Proposal for an RCE-based DAQ system for LBNE

M. Convery, M. Graham, G. Haller, R. Herbst, M. Huffer

SLAC National Accelerator Laboratory, Menlo Park, CA 94025

(Dated: February 12, 2013)

ABSTRACT

This document presents a proposal to use the SLAC-developed DAQ toolkit

1 Introduction

The main purpose of the LBNE DAQ system is to read the raw data from the Front End Boards (FEB), which are mounted on the Anode Plane Arrays (APA) inside the cryostat, to build events from the different parts of the detector and to pass these events on to long term storage. The Level 3 requirements for this system include[1]:

- LArFD-L3-DAQ-3: The DAQ shall be capable of receiving raw data from a freely running readout from all detector systems.
- LArFD-L3-DAQ-7: The DAQ shall be designed to collect data continuously
- LArFD-L3-DAQ-8: The DAQ shall perform prompt processing of data

The DAQ-7 requirement is relevant mainly to non-beam physics. As such, it was left out of the requirements at the time of CD-1, which assumed a surface-located Far Detector (FD). Nonetheless, continuous readout remains a valuable goal that would be desirable to have in the final LBNE DAQ system.

The key electonics module that needs to be provided for the back-end DAQ is one that is capable of reading the data streams from each of the APA's and concentrating the data down to a smaller number of high-bandwidth data streams that are then passed to an event-building network. This is commonly needed function in modern HEP experiments that has frequently been addressed with custom modules built explicitly for a single experiment. This may require significant development time. However, the modules produced quickly become obsolete, as available networking technology progresses. This limits the desirability of reusing such modules in subsequent experiments.

The SLAC Research Electronics Group (REG) has developed a solution to this obsolescence problem by producing a set of modules, together with firmware and software, that can be adapted for use in multiple experiments. The development costs are then leveraged over multiple experiments, allowing each of them to benefit from the latest networking hardware, at a significant reduction in development costs. This "DAQ toolkit" uses the modern Advanced Telecommunications Architecture (ATCA)for its physical structure. The key element of the system is the Reconfigurable Cluster Element (RCE), which is based on a Virtex 5 "System on a Chip". A single board combining several of these RCE's can handle very

high bandwidths measured in the 100's of Gigabits/second. This system has been adopted in several HEP experiments already, and will likely be adopted by more in the future. The REG is continuing to develop and support new generations of the toolkit to take advantage of new networking equipment as it becomes available.

We are proposing to make use of this toolkit in the DAQ systems to be produced for the LBNE 35 ton prototype and the full Far Detector. The bandwidth available in the current generation of the toolkit far exceeds that of the baseline system based on the Nova Data Concentrator Module (DCM). The increased flexibility afforded by this extra bandwidth may be highly valuable to ensuring LBNE success. Furthermore, leveraging the work already done by the REG as well as benefitting from their support in the future, will provide many benefits to LBNE and may reduce the development costs.

2 The Data Acquisition Toolkit

2.1 Advanced Tele-Communication Architecture and the ATCA Shelf

In this document any such usage which employs a specific standard will be referenced as a Platform. For example, VME would constitute one such platform. For LBNE that platform is based on an existing standard developed by the PCI Industrial Computer Manufactures Group (PICMG) commonly referred to as the Advanced Tele-Communication Architecture, or ATCA, whose current revision is referred to within that consortium as PICMG 3.0. As a platform ATCA is now quite mature, having been in existence for more than ten years, with a broad design base and a wealth of equipment deployed in the field as well as a burgeoning eco-structure within the telecommunication and defense industries.

ATCA usage by LBNE will be entirely compliant with the PICMG 3.0 specification. That specification is described in [6] with an introduction available from [5]. However, the remainder of this section is intended to provide sufficient background to gain a thorough understanding of the physical design description.

The ATCA shelf is known historically as the chassis and is by analogy, equivalent to a VME crate. Shelves house the Front-Boards and RTMs described below (see Sections 2.1.3 and 2.1.4). They contain, from front to rear, pairs of slots with each pair housing a Front-

Board in the front and the Front-Board's corresponding RTM in the rear. The shelf allows for hot-swap of any board in any slot. Depending on form factor the number of its slot pairs varies from two (2) to sixteen(16). The orientation of those slots also varies, as shelves are offered with either horizontal or vertical orientation. In turn, that orientation affects the flow of air; from either left to right (horizontal), or top to bottom (vertical). Broadly, the shelf is composed of a sub-rack, backplane, filters and cooling devices (fans). The subrack provides the infrastructure to contain the Front-Boards and RTMs described below. This includes guide rails, ESD discharge, alignment, keying, and backplane interface. Backplanes are passive circuit boards which carry the connections between slots. Although somewhat more complicated in detail, for this document, those connections can be partitioned into three logical groups: power, control and differential data pairs. The topology for both power and control connections is invariant of backplane. However, in order to accommodate different applications the connection topology of data pairs can vary. Two commonly used topologies are the dual star and full mesh.The backplane (and ATCA) is protocol agnostic with respect to the usage of these differential pairs with the choice delegated to the shelf's specific Front-Boards.

