management and error reporting are handled by additional subsystems which are loosely coupled to data flow. Each sub-detector readout system is controlled by a sub-detector computer which configures and monitors data flow hardware and embedded processors but does not participate in data flow itself.

IV. The DT16/PCI interface

The link from the sub-detectors to the PC farm is the only part in the PC farm for which custom-made interfaces are used. During the planning of the system the collaboration had the occasion of testing the S-LINK[3], an high speed link currently being developed at CERN for LHC experiments. The S-Link is a new concept that should provide the benefits of standardization without the limitation of choosing one technology for the physical link or the overhead of conventional network protocols.

In fact the S-Link specification does not define the physical layer of the link, but a simple FIFO-like user interface on which the use of the signals remains independent of the technology used to implement the physical link. The mapping of the S-Link signals to the protocol used on the physical link is left open to the link designer, allowing to map it in the most suitable way onto the underlying link technology.

Several types of link source and destination cards already exist or are under development which use a variety of technologies like FibreChannel, OPTOBUS, MATCH, etc., for data transfer speeds between 100 and 160 MByte/s.

Due to the very tight time schedule for the upgrade of the NA48 event builder and to the different need for data transfer speed (the link used to transfer the data to the old event builder had a speed of 10 MByte/s), the collaboration decided to adopt a lower performance technology for the physical link. The choice has been the DT16-bus, an adapted version of the DT32-bus originally developed for the Eurogam Project[4].

The DT16-bus is a 16-bit wide parallel bus implemented with differential ECL signals over a 20 pair twisted-pair cable. The maximum speed is 33 MByte/s, transferring 1 word of 16 bits every 60 nsec. Like the original DT32, it is possible to have multiple sources on a single DT16 bus, arbitration being handled by a simple daisy chain mechanism. This allows the read-out of multiple data sources with a single PC.

To use DT16 to connect sub-detectors to PCs, two custom-made boards have been developed. On the PC side, a DT16-to-SLink (DT2SL[5]) board has been designed. This board is a mezzanine board which fits on an existing S-Link-to-PCI interface[6], and acts as a master on the DT bus. On the detector side a DT16-I/O board has been used in conjunction with the Data RIO modules used by all sub-detectors (DRIO modules based on CES RIO8260[7] with custom I/O interfaces[8]).

The driver for the DT16/PCI board has been written by the collaboration based on a similar development [9]. It is a polling user-space zero-copy driver which maps the S-LINK to PCI card into memory and then sets up the DMA transfers from the card into the PC memory. This requires no interrupts and a data rate of 117 MByte/s has been achieved with an infinite packet size, i.e. the S-LINK card was sending a test pattern of data continuously with no gaps for protocol.

V. The PC farm

All machines in the farm are industry-standard PCs equipped with Intel PentiumII processors. The operating system is Linux. To simplify the management of the farm, a standard RedHat 5.0 [10] distribution has been modified to support BOOTP/TFTP booting and diskless operation over NFS.

Currently there are 24 PCs: 11 PCs are connected to the sub-detectors (Sub-Detector PCs or SDPC), 8 PCs do the event building (Event Building PC or EBPC), 4 PCs are responsible for the IP routing to the Central Data Recording facility of the laboratory (CDRPC) and 1 PC acts as boot/file server and farm controller.

The 11 SDPCs are single processor machines (266 MHz PentiumII) equipped with 128 MByte RAM for data buffering, a DT16/PCI interface card and a Fast (100 Mbit/s) Ethernet adapter. The 8 EBPCs are dual-processor machines, with 192 MByte RAM, 18 GByte SCSI hard disks and a Fast Ethernet card. The 4 CDRPC have 128 MByte RAM and a Fast Ethernet card as well as an FDDI adapter. The interconnection

