SLAC-PUB-16182

Design of the SLAC RCE Platform: A general purpose ATCA based data acquisition system.

*R. Herbst, R. Claus, M. Freytag, G. Haller, M. Huffer, S. Maldonado, K. Nishimura, C. O'Grady, J. Panetta, A. Perazzo, B. Reese, L. Ruckman, J. G. Thayer, M. Weaver

Abstract - The SLAC RCE platform is a general purpose clustered data acquisition system implemented on a custom ATCA compliant blade, called the Cluster On Board (COB). The core of the system is the Reconfigurable Cluster Element (RCE), which is a system-on-chip design based upon the Xilinx Zynq family of FPGAs, mounted on custom COB daughter-boards. The Zynq architecture couples a dual core ARM Cortex A9 based processor with a high performance 28nm FPGA. The RCE has 12 external general purpose bi-directional high speed links, each supporting serial rates of up to 12Gbps. 8 RCE nodes are included on a COB, each with a 10Gbps connection to an on-board 24-port Ethernet switch integrated circuit. The COB is designed to be used with a standard full-mesh ATCA backplane allowing multiple RCE nodes to be tightly interconnected with minimal interconnect latency. Multiple shelves can be clustered using the front panel 10-gbps connections. The COB also supports local and inter-blade timing and trigger distribution. An experiment specific Rear Transition Module adapts the 96 high speed serial links to specific experiments and allows an experiment-specific timing and busy feedback connection. This coupling of processors with a high performance FPGA fabric in a low latency, multiple node cluster allows high speed data processing that can be easily adapted to any physics experiment. RTEMS and Linux are both ported to the module. The RCE has been used or is the baseline for several current and proposed experiments (LCLS, HPS, LSST, ATLAS-CSC, LBNE, DarkSide, ILC-SiD, etc)

I. Introduction

Every physics experiment deployed requires some form of a data acquisition system. It is very common to adapt an existing design when building a data acquisition system for a new experiment. Although this approach is attractive in that it reduces risk and minimizes the learning curve, it results in the experimental group not employing the latest available technologies and lagging behind the technology curve. Very often the computational requirements of the new experiment exceed the abilities of the antiquated technology, causing the experiment to make compromises in its data processing or start a new design specific to the experiment

Attempts to employ commercial equipment are met with limited success. Although commercial equipment can be found to solve individual problems, the integration of these various components, usually from different manufactures, proves awkward and often results in a complex system which is often more expensive.

SLAC has embarked on a research project intended to package the common requirements of a data acquisition system into a generic platform which can easily be adapted for the specific needs of existing and future experiments. By defining a clear set of interfaces, both at the hardware and software level, the common system can be upgraded and modernized as necessary without requiring the experimenters to struggle with a new unfamiliar system.

The SLAC RCE Platform is built upon a number of leading edge commercial electronic integrated circuit components packaged into a commercial crate technology. It is architected to be flexible enough to be adapted to the specific needs of a physics experiment, while providing a turnkey base platform.

II. RCE Platform Architecture

The core of the RCE platform is the Reconfigurable Cluster Element (RCE), which is a system-on-chip design based upon the Xilinx Zynq family of integrated circuits. The Zynq architecture couples a dual core ARM Cortex A9 based processor with a high performance 28nm Field Programmable Gate Array (FGPA).

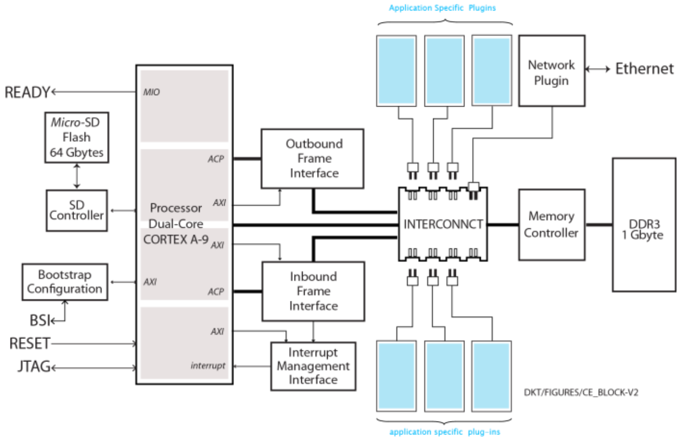

Fig. 1: Block diagram of the RCE.