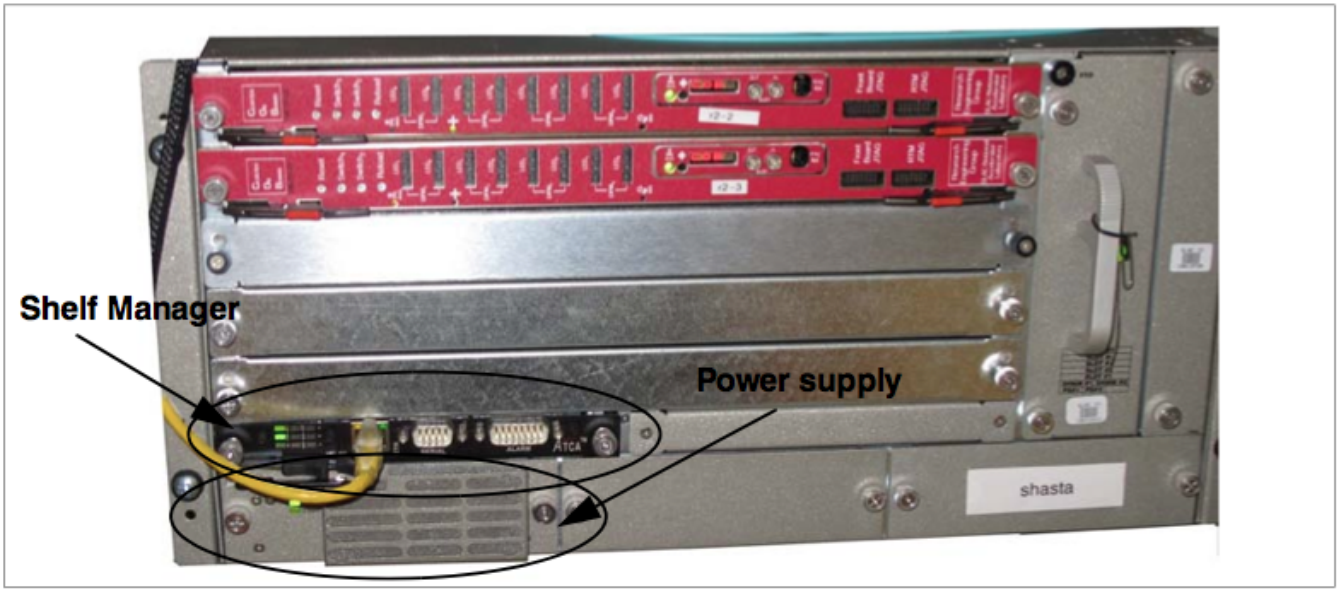

FIG. 1: Front view of 5-slot ATCA shelf.

This photograph is of a COTS1 shelf purchased from ASIS [7]. It has a horizontal orientation within its corresponding rack with airflow from left to right. It contains a replicated, full mesh backplane. Two of its five front slots are populated with Front-Boards,

while its unused slots are populated with dummy air baffles. Note the RJ45 connector located on the front-panel of its Shelf-Manager (ShMC). This provides the shelf manager access to the Ethernet from which control and monitoring (through IPMI) of the shelf would be accomplished. Further, note the integral power supplies. These supplies are not required by the ATCA standard, but are provided by ASIS as a convenient feature for bench-top usage. The same shelf viewed from the rear is illustrated in Figure 2.

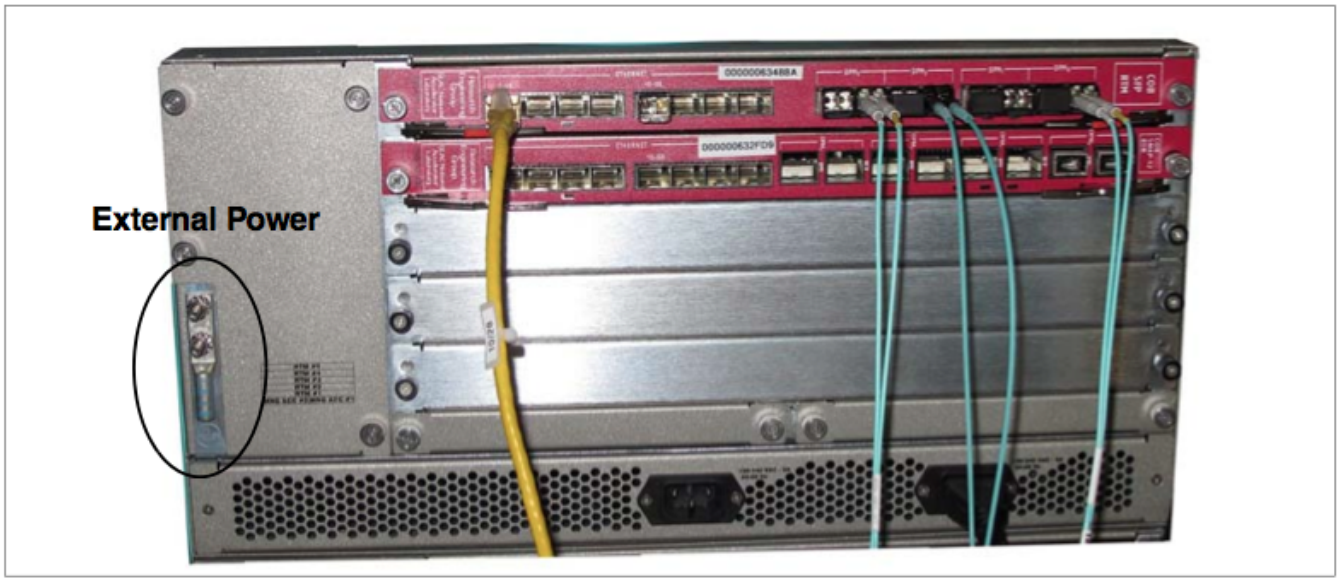

FIG. 2: Back view of 5-slot ATCA shelf.

As was the case for the front, two of its five rear slots are also populated, however, with RTMs (see Section 2.1.4) rather than Front-boards. Note, as was the case for the front slots, unused slots are populated with dummy air baffles. Further, note the power pins provided for external input of shelf power (+48 VDC). Last, Figure 4 provides an identical view, although now unpopulated, offering an unobstructed view of its backplane. Note the open area on the right allowing access to the P3 zones between Front-boards and corresponding RTM.

2.1.1 Shelf Power

An ATCA shelf does not have any requirement for provision of its own power. Further, a shelf also has no explicit requirement for the control and monitoring of that power independent of source. Instead, its minimum requirement is to simply support external connections

for both primary and redundant supplies. Those supplies must provide +48 VDC. In a large scale installation this feature allows for rack aggregation of power over many shelves. Power for the LBNE shelf is discussed in Section 2.4.

