Design and Implementation of the STAR Experiment's DAQ[a]

A. Ljubicic, Jr.[1], M. Botlo[1,b], F. Heistermann[1], S. Jacobson[2], M.J. LeVine[1], J.M. Nelson[4], M. Nguyen[1], H. Roehrig[5], D. Roerich[5], E. Schaefer[6], J.J. Schambach[3], R. Scheetz[1], D. Schmischke[5], M. Schulz[1] and K. Sulimma[5]

[1]Brookhaven National Lab., Upton, NY 11973, USA

[2]Lawrence Berkeley National Lab., Berkeley, CA 94720, USA

[3]University of Texas at Austin, USA

[4]University of Birmingham, Great Britain

[5]University of Frankfurt, Germany

[6]Max Planck Institute for Physics, Munich, Germany

Abstract

The STAR experiment is one of the two large detectors currently being built at the Relativistic Heavy Ion Collider (RHIC) at Brookhaven National Laboratory, Upton, U.S.A.

The major issue of STAR's DAQ is the large amount of data that has to be processed as fast as possible. The required data rate is of the order of 90 Gbits/s which has to be processed and scaled down to about 15 MBytes/s and stored to tape or other permanent archiving media. To be able to do so the STAR DAQ uses a custom built ASIC which preprocesses the raw data for later use by a software Level 3 trigger.

The Level 3 trigger selects events to be archived depending on physics criteria based upon the particle track information extracted during Level 3 processing.

The design presented is a massively parallel multi-processor system which consists of front end microprocessors hierarchically organized within a VME crate system. Each VME crate contains 6 custom made Receiver Boards with 3 Intel I960HD processors per board for a total of 18 processors per crate. The STAR's TPC detector uses 24 such crates and the SVT detector will use 4 crates for a total of 504 microprocessors.

I. INTRODUCTION

A. The STAR Experiment

The STAR experiment (Solenoidal Tracker At RHIC) [1] is one of the two major experiments being built for the RHIC accelerator at BNL, USA. The RHIC accelerator itself is a heavy-ion collider which will be able to accelerate various species of heavy ions up to the energy of 100 GeV/nucleon. Both the experiments and the accelerator will begin commissioning in June 1999.

Because of the high energy and large mass of the ions the multiplicity of particles produced in a central collision will be in the thousands. The philosophy of STAR was to use one large TPC (Time Projection Chamber) detector which will be able to track all charged particles in the ±1 pseudorapidity range. Additional detectors which will come online after the first year of running include the SVT (Silicon Vertex Tracker).

B. Detector Subsystems

The baseline detector of STAR is the TPC and it is the only detector (apart from the various trigger detectors) that will be online at experiment startup (1999.). The second detector coming online will be the SVT but due to the similarities of the readout scheme the peculiarities of the SVT are already included in the design of the custom boards as well as in the software.

The TPC detector is organized in geometrical sectors with groups of 12 sectors on each side of the detector for a total of 24. Each sector delivers the data in digital form through 6 readout boards each equipped with a fiber optic link with a 1.5 Gbit/s bandwidth.

The readout boards handle a total of 1152 pads where a pad is sampled in 512 timebins. The raw data coming down the fiber has 10 bit resolution, for a total data size of 810 Mbits for each event.

The SVT is organized in 24 readout cards with each card responsible for 4320 anodes sampled in 128 timebins so a similar calculation gives 128 Mbits worth of raw data. The SVT uses the same fiber optic link and the same 10 bit resolution.

Other possible detectors that would eventually become part of STAR would be the EMC (Electromagnetic Calorimeter) and a TOF wall which both have different requirements to the DAQ system from the TPC but due to their negligible data volume compared to the TPC they do not pose a serious challenge. Another possible detector, the Forward TPC (FTPC), would use the same readout scheme (and readout boards) as the main TPC and would require the same organization. This detector would add about 20% to the data rate and is again not seen as a problem. The requirements of these detectors are not evolved enough at this time so they are not discussed further.

[a]This work is supported in part by the U.S. Department of Energy under Contract No. DE-AC02-76CH00016

[b]Currently at Morgan-Stanley, Inc

C. DAQ requirements

The main driving force behind the DAQ system is the experiment's requirement that the DAQ should be able to process data in central heavy-ion collisions at 100 events per second for both the SVT and TPC detectors. This would amount to a processing data rate (or throughput) of 92 Gbits/second.

