

A Gigabit/s Data Acquisition System

Andrei Sukhanov, Piotr Kulinich, Pradeep Sarin

Abstract- A fully pipelined data acquisition system has been developed for the PHOBOS experiment at RHIC in BNL. The system began operation in 1998; after the most recent upgrade in 2004, it reached a trigger rate of 575 Hz, processing speed of 200 MB/s, and maintained a sustained data-sinking rate to HPSS of 50 MB/s. The system receives data from 135,168 silicon pad detectors (12-bit amplitudes) and from 2,000 scintillating detectors (time and amplitudes). The data from the silicon detectors are transmitted through two 120 MB/s fiber links to a 24-node processing system (Mercury Computer System, Inc.) embedded into a VME crate and interconnected using a RACEWAY switching fabric. Each node (PPC750 300 MHz) compresses a part of the data stream using lossless Huffman algorithm. The four-fold compressed data are sent directly from the nodes to a model TK200 data recorder (Conduant Corporation) through an FPDP link. The event building is done on the fly by adjusting offsets and lengths of a scatter-gather DMA transfer. The TK200 has a disk capacity of 3.2 Tbyte and aggregate bandwidth of 200 MB/s, and is connected to the HPSS over a single Gigabit fiber connection. Under laboratory conditions, the TK200 was able to handle 120 MB/s of recording and 70 MB/s of reading simultaneously.

The upgraded system was used for data taking during RHIC run 2005 and recorded over 0.5 billion (50 Tbyte of data) of heavy-ion collisions over a period of two months.

The data acquisition system (DAQ) at PHOBOS [1] receives data from 135,168 silicon detectors and 2000 scintillation detectors. It has following components :

- Processing farm with 24 processing nodes (CE) with MPC750 (300 MHz processor, 1 MB of L2 cache SRAM, 32 MB of DRAM). The farm also contains 2 input ports (RINT) and one output port (ROUT). The nodes and ports are located on 8 VME motherboards (PCH6), all connected to a RACEway switching fabric [2][3][4].

- Model TK200 data recorder with disk capacity of 3.2 Tbyte [5] connected to the output port of the farm and over Gigabit interface to a central storages system (HPSS).
- Several programmable logic modules based on a custom VME board PCDP [6] for triggering and synchronization.
- RS485 link, also based on PCDP, for taking data from FASTBUS and VME crates with ADCs and TDCs.
- Trigger monitor: MVME-2306 board with PPC 604 processor, VxWorks OS
- Host CPU (Themis USPIIe with UltraSPARC 400 MHz CPU) for monitoring and code development for the farm.

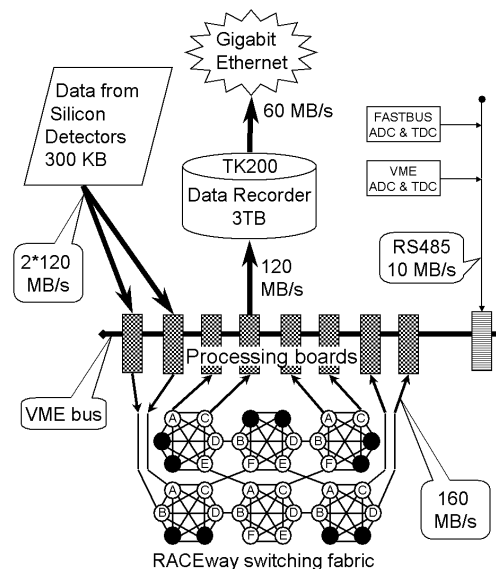


Fig1. Structure of the PHOBOS Data Acquisition System.

The whole system (excluding TK200 data recorder) occupies a single VME crate (DAQ crate).

The signals from the silicon sensors are digitized in Front End Controllers (FEC), which are located in the experimental area [7]. Each sensor is digitized using a 12-bit ADC but samples are stored as 16-bit words. The multiplexed data are transferred to the PHOBOS

DAQ crate over two optical fibers using FPDP (ANSI/VITA standard 17-1998) synchronous data flow protocol to two RINT modules.

The signals from the scintillating detectors are digitized in one FASTBUS and one VME crates. Each crate contains a PowerPC based single board computer running VxWorks and a PCDP board configured as trigger receiver and data transmitter (TRDT). The TRDT receive an event token (last 8 bits from the global event number) from an Event Manager Module located in the DAQ crate. The RS485 ports of the TRDT boards are configured as output ports and are connected to a front panel flat-cable bus which is connected to a receiver (another PCDP board located in the DAQ crate).

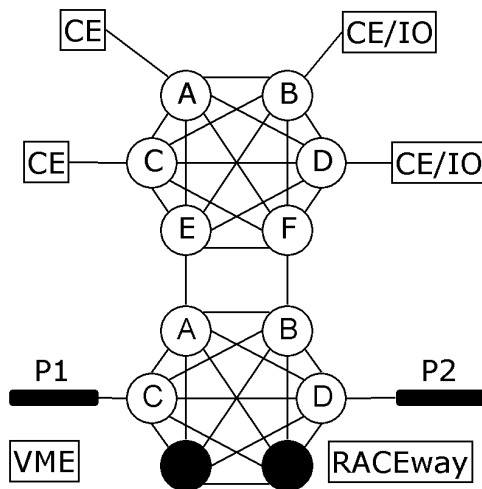


Fig2. Structure of the processing board. It consist of four processing nodes (CE), two CE could be replaced by single IO module. The CE are interconnected using a crossbar switch and connected over second crossbar switch to VME backplane and RACEWay switching fabric.