2.1.2 The Front-board

The Front-Board constitutes the heart of the ATCA eco-system. From a shelf's perspective that board is simply a PCB board, 8U wide x 280 mm deep and which plugs into one of its front slots. That board, although following ATCA mechanical and electrical interface standards, contains logic which is application specific. And from that logic's perspective the shelf exists simply to provide a platform to serve its application specific content. On its near side the board's front-panel contains a hot-swap handle as well as four ATCA defined LEDs to help direct an operator in board insertion and removal. The remainder of the panel is considered application specific. The board's rear side contains three logical Zones. Zones 1 and 2 connect directly to a shelf's backplane. Zone 1 provides access to shelf power (+48 VDC) as well as the I2C communication channels which the board uses to communicate with its shelf manager. Zone 2 provides access to the high-speed, differential pairs connecting boards together. The area encompassed by Zone 3 is application defined, but reserved for connections to the board's RTM. PICMG defines an extension to the standard which allocates that area. This standard is PICMG 3.8 (ATCA for physics, [23]), which follows the convention of Zone 1 and 2 and partitions its area into two zones, one for power/control and the other for signals. The connector used for signals allows for allocation of up to 120 differential pairs between board and RTM. Any and all boards used by LBNE adhere to PICMG 3.8. Note, that independent of any allocation scheme for Zone 3 the power for an RTM, if defined, must go through the Front-Board. Each board must also contain a local controller called its IPM Controller, or IPMC. The IPMC manages the board's activation/deactivation policy as well as monitors its health and safety. It serves as a proxy to the board's shelf manager and communicates using the I2C channels on Zone 1. The IPMC, as was the case for the board itself, must satisfy ATCA interface standards, but its implementation is also, necessarily, application specific. The standard specifies that the sum of the power drawn from a Front-Board and its corresponding (if any) RTM must not exceed 300 Watts. LBNE requires one application specific Front-Board. That board is de-

scribed in Section 2.6. A photograph of a representative Front-Board, showing connectivity to an RTM (using PICMG 3.8) is illustrated in Figure 5:

2.1.3 Rear Transition Module

The RTM (Rear-Transition-Module) is simply a PCB board, 8U wide x 70 mm deep which is used to extend a Front-Board (see Section 2.1.3). Although not required, that extension is typically found necessary for two reasons: First, to increase the useful footprint of the Front-Board and second, to house a board's external, I/O interface. The RTM shares the same hot-swap model as the Front-Board and specifies an identical pitch (1.2"). This allows the RTM to reuse the same panel, handle switches, and LEDs as its Front-board. The RTM connects to its Front-board through Zone 3. The form of that connection is application specific. However, if power for the RTM is necessary, it must be provided by the Front-board and must be brought through Zone 3. The ATCA specification is somewhat ambiguous with respect to the maximum power drawn by an RTM. A shelf is required to provide at a minimum 15 watts of cooling, but is, however, free to provide more. This is typically the case for all shelf manufacturers with maximum numbers more in the 40 to 70 watt range. The RTMs employed by LBNE standardize the usage of Zone 3 by application of PICMG 3.8 [17]. That standardization allows such an RTM to "plug and play" with LBNEs Front-Board (see Section 2.6). PICMG 3.8 populates Zone 3 with two connectors, one for power and one for signal. Power provided through the power connector is +12 VDC and that connector also contains pins for JTAG as well as I2C support. The I2C channel is expected to be used by the Front-Board for control of the RTM's hot-swap switch as well as its front panel LEDs. The signal connector provides up to 120 differential pairs. How those pairs are assigned between Front-Board and RTM is considered application specific. However, for LBNE's Front-Board, each one of its four DPM bays is assigned 1/4 of those pins or thirty (30) pairs (see Section 2.6). The two types of RTMs contained in LBNE are described in Sections 2.8 and 2.9. Figure 6 illustrates a representative RTM showing its PICMG 3.8 interface connected to a Front-Board:

2.1.4 IPMI and Shelf Manager

ATCA adapts a somewhat locally autonomous philosophy with respect to environmental control and monitoring. As part of this model, each shelf has associated with it a single entity responsible for maintaining the health and safety of its infrastructure. That entity is called the Shelf Manager (ShMC). Front-Boards, through their own local controller (or IPMC) negotiate both individually and independently with their shelf manager for their own activation or deactivation. They do so by publishing changes to their state through dedicated I2C channels on the backplane. The shelf manager determines, based on hot-swap interface, when a board requires activation or deactivation. Power levels are negotiated based on both a board's request and the shelf's total available power. Shelf temperatures are maintained at safe levels autonomously by the shelf manager using information published by each board and adjusting power levels and fan-speeds accordingly. In short, once a shelf's power is applied and while its shelf manager is active, no external monitoring or control is necessary to maintain the shelf's health and safety. Although the health and safety of its shelf is maintained autonomously, the shelf manager still has provision for an external interface. Through this interface any information published to the shelf manager can be exported and the shelf manager can itself be configured. That physical interface is Ethernet and the shelf manager contains a TCP/IP Stack through which external communication is maintained. The logical interface for control and monitoring of the shelf is IPMI [18] and a wealth of tools exist, which interact with this interface.

2.2 Cluster-on-Board

The COB (Cluster-On-Board) is an 8U, ATCA compliant Front-Board (see Section 2.1.3) with a PICMG 3.8 Zone 3. Functionally, the COB serves as a carrier board for the RCEs hosting the firmware and software developed for LBNE (see Section 2.7). Those RCEs are mounted on mezzanine boards (see Section 2.7.4), which in turn plug into Bays on the COB. Bays are connected to the COB's two separate, independent Interconnects as well as its Zone 3 connectors. Interconnects provide arbitrary, high speed communication paths between the elements contained on the bay's mezzanine boards, both (it is important to note), inter and intra COB. Although rated up to 300 watts, when fully populated with five

mezzanine boards, a COB draws closer to 120 watts. This board is one deliverable from SLAC's R & D program on high-speed DAQ. As such, LBNE simply purchases this board and from its perspective, that board consequently requires neither design nor development. A photograph of that COB (in preproduction form) with its five bays occupied is shown in Figure 8:

The COB contains five (5) bays; one (1) DTM bay (see Section 2.6.1) and four (4) DPM (see Section 2.6.2) bays. Although all bays share identical form factors and connectors (see Section 2.7.4), they can be differentiated, primarily by how they connect to Zone 3, with the DTM connecting only to its power connector and the DPM only to its signal connectors. In turn, those connections determine the function of their corresponding mezzanine boards. The DTM, interacting with its shelf manager, manages the health and safety of both COB and RTM, while DPMs acquire and process data originating from the RTM. Those data, their interface, acquisition and processing are all intended to be application specific. The mezzanine board plugged into the DTM (Data-Transport-Module) bay contains one RCE as well as the COB's IPM Controller (IPMC). The IPMC is the element responsible for monitoring the underlying health and safety of the COB as well as its corresponding RTM. It is also responsible, in conjunction with its corresponding shelf manager, for board and RTM activation/deactivation. It performs all these activities by interacting with various components on the COB, specifically with the RCEs contained within the COB's five bays. That interaction is accomplished through dedicated, local I2C busses. The IPMC is a SOC (System-On-Chip), containing a dedicated ARM based (M3) processor. That processor runs de-facto, industry standard Pigeon-Point IPMC firmware and software [41], suitably modified to control and monitor the specific functionality of the COB. Although in capability and form no different than any other RCE, the DTM's RCE has the fixed, dedicated responsibility for managing both of the board's interconnects. For this purpose it contains specific firmware and software. For example, as one responsibility, it must maintain the configuration and supervise the 10G-Ethernet switch contained within the fabric interconnect. That switch's management interface is a single lane PCIe. To communicate with this switch, the RCE contains a PCIe Protocol-Plug-In (firmware, see Section 2.7.1) as well as the tools (software) to configure and monitor that switch. Note, however, that while the DCM's RCE has predefined, base responsibilities it also remains accessible for user applications. For example, LBNE uses this RCE as a trigger simulation and that RCE has the

capability to drive TTC protocol to not only the elements of its own board, but also to the elements of the entire shelf (see Section 3.3.3). The mezzanine board plugged into a DPM (Data-Processing-Module) bay contains two (2) RCEs. Each DPM provides connections to thirty (30) differential pairs originating from the RTM, but carried through the COB's Zone 3 signal connector. The mapping of those thirty pairs to the mezzanine board's two RCEs is arbitrary and determined by application. The function of either RCE is determined not only by the mapping of those thirty pairs, but by the firmware and software it contains. For LBNE, that function will be either as a Feature Extractor or as a Formatter (see Section 2.2). For LBNE, each RCE on the DPM is connected to eight (8) differential pairs of the fabric interconnect and four (4) pairs on the base interconnect. For the fabric interconnect, although those eight pairs can be configured a variety of ways, they will, for LBNE be configured as one (1) channel of 10G-Ethernet (XAUI). For the base interconnect two pairs receive TTC (one primary and one redundant) and two pairs transmit BUSY (one primary and one redundant).

2.3 Interconnects

The Fabric interconnect contains, as its principal feature, a local, 10-Gigabit Ethernet (10-GE). Packets are switched on that network using a commercial, 1163 ball ASIC [38]. That ASIC is a fully compliant Layer-2, 10G-Ethernet switch. Although fully provisioned for buffered transfer, switch operation is, by default, cut-through with an ingress/egress latency of less than 200 Nanoseconds. It is also a fully managed switch with a PCIe interface connected to the DTM's RCE. Through its interconnect the COB's RCEs appear as nodes on that Ethernet. The interconnect allows its physical network to be extended to both nodes and networks external to the COB. Those networks could be, for example, other COBs residing in the same shelf, or even nodes physically disjoint from both COB and its shelf.

Internal to its shelf, the interconnect extends its network through its connections to Zone 2 of its backplane, specifically those connections to that backplane's fabric interface. The interconnect has individual connections to each of the thirteen slots of the shelf's backplane. With a full mesh backplane, this allows each network of every COB to be connected to each network of every other COB. External to its shelf the interconnect extends its network

through its connections to the COB's fiber-optic transceiver bay. That bay can contain up to eight (8) SFP+ transceivers [20]. The interconnect's switch is organized in units of Ports. Each port is composed of four lanes and each lane is constructed from two differential pairs. Each lane forms a full-duplex channel with one pair allocated for transmission and one pair for reception. Each lane of each port is capable of operating independently at a fixed set of speeds ranging from 1.0 Gigabits/second up to 12.5 Gigabits/second. Lanes may also be bound together to form a single Ethernet channel which operates at four times the speed of any one lane. For LBNE, which carries 10-GE, the switch is configured to run XAUI, requiring four lanes, each operating at 3.125 Gigabits/second. The switch contains twenty-four (24) ports. Those twenty-four ports are allocated to the fabric interconnect as follows:

- One (1) port connected to the DTM bay (one RCE).
- Eight (8) ports connected to the four DPM bays (two per bay, one for each RCE).
- Two (2) ports are connected to the SFP+ transceiver cage.
- Thirteen (13) ports are connected to the fabric interface (P2).

In short, within a shelf, the fabric interconnect allows for the formation of a uniform Ethernet populated with a flat space of RCE nodes.

The base interconnect's principal function is to manage and distribute synchronous timing to the COB's five bays. Note that unlike the fabric interconnect the protocol distributed over this interconnect is application specific. In further contrast to the fabric interconnect which functions identically independent of the shelf slot it occupies, the base interconnect has slot dependent responsibilities. This is a consequence of the fact that while the fabric interconnect uses ATCA's fabric interface, the base interconnect uses its base interface. That interface employs a backplane topology that is fixed by the standard at dual-star. ATCA refers to slots at its roots as Hub slots and slots at its leaves as Node slots. Necessarily, the behavior of a board, specifically its base interconnect, must vary depending on whether it occupies either a hub or a node slot. While boards in node slots need only distribute timing locally, boards occupying node slots must distribute timing not only locally, but also to other boards occupying its shelf. In short, while occupying a hub slot the base interconnect drives

its base interface, but while occupying a node slot receives timing. The distribution model for the base interconnect allows timing to originate from one of three potential sources:

- Internal, where the source is the base interface.
- External, where the source is the COB's Front-Transition-Module (FTM).
- Local, where the source is the COB's DTM.

