A Data Formatter for the ATLAS Fast Tracker

Jamieson Olsen, Ho Ling Li, Ted Liu, Yasuyuki Okumura, Bjoern Penning

***Abstract*- The Fast TracKer (FTK) is an upgrade to the ATLAS level-2 trigger. The FTK system will reconstruct tracks using data from the inner Pixel and SCT silicon detector modules at trigger rates up to 100 kHz. We present an overview of the Data Formatter system, which is designed to remap, share and reformat the Pixel and SCT module data to match the geometry of the FTK trigger towers.**

I. Introduction

CROSSINGS in the LHC occur at the nominal rate of 40 MHz with a design luminosity of 1×10^{34} $cm^{-2}s^{-1}$ with approximately 25 overlapping proton-proton interactions. The ATLAS detector trigger system must reject a vast majority of these events as only 200 events per second can be stored for later analysis. Instantaneous luminosity is expected to increase to 3×10^{34} $cm^{-2}s^{-1}$ with an average of 75 proton-proton interactions per crossing. Under these conditions the existing ATLAS trigger is strained and the need for a tracking trigger is clear.

The Fast Tracker (FTK) processor is an upgrade which adds a hardware-based level-2 track trigger to the ATLAS DAQ system [1]. The FTK system includes a Data Formatter to remap the ATLAS inner detector geometry to match the FTK η-φ trigger towers. The Data Formatter system also performs pixel clustering and data sharing in overlap regions. Based on the current design requirements and the need for future expansion capabilities, a full mesh Advanced Telecom Computing Architecture (ATCA) backplane interconnect is a natural fit for the Data Formatter design. Our baseline design also works well as a general purpose FPGA-based processor board. The Data Formatter may prove useful in scalable systems where highly flexible, non-blocking, high bandwidth board to board communication is required.

II. The Fast Tracker

The FTK system finds tracks using data from the ATLAS inner detector Pixel and SCT modules shown in Fig. 1. In response to a level-1 accept chains of Pixel and SCT modules are read out through front end electronics (radiation hardened ASICs) and Readout Driver (ROD) crates. The ROD outputs are duplicated using a new SLINK transmitter mezzanine board and these extra output links are used by the FTK system. In total the FTK system receives 222 gigabit fiber SLINKs from the Pixel and SCT RODs.

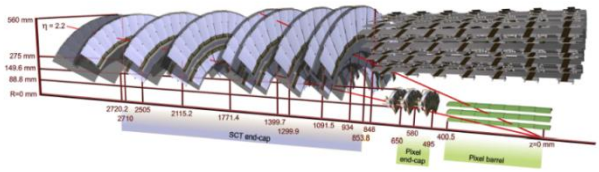

Fig. 1. ATLAS Pixel and SCT silicon detector modules. The Pixel sub-detector is composed of 1,744 modules arranged in three barrels and six end-cap disks. The SCT sub-detector is composed of 2,112 modules arranged in four barrels and 1,976 modules arranged in 18 end-cap disks. In total the Pixel and SCT modules contain over 90 million silicon detector elements.

The arrangement of the inner detector Pixel and SCT modules does not match the geometry of the 64 FTK η-φ towers. An additional hardware layer is needed to intercept the ROD output links and remap, share and reformat the Pixel and SCT module data prior to transmission to the FTK hardware. This hardware layer is the Data Formatter system. The FTK system is shown in Figure 2.

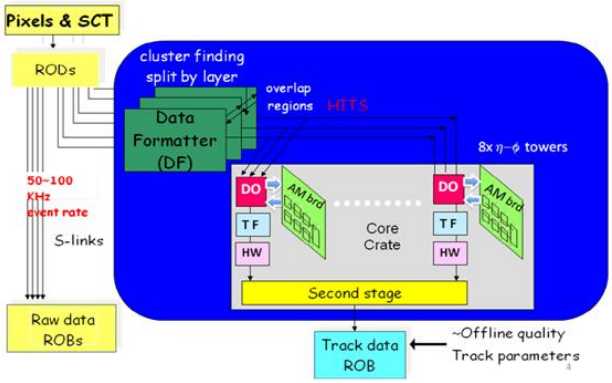

Fig. 2. The FTK system receives copies of the ROD outputs. The Data Formatter boards remap, share, and reformat the ATLAS Pixel and SCT inner detector module data so that it matches the FTK system geometry.