The RCE is a synergy of software and firmware. Together they combine higher level programming with low level firmware data processing, with a flexible boundary between the two. An interface is defined at the software to firmware boundary, defined here as the "Protocol Plug In". A common socket exists as a firmware interface in the FPGA and an API at the software layer. This socket is defined to be hardware independent, protecting the application firmware and software code from changes in the underlying architecture. Application software and firmware exists as a "plug" connecting to the standard socket. A set of common blocks exist which interface

[*] First author: rherbst@slac.stanford.edu
The authors are with the Research Engineering Division, SLAC National Accelerator Laboratory, Menlo Park, CA.

This material is based upon work supported by the U.S. Department of Energy, Office of Science, Office of Basic Energy Sciences, under Contract No. DE-AC02-76SF00515.

to these sockets, accelerating the development cycle of the end user.

To provide for the most common use cases, a set of plug-ins are provided by default by the RCE. The first of these is a network interface, supporting both a 1Gbps and a 10Gbps Ethernet interface. Secondly, a bootstrap interface is provided which interfaces to the overall shelf management system. In addition to allowing the RCE to communicate its status to the central shelf manager, this interface allows the RCE to be gracefully shutdown.

Common software architecture across multiple experiments is very important so new experimenters, or collaborators and users, can program and use the system easily. The development of a stable base software platform is often underestimated or not implemented which in turn can make the system difficult to use.

The RCE software and firmware is stored locally on a removable micro-SD card. When the system is powered up, the boot loader reads the FPGA firmware from this SD card and loads it to the programmable fabric. Upon successful load of the firmware the chosen operating system is then loaded. Currently two operating systems are supported by the SLAC RCE platform: Linux and RTEMs.

For Linux, SLAC uses the ArchLinux distribution combined with the 3.x Linux kernel provided by Xilinx. ArchLinux was chosen for its capability with the current Xilinx Kernel as well as its rolling upgrade approach, easing the process of keeping the distribution up to date. Most importantly, ArchLinux has a strong user base which supports the ARM architecture.

RTEMs is provided as an alternative to Linux, for applications which require a real time operating system. RTEMs is an open source real-time operating system, designed for embedded systems. RTEMS supports various open API standards including POSIX, providing compatibility with the Linux operating system. A common software API is provided for the protocol plug in layer, allowing application layer software to be easily ported between operating systems.

III. Packaging And Data Flow

An RCE node can be used as a standalone data processor or as port of a larger system. When used alone, it typically sits on a carrier board which contains the required support circuitry. This board will provide power and external interfaces for the RCE. This paper focuses on the applications which use the RCE in a larger clustered platform.

A total of 9 RCE nodes are supported on a single Cluster On Board (COB) ATCA blade. 8 of these nodes are designated for data processing while the remaining node handles interconnect management and timing distribution. The 8 processing nodes are organized in pairs on daughter boards known as the Data Processing Modules (DPMs), with a total of 4 DPMs supported on a single COB. A Data Transport Module (DTM) daughter board contains the remaining RCE node.

Data from physics experiments is received by each RCE node over one or more high speed serial links through a Rear Transition Module (RTM). Each RCE node supports 12 bi-direction high speed links with a total of 96 high speed links supported by the COB. The 12 serial links on each RCE are extremely flexible supporting a wide range of speeds (<100Mbps – 12Gbps) and protocols. While the total aggregate bandwidth available to each of the 8 RCE nodes on a COB is 144Gbps, it is expected that most applications will utilize a small number of high speed links or a large number of slow links. The specific technology used to transport these serial links between the RCE and the front end electronics is determined by the design of the RTM. The RTM is usually experiment specific, basically just adopting the connectors and signal levels between the detector and the DAQ. Most experiments perform some sort of optical to electrical conversion for the serial links. Analog-to-Digital conversion has also been implemented on the RTM for an experiment.

