High Bandwidth DAQ R&D for ATLAS Upgrade

Version: V1-3-0

Updated: 11 February 2011

Authors: Rainer Bartoldus, Ric Claus, Andy Haas, Gunther Haller, Ryan Herbst, Michael Huffer, Martin Kocian, Emanuel Strauss, Su Dong, Matthias Wittgen (SLAC)
Erik Devetak, David Puldon, Dmitri Tsybychev (Stony Brook)
Bob Blair, Jinlong Zhang (ANL)

The expected increase of LHC luminosity over time will provide a continuous challenge to the ATLAS TDAQ system. While the TDAQ system and detector subsystem readout are certainly expected to undergo major upgrades for the High Luminosity LHC (HL-LHC) at 5 times the original LHC design luminosity, there are already a number of detector upgrades planned for Phase-1 (2016) or earlier that will require new Read Out Drivers (RODs). In addition, operational experience with the High Level Trigger (HLT) suggests that significantly improved readout throughput at the Read-Out System (ROS) stage of the TDAQ system could enable a number of very desirable triggering options with the HLT, options not otherwise viable in the current ROS implementation. The concept of the Reconfigurable Cluster Element (RCE), based on an R&D program on large-scale, high-bandwidth DAQ systems at SLAC offers an attractive, generic option for all these needs. The existing RCE hardware implemented on the ATCA industry standard and its associated software have already been adapted to serve the test stand and test beam DAQ needs for the pixel Insertable B-layer (IBL) project and have demonstrated the suitability and effectiveness for meeting the ATLAS upgrade needs. Due to the modular nature of these readout systems realized through the RCE concept, applications to various detector/TDAQ upgrades could be incremental, involving only specific subsystems, or even portions of a single subsystem, and would be plug-compatible with the existing TDAQ architecture. This document is intended to introduce this R&D effort and seed discussions towards a broader collaborative effort to serve the ATLAS upgrade needs with substantially improved capacity, as well as increased regularity and simplicity.

Introduction and motivation

The ATLAS detector, and in particular the detector readout, trigger and data acquisition, are expected to continuously evolve as accelerator luminosity increases in order to continue to maximize the physics output of the experiment. While the readout upgrade for the HL-LHC era (Phase-2) is inevitable given the expected higher volume and higher data speed, we believe there is also a viable path for ATLAS to already take advantage of modern technologies for many detector readout and TDAQ needs at Phase-1 or even Phase-0 upgrades, and to leverage the corresponding performance benefits much earlier. Some of the motivating factors and design considerations for pursuing this R&D program are:

- The RODs for the various subsystems serve rather similar functions; however, they were designed largely separately, resulting in several *different flavors* of ROD. While understandable from a historical perspective, the replicated effort from design, construction and commissioning of different types of RODs carries further into replicated burden and risk for commissioning, maintenance and operation. This burden has already caused some serious concerns in the struggle to bring all present subsystem readout components to their required performance, with some narrow escapes and with still not all subsystems fully compliant and stable. The upgrade presents a unique opportunity to reestablish the original ideal to concentrate resources towards a more uniform system that will allow smoother commissioning and easier maintenance in the future. The existing R&D effort has therefore placed high importance on providing generic solutions, aiming for a common readout system to serve as many different upgrade needs as possible, yet with built-in flexibility to allow any specific system to optimize its readout for its individual needs.
- The ROS PC *request rate limitation* is one of the most sensitive performance issues for the HLT. A single ROS PC can contain as many as four (in some cases five) ROBINs [1] and each ROBIN contains up to three Read Out Links (ROLs), implying a potential input rate of almost 2 Gigabytes/s. On the other hand, the PCI bus and the four network interfaces limit its output rate to no more than 533 Megabytes/s[1]. A more constraining restriction is the roughly 20 kHz request rate limit of the ROS PC. Any alternative architecture that can offer much higher access bandwidth would significantly extend the capability and versatility of the current HLT system.

[1] 125 Megabytes/s per card if one uses the ROBIN's onboard NIC.