III. The Data Formatter

The Data Formatter is an 8U×280mm ATCA board which supports up to four mezzanine cards and two Kintex-7 FPGAs. These FPGAs connect directly to the full mesh fabric and fiber optic transceivers on a rear transition module (RTM). The Data Formatter block diagram is shown in Fig. 3 and the prototype board layout is shown in Fig. 4.

Manuscript received June 15, 2012.

Jamieson Olsen is with the Fermi National Accelerator Laboratory, Batavia, IL 60510 USA (telephone: 630-840-2779, e-mail: jamieson@fnal.gov).

Ho Ling Li is with the Department of Physics, University of Chicago, Chicago, IL 60601 USA (telephone: 773-702-8097, e-mail: hlli@uchicago.edu).

Ted Liu is with the Fermi National Accelerator Laboratory, Batavia, IL 60510 USA (telephone: 630-840-6675, e-mail: thliu@fnal.gov).

Yasuyuki Okumura is with the Department of Physics, University of Chicago, Chicago, IL 60601 USA (telephone: 630-840-6675, e-mail: yasuyuki.okumura@cern.ch).

Bjoern Penning is with the Fermi National Accelerator Laboratory, Batavia, IL 60510 USA (telephone: 630-840-6623, e-mail: penning@fnal.gov).

Operated by Fermi Research Alliance, LLC under Contract No. DE-AC02-07CH11359 with the United States Department of Energy

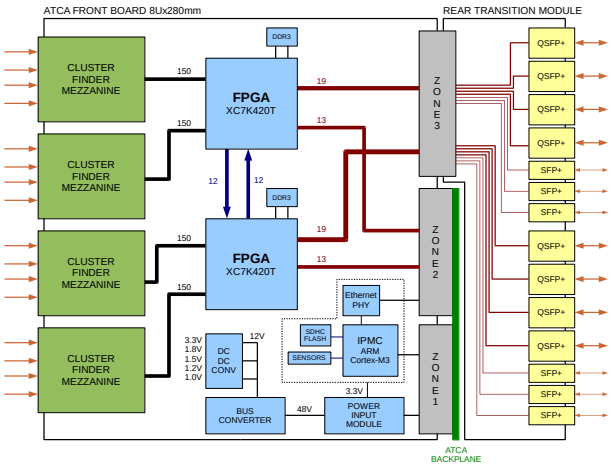

Fig. 3. The Data Formatter board with the rear transition module (RTM).

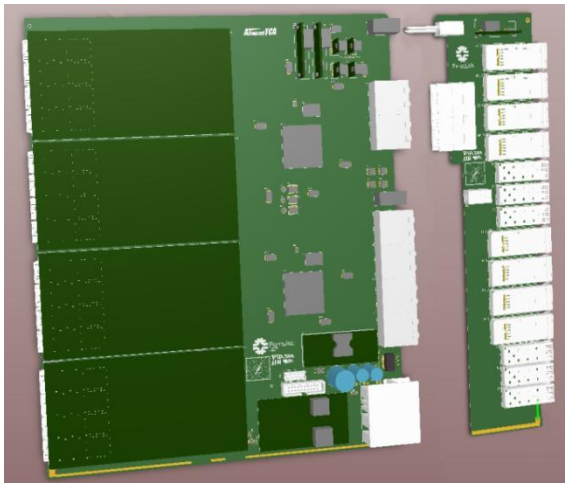

Fig. 4. The prototype Data Formatter and RTM boards in layout. The high speed serial connections from the FPGAs to the backplane fabric interface and RTM connectors are short and direct, simplifying board layout and helping to improve signal integrity for maximum performance. The printed circuit board is ten layers.

A. *Mezzanine Cards*

The Data Formatter board supports up to four mezzanine cards. These mezzanine cards are similar to PMC mezzanine cards but use different connectors. A single PMC style connector is provided for backwards compatibility with an SLINK receiver mezzanine card (HOLA). The other mezzanine card connector is based on the HSMC specification [2] and supports high speed LVDS data rates of up to 64 Gbps.

A multi-channel SLINK receiver and cluster finder mezzanine card is in development [3]. This mezzanine card receives up to four gigabit fiber SLINKs and contains FPGAs which are used to implement 2D cluster finder algorithms. The cluster finder mezzanine cards will use the HSMC connector to communicate with the Data Formatter FPGAs.