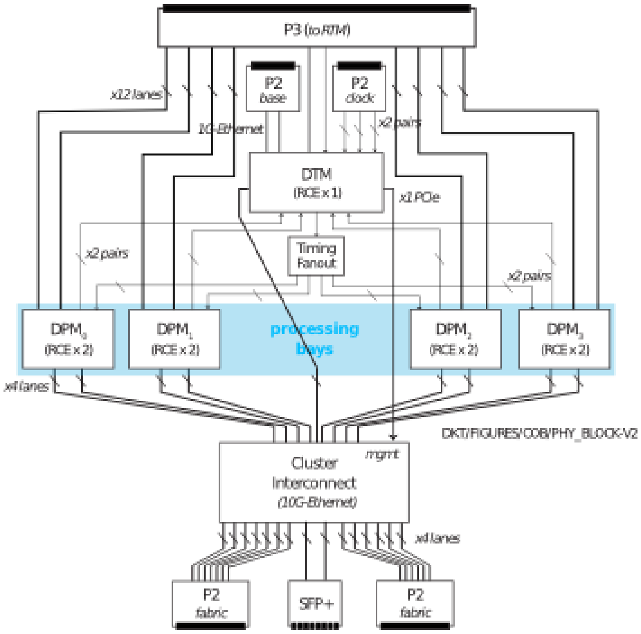

Fig. 2: Block diagram of the RCE platform.

IV. Data Processing

A common data processing step in physics experiments is to receive raw data from a front end and perform some form of data reduction. This data reduction process normally requires a set of computations which adjust for each channel's baseline, filter out low frequency baseline drift and then perform some form of threshold comparison on the incoming data. Secondary processing may then include waveform fitting and/or peak detection. The FPGA logic which forms the core of the RCE node is well suited to perform these types of calculations. DSP elements within the FPGA fabric can perform complex calculations on this incoming data before it is passed to the software running on the ARM processor. This allows for increased bandwidth between the experiment front-end and each RCE processing node.

The RCE platform provides a number of firmware libraries to support easy development without specific knowledge of the underlying platform. Similarly a number of pre-canned modules exist to facilitate the usage of the external high speed serial links. The most commonly used library is SLAC's own PGP (Pretty Good Protocol) which supports high speed data transport along with accurate fixed latency trigger and timing distribution. Additional protocols in our library include G-LINK, SATA, PCI-Express and Ethernet.

Timing delivery in the system is supported with a number of firmware cores. For most applications the standard timing delivery firmware can be used. This firmware supports clock, trigger pulse and timestamp delivery. An advanced user can make use of the lower level libraries and define their own timing delivery system, using the COB and ATCA resources in a customized fashion.

One of the biggest challenges in large scale firmware development is integrating a number of firmware blocks written by many different groups. This problem is amplified in environments where the application developer does not have a background in firmware development. This is often the case in physics experiments where most of the analysis and data processor work is done by physicists. Xilinx partial reconfiguration reduces this complexity by allowing regions of the FPGA to be independently configured. This allows the base portion of the RCE firmware to be compiled separately from the application firmware.

Similar application libraries are provided on the software side, including full GNU debugger support. Application libraries are available for common computing and data reduction processes performed on event data. Makefile systems for both software and firmware are provided in the software development kits available from SLAC.

IV. Clustering

In a number of experiments it is necessary to include data from neighboring detection sensors within a detector when implementing data reduction algorithms. For a given section of the detector, a block of these sensors are processed by the same RCE node, allowing this calculation to be performed locally. In some cases these neighboring sensors will exist in separate detector groups, processed by different RCEs. The RCEs processing these neighboring channels must share information in order to complete their individual computations. A low latency interconnect is required for the RCE nodes to share this information for each received event.