- The *longevity* of the existing RODs for the time frame beyond 2016 is of concern. By that time, the current readout hardware will be ~10 years old. Many of the components used in their implementation are already obsolete today and the expertise needed to maintain these RODs will be increasingly difficult to come by. Continuing to build old-style RODs for 2016 and beyond will not be as simple as `just building more spares' might sound today.
- The *backward compatibility* to the current TDAQ architecture is a necessary concern for upgrades before the major Phase-2 HL-LHC era. However, the naïve assumption that this automatically implies one has to stay with the VME technology for the RODs until Phase-2, without other choices, is based on a misconception. The usage of the VME protocol in the current ATLAS detector is mainly for TTC communication, detector configuration, and sometimes calibration data collection, while the main DAQ functions are already avoiding the VME backplane due to its very limited bandwidth at 40 MB/s. Current operational experience also already indicated some performance inadequacy of the present architecture of a single board computer (SBC) in conjunction with the slow VME backplane for detector configuration and calibration data collection in some cases. It would be very unfortunate if any intermediate upgrade would continue to inherit these known inadequacies. New technologies for any ROD upgrade are just as viable as long as they continue to satisfy all the required input and output interfaces to ATLAS, including the TTC interface, specific detector interface and TDAQ software interfaces. This point will be illustrated in the discussion of the case study of the pixel IBL upgrade.
- The *technology choice* to stay with VME might be perceived as a conservative default, but moving forward to modern technology such as, for example, ATCA (*Advanced Telecommunication Computing Architecture*) is not as risky or speculative a choice as one might imagine. While the HEP detector community prides itself as leading edge pioneers of electronics, the rapid growth of the telecommunication industry has left us behind by some considerable margin in recent years. Moving away from legacy hardware in favor of more modern, mainstream technology is more in the nature of catching up with common industry practice, which already has a wealth of operational experience with these technologies. For example, ATCA is widely deployed in both the telecommunication and military industries.
- *Scalability*. To cope with the readout demand at increasing luminosity by simply replicating more ROD/ROSes with the current technology will also become problematic from an infrastructure perspective, in terms of both cost and maintenance. The electronic rack space in USA15 is close to saturation. Modern technologies enabling implementation of the same functionalities on a much smaller physical footprint can offer more flexible and viable installation and commissioning paths.
- While there are legitimate concerns regarding electronics infrastructure *maintenance* by the introduction of anything other than VME-based equipment, modern industry practice has evolved considerably with respect to maintenance practices since the advent of the VME standard in the 1980s. For example, ATCA's telecommunication pedigree places special emphasis on reliability and uninterrupted operation. Its focus on these requirements will considerably ease these concerns through a lower probability of failure, while its extensive tools for monitoring and recoverability imply, in the face of a failure, a much faster mean time to recovery.

The upgrade R&D platform outlined in this document offers a wide range of possibilities to explore various upgrade schemes involving the entire existing ROD/ROL/ROBIN/ROS plant, based on the RCE (*Reconfigurable Cluster Element*) and CI (*Cluster Interconnect*) as its fundamental building blocks. These two components are physically packaged together using ATCA (*Advanced Telecommunication Computing Architecture*). The RCE, CI and ATCA usage

were developed out of the current SLAC research program designed to study the needs for new generations of high-speed DAQ systems. In the following sections, a description of those building blocks as well as ATCA is provided, followed by case studies that detail how these blocks could be combined in applications to satisfy many of the ATLAS readout/DAQ/trigger upgrade needs.

The RCE Concept and ATCA

Current generations of HEP Data Acquisition systems either in production or development are differentiated from DAQ systems used in other disciplines by the significant amounts of data they must both ingest and process, typically at very high rates. Future generation systems will require even greater capability. In practice, this has resulted in the construction of systems that are in fact massively parallel computing systems. They are distinguished from commercial systems by the significantly greater amount of I/O capability required between computational elements, as well as the unique and disparate I/O requirements imposed on their interfaces. Given their unique requirements, traditionally, such systems have been purpose-built by individual experiments. However, it has long been recognized that all these systems share a large degree of architectural commonality. To support future experimental activities, SLAC has embarked on a research project intended to capture this commonality in a set of generic building blocks, as well as an industry standard packaging solution. It is intended that these blocks plus their corresponding packaging solution could be used in the construction of arbitrarily sized systems and which may also be used to satisfy a variety of different experimental needs. Systems constructed using these three concepts will share the desirable property of being able to readily scale to arbitrary sizes. These components are already deployed for Photon Science experiments at the LINAC Coherent Light Source (LCLS) at SLAC. Planning of their use for future experiments, such as LSST and ATLAS upgrade, as well as future computing initiatives, such as the *PetaCache*, are well underway. The relative ease in applying this concept to the designs of these very different projects has provided significant validation as to the correctness of this approach.