The first step in the reduction of this data amount is the 10-to-8 bit translation since it was shown by simulations and previous experience that the analysis doesn't need the full linear 10 bit dynamic range but can use a non-linear 8 bit translation scheme.

The second step in the reduction scheme relies on the fact that the occupancy (the number of channels above the pedestal level) in these experiments and detectors is in the 5% - 20% range. This means that a compression down to 15% of the original (on average) can be done with zero-suppression, keeping only the data that lies above some low threshold.

After these two steps the data rate is still about 1.3 GBytes/second. Since there is no likely single storage system that can run with such a data rate in the next few years one approach would be to partition this stream into many smaller streams running in parallel. This would be the "classical" approach taken by many other experiments.

This approach is impractical for many reasons. Just managing and staging the large number of small tape drives would be very complicated and error prone, i.e. would require more than 250 tapes running at a currently viable rate of 5 MB/s. Even if we would use the current state-of-the art tape systems this number would still be in the 50's.

Perhaps even more important is the effort to analyze these tapes later in the offline scheme - this would require an additional (comparably) large number of tape drives just to keep the analysis rate equal to the data taking rate.

It is more practical to add an additional level of triggering (Level 3) which will select events at a low rate for archiving to tape. This is a purely software trigger that allows us to reduce the archiving data rate by a factor of 100 thus making it easily manageable even with a single tape unit. This large compression factor cannot be achieved without the knowledge of the "physics" of an event but has to be based on physically relevant variables such as particle track parameters in either or both the TPC or SVT. This required devising a way to do track recognition online for purposes of triggering in the Level 3 scheme. One should point out that the quality of this tracking need not be as high as the one used in further offline analysis but has to be only sufficiently precise to enable a fast trigger decision. This decision could be made i.e. by just track counting, by cuts in the various track-based histograms (p_t vs. rapidity etc...), by selecting events with one or more high momentum tracks (or lacking them...), by some topological qualifications etc.

The hierarchical nature of the system (the detectors organized in geometrical sectors which contain readout boards) lends itself to a parallel processing approach with details in the following chapters.

Additional performance issues like local throughputs, network bandwidths, system latency will be handled in a future publication.

II. The Architecture of the DAQ System

A. Overview

The STAR DAQ is a hierarchical system of VME crates interconnected with 2 networks (Figure 1). The slower Ethernet is used for system control, booting of the VME processors, error message passing and system monitoring. The fast SCI network is used for event building, inter-crate communication/synchronization and for data passing from the Sector Level 3 trigger to the Global Level 3.

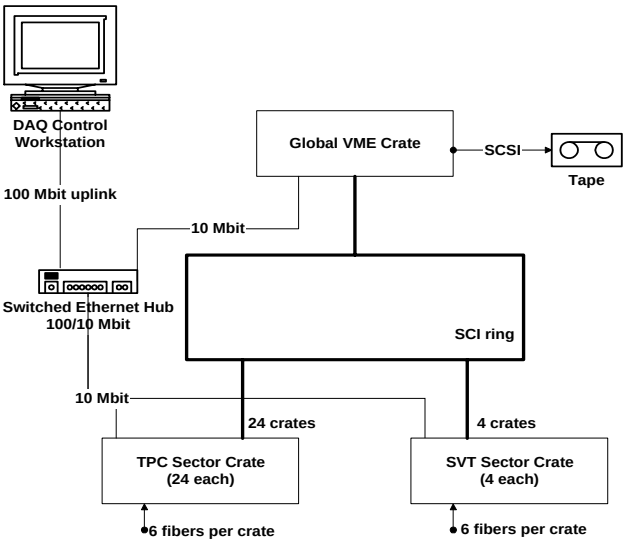

Figure-1 - Global view of the STAR DAQ system

At the top of the hierarchy sits the Global Crate (Figure 2) which is a 6U VME crate that houses the Event Builder processor(s) (EVB) and the Global Level 3 Processor(s) (GL3). The main EVB processor controls a fast tape unit via a fast-wide SCSI connection.

The processors in the Global Crate are controlled directly via the DAQ Manager Workstation which is currently running a normal Unix OS and which communicates with the rest of the DAQ hierarchy via TCP/IP over Ethernet.