The 9 RCE nodes on the COB are locally interconnected by the Cluster Interconnect (CI). The CI is a low latency, Layer-3 compliant, 24-port 10Gbps Ethernet switch integrated circuit. Each of the 8 data processing RCEs has a 10Gbps line connected to the switch while the data transport RCE has a 1Gbps connection. Two of the switch ports are routed to the COB front panel, one supporting an SFP+ 10Gbps connection and the other supporting an SFP 1Gbps connection. The remaining Ethernet switch ports are routed to the ATCA backplane. These connections, combined with a full-mesh ATCA backplane, allow a high speed, low latency (< 300 ns per hop) cluster to be formed which includes up to 14 COB blades within a single ATCA crate. The front panel Ethernet connections provide an output path for the processed data, in additional to allowing multiple crates to be networked together to form a larger computing cluster.

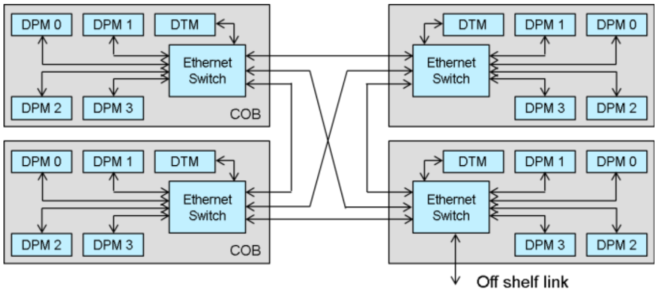

Fig. 3: ATCA Shelf Ethernet Clustering

This low latency clustered interconnect allows both software and firmware intercommunication between RCE processing nodes. Software intercommunication is provided through standard network protocols, including UDP multicast. Firmware intercommunication is provided using low level Ethernet.

V. Timing Distribution

A typical physics experiment requires that all of its elements, including the front end data collectors and the back end data acquisition nodes, be tightly synchronized. A central distribution system provides clock and trigger information while accepting busy feedback from each sub-system's data acquisition nodes. The data acquisition hardware designed for each experiment must contain logic to interface to this externally provided timing information, forwarding clock and trigger to the front end electronics. All of this functionality is included in the SLAC RCE system.

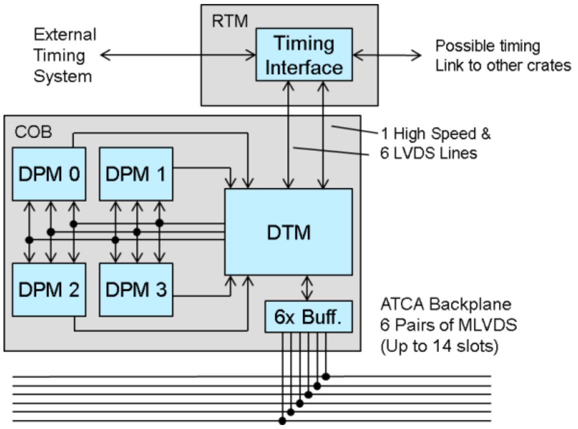
RTM
External Timing System
Timing Interface
Possible timing Link to other crates
COB
1 High Speed & 6 LVDS Lines
DPM 0
DPM 1
DTM
DPM 2
DPM 3
6x Buff.
ATCA Backplane 6 Pairs of MLVDS (Up to 14 slots)

Fig. 4 RCE Platform Timing Distribution

The DTM, along with experiment specific hardware on the RTM, serves as the interface between the external timing system and the internal nodes within the RCE platform. A number of signals are provided between the DTM and RTM to support a wide variety of timing interfaces. A single high speed serial link supports timing systems which are based on serial optical links. 6 bi-directional pairs can be used either as differential or single ended signals. The first three of the differential pairs are routed to global clock inputs on the DTM FPGA.

The DTM extracts the clock and trigger information from the external signals and then distributes them to the RCEs located both on the local COB. Three differential signals from the DTM are replicated and fanned out to each of the DPMs. One of these three signals is routed to one of the GTX reference inputs on the DPM FPGAs. The other two are connected to global clock inputs on the DPM FPGAs. A single pair from each DPM FPGA is routed back to the DTM. These signals allow the DPMs to provide busy, trigger acknowledge or other feedback to the DTM.