Out of this research, the need for two types of building blocks has been identified:

- A generic computational element. This element must be capable of supporting different models of computation, including arbitrary parallel computing implemented through combinatoric logic or DSP style elements, as well as traditional procedural-based software operating on a CPU. In addition, the element must provide efficient, protocol-agnostic mechanisms to transfer information into and out of the element.
- A mechanism to allow these elements to communicate with each other both hierarchically and peer-to-peer. This communication must be realized through industry standard, commodity protocols. The connectivity between elements must allow low latency, high-bandwidth communication.

In turn, this research has satisfied these needs with the *Reconfigurable Cluster Element* or RCE and the *Cluster Interconnect* or CI. Packaging of these blocks would be provided by a maturing *telecommunication* standard called ATCA. This packaging solution as well as the two building blocks is described in the sections below.

ATCA

ATCA is a communication and packaging standard developed and maintained by the PCI Industrial Computer Manufacturers Group (PICMG). This specification grew out of the needs of the telecommunication industry for a new generation of “carrier grade” communication equipment. As such, this standard has many features attractive to the HEP community, where “lights-out”, large-scale systems composed of multiple crates and racks are the norm. This specification includes provision for the latest trends in high speed interconnect technologies, as well as a strong emphasis on improved system Reliability, Availability and Serviceability (RAS) to achieve lower cost of operation. While a detailed discussion of ATCA is well beyond the scope of this paper (See [2]), these are its most pertinent features:

- A generous board form factor (8U *x* 280 mm with a 30.38 mm pitch). The form factor also includes a mezzanine standard (AMC or the *Advanced Mezzanine Card*) allowing construction of substrate boards.
- A chassis-packaging standard, which allows for as few as two boards and as many as sixteen boards.
- The inclusion of hot-swap capability.
- Provision for *Rear-Transition-Modules* (RTM). This allows for all external cabling to be confined to the backside of the rack and consequently enables the removal of any board without interruption of the existing cable plant.
- Integrated “shelf” support. Each board can be individually monitored and controlled by a central shelf manager. The shelf manager interacts with external systems using industry standard protocols (for example RMCP, HTTP or SNMP) operating through its Gigabit *Ethernet* interface.
- By default, external power input is specified as low-voltage (48 V) DC. This allows for rack aggregation of power, which helps lowering cost of power distribution for a large-scale system.
- It defines a high speed, protocol-agnostic, *serial* backplane. The backplane does *not* employ a data-bus; rather it provides point-to-point connections between boards. A variety of connection topologies are supported, including dual-star, dual-dual star as well as mesh and replicated mesh.

The Reconfigurable-Cluster-Element (RCE)

The RCE is a generic computational building block based on *System-On-Chip* (SOC) technology capable of operating from 1 to 24 lanes of generic high-speed serial I/O. These lanes may operate individually or bound together and may be configured to operate arbitrary protocol sets running at lane speeds anywhere from 1 to more than 40 Gigabits/s. The baseline configuration includes the ability to instantiate one or more channels of *Ethernet*, each channel operating at three selectable speeds of 1, 2.5 and 10 Gigabits/s. The present generation (“Gen-I”) RCE provides support for three different computational models:

- A 450 MHz *PowerPC* processor configured with 128 or 512 MB of RLDRAM-II as well as 128 MB of configuration memory. Standard GNU tools are available for cross-development.
- Up to 192 Multiple-And-Accumulate (MAC) units. Each MAC is capable of one cycle, 18 *x* 18 fixed-point multiplication summed into a 48-bit accumulator. MACs may be operated either independently or cascaded together.

— Generic combinatoric logic and high-speed block RAM.

An application is free to distribute their specific computation over these three mechanisms in any appropriate fashion. Independent of the mechanism used, all three mechanisms have access to data transmitted and received using the built-in serial I/O. In the case of the processor, a generic set of “DMA-like” engines are incorporated to transfer information between processor memory and serial I/O. These engines are optimized for low latency transfers and the underlying memory subsystem is able to sustain a continuous 8 Gigabytes/s of I/O. The processor is running under the control of a Real-Time kernel called *RTEMS*. RTEMS is an *Open Source* product, which contains, along with the kernel, POSIX standard interfaces as well as a full TCP/IP protocol stack.