One of the processing nodes is dedicated Master, its main role is to control data flow from RINT modules. Another dedicated mode is a Collector; it controls the ROUT and receives data from the RS485 link. The rest nodes are Workers, they all are running the same code.

The Master forwards data in a scatter-gather manner from both RINTs to the Workers' local memories. When a Worker detects arrival of the last word of a dedicated data buffer (frame) it starts processing this frame. The Worker keeps a queue of processed frames in its shared memory buffer (SMB). After finishing with the current frame it updates entry in a global table of output frame descriptors (TOFD).

Event building

The event building is done on the fly. The Collector prepares a scatter-gather DMA transfer from Workers' SMBs directly to the ROUT and subsequently to the disk buffer of TK200. This method makes it possible to eliminate the traffic jams when the big volumes of data coming in and going out from the same CE, effectively halving the system bandwidth. The Collector modifies the sizes and offsets for a predefined transfer using information from TOFD and initiates the transfer on a DMA engine associated with the ROUT. The Mercury system provides extensive tools for traffic shaping but for our applications we use only hardware handshaking using SUSPEND line of the FPDP link.

Compression of the silicon data

The silicon pad detectors have substantial common mode noise and cross talk. Signals from empty channels are used in offline analysis for common mode noise correction and for base line shift compensation. Therefore, the data compression is done using lossless Huffman algorithm [8], its main advantage is a high encoding speed and compression ratio close to the boundary of Shannon's entropy [9]. Pedestals are subtracted from all channels before the compression, resulting in approximately the same mean amplitude over all channels; this makes it possible to use single coding tree to encode data from all channels. The pedestals and the coding tree are periodically updated and written to the data stream. The typical combined signal distribution is shown on Fig. 4. The entropy of the distribution is 4.1. The achieved entropy (average number of bits per data channel) of the compressed data is 4.0; this corresponds to the compression factor of 4.0.

The RISC architecture of the MPC750 CE is very suitable for the Huffman encoding mainly because it has a large set of internal registers (32) and fast L1 cache which is well balanced with large external L2 cache. The important requirement is that the Huffman table (8 KB in the case of 12-bit data) fits into the 32 KB L1 cache. The measured performance of the compression algorithm is 16 MB/s of input data per Worker and it does not depend on frame size.

Data recorder

For temporary data staging we use the model TK200 data recorder from Conduant Corporation [5]. The TK200 consist of a commercial motherboard with dual Xeon 2.4 GHz processors, 4 GB of RAM and an interface card to an array of 16 IDE disk drives working in parallel. The TK200 can handle 200 MB/s of raw, sustained recording or playback. It has a total disk capacity of 3.2 Tbyte, which is sufficient to accumulate data for 22 hours at maximum input rate of 575 Hz. The main advantage of using TK200 is that it writes data directly from the FPDP port to disk array; therefore there is no need for intermediate buffering. Another advantage is that the recording is completely decoupled from the CPU. The TK200 runs under Linux OS and is delivered with extensive API. Originally the TK200 was designed for one-way operation – record first, playback later. For our application the company developed the FIFO mode, which allows simultaneous recording and read. Under laboratory conditions, the TK200 was able to handle 120 MB/s of recording and 70 MB/s of reading simultaneously. Under real-life experimental conditions the device was working very reliably for months, the average recording rate was 50 MB/s, and the reading rate was 128 MB/s.

References

- [1] B. B. Back, M. D. Baker, D. S. Barton, S. Basilev, R. Baum, R. R. Betts, et al. "The PHOBOS detector at RHIC," *Nucl. Phys.*, vol. A698, pp 416-419, 2002.
- [2] (2001, March). RACE++ Series PowerPC 750 Daughtercards. Mercury Computer Systems, Inc., MA. [Online]. Available: http://www.mc.com/literature/literature_files/ppc750-dc-ds.pdf
- [3] B.C. Kuszmaul, (1995, Apr). The RACE Network Architecture. Mercury Computer Systems, Inc., MA. [Online]. Available: http://www.mc.com/literature/literature_files/netwrkarc-h-wp.pdf
- [4] (2001, March). RACEway Interlink Functional Specification. Mercury Computer Systems, Inc., MA. [Online]. Available: http://www.mc.com/literature/literature_files/racewayin-trlnk-spec.pdf
- [5] <http://www.conduant.com>
- [6] P. Kulinich, P. Sarin, A. Sukhanov. The DAQ system with a RACEway switch for the PHOBOS experiment at RHIC. *IEEE Trans. on Nucl. Sci.*, vol. 49, pp 2455-2458. Oct 2002
- [7] P. Kulinich, "Silicon DAQ Based on FPDP and RACEway," in *Proc of the 6th Workshop on Electronics for LHC Experiments*, 2000, pp 479-482.
- [8] D. A. Huffman, "A Method for the Construction of Minimum Redundancy Codes," *Proc. of the IRE*, Vol. 40, pp. 1098—1101, 1952.
- [9] C. E. Shannon. "A Mathematical Theory of Communication," *The Bell System Technical Journal*, vol. 27, pp. 379–423, 623–656, July, October 1948.