The Sector Crates (Figure 3) house 6 Receiver Boards, each of which is connected to the detector front-end via a fiber optic cable. The Sector Crate contains one master processor (the Sector Broker) which is the main interface of this crate to the rest of the system for control purposes. Apart from the receiver boards the Sector Crate contains one or more Sector Level 3 (SL3) processors which will either be fast VME processors directly plugged into the crate (as shown in Figure 3) or will reside in a different housing and be connected to the Sector Crate via a dedicated network.

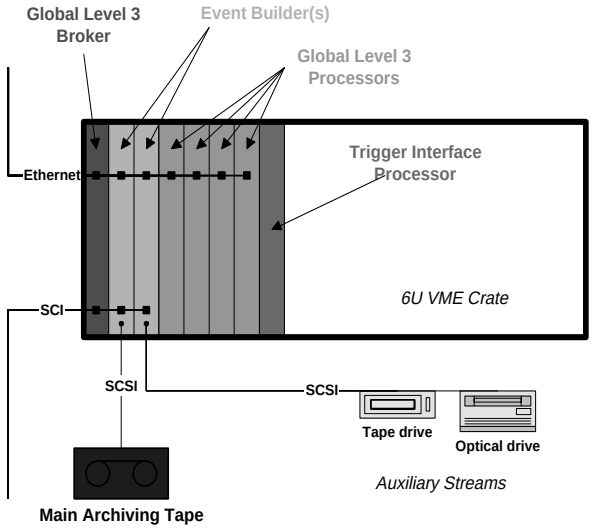

Figure 2 - Schematic view of the global VME crate

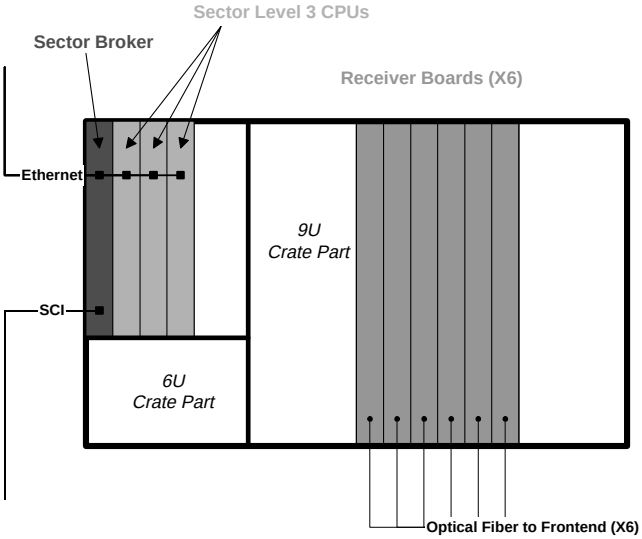

Figure 3 - Schematic view of a Sector Crate

B. Sector Crates

The Sector Crates are DAQ's front-end to the data stream. The whole system consists of 24 VME crates for the TPC detector and 4 VME crates for the SVT.

The physical crates are custom made 9U high crates but with a number of slots made for 6U VME cards. These slots are occupied by off-the-shelf VME processors like the Sector Broker and Sector Level 3.

a) Sector Broker

The Sector Broker is a VME processor whose main job is to be the central controller of the Sector Crate. The software running in the Sector Broker "knows" all the details of the crate thus simplifying the software interface with the rest of the system and isolating one crate from the other.

The current choice is a Motorola MVME2604 with 32 MB of RAM running at 200 MHz and has a single PMC slot which is used for the VMETRO/Dolphin SCI network card.

The main tasks of the Sector Broker are to:

- provide a boot server for the mezzanine kernels. The mezzanines have no network connection so the SB acts as a "gateway" to their boot code and data.

- provide the only contact/command point the mezzanines see in the crate. All the commands given to the crate and the mezzanines are passed from the global DAQ control to the Sector Broker first. The SB in turn distributes/gathers this information to/from the mezzanines in the system

- provide configuration information for the mezzanines and the Sector Level 3 CPUs. The configuration information originates from the DAQ Manager (and in turn from the experiment run-control).

- provide a "sink" for the error messages originating from the mezzanines. The SB will in turn pass these messages to the DAQ Manager for display.