In order to develop the RCE, the Gen-I RCE has been realized on an ATCA node (front) board. This board contains two RCEs. The RCE contains two dual-homed *Ethernet* interfaces. One is connected to the backplane’s base interface and one to its fabric interface. The base interface operates at either 1 or 2.5 Gigabits/s while the fabric interface operates at 10 Gigabits/s. This board along with its corresponding RTM is shown in Figure 1:

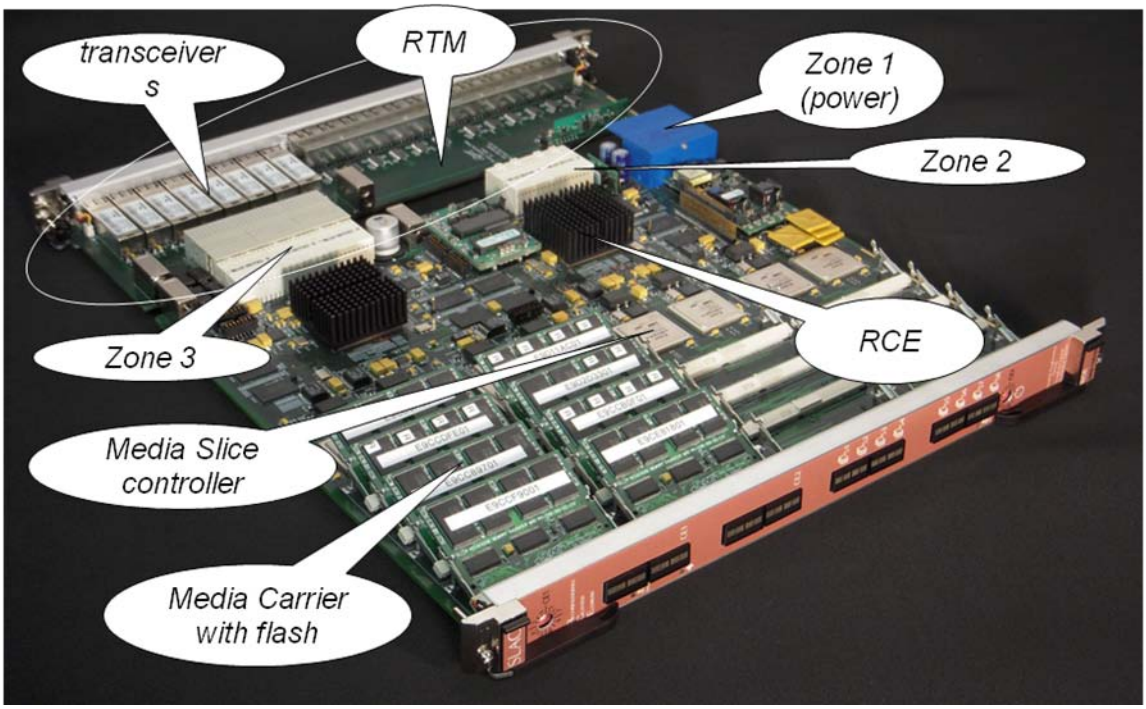

Figure 1 The RCE board. (The Media Slice controller and Terabyte of flash memory are only relevant to the PetaCache project.)

The RCEs are hidden behind their black heat sinks. The figure connectors used within the three ATCA zones are also called out. The connector used in Zone-3 (P3) is used to connect signals between the node board and its corresponding RTM. For this particular board the P3 connector carries the signals for eight serial links. Four of these links are connected to one RCE and the other four to the other RCE. The RTM simply contains eight optical transceivers (operating at 3.125 Gigabits/s), each converting a serial link from copper to fiber and fiber to copper. The fiber-optic links would typically be used to carry input data from front-end electronics systems. For example, in the case of ATLAS pixel readout applications, the RTM will contain TX/RX plug-ins for optical links connecting to the on-detector opto-boards for communicating with the pixel front-end electronics for clocks, configuration commands and DAQ data reception.

