The NA48 online and offline PC farms

Andreas Peters representing the NA48 Collaboration

Institute for Physics, University of Mainz, Germany

Abstract

The NA48 experiment at CERN SPS aims to measure the CP violating parameter Re(ϵ'/ϵ) in the neutral kaon system with an accuracy of 0.2 permille. Trigger and DAQ system have to handle event rates up to 10 kHz and data rates up to 150 Mbyte/s during the 2.5s beam burst. In 1998 a farm of 24 Intel PentiumII based PCs running the LINUX operating system was designed and built as a memory buffered realtime data acquisition system based on *Fast Ethernet* switching technology to cope with an average rate of 32 Mbyte/s.
From the farm raw data are transferred via an optical gigabit link to the CERN computer center, situated 7 km from the NA48 experiment. Arriving data is processed in real-time with the Level 3 software trigger/filter running on a PC farm using 42 dual PentiumII PCs, 3.5 Tbyte disk space and StorageTek taperobots.
The Level 3 software is used as a 'step one' online filter performing full reconstruction and physics analysis with separation of data into several streams in a raw and a compressed physics data format suitable for analysis.
The raw data volume processed with online and offline PC farms in 1999 was about 100 Tbyte collected in 125 days, the physics data output volume in 1999 was on the scale of 3 Tbyte.

Keywords: NA48 PC farms

1 Introduction

The NA48 online PC farm was successfully installed in 1998 to replace an existing hardware data merger which was limited to a data processing rate of 220 Mbyte/burst at 86.5% efficiency. A detailed presentation has been given at ICHEP conference, Chicago 1998 [1].
The new data acquisition system was able to cope with a nominal rate of 25 Mbyte/s, the I/O and processing speed of the Meiko CS2 computer[1] running the Level 3 real-time reconstruction filter and the *Central Data Recording System* was limiting the data rate to 15 Mbyte/s.
In 1999 the Meiko CS2 computer system was replaced by the offline PC farm. The Level 3 reconstruction software has been ported to the Linux platform.
The link between online PC farm and CERN computer center was upgraded to a real gigabit to gigabit connection with 110 Mbyte/s maximum bandwidth.
The online PC farm stability and data throughput has been improved by changes in the software model and by adding new supervision tasks.

[1] 128 processor parallel computer with 2.5 Tbyte parallel filesystem

2 The online PC farm

The online PC farm consists of 24 PentiumII PCs running the Linux operating system. 11 PCs (*subdetector* PCs) with DT16-to-PCI interfaces ensure the connectivity to the DT16 bus of the readout systems [2].
Data from the detector readout systems is received during the 2.5s long beam burst repeated every 14.4s [2] and written into the *subdetector* PCs memory using *Direct Memory Access* at a maximum speed of 30 Mbyte/s. A typical burst comprises 17.000 trigger and fills a data volume of 260 Mbyte. To allow for additional calibration triggers outside the beam time window, the readout→PC transfer time is expanded to 5.4s. The remaining 9s before the next beam cycle are reserved for event building, performing the conversion of parallel subdetector data into sequential single event data. Subdetector data is dispatched to at most 12 *event building* PCs using TCP/IP connections via a *Fast Ethernet* switched network.