- orchestrate the building of an event at the sector level

- orchestrate the interaction of the Sector Level 3 CPUs with the mezzanines

a) Receiver Boards

The RBs are custom built VME boards. They receive the raw data over the fiber from the detector front-ends. They also provide intermediate storage for the event in buffers which have enough capacity to store 12 full events in VRAM thus adding some elasticity to the overall DAQ pipeline.

Enough processing power is provided in the RB to format their part of the event as appropriate if the trigger logic decides it wants to accept and store the data.

The other important task of the RBs is to do cluster finding on their portion of the raw data (with the help of STAR DAQ ASICs [2]) thus preparing the data for further Level 3 processing in the Sector Level 3 processors.

More details will be given in a following chapter.

a) Sector Level 3

The Sector Level 3 processors are the processors that do the actual track finding at the sector level of the TPC or SVT. The input to their track-finder algorithms are the clusters already found in the Receiver Boards and their output is a list of track parameters that is shipped to the Global Level 3 processors which in turn make the final trigger decision.

The most important requirements are:

- cheap processing power (especially floating-point)

- good connectivity with the rest of the system and the mezzanines

Commercial, off-the-shelf VME CPUs fill these requirements. While this can easily be done a more attractive idea from the point of view of the cost is to connect cheap PCI motherboards already equipped with CPUs to the Sector Crate VME bus via a fast network.

A typical PCI motherboard is currently much cheaper than a VME SBC thus making this idea attractive in terms of price/performance. However, the first problem encountered was how to connect a PCI motherboard transparently to a VME system but due to the availability of SCI network cards in both the PCI (edge-connector) and PMC form factors (which are well suited to VME) this may be a viable option. This topology is currently under investigation

C. Global Crate

The main VME crate in the hierarchy is the Global Crate. All the Sector Crates are connected to the Global Crate via the SCI network for data passing, event building and event-related control.

This crate also houses the Trigger Control CPU which provides the connection to the Level 0,1 and 2 Trigger subsystem and passes information about the trigger detectors either for use by the Level 3 scheme or for later storage to tape.

a) Event Builder

The Event Builder (EVB) is responsible for gathering the event fragments from all of the sectors of all the detectors and storing them to the archiving media.

The EVB consists of two parts: the Main EVB which handles the main archiving stream and the Auxiliary EVB which handles the auxiliary streams.

The auxiliary streams are connected to slow, cheap and readily available storage devices and their main use will be during the debugging, testing and commissioning phase of various detectors.

STAR's detectors will become available in a time period of a few years so a way must be provided for these detector groups to gather their test and setup data away from the main data taking which would be in progress.

The auxiliary streams are of secondary importance; they can be started/stopped in the middle of the main data run, they can be taken offline without affecting the main stream, etc. Care will be taken that they don't appreciably affect the performance of the main data stream.

a) Tape Storage

The global RHIC Computing Facility (RCF) plans to provide storage for all the RHIC experiments through a Hierarchical Storage Manager facility. Each experiment would transfer the data to RCF via an optical fiber on an event-by-event basis. The data would physically reside in the RCF for further use by the offline systems of the various experiments.

Although the EVB's output stage is easily adapted to handle this type of output the exact interfaces (both hardware and software) are not specified by RCF at this time. To provide a working system during commissioning and tests a "classical" tape drive is used which can be then later be used as a backup whenever RCF becomes available.

The main requirements are the throughput and the storage capacity since the required throughput is of the order of 12-15 MBytes/s on average which is also the size of our event.

At this moment there are several different tape systems on the market which satisfy our requirements, all of them using fast-wide SCSI as their interface. The current DAQ architecture provides support for the SCSI interface but the purchase of a particular tape system is deferred until a later time.

a) Global Level 3

The Global Level 3 is the processing point that gathers all the track information from the Sector (i.e. Local) Level 3 processors in the system. It then assembles all the tracks of a given event, processes the event as a whole and finally makes a decision to either build and store the event or discard it.

The algorithms running in this system of processors will be refined through time as we understand the nature of the physics in these collisions.

a) Trigger Control CPU

The main part of DAQ's communication with the Trigger Group will be done through this board via VME shared memory and VME interrupts.

The Trigger provides a unique "token" for every Level 0 trigger (event) issued thus uniquely labeling this event throughout its lifetime in the STAR DAQ. The token will be reused after the event is either discarded by the higher level triggers or stored to tape.