Internal timing was described above. External timing allows the timing source to originate off the shelf. The FTM is a bay which contains an application specific, small PMC-like daughter board. Logically, the FTM serves the same role on the front of the COB as the RTM does on its rear, that of media adaptation. Eight (8) differential pairs from this daughter board connect directly to the base interconnect and eight (8) differential pairs connect to the DTM's RCE. Those eight pairs are intended to allow that RCE supervision of the FTM. Local timing allows the board to operate either stand-alone or perhaps more usefully provide a simulation of timing which would normally be sourced either internally or externally. LBNE has purpose built versions of both FTM and base board. Those version are described in Sections 2.6.5 and 2.6.6:

2.4 Reconfigurable Cluster Element

The RCE (Reconfigurable-Cluster-Element) is a bundled set of hardware, firmware and software components. Together, those components form a generic computational element targeted to process efficiently, with low latency, those kinds of data found passing through HEP DAQ systems. Those data have in common three features which make specific, somewhat, competing demands on the functionality of any such element. Those features are:

- Highly parallel: Data which are massively parallel are most naturally also processed in parallel, requiring computational elements which scale in cost, footprint and power. Those elements, in order to manage the flow of their data both efficiently and coherently, communicate together. This necessitates a communication mesh which shares the same scaling properties as the elements themselves.
- Inhomogeneous: As those data typically originate with their corresponding detector they are carried necessarily over a variety of media employing various inhomogeneous

protocols. The element's I/O structure, must support, naturally, without sacrifice of performance that diversity.

- Transient: Transient data arrive at an element once, to be either transformed or reduced before immediately exiting the element. Such data are not typically amenable to caching strategies and require elements whose optimal computational model emphasizes a performanent efficient I/O structure, coupled strongly to a large, low latency memory system over raw processor speed.

The RCE is optimized for those three features. Physically, one element can be contained in a footprint of less than 32 cm2, typically draws less than eight (8) watts, costs (in small quantities) around $750 and contains a native 10-Gigabit Ethernet interface. Elements are connected through a commercial, commodity ASIC containing a 64 channel, Layer-2, cut-through1, Ethernet switch [38]. The combination of elements and switch define a Cluster and the nature of ethernet as well as functionality within that switch allows for the composition of arbitrary numbers of cluster hierarchies. For example, from the RCE perspective, the COB (see Section 2.6) represents a single cluster of nine (9) RCEs and its ATCA shelf is simply a container for a single level hierarchy of up to fourteen (14) nine node clusters. A block diagram of the major physical features of the RCE is illustrated in Figure 11:

The principal implementation feature of the RCE is in its reuse of System-On-Chip (SOC) technology, specifically, member's of Xilinx Virtex-5 FX family [55]1. As such, the RCE is neither processor, FPGA or DSP. Instead, it can be simultaneously any combination of the three. Within its fabric the FPGA contains both soft (user defined) and hardened (manufacture defined) silicon. That fabric is configured automatically on POR (Power-On-Reset) and is either downloaded directly from images previously stored on the FPGA's configuration (platform) flash, or indirectly through the RCE's JTAG interface. Note also that the platform flash is itself programmed through the RCE's JTAG interface. The RCE employs standard Xilinx tools and software to program the FPGA. Xilinx refers generically to its set of different, hardened silicon as resources. Among the more important of those resources are high speed serializers/deserializers, I/O adapters, DSP tiles, dual-port RAM and of course, its processor. The RCE allocates the processor as well as a modest number of additional resources and soft silicon for its CE (Cluster-Element). The CE has exclusive use of, but interfaces indirectly with (see Section 2.7.2) its external DDR3 memory and micro-