B. *FPGAs*

The heart of the Data Formatter board is a pair of Xilinx Kintex-7 FPGAs. Each FPGA implements the core logic for a single FTK η-φ tower. The prototype Data Formatter board will use XC7K325T devices and the production board will use the larger XC7K480T devices when these parts become available later this year. Kintex-7 FPGAs feature up to 480k logic cells and up to 32 12 Gbps serial transceivers.

Each FPGA has an external 256MB memory chip which may be used for diagnostic spy buffers or other general purpose data storage. The memory chip is a DDR3-800 device with a 16-bit interface. Data rates on the order of 1 GBps are possible when accessing this external memory.

The FPGAs share data over a high speed LVDS local bus which supports data rates up to 20 Gbps in each direction. The FPGA high speed SERDES transceivers (GTX) connect directly to the ATCA full mesh backplane and RTM fiber transceivers. Since the FPGAs connect directly to the fabric interface no external cross-point switch chips are required and the board layout is simplified.

C. *ATCA Backplane*

The ATCA full mesh backplane consists of up to four bidirectional ports per channel, where each port is rated for up to 10 Gbps. In a 14 slot crate 13 channels are provided for each slot. The Data Formatter top FPGA connects to port 0 and bottom FPGA connects to port 1 of each channel. This means that within a crate all top FPGAs are directly connected and all bottom FPGAs are directly connected through the full mesh *fabric interface*. Any data protocol which uses low voltage 100 ohm differential signaling is allowed on the fabric interface.

A dual star Gigabit Ethernet *base interface* is also provided on the ATCA backplane. This interface is intended to provide processor boards with a medium-speed network connection via a switch or hub board installed in logical slots 1 or 2.

D. *Intelligent Platform Management Controller*

High availability is achieved through redundancy and a robust hardware management scheme. Every active component in an ATCA crate is expected to monitor its health and communicate with the shelf manager boards over the Intelligent Platform Management Interface (IPMI). ATCA boards use the IPMI protocol to report various sensor readings and coordinate the hot-swap power sequencing through the shelf manager. The Data Formatter boards use an inexpensive ARM Cortex-M microcontroller to implement the IPMI controller (IPMC) functions.

The Data Formatter microcontroller is also used for slow controls and managing FPGA firmware image files, which are stored locally on an SDHC flash memory card. Firmware image files may be downloaded to the flash memory over a 100BASE-T Ethernet interface which connects to the backplane *base interface*. After the FPGAs are configured the microcontroller may then access firmware registers over a dedicated SPI serial bus.

E. *Rear Transition Module*

The Data Formatter FPGAs connect to fiber transceivers located on the RTM. Quad small form factor pluggable (QSFP+) fiber optic modules support data rates up to 40 Gbps

and are used to send data downstream to the FTK system. Small form factor pluggable (SFP+) transceivers support data rates up to 10 Gbps and are intended for inter-crate sharing. The CERN SLINK protocol will be used on all input and output fiber links.

The RTM supports hot-swap power sequencing and is managed by the Data Formatter IPMC microcontroller. The Data Formatter RTM will conform to the new ATCA RTM standard [4].

F. Power

The ATCA backplane supplies dual 48VDC busses to each slot. Board power inputs are fused and connected to an ATCA-specific power input module (PIM) which contains filters and a small isolated DC-DC converter used to power the IPMC. Upon board insertion the IPMC powers up and negotiates board power requirements with the shelf manager. The shelf manager then grants permission for the board to enable its isolated 48V to 12V DC-DC bus converter, which supplies power to the main board, mezzanine cards and RTM. Small non-isolated DC-DC converters are used to generate the low voltage power rails for FPGAs and other components.

Power consumption on the Data Formatter board is estimated to be less than 30W (not including the mezzanine cards). When fully loaded with fiber transceivers the RTM power is estimated to be on the order of 20W. The ATCA crate is rated for up to 200W per slot using forced air cooling. Hearing an ATCA crate running with the fans set to maximum speed leaves little doubt as to its substantial cooling capability.

IV. Data Sharing

Data Formatter inputs consist of 222 gigabit SLINKs from the Pixel and SCT RODs. Each input SLINK supplies data from 7 to 48 modules depending on the sub-detector type and geographic location. The mapping between modules and input links is somewhat irregular and asymmetric (e.g. in some cases a single input link contains modules from a large φ region). Such input links pose a challenge because their data must be shared with several FPGAs.