The Cluster-Interconnect (CI)

The CI consists principally of a 10-Gigabit *Ethernet* switch. The number of ports per switch may vary from eight (8) to twenty-four (24), and the physical interface for each port is *XAUI*, KFR, or KR4, or SM-II (100, 1000 Base-T). Physically, the switch is an ASIC based on the *Fulcrum Microsystems* FM22xx or FM6xxx chip families. To manage the switch, the CI also includes an RCE. Used in this role, the RCE is called the *Switch Management Unit* (SMU). The SMU communicates with its corresponding switches through a Fulcrum-defined, private management bus. Software interfaces operating on the SMU allow the switches to be fully configured, monitored and controlled externally.

As was the case for the RCE, the existing version of CI has been realized on an ATCA board. This board is used extensively with no changes for all projects involved. Its primary function is to network together RCEs. This board (with a representative RTM) is shown in Figure 2:

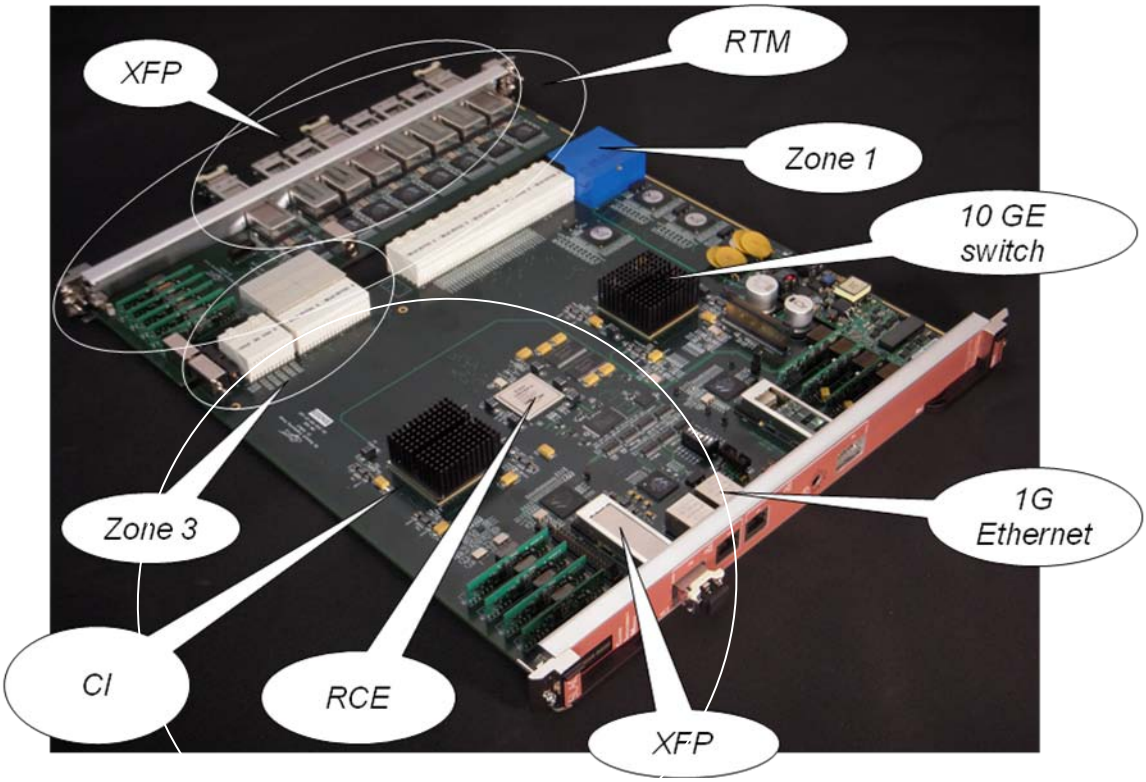

Figure 2 The CI board.

Evolution of the RCE

Based on the experience of the various existing applications, a significant upgrade to the RCE is now underway leading to a second-generation (“Gen-II”) part. This version will continue to have all the features associated with the current generation, however with significant improvements in interface, function and performance. Many of the changes are designed with ATLAS in mind. While somewhat different in functionality and in its hardware interfaces from its Gen-I counterpart, this part is intended to be fully software compatible with Gen-I, allowing for a relatively straight-forward migration path for application software. Some key features and improvements are:

- The RCE core will be based on *Virtex*-5 technology (Gen-I uses *Virtex*-4), which provides an integrated crossbar used to interface both memory and I/O around its dual-core 550 MHz *PowerPC* 440 processor. Significant design implementation changes were made in order to free up more firmware resources for user applications. The design is future-proofed to support the migration to the ARM-based processors in the upcoming generation of *Xilinx* devices, so that there is continuing good prospect for significantly enhanced CPU processing power in the near future.
- Each RCE will have access to a much larger associated memory system of up to 4 GB of either DDR-2/DDR-3 housed in commodity SO-DIMM. Both processor and I/O will share eight (8) Gigabyte/s bandwidth into that memory.
- The RCE's *Ethernet* interface will support a variety of physical interfaces, including not only its present XAUI (at 10 Gigabits/s), but also support for KR-4 (40 Gigabits/s) as well as SM-II (100 – 1000-BaseT).
- The RCE interface specification will no longer depend on ATCA concepts. This will allow implementations on purpose-built mezzanine boards, or for that matter, any physical packaging standard. In turn, this will allow easy application evolution over a variety of different physical board types, where the full functionality of an ATCA eco-structure is not required and thus its expense may be eliminated.
- The TTCrx ASIC, widely used in LHC experiments for timing and trigger, is an integral part of the new RCE mezzanine card (see below). It provides a generic timing control interface, allowing full compatibility with the current ATLAS TTC system. This will allow each RCE to receive the original standard TTC distribution from the LTP/CTP system, eliminating the need for intermediate interpretation of TTC information, as in, for example, the TIM modules used by the current VME-based muon and silicon subsystems. In addition, this functionality will allow an RCE to act as an essentially unlimited (in time), standalone trigger and timing emulator.

Cluster on Board (COB)

For ATCA board-level integration, a new architecture is now favored with each front-board in a shelf managing its own, individual cluster. These clusters are themselves clustered together using a full-mesh backplane. Thus, one shelf is an integrated set of up to fourteen (14) clusters making an integrated *Ethernet* of up to 112 RCEs. This board is called the *Cluster-On-Board* or COB. Purpose-build software and firmware of the RCEs contained on each board allows the board to be configured differently to serve the various applications. The block diagram of the COB is shown in Fig. 3

The COB would service data from up to forty-eight (48) serial channels from a detector's Front-End-Electronics. These data would be processed by up to eight RCEs with each RCE servicing as many as 6 serial channels of data operating at up to 10 Gigabits/s per channel. Reduction of input data would be implemented by subsystem specific code running on the RCEs. The physical interface for the input data would be provided by the RTM, allowing each subsystem to tailor its input interfaces through its own specific RTM while still preserving the commonality of the COB. For example, if a subsystem's FEE transports its data via S-links, its corresponding RTM would contain S-link transceivers. The back-end of the RTM goes to a common Zone-3 connector on the COB carrying its serial data to the COB and subsequently to the appropriate RCEs. For output, one *Ethernet* port (up to 40 Gigabits/s) from each RCE goes to one port on the DTM's switch, which is connected to the full mesh backplane fabric (P2) provided by the ATCA shelf. This full mesh backplane is organized such that each COB has a data communication bandwidth of 40 Gigabits/s with any other COB of its shelf (crate) simultaneously. In addition to its normal

responsibilities for switch management, the DTM would processes the L1 timing and control signals sourced from the backplane and fan those signals out to the board's RCEs.