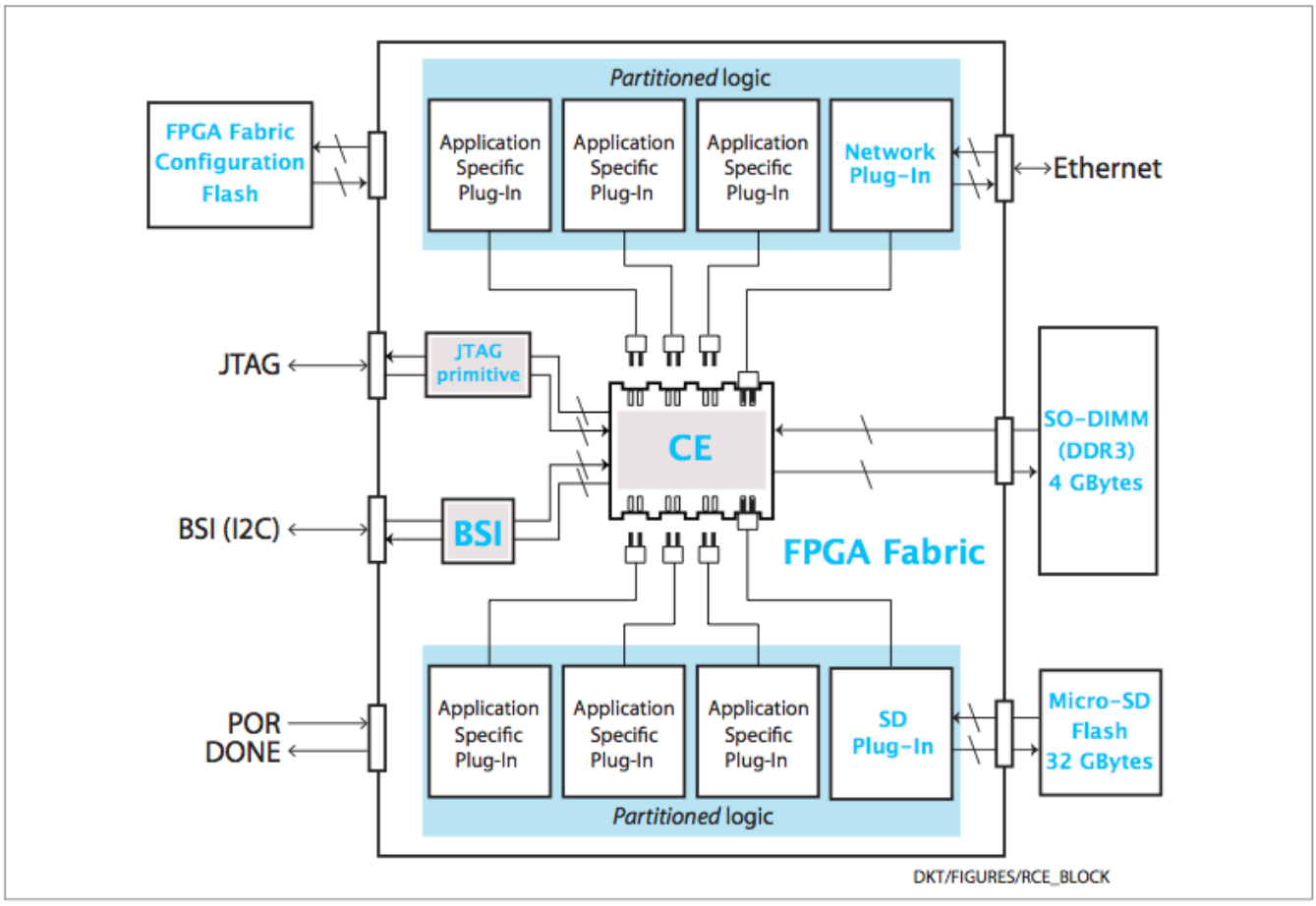

FIG. 3: Block diagram of an RCE.

SD flash system. Memory is packaged as SO-DIMM and the micro-SD flash is removable, allowing its capacity to be determined by user application. The BSIs (Boot-Strap-Interface) principal function is to reset the CE. However, it also contains the initial configuration information necessary for the CE's bootstrap loader to boot its processor. The BSI is outside the CE so that its configuration may be retained over resets of the CE. External to the FPGA the BSI appears as a standard I2C device and receives its command and control through that interface. Note, for the COB, that device is controlled and monitored through its IPMC (see Section 2.1.3). To provide isolation between system and user firmware and insure reproducible behavior, system firmware is partitioned [53] away from application specific logic. System firmware is defined as the CE, the BSI, JTAG support and both Network and SD Plug-Ins. The CE, which is both at the heart of the entire RCE and contains a significant fraction of the user's intellectual investment is described in Section 2.7.2. The remainder of the fabric, both hardened and soft silicon is reserved for application

specific logic. That logic and its relationship with the CE is described below in Section 2.7.1.

Although both user defined and implemented, any application specific logic, does of course require information exchange between it and its CE1. The interface model which allows such exchanges is the plug and socket. To follow that model, the user wraps their implementation specific logic with a thin veneer of system provided firmware. That wrapper is the plug and the combination of user logic and its plug is called a Protocol-Plug-In or PPI. When wrapped, that logic is now capable of being plugged into any of the eight predefined sockets on the CE. And once plugged in, both PPI and CE are now able to exchange information. Although bundled with its base system the RCE itself takes advantage of this model to glue its Ethernet and SD interfaces to the CE. Both are good examples of one class of PPIs which must interface outside their FPGA. Such PPIs when plugged into their CE have as their closest analogy the classic I/O device and processor model. However, unlike that model the PPI model coupled with the resources offered by the FPGA fabric provides an essentially unlimited way to either customize or mold the CE to arbitrary devices and protocols. Of course, the user is not limited to using the fabric and its resources solely for I/O. One can define PPI whose sole purpose is to take advantage of the DSP tiles and combinatoric logic of the FPGA to process rather than transfer data. LBNE uses this functionality to its advantage in performing its feature extraction (see Section 3.2.2).

The essential function of the CE is as a platform which serves as an application specific nexus for the data both received and transmitted through the RCE's application specific PPIs (see Section 2.7.1). As such, the CE can be considered as both a hardware1 and software platform. As a hardware platform its principal blocks are illustrated in Figure 12. As a software platform its corresponding services are described in Section 2.7.3.

Its principal implementation blocks are its Memory Controller, Crossbar and Processor: The Memory Controller: Interfaces the RCE's external memory with the CE's Crossbar. It is a soft controller, derived from an existing Xilinx DDR2 design, but tailored for usage of low latency, DDR3 memory. The controller allows addressing of up to four (4) Gbytes of memory. It is clocked at 320 MHZ, has separate, internal, 64-bit, read and write datapaths providing roughly 5 Gbytes/second of either read or write bandwidth.

- The Crossbar: The Crossbar interconnects memory controller, processor1, and up to eight (8) PPI sockets allowing for autonomous, concurrent transfers between all three types of entities and providing arbitration for when those transfers might collide. The

crossbar is clocked at the same rate as its memory controller (320 MHZ) and contains internal, separate, 128-bit, read and write datapaths. Its core is hardened silicon [56], but suitable enhanced with purpose built firmware which glues the eight PPI sockets to that core.

- The Processor: A 32-bit, PowerPC-440, superscaler, single core, RISC processor with separate 32 Kbyte data and instruction caches [56]. It is clocked at 475 MHZ. In addition to the three busses connected to the crossbar, the processor contains another, separate, independent, 128-bit wide bus called its APU bus [56]. One side is connected to the processor and its other side is an interface to the FPGAs fabric. This bus is unique in that it interacts directly with the processor's registers and data cache, bypassing its memory completely. Essentially, it allows the user to extend the processor's instruction set with application specific logic implemented in its fabric. Taking advantage of this feature, the CE uses the APU to control and manage its PPI sockets through a set of instructions which transfer data into and out of a socket directly from either registers or cache. This provides a very effective, low latency, performanent mechanism to transfer small amounts of data between processor and PPI. A similar mechanism is used for large data transfers, where data, rather than passed to and from the socket by value, are now passed to and from by reference. The socket autonomously takes care of transferring the data pointed to by that reference either to or from the PPI. Arbitrary transactions which interleave data by both value and reference are supported.