This interface also enables the DAQ to gather the data of the trigger detectors which can assist the Level 3 trigger decision and/or will be stored to tape together with the event data for further offline analysis.

D. DAQ Manager Workstation

The DAQ Manager is the main control workstation. It runs a version of the Unix OS without any special real-time options.

Its main purpose is to be the interface to the outside world and will provide the following:

- a hard disk that will be shared through NFS with the various VME processors in the system
- a common boot platform for the VME processors
- an error logging and display facility
- a DAQ monitoring and display facility
- a common database for various system uses
- a run control interface during debugging and testing

The workstation will additionally be equipped with a dedicated 100 Mbit Ethernet interface and will act as a firewall towards the rest of the DAQ since the DAQ VME system shouldn't be accessible from the outside world.

E. Networks

The two networks in the system are Ethernet and SCI (Scalable Coherent Interface).

Ethernet is the classical 10 Mbit Ethernet running through switched hubs with a 100 Mbit/s uplink towards the DAQ Manager for performance reasons. It is only used for "slow" commands: setup, booting, run-control, shared disks etc.

SCI is a very low latency (3-5 us), high bandwidth (800 Mbits/s) network which is well suited for our needs: low latency for “fast” command messages of small size and, at the same time, high throughput for the data chunks during event building.

A single loop topology is currently used but if necessary it can easily be subdivided to a multiple loop setup or a SCI switch may be added which is currently also available.

Although SCI could be used instead of Ethernet, Ethernet (particularly TCP/IP) gives us the benefit of a well proven technology with all the existing software without additional costs.

On the other hand, the SCI network is very well suited for an extremely light-weight protocol which is custom developed and used for message passing throughout the DAQ system.

III. Receiver Boards

The Receiver Boards are custom designed 9U VME boards that plug in the Sector Crates. Each Receiver Board (Figure 4) is additionally connected to a front-end Readout Board via a dedicated optical fiber operating at 1.5 Gbits/s.

The board itself consists of a 9U motherboard and 3 custom mezzanines that plug face down on the motherboard. Each mezzanine holds the 6 custom built STAR DAQ ASICs [2].

A. The Motherboard

The main purposes of the motherboard are to provide a data path from the front-end electronics to the processing intelligence (the 3 mezzanines) and to provide a data path for the processed data to the Sector Broker (and onward to the Event Builder) and the Sector Level 3 Processors.

The board contains a HP Glink/Methode combination to provide the fiber optical interface which is coupled to a multiplexer that feeds the 3 daughterboards. The multiplexer logic also strips the header from the front-end data streams and places it in a VRAM device for further use by the mezzanines. This front-end interface and multiplexer is implemented in programmable CPLDs and is running at 60 MHz.

The data path to/from the VME utilizes a local PCI bus connected to VME through a Tundra Universe PCI-to-VME bridge chip. The other PCI device on the board is a PLX PCI 9060SD bridge chip which provides PCI access to the header VRAM, 8 MB of general purpose SDRAM and some other registers used for control of the board. The board also provides 3 PMC PCI slots for the 3 mezzanines. The PCI bus is 32 bits wide and runs at 33 MHz.

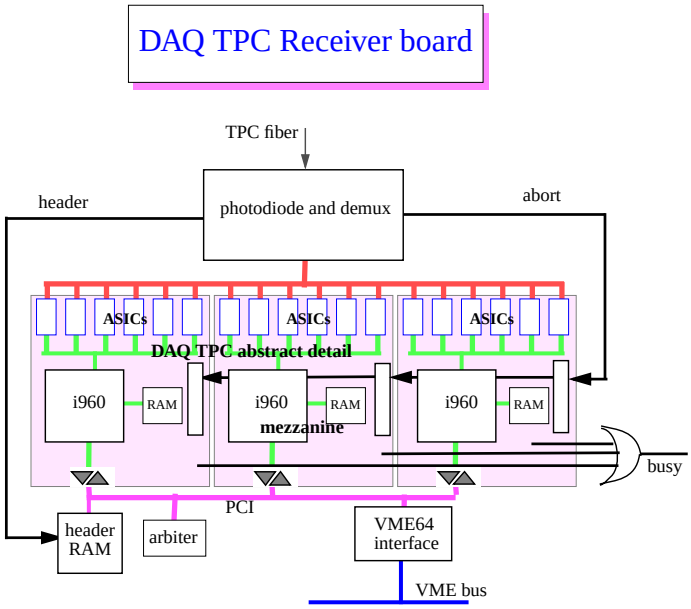

Figure 4 - The DAQ Receiver Board

B. The Mezzanine Boards

The mezzanine boards (Figure 5) are PCI boards with PMC connectors for the PCI side and separate auxiliary connectors for the front-end data. Additional power is also provided through these separate connectors. The boards are 4 inches wide and 14 inches long.