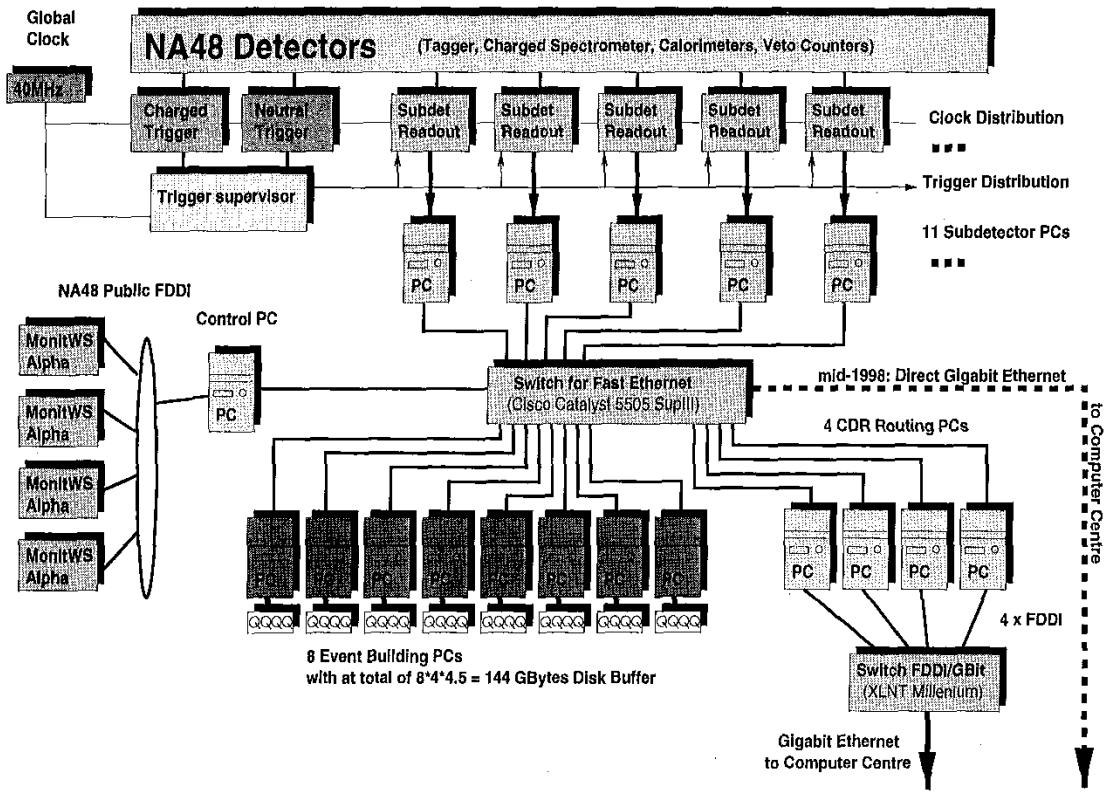

Figure 1: : Block diagram of the connections between the detector, the PCfarm and the Central Data Recording facility.

between the PCs is handled by a Catalyst 5505/SupervisorIII Fast Ethernet switch made by Cisco[11] with a single 24-port 100 Mbit/s module providing a total bandwidth of 1.2Gbit/s for event building and data recording. The switch is of a *store-and-forward* type which means that incoming packets are stored in a buffer at the input and then forwarded to the required output port when the backplane is available. The technology used to connect the PCs to the switch is 100 Mbit/s Ethernet running over unshielded twisted pair cables (100baseTX). Very long connections are implemented using fiber-optic repeaters. Figure 1 shows a scheme of the connections.

During an SPS burst the SDPCs simply receive data through the DT16/PCI interface and store them in memory. After the burst has completed, each SDPC checks the received data block for consistency and sends the number of event fragments received to the Farm Control Program (FCP). The FCP then partitions the event into M blocks, each block being assigned to an EBPC. A dynamic load balancing algorithm equalizes the data load on the EBPCs.