2.1 Event worker model for event building

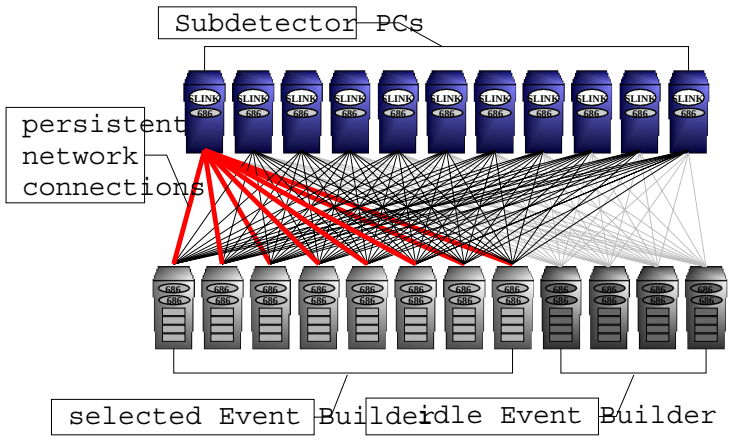

Figure 1: Event worker model for the online PC farm

The dispatching of data from *subdetector* to *event building* PCs is steered by a control task running on a dedicated *Control* PC. Before any data is transferred the control task checks the status of all *event building* PCs and selects 8 out of 12 PCs for event building in a round-robin way. This *event worker model* treats four PCs as 'hot spares' (figure 1).
The system guarantees that at maximum 1/8 of one burst is lost in case of an *event building* PC crash. As long as a crashed PC is not rebooted successfully, it is not selected by the control task.
Event building PCs can be masked or unmasked during run time to remove or add them from the event building process. All necessary TCP/IP connection from *subdetector* PCs to *event building* PCs are automatically rebuilt.
The number of selected *event building* PCs can be dynamically configured according to the required data throughput. With the standard setup of 8 selected *event building* PCs the system can deal with 400 Mbyte/burst. The system performance does not scale with the number of *event building* PCs. A configuration with 12 selected *event building* PC leads to a maximum throughput of 440 Mbyte/burst. This is due to the unequal data sharing on the *subdetector* PCs and the maximum TCP/IP transfer speed of 11 Mbyte per PC.

[2] The timing is given by the *SPS* (Super Proton Synchrotron) beamcycle

2.2 Treatment of event building network scheduling problems

An *event building* PC has to deal with 11 incoming connections (one from each *subdetector* PC) and one outgoing connection to the CERN computer center, each handled by a different process in 1998.
Using the newest stable Linux kernel version of February 1999 (V. 2.2.3), the operating system couldn't provide an equal input and ouput rate for an unbalanced number of input and ouput processes. A typical value was 11 Mbyte/s input rate and 2 Mbyte/s output rate.
For this reason the event building software was rewritten such, that *one* process handles all input connections and a separate process handles the output. The input process does a round-robin read operation on all input connections. The scheduling between the 11 input connections is done in a complex way which takes into account the availability of data and the time intervals between two read operations.
With this change 10 Mbyte/s continuous input and 6-8 Mbyte/s output rate have been reached during the parallel *full duplex* operation.

2.3 Raw Data Format Supervision

During the data transfer from *subdetector* to *event building* PCs, the individual subdetector data is checked in parallel on the *subdetector* PCs for format errors. Errors for each burst are logged in a file and displayed in detail in a X11 GUI for the shift crew.
This raw data supervision system allows fast error detection of the readout systems and improved the data quality in 1999.

2.4 PC Slow Control and Self Recovery System

On the *Control PC* a daemon process checks in short intervals, whether all online PCs are reachable via a network connection[3]. If a PC seems to be unreachable, it is automatically rebooted by the daemon process. The 'power' and 'reset' of all PCs can be switched using a *Slow Control* PC using a multi I/O card with 48 channels connected to the power and reset pins of the PC mainboards.
The event worker model prevents a downtime of the DAQ system in case of an *event building* PC crash. A *subdetector* PC crash produces an inevitable DAQ downtime as long as the PC is not rebooted. A typical downtime due to a *subdetector* PC reboot is 2 minutes and the average reboot frequency for a subdetector PC is one out of 11 PCs per day. Crashes are mostly due to the DT16-to-PCI interface card.

3 The offline PC Farm

The offline PC farm consists of 42 dual PentiumII computers running Linux as operating system and four Sun 450 (multiprocessor) server machines with 3.5 Tbyte SCSI disk arrays. All disks are NFS mounted on the Linux PCs, but data is read for processing using the *rfio* [4] protocol. The disk server I/O load reached 120 Mbyte/s in 1999 at peak times.
The network structure of online and offline farm is shown in figure 2. The main task for the offline PC farm is to run the Level 3 reconstruction program as a 'step 1' real-time prefilter during the runtime of the experiment and as a 'step 2' offline filter for reprocessing of prefiltered data.

[3] using the 'ping' command as handshake mechanism
[4] *rfio* allows file I/O on a remote host using a direct point-to-point TCP/IP connection

