An Overview of the STAR DAQ System

J. M. Landgraf[1], M. J. LeVine[1], A. Ljubicic, Jr.[1], J. M. Nelson[2], D. Padrazo[1], M. W. Schulz[3], for the STAR Collaboration

[1]*Brookhaven National Laboratory, Upton NY 11973*

[2]*University of Birmingham, Birmingham B15 2TT, United Kingdom*

[3]*Universtitat Heidelberg, D-69120 Heidelberg, Germany*

Abstract

We describe the STAR Data Acquisition System. STAR is one of four experiments commissioned at the Relativistic Heavy Ion Collider (RHIC) at BNL in 1999 and 2000. DAQ combines custom VME-based receiver electronics with off-the-shelf computers in a parallel architecture interconnected with a Myrinet network. Events of size 200MB are processed at input rates up to 100Hz. Events are reduced to 10MB by zero suppression performed in hardware using custom designed ASICs. A Level 3 Trigger reconstructs tracks in real time and provides a physics-based filter to further reduce the sustained output data rate to $\sim$30MB/sec. Built events are sent via Gigabit Ethernet to the RHIC Computing Facility and stored to tape using HPSS.

STAR[1] is a large detector at the Relativistic Heavy Ion Collider (RHIC) located at Brookhaven National Laboratory. The design and implementation of the STAR DAQ system[2,3] was driven by the characteristics of STAR's main detectors, a large Time Projection Chamber (TPC)[4], and to a lesser degree two smaller Forward TPCs (FTPC)[5] and a Silicon Vertex Tracker (SVT)[6]. Together, these detectors produce 200MB of data per event and are able to read out events at 100Hz. The RHIC Computing Facility (RCF) manages the storage of raw data for all of the RHIC experiments using an HPSS hierarchical storage system. By balancing the expected rate of offline data analysis with the rate of data production, resources were allocated to STAR to support sustained raw data rates of up to 30MB/sec for steady state operation. The central task of the STAR DAQ system is then to read data from the STAR detectors at rates up to 20,000MB/sec, to reduce the data rate to 30MB/sec, and to store the data in the HPSS facility. In addition, STAR has numerous other detectors with much smaller data volumes. These include several trigger detectors[7] as well as a Ring Image Cherenkov (RICH)[8] detector, barrel

[1] This work is supported in part by the U. S. Department of Energy under contract No. DE-AC02-98CH10886.