All SDPCs then distribute their data to the EBPCs according to the FCP's partition decision by means of sender processes. Each sender process maintains a logical connection to a receiver process on an EBPC. If there are N SDPCs and M EBPCs, this results in NxM logical connections. TCP/IP is used as protocol, since it handles all issues of flow control and ensures data integrity and completeness. Note that at this stage the event structure is invisible and there is no correspondence between IP packets and event fragments.

When a receiver gets data, it stores them in memory. An event builder task running in every EBPC searches through the received data blocks and pulls out a list of pointers for each event which identifies its component fragments. The pointers are stored in a pointer table in shared memory. After the event building stage, it is then possible to apply further data integrity checks or fast filter algorithms to the data stored in memory. A disk writing process then takes sets of complete pointers from the table and writes the complete events to the local hard disks. Each EBPC writes its share of the burst data. During the event building and processing the chronological order of the events is retained.

Central Data Recording processes then move completed disk files to the computer center. Data are simply sent back through the switch (again using TCP/IP over Fast Ethernet, FDDI and Gigabit Ethernet) to the computer center which is ca. 7km away from the experimental area. As soon as a burst fragment file has been transferred successfully it is deleted from the EBPC's disk

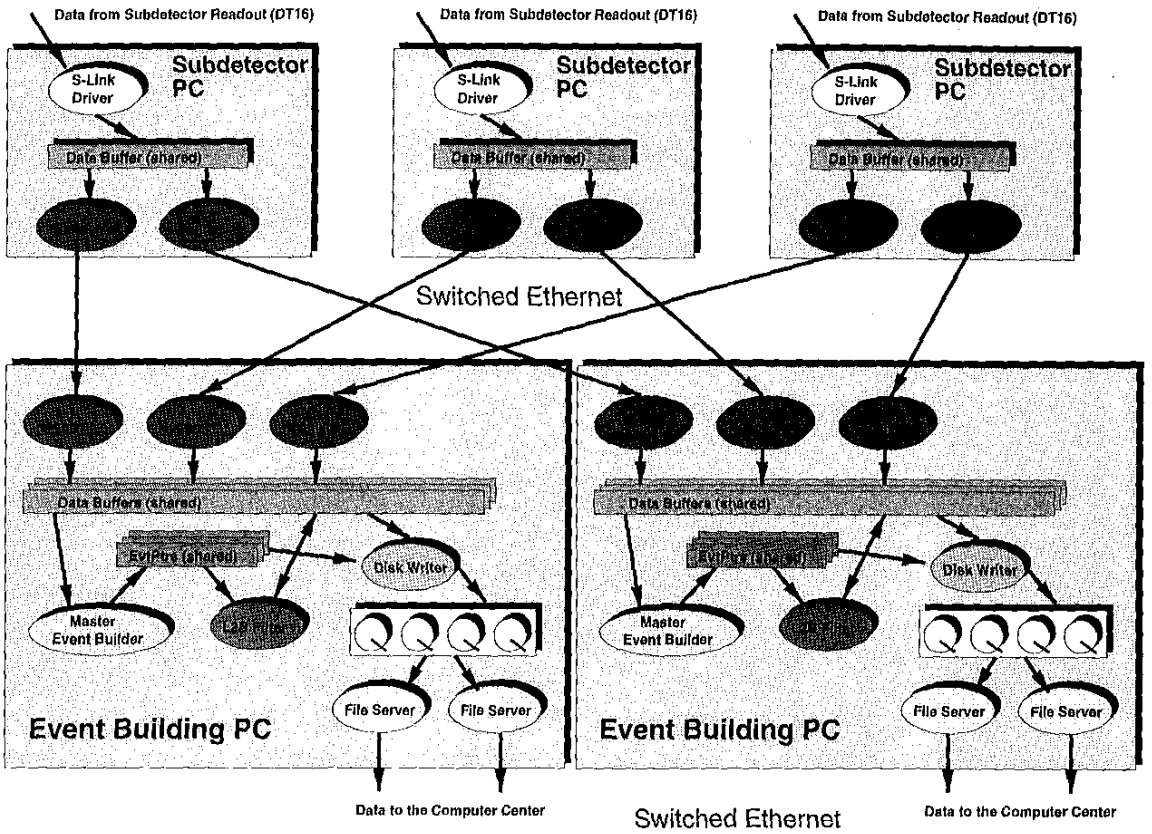

Figure 2: : Simplified diagram of event building processes and interprocess connections with 3 SDPCs and 2 EBPCs.