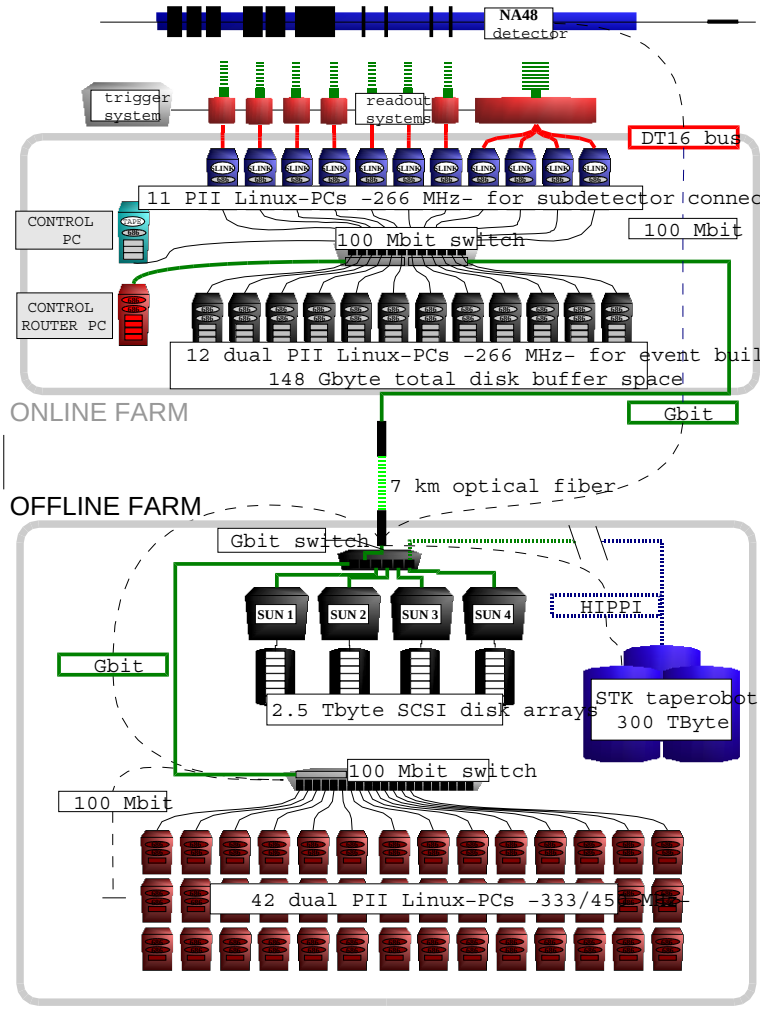

Figure 2: The NA48 online and offline PC farms

3.1 Level 3 filter facilities

The Level 3 software ensures a full 'real-time' decoding/reconstruction with the best known calibrations. The standard offline reconstruction algorithms are called in an event-driven way. Typical processing times are 50 ms per event per process.

Events are classified and written into different output streams as chosen by a configuration file. For online processing the configuration file is extracted from the *online database* depending on experimental conditions.

Output streams can be written in the original *RAW* data format or in the compressed physics data format *COmPACT* . For the online operation a *gold RAW*[5] stream is written with a reduction to 1/8 of original data volume. This *gold RAW* stream is used as input for the step 2 offline filter reprocessing to produce the *gold COmPACT* stream with a typical reduction to 25% of the input data volume[6]. With this procedure the 100 Tbyte *RAW* data volume in 1999 was reduced to 3 Tbyte *gold COmPACT* data, which are the bases for the data analysis.

[5] the gold RAW stream contains all interesting candidates for analysis

[6] the output size is 15% but the stream is enlarged by adding events with overlayed noise for accidental studies

3.2 Steering of Level 3 processing and storage of processed data

All Level 3 processes on the offline PC farm are connected via TCP/IP connections to the Level 3 *Control Pool Daemon* running on a SUN disk server. This daemon dispatches filenames of unprocessed data bursts to idle Level 3 processes. For the reprocessing procedure an input file list is generated 'manually'. For online processing the server data disks are automatically scanned for new unprocessed data. Sucessfully processed bursts are put to tape by a tape daemon using the *Central Data Recording Service* of the CERN IT center which provides StorageTek taperobots with 50 Gbyte tapes and 300 TByte total staged storage space [3].
All necessary synchronisation between tape storage and processing is made with *tagfiles* on the SUN disk servers using *csh* scripts and C or C++ programs. All tasks are as autonomous as possible.

4 Conclusions

Online and offline PC farms have been successfully designed in a way that they are scalable and dynamically adaptable to changing requirements on the DAQ system performance.
Most important design feature for the online PC farm was to guarantee a 24 hours continuous operation for event building using a recovery system for standard 'operational' problems without manual interaction. This has been reached with 99% efficiency during the run period 1999.
The main task of the offline farm is to run the Level 3 reconstruction and filtering with the ability to access several Tbytes of diskspace using a centralized disk server concept.
Two hours disk buffering during data taking on the online farm ensure a sufficient decoupling between the online operation and possible Level 3 processing or tape storage problems.
During 125 days of operation in 1999 the system demonstrated the possibility to run the DAQ system with high efficiency at an average rate of 10 Mbyte/s and a maximal rate of 20 Mbyte/s.
The 1999 improvements of the online farm and the replacement of the CS2 computer by the offline farm were responsible for an increase of 60% in data throughput compared to 1998.
For future rare decay programs it is forseen to double the trigger rate by using a 32-bit bus system for subdetector connectivity and moving parts of the zero suppression from the readout hardware to *subdetector* PCs.

References

1 S. Luitz, "THE NA48 DATA ACQUISITION PC FARM", CHEP'98, Chicago, Autumn 1998.

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

3 B. Panzer-Steindel "Central Data Recording for high data rate Experiments at CERN", CHEP'98, Chicago, Autumn 1998.