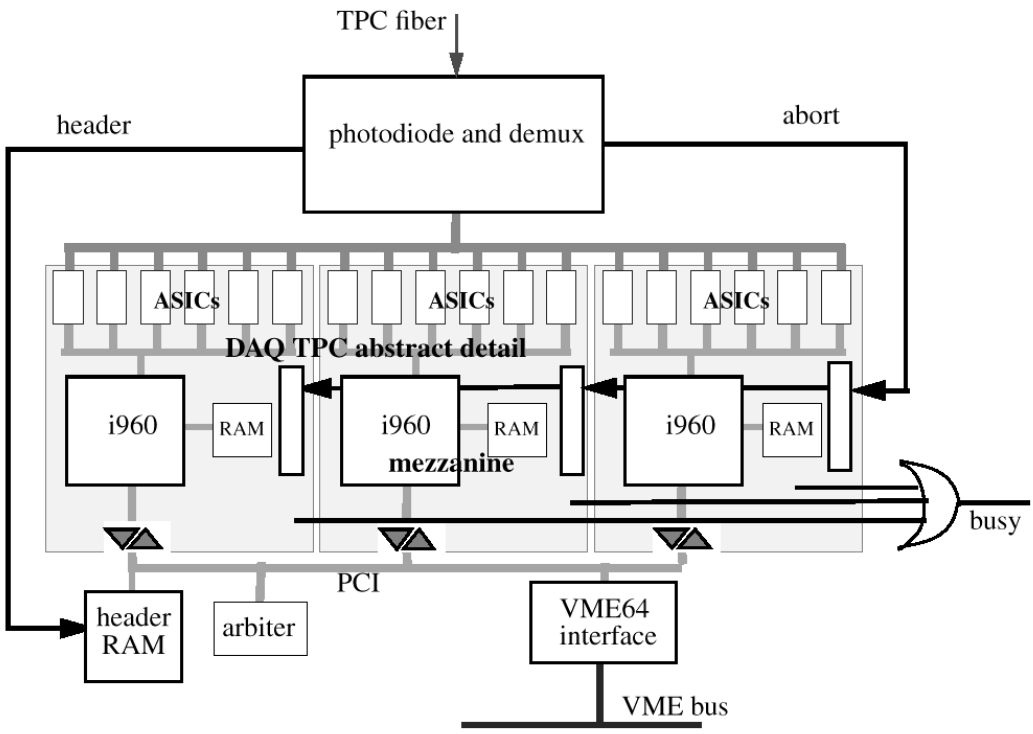

Fig. 1. Block Diagram of the TPC, SVT and FTPC receiver boards.

and end-cap Electro-Magnetic Calorimeters (EMC)[9], a Photon Multiplicity Detector (PMD), and a Time of Flight prototype (pTOF) as well as the Forward Π^0 detector. A further requirement of the STAR DAQ system was a modular design able to conveniently incorporate these additional detectors (as well as unknown upgrades) into the system.

The large input data rate to the DAQ system demands parallel processing at the DAQ front end. Multiple Receiver Boards[10] (144 for the TPC, 20 for the FTPCs, and 24 for the SVT) receive data in parallel on separate optical fibers from the detectors. The Receiver Boards (RBs) are grouped together in VME crates. Each crate is controlled by a Detector Broker CPU (DET). There are 12 DETs for the TPC, two each for the SVT and FTPC, and one for each additional detector. We use two strategies to reduce the data volume. First, we zero-suppress the data to reduce the event size to 10MB for central events. Secondly, we apply a physics-based filter, the Level 3 Trigger (L3)[11], to choose which events to write to tape. The L3 must find on the order of 1500 tracks in the TPC and make trigger decisions based upon them within 200ms. This limits the time available for DAQ front-end processing, and creates the need for a farm of $\sim$50 CPUs integrated within DAQ dedicated to tracking. In addition, the delay between receiving the event and receiving the build/reject decision from the L3 trigger makes it necessary for the system to manage multiple events at the same time, in various stages of completion.

For the TPC, SVT and FTPCs front end processing electronics resides in Sector

Crates. Initial processing: 10 to 8 bit conversion, zero suppression, and data formatting are performed on the three Mezzanine Boards (MZ) contained on each Receiver Board (figure 1). The Mezzanines are each driven by an i960 CPU. Each Mezzanine also contains 6 ASICs[12] which handle the 10 to 8 bit compression and gain correction, pedestal subtraction and one-dimensional cluster finding. This frees the i960 CPU for event formatting and two-dimensional cluster finding. The cluster finding algorithm uses a charge-weighted average of adjoining detector pad hits to produce a list of particle hits with space coordinates for consumption by the L3 tracking algorithm.

The EMC detector uses simplified versions of the Receiver Board, which also receive data over optical fiber using a private protocol but without mezzanine cards, to receive the data from the detector. In this case, data formatting is done by the Detector Broker.

Events from the other small detectors are read into Motorola MVME processors over private, point-to-point 100Mbit/sec ethernet. In every case, the data from each input channel from each detector is ultimately controlled by a Detector Broker. Although the internal implementation of the Detector Broker is different for each detector, the Detector Brokers present a detector-independent interface to the DAQ network.

The components of the DAQ system are connected using Myrinet[13,14], a low-cost, high-performance, low-latency commercial network from Myricom [15]. The network consists of PCI/PMC network cards connected by 1.28 Gbit/sec full duplex links. Myrinet is supported for a large number of systems, including the platforms used in the STAR DAQ: VxWorks nodes running on MVME processors, Linux, and Solaris workstations. We use Myrinet for messaging and data transfers. In the latter case, we make use of the directed send capability of Myrinet which allows data to be sent into a specified memory location in the destination node using a DMA engine resident on the network card, without the intervention of the receiving node.