Clocking, trigger pulses and timing information are forwarded from the RCEs to the front end logic over the serial links between them. The serial transmitter in the RCE supports a mode which ensures a specific alignment between the input clock and the serial encoded protocol. Similarly the serial receiver is able to extract this block with a fixed alignment to the received serial data. This fixed latency transport ensures a fixed alignment between the clock on the RCE and the clock in the front end. The phase difference between the two is a function of the length of the physical medium over which the data is transported.

It is a requirement of most physics experiments to ensure that all of the front ends are aligned in phase. Logic on the RCE has the ability to measure the round trip latency of the serial link to a front end. It does this by sending down a special trigger code which gets echoed back from the front end along with a re-encoded copy of the clock. Logic contained in the RCE firmware measures this round trip phase. System level software can then adjust the delay of each serial link and properly align all of the front ends.

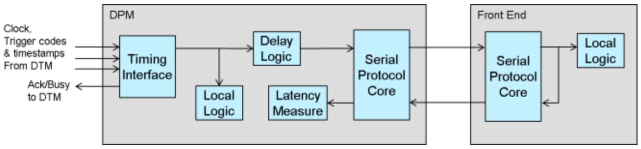

Fig. 5: Timing Distribution To The Front End.

VI. Management

Platform management is an important part of a crate based data acquisition system. This entity is responsible for controlling the power up of the system, monitoring the health of its individual components and reporting this information to the central run control system. Each component in an ATCA shelf must comply with a common standard.

The COB includes an integrated ATCA management client, IPMI (Intelligent Platform Management Interface), controller which communicates with a shelf manager contained within the ATCA crate. The combination of the COB IPMI controller and the central shelf manager provide thermal and power budgeting, system status monitoring and power control. The local COB IPMI controller also assists in the RTEMS and Linux boot loading process.

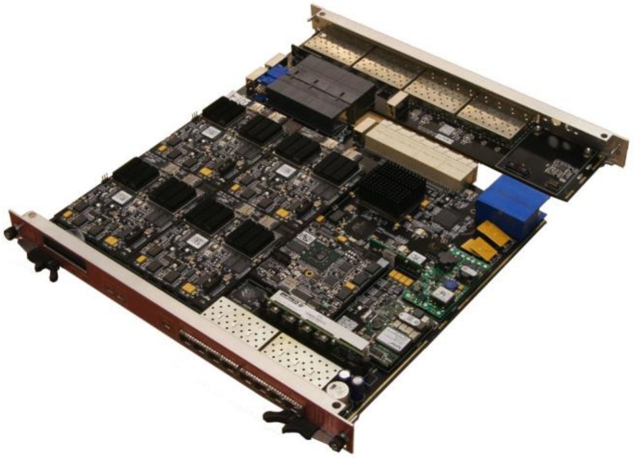

Fig. 6: COB populated with 4 DPMs and a DTM.

A software development kit provides a number of utilities for accessing and managing the RCE nodes and the RCE cluster. These applications can be compiled in both Linux and RTEMs for use on the RCE, as well as Linux for use on a host server. The included applications provide mechanisms for configuring and monitor the shelf through the shelf manager.

VIII. Applications

Prototype versions of the RCE platform have been successfully deployed in a number of SLAC projects. Most recently it was deployed as the core of the Silicon Vertex Tracker data acquisition system for the Heavy Photon Search (HPS) apparatus during 2012 with a second run scheduled for March 2015. The RCE has also been used for X-Ray detector control and readout in FEL (LCLS) experiments.

The current production version of the RCE platform has been deployment in the LHC ATLAS CSC detector, replacing the current readout electronics as an upgrade to achieve trigger rates above 100KHz.

The RCE platform has been selected as the data acquisition node for the Large Synoptic Survey Telescope where it will serve double duty as a data acquisition node as well as a short term data storage and processing platform.

Additional experiments where the RCE Platform has been selected include the LBNF, DarkSide and the proposed ILC SiD detector.