2.4.1 Software Services

The RCE includes bundled software to accelerate and leverage the development of application specific code for the CE. Some set of this software is linked to and executes with those applications (system resident software), while a subset is in the form of tools that operate cross-platform. Any and all system resident software is distributed with each RCE and if used, is dynamically linked to its corresponding applications. Remote tools and any software updates have a well defined release and distribution mechanism. JIRA is used for a bug-tracking and reporting system. Here is a summary of the software services bundled with the RCE:

- Bootstrapping: A generic bootstrap loader which allows, on reset, transfer to arbitrary code based on an externally controlled configuration parameter called its current vector (contained within the BSI, see Section 2.7). The code loaded and executed by the loader is assumed stored in the RCE's micro-SD device. The code pointed to by any specific vector is called a bootstrap. Bootstraps may be either standalone code or Version/Issue: 1.1/1 code which loads and transfers control to other code (a secondary loader). The CE may contain and transfer control to an arbitrary number of different bootstraps. For LBNE, on reset, control is transferred to a secondary bootstrap which starts up RTEMS (see below).

- Operating/System: Although the CE is itself O/S agnostic, its system resident software is not and depends on functionality best provided by the services of an underlying O/S. In order to not compromise the RCE's innate performance a Real/Time (R/T) kernel offered the best compromise in satisfying that functionality. That kernel is RTEMS. RTEMS has a fully provisioned set of multi-tasking services as well as being both compact and efficient. It also maintains POSIX compliant interfaces, easing the burden of porting third-party software. However, perhaps most importantly, it is an Open-Source product with no licensing issues. RTEMS is described in additional detail in [31].

- Persistency: Access to micro-SD based media using its bundled PPI. That media is formatted as FAT-16 and is used by the CE for storage of system code and configuration (see bootstrapping above). However, that media is available directly to applications for storage of their own application specific code and configuration.

- Networking: Includes a complete TCP/IP stack. The stack's MAC layer is satisfied by the RCE's bundled 10G-Ethernet PPI. The user interfaces to that stack are POSIX compliant.

- Linking: The same dynamic linker used to bridge system and user code.

- PPI support: Interrupt and reset support for an application's PPI.

- Debugging: Support for both local and remote debugging. Local debugging (SMD) interfaces to JTAG through standard Xilinx tools. Remote, network based, debugging

uses the GNU interface.

- Diagnostics: Built-in self-tests as well as diagnostics. These are included on the CE as an alternate boot image providing the ability to rescue or repair inadvertent burns of the micro-SD media. Development employs the GNU cross-development environment [34].

2.4.2 Mezzanine Board

The mezzanine board is one physical implementation of the abstract RCE described above in Section 2.7. It is a PCB board (100 mm x 80 mm) which hosts either one or two elements of RCE. A mezzanine board plugs into any one of the five bays contained on a COB (see Sections 2.6.1 and 2.6.2). Power (+6 VDC) to this board is applied using two separate, but identical connectors. One connector is assigned to each element of the board. Those connectors provide, in addition to power, a presence sense pin as well as an enable pin for that power. The board's two, internal PDS (Power-Distribution-Systems) takes that input voltage, divides it down and distributes the necessary, well regulated voltages to each element. Each PDS can source 25 Watts.

A high-speed, high density, differential connector carries signals between the COB and the elements of the mezzanine board. Those signals include:

- To and from RTM (thirty pairs). See Section 2.6.2.
- To and from the Fabric interconnect (sixteen pairs) See Section 2.6.3.
- To and from the Base interconnects (eight pairs). See Section 2.6.4.
- JTAG.
- To and from the IPMC (I2C); one per element.

On each of its two I2C channels the board contains, in addition to the element's BSI (see Section 2.7) various I2C devices which provide the following information:

- PDS status.
- Board and die temperatures.

- Element serial number (64 bit).
- Persistent, configuration information (MAC addresses, element wiring, etc.)

The COB's IPMC uses that information to plug and play with its bays, including their activation as well as in the monitoring of their health and safety. To illustrate both mezzanine concept and its relationship to the RCE, a photograph of the prototype (single element) GEN-II RCE, mounted in a mezzanine board is shown in Figure 13:

3 Implementation of RCE-based DAQ for LBNE

The elements of the DAQ-toolkit described in the previous section can be easily applied to the LAr TPC for LBNE. The block diagram of a possible configuration is shown in Fig. 4. We define the "front-end DAQ" as everything between the (cold) FPGA and the ATCA shelf; from the ATCA-shelf onward is referred to as the "back-end DAQ". The primary goal of this document is to propose a solution for the back-end DAQ and so, for this purpose, we will assume that the signals come into the back-end DAQ from the output of the front-end board (FEB) FPGA, each of which collects the output of $8 \times 16 = 128$ TPC wires.

The basic structure of the back-end RCE-based DAQ is fairly straightforward. The data from the ADCs is encoded (possibly using the PGP protocol; see Appendix **??**) in the FEB FPGA and driven out of the cryostat to a "transition board" which converts the electrical signal to an optical signal. The transition board is an optional step, but one which allows the back-end DAQ crates to be conveniently placed without worrying about signal degradation. The optical signal is then sent to the RTM which interfaces with the COB. RTM designs with up to 48-channel fiber optic inputs exist and are currently in use by LCLS and for LSST development (???this is made up???). The RTM also incorporates the output to the DAQ PC farm via 8 x 10 Gbps ethernet. The RCEs on the COB can be used to perform event building or even some level of pattern recognition.

Below, we discuss some specific issues with the implementation in the 35t prototype and full LBNE, as well as some ideas for the front-end DAQ.