The management of events within the DAQ system (Figure 2) can be described in two phases according to whether the build decision for that event has been made by L3. Before the decision, the Global Broker (GB) handles the overall management of the event. At the same time as the data are read from the detectors into the DETs, the GB receives a token and trigger detector data from the Trigger/DAQ Interface (TDI) via the Myrinet network. The GB assigns L3 processors to analyse the event and waits for an event decision. If the event is rejected by L3, GB instructs the DETs to release the buffers associated with the event and returns the token to TDI for re-use. If the event is accepted by L3, responsibility for the management of the event is transferred to the Event Builder (EVB). The EVB collects and formats all of the contributions. At this time, EVB instructs the DETs to release the buffers associated with the event and passes the event to a Spooler (SPOOL) which handles the writing of the event to RCF. When the event is written, EVB returns the token

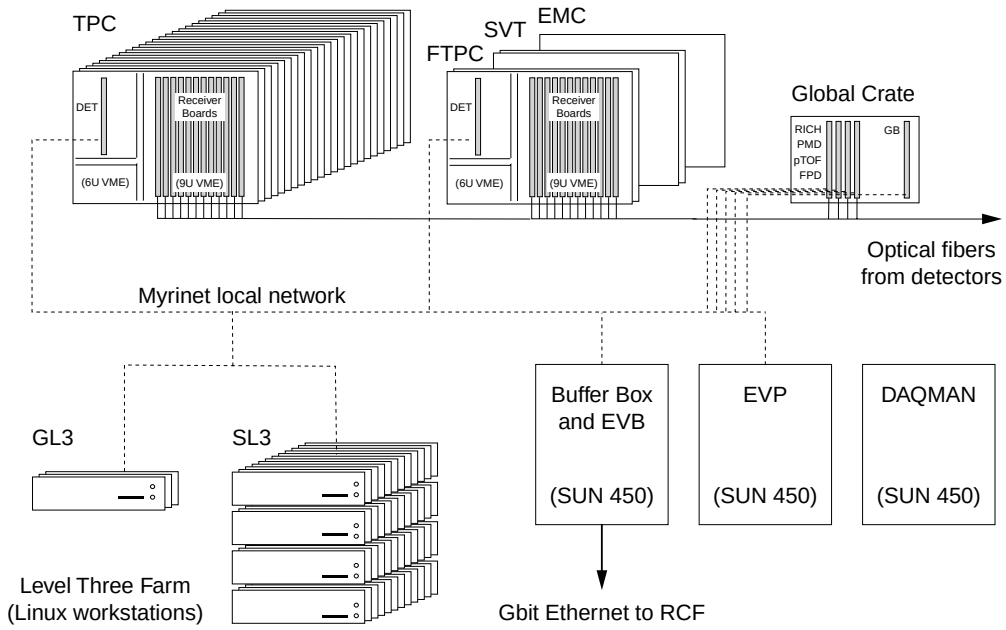

Fig. 2. Schematic Overview of the STAR DAQ.

to the TDI.

The Level 3 system is implemented on a farm of ∼50 Linux nodes connected to DAQ on the Myrinet network. Level 3 tracking is handled by Sector Level 3 (SL3) processors. For each event one SL3 node from a pool of 48 is assigned to each DET. For the TPC, the area assigned to each SL3 node corresponds to two TPC sectors. When tracking is finished the SL3 nodes pass track data to a Global Level 3 (GL3) which runs a series of algorithms to make the event decision.

Event Building is performed on a Sun Solaris 450 workstation called the Buffer Box (BB) because it has 140GB of disk buffer for use in case the HPSS system becomes temporarily unavailable. All process memory is locked to prevent swapping. Communication between processes is handled using a pipe-based message queue library. Events are built in a large (1.5GB) shared memory segment. A separate process, the Spooler, prepares lists of events to be taped, and passes lists to one of two taping tasks. These tapers write to the HPSS system at RCF using a version of parallel FTP (PFTP) which we modified to write data using a descriptor list directly from shared memory. The reason for two tape streams is that the data throughput for a single stream is limited to ∼30MB/sec by the rate of writing to the a buffer disk on the HPSS system. Opening a second data stream allows the HPSS system to write data to multiple disks at the same time, leading to DAQ throughput as high as 50MB/sec, as long as there is free buffer space on the HPSS system. When the 512GB disk buffer space allocated to star is exhausted, disk contention arises between the writing streams and the migration of data to tape which reduces the throughput to ∼30MB/sec. The increased 50MB/sec throughput for the first 2.5

hours of operation should improve data taking efficiency during periods of unsteady collider/detector operation.