The boards house 6 STAR DAQ ASICs with each ASIC connected to a 128 k X 8bit SRAM (512 k X 8 bit for the SVT case). The SRAMs hold the threshold values for each channel as well as the results of the ASIC specific processing - the cluster-pointer pairs which are explained in the next section.

The master of the board is an Intel I960HD processor running at 33 MHz bus speed and is acting as the processor/controller of the board.

The board contains a PLX PCI9060 bridge chip which bridges the PCI bus and the local Intel I960 bus running at 33 MHz.

The board additionally houses 4 MB of SDRAM memory which is the working RAM for the CPU and 4 MB of VRAM organized in 2 banks of 4 chips each. The VRAMs are used to store the raw data as it comes through the ASIC and also to store the cluster-pointer pairs at the end of event processing.

The 4 MB size was chosen to be large enough to accommodate and buffer 12 events.

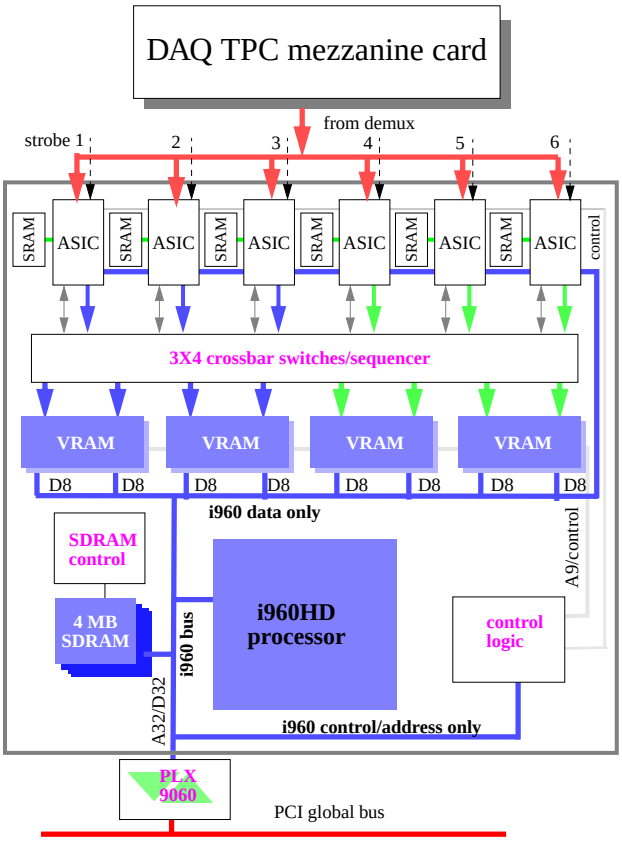

Figure 5 - The DAQ Mezzanine Board

The sequential side of the VRAMs is connected to the data outputs of the ASIC through a digital switch which is in turn controlled through a programmable CPLD. This setup allows data reordering and address translations thus reordering the data residing in VRAM so that it is presented to the CPU in such a way to minimize the subsequent work of the CPU.

C. The STAR DAQ ASIC

The ASIC has two main tasks:

1. It does the 10-to-8 bit translation of the incoming data

2. It performs thresholding (on a per channel basis) of the raw data and subsequent cluster finding “on the fly” and stores the beginning and ending time sequences of the clusters in separate SRAM which is associated with each ASIC. These pairs of pointers are 16 bits wide and are referred to as Cluster Pointer Pairs (CPP). They are stored in the associated SRAM during data processing and are transferred to the VRAM at the end of an event upon a command from the VRAM controller.

The ASIC is covered in more detail in [2].

VI. References

[1] STAR Collaboration, "Conceptual Design Report," *Lawrence Berkeley Lab., University of California*, PUB-5347, June 1992.

[2] M. Botlo et al, “The STAR Cluster-finder ASIC”, *This publication*