Since each SLINK must ultimately connect to a single FPGA the "best" input link assignment minimizes the data sharing between FPGAs. After optimizing the input link assignments data sharing between the 64 FPGAs was analyzed and the resulting matrix is shown in Fig. 5. In this matrix the non-zero elements represent Pixel and SCT modules which must be shared between FPGAs. The four red boxes represent crate boundaries. FPGAs located on Data Formatter boards within these crate boundaries will share data directly over the full mesh backplane. Non-zero matrix elements outside the crate boundaries represent modules which must be shared between FPGAs in different crates using dedicated fibers on the RTM.

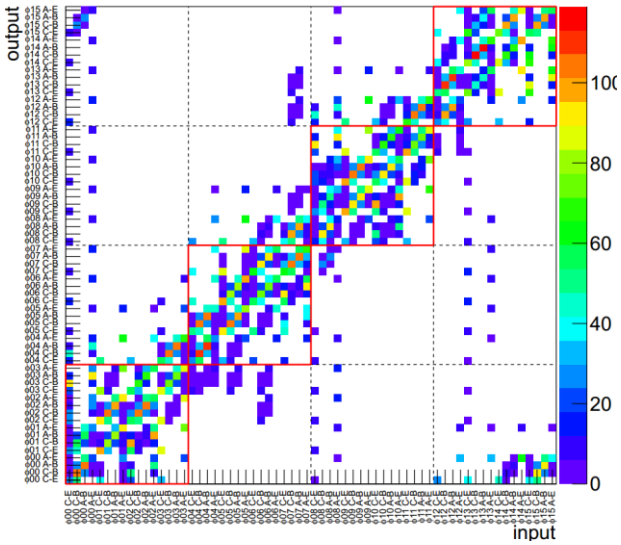

Fig. 5. This 64×64 matrix illustrates data sharing between Data Formatter FPGAs. Off diagonal elements represent the number of inner detector modules which must be transferred between FPGAs. The red boxes represent crate boundaries.

A. Data Flow Analysis

A software model of the Data Formatter has been created so that data flow in the system may be simulated using actual event records obtained from the RODs. For this analysis we assume that the input mezzanine cards do not perform any clustering (e.g. no data reduction) and the level-1 accept rate is 100 kHz. The ROD output data format is assumed to be the same as that which was used in 2011 and 2012 data taking. We also assume that whole modules are shared between FPGAs. (Sub-dividing modules to more accurately match η-φ tower boundaries will result in a data volume reduction.)

Fig. 6 and Fig. 7 show the results obtained from 1300 events from a run in 2011 with center-of-mass energy of 7 TeV and ten interactions per bunch crossing (pileup $<\mu>=10$). The results from our simulation agree well with the FTK technical proposal.

Table I summarizes the bandwidth requirement evaluated with 2011 data for output and internal communication links respectively, as well as requirements extrapolated to 14 TeV and $<\mu>=80$. For the extrapolation, data volume is assumed to be proportional to $<\mu>$ and 15% correction is applied by considering the track multiplicity difference between 7 TeV and 14 TeV.

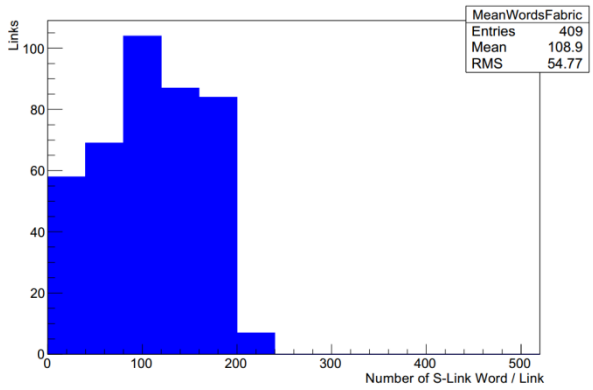

Fig. 6. Data volume distribution on the ATCA full mesh backplane links. The high traffic links transfer 240 words per link per event, corresponding to a data rate of 0.64Gbps.