The EVB also writes a fraction of its data to the Event Pool (EVP), implemented on another Sun 450, which provides a pool of recent events which are used by online monitoring software to provide quality control for the operation of the STAR detector.

In addition to writing the main data stream, the BB also writes summary data to the tag database, which contains a short descriptor record for each event, as well as summary information for each run and for each data file. The SPOOL task writes database records directly to local disk files, which are in turn read by separate database writing processes. This effectively decouples the performance of the database system from that of DAQ.

Finally, another Solaris node, DAQMAN, is used for booting the VxWorks nodes and for running the run control handler, which handles the distribution of run control commands to the the various run-time systems for the Run Control GUI. This node also contains a centralized logging system, and a monitoring system for all the nodes.

References

[1] STAR Collaboration, 'STAR Conceptual Design Report', LBL-PUB-5347, June (1992).

[2] A. Ljubicic Jr., M. Botlo, F. Heistermann, S. Jacobson, M. J. LeVine, J. M. Nelson, M. Nguyen, H. Roehrich, E. Schaefer, J. J. Schanback, R. Scheetz, D. Schmischke, M. W. Schulz and K. Sulimma, 'Design and implementation of the STAR experiments DAQ', *IEEE Trans. Nuc. Sci.*, **45, No. 4**, pp. 1907-1912 (1998).

[3] A. Ljubicic Jr., J. M. Landgraf, M. J. LeVine, J. M. Nelson, D. Roerich, J. J. Shamback, D. Schmische, M. W. Schulz, C. Struck, C. R. Consiglio, R. Scheetz and Y. Zhao, 'The STAR experiment's data acquisition system', *IEEE Trans. Nuc. Sci.*, **47, No. 2**, pp. 99-102 (2000).

[4] M. Anderson, *et al.,* 'The STAR Time Projection Chamber', (this volume).

[5] K.H. Ackerman, *et al.,* 'The Forward Time Projection Chamber in STAR', (this volume).

[6] R. Bellwied, *et al.,* 'The Silicon Vertrax Tracker', (this volume).

[7] F. S. Beiser, *et al.,* 'The STAR Trigger', (this volume).

[8] A. Braem, *et al.,* 'Identification of High Pt Particles with the STAR RICH Detector', (this volume).

[9] T. LeCompte, *et al.,* 'The STAR Barrel Electromagnetic Calorimeter', (this volume).

[10] M. J. LeVine, A. Ljubicic Jr., M. W. Schulz, R. Scheetz, C. Consiglio, D. Padrazo and Y. Zhao, 'The STAR DAQ receiver board', *IEEE Trans. Nuc. Sci.*, **47, No. 2**, pp. 127-131 (2000).

[11] 'The STAR Level 3 Trigger System', C. Adler *et al.* (this volume), *and* C. Adler, J. Berger, M. Demello, D. Flierl, J. M. Landgraf, J. S. Lange, M. J. LeVine, V. Lindenstruth, A. Ljubicic Jr., J. M. Nelson, D. Roehrich, E. Schaefer, J. J. Schaefer, J. J. Schanback, D. Schmischke, M. W. Schulz, R. Stock, C. Struck and P. Yepes, 'The proposed level-3 trigger system for STAR', *IEEE Trans. Nuc. Sci.*, **47, No. 2**, pp. 358-361 (2000).

[12] M. Botlo, M. J. LeVine, R. A. Scheetz, M. W. Schulz, P. Short, J. Woods and D. Crosetto, 'The STAR cluster-finder ASIC', *IEEE Trans. Nuc. Sci.*, **45, No. 4**, pp. 1809-1813 (1998).

[13] J. M. Landgraf, C. Adler, M. J. LeVine, A. Ljubicic, Jr., J. M. Nelson, M. W. Schulz and J. S. Lange, 'The implementation of the STAR data acquisition system using a Myrinet network', *IEEE Trans. Nuc. Sci.*, **48, No. 3**, (2001).

[14] N. J. Boden, D. Cohen, R. E. Felderman, A. E. Kulawik, C. L. Seitz, J. N. Seizovic, and Wen-Ken Su, 'Myrinet — a gigabit-per-second local-area network,' *IEEE Micro*, **15, No. 1**, pp. 29-36 (1995).

[15] http://www.myri.com.