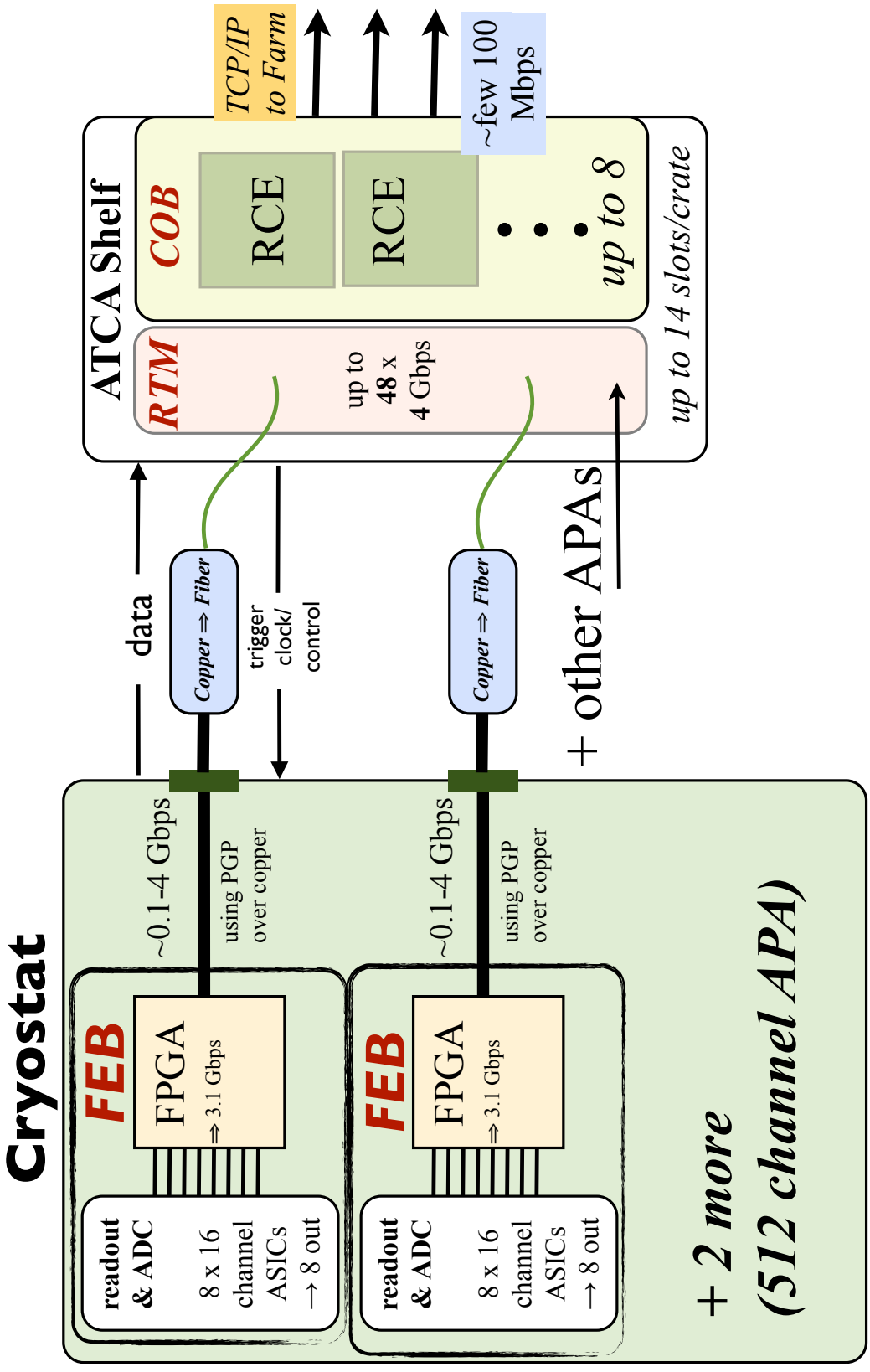

FIG. 4: Block diagram of the RCE-based DAQ for a single TPC APA.

	35t	Full LBNE
Total Channels	∼2.3k	∼307k
Number of APAs	4 (?)	120
Number of FEBs	18	2400
Transition Boards	2(???)	200(????)
RTM+COB Boards	1	50
ATCA Crates	1	4 (14-slot)

TABLE I: DAQ-related quantities for the 35t and full LBNE (as of Jan. 2013 design).

3.1 DAQ Layout for 35t Prototype

The 35t prototype TPC will have $\sim 100\times$ fewer channels than the full LBNE TPC and additionally will be externally triggered to observe cosmic rays.

.... timing and triggering ... configuration ....

	35t	Full LBNE
Total Channels	XXXX	XXXX
Number of APAs	4 (?)	120
Number of FEBs	16	2400
Transition Boards	16(???)	2400(????)
RTM+COB Boards	1	50
ATCA Crates	1	4 (14-slot)

TABLE II: Data rates etc ... (as of Jan. 2013 design).

..."transition boards" are the copper-¿fiber boards...maybe these are in the flange itself...for full LBNE, would make sense to do some multiplexing here (maybe 20:4 ... go from an APA, single cable/FEB to a 4-fiber cable???)

—-¿ from mike: SNAP-12 has 12 fibers/connectors, so this is a good number to use...

3.2 Full LBNE

.... assumptions, schematic of DAQ chain, summary of what/how many of each component we need ....

3.3 Comparision of RCE-based vs DCM-based Backend DAQ Systems

3.4 High-speed Data Links From Cold FPGA to Backend DAQ

...possibilities and our plans on this ...

3.5 DAQ Test-stand

3.6 High-speed Data Links From Cold FPGA to Backend DAQ

...possibilities and our plans on this ...

4 Schedule and Budget

... show both 35t and 35t+full lbne? ...

5 Conclusions

... why there is no choice be to go with us ...

6 References

[1] B.Baller *et al.*, LBNE Document 3747-v5,"LAr-FD Level 2 Programmatic and Scientific Requirements and LAr-FD Level 3 Requirements"