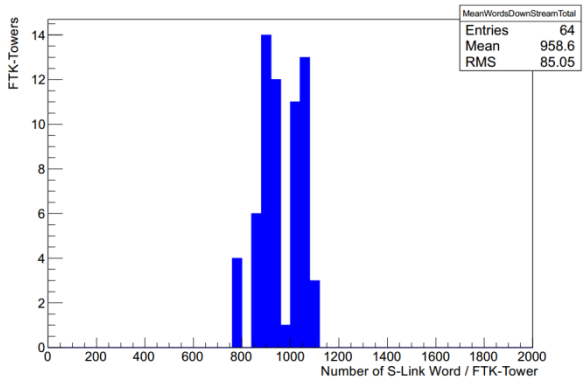

Fig. 7. Data volume distribution on the Data Formatter output links. The busiest links transfer 1100 words per link per event, corresponding to a data rate of 3.5 Gbps.

TABLE I. DATA RATE SUMMARY

Link	7 TeV <μ>=10	14 TeV <μ>=80
Output to FTK AUX cards	2.9 Gbps	26.7 Gbps
Output to Second Stage Boards	0.6 Gbps	5.9 Gbps
Local Bus	2.4 Gbps	22.1 Gbps
Fabric Channel	0.6 Gbps	5.9 Gbps
Inter-Crate	0.7 Gbps	6.6 Gbps

B. *Data Re-Routing Techniques*

Our data flow analysis reveals that some full mesh channels are either not used or are lightly used while other channels are have a high data volume (Fig. 6). If the FPGA data packet routing firmware supports re-transmission then it is possible to divert traffic from high data volume channels and re-route the data over low data volume channels on the full mesh backplane. Data re-routing and re-transmission is attractive because it has the potential to increase the effective bandwidth of the system, simultaneously reducing latency and increasing throughput. The performance and logic resource requirements of an advanced data packet routing scheme (Fig. 8) are currently being investigated.

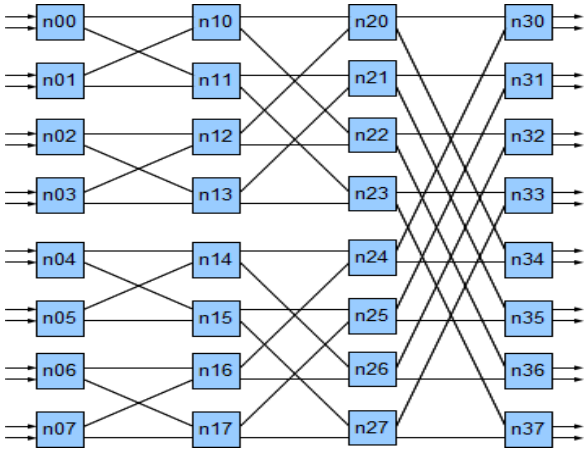

Fig. 8. A 16×16 port internally-buffered banyan switch. This type of switch may be present in each Data Formatter FPGA and could support data packet re-transmission, thereby taking advantage of under-utilized alternate data paths in the full mesh backplane.

C. *A Scalable and Flexible Architecture*

The Data Formatter system will consist of 32 boards in four ATCA crates. Each crate will have six available slots for future expansion. Additional Data Formatter boards may be installed as more input links are added to the system. These extra Data Formatter boards will however not use an RTM as they will route their data packets over the full mesh fabric.

Inter-crate bandwidth may be increased by simply adding more parallel fibers between crates and updating the data packet routing tables. Ether the SFP+ (up to 10 Gbps) or QSFP+ (up to 40 Gbps) transceivers may be used for inter-crate data transfers.

V. CONCLUSION

A "bottom up" approach was employed when designing the Data Formatter hardware. Based on the input/output requirements and data sharing analysis we determined that the ATCA full mesh backplane is a natural fit for the Data Formatter system. Our data flow analysis suggests that all data rates in the Data Formatter system are within the bandwidth limits imposed by the FPGAs, fiber optics, and full mesh backplane channels. Prototype boards are in layout now and we anticipate fabricating the first boards later this year.

REFERENCES

[1] A. Andreazza, "The FastTracker Real Time Processor and Its Impact on Muon Isolation, Tau and b-Jet Online Selections at ATLAS", Trans.Nucl.Science, Vol 59, Issue 2.
[2] High Speed Mezzanine Card, 1.7, 2009.
[3] A. Annovi and M. Beretta, "A Fast General-Purpose Clustering Algorithm Based on FPGAs for High-Throughput Data Processing", submitted to NIMA, arXiv:0910.2572v1.
[4] Advanced TCA Rear Transition Module, PICMG 3.8, 2011.