buffer. An automatic retry mechanism asynchronously takes care of failed transfers.

In the computer center the burst fragment files are combined to complete bursts, processed by a software filter and fed to the online reconstruction program. Eventually, raw or filtered burst files and the results of the online reconstruction pass are written to Redwood STK tapes using the computer center's tape robot.

Figure 2 illustrates the event building software and connections in a simplified setup with 3 SDPCs and 2 EBPCs. The NA48 setup during the 1998 running period was made of 11 SDPCs and 8 EBPCs.

VI. Conclusion and Performance

During the winter 1998 shutdown period the NA48 collaboration has upgraded its custom-built event builder hardware to an online PC farm to increase the available data bandwidth and to add filtering and monitoring capabilities. The new system consists of 24 commercial PCs running the Linux operating system. The PCs are interconnected with full-duplex 100 Mbit/s Ethernet using a switch.

Event fragments are transferred from the readout front-ends into dedicated PCs using 33 MByte/s parallel ECL links interfaced to PCI. Taking advantage of the large total bandwidth of the switched Ethernet, the event fragments are then combined to complete events by a distributed parallel event building software, written out to disk buffers and sent to the computer center via a 7 km long distance Gigabit Ethernet link for further processing and storage.

After an installation and setup period of about a month the system has been continuously taking data since April 1998, and has proven to have high performance and is stable and reliable at data rates of more than 250 MByte per beam burst, corresponding to more than 16 MByte/s (about 1 GByte/minute) continuous data load. In the 1998 running period approx. 70 Terabyte of physics events have been assembled and sent to the CERN computer center.

VII. Acknowledgements

We would like to thank all our colleagues and technical staff in the collaboration for all their efforts, help and commitment during the upgrade of the system and 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.

VIII. REFERENCES

[1] F. Bal et al., "The NA48 Data Acquisition System", IEEE Trans. Nucl. Sci. vol. 45 (1998) 1889-1893.

[2] G. D. Barr et al., CERN/SPSC/90-22/P253

[3] O. Boyle et al., "The S-LINK Interface Specification", ECP-Division CERN, Geneva, Switzerland, 27 March 1997, http://www.cern.ch/hsi/s-link/spec.
H. C. van der Bij, "S-Link, a Data Link Interface for the LHC Era", *IEEE Trans. Nucl. Sci.* vol. 44 (1997) 398-402.

[4] J. Alexander, "Eurogam Project: DT32-Bus Specification", Nuclear Physises Support Group, CLRC, Daresbury Laboratories, Daresbury, Warrington, Cheshire, UK WA4 4AD, United Kingdom.

[5] F. Bal, A. Lacourt, "DT2SL, DT16 to S-Link PCI Interface User Manual", EP-Division internal note, CERN, Geneva, Switzerland.

[6] Incaa Computers, P.O. Box 722, 7300 AS Apeldoorn, The Netherlands.

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

[8] F. Bal, A. Lacourt, "RIO Tic, RIO Dat Interfaces", ECP-Division internal note, CERN, Geneva, Switzerland.

[9] A. Cisternino, "Generic AMCC S5933 Linux Device Driver", http://pcape1.pi.infn.it/ acister/dev/driver.html.

[10] Red Hat Software Inc., P.O. BOX 13588, RTP, NC 27709, USA.

[11] Cisco Systems Inc., 170 West Tasman Dr., San Jose, CA 95134, USA.