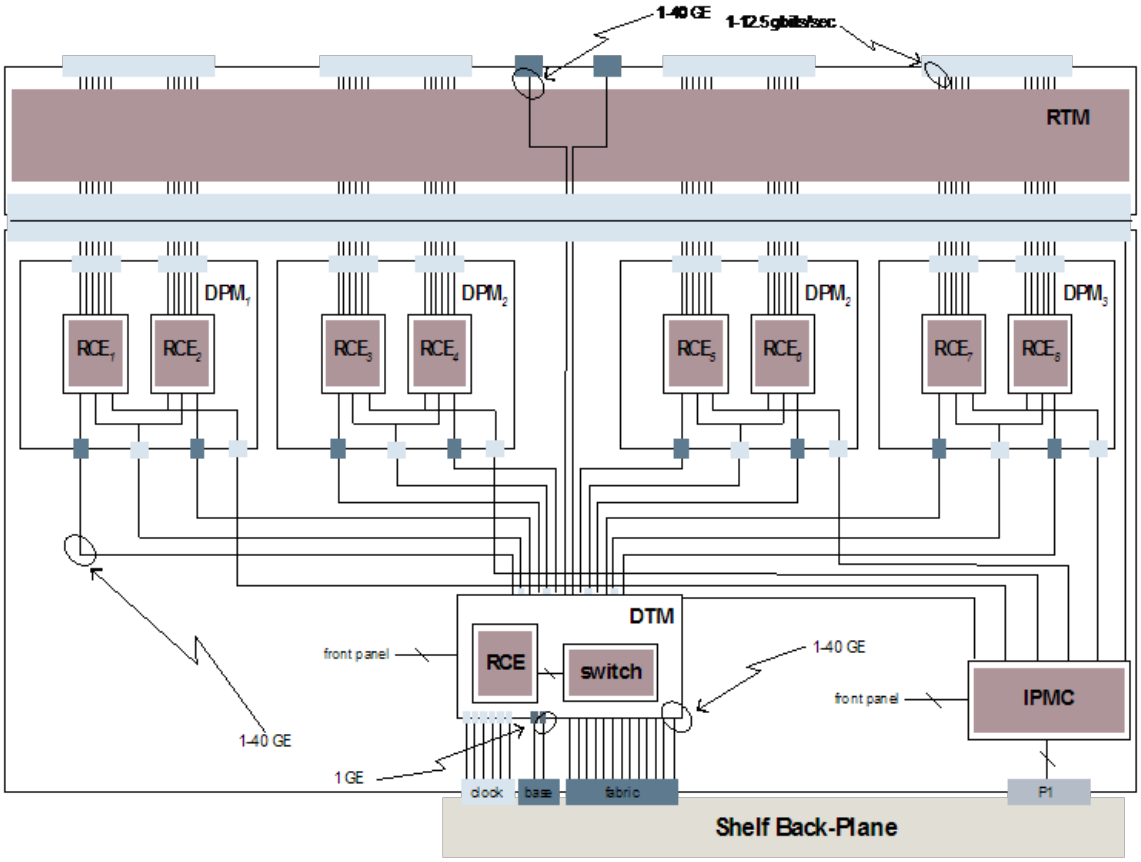

Figure 3: Block diagram of the Cluster-On-Board (COB). Each Data Processing Module (DPM) contains two RCEs. The Data Transportation Module (DTM) contains a single RCE, capable of functioning as a TTC emulator. The IPM Controller (IPMC) manages and monitors the board's configuration.

As the COB has encompassed most of the resources for common ROD functionalities, the RTMs are typically simpler boards that will carry the detector-dependent interfaces. The simpler demands on the RTMs in turn gives the flexibility to design different versions of the RTMs to serve different applications with less concern of complexity and compatibility as they are easier to replace for future evolution.

HSIO Module

Another auxiliary component already widely used in ATLAS is the High Speed I/O (HSIO) [3] module. This module was originally designed at SLAC for LCLS applications, but with a broad range of applications in mind by including large number of I/O channels of many different types of commonly used I/O interfaces, in conjunction with a *Virtex*-4 FPGA for processing. This was favored by the ATLAS silicon strip upgrade community as their stave test DAQ driver, and over 20 modules from a recent production have been distributed to many groups around the world. This board is also used as an interface between the RCEs and front-end for all the pixel applications described later. The standalone use of the HSIO has not been very easy due to the threshold of programming a large FPGA in firmware from scratch with little common utility. The pixel test experience has led to the plan to add an RCE mezzanine card to the next version of the

HSIO to benefit from the RCE software and firmware base for faster user application development.

ATLAS Upgrade Connections

While the core RCE has been a generic DAQ R&D effort at SLAC aimed at serving many different projects, some of the key RCE core personnel have contributed a significant effort in the existing ATLAS detector readout effort. In particular, Mike Huffer (lead of the RCE project) has managed the team from SLAC and UC Irvine in the recent crash-course major overhaul of the muon CSC ROD. This effort allowed the CSC to join the ATLAS data taking just in time for first collisions, and its overhaul has kept up with the increasing L1 trigger rate without any loss of CSC data. Through the experience of bringing a system that could barely run a few events before to today's rate capability of more than 70 kHz, a full set of knowledge on the ATLAS readout protocol and compliance was gained.

Some milestones of the RCE-related ATLAS upgrade activities:

- Feb-Mar/2009: An initial discussion of possible use of the RCE/ATCA R&D for ATLAS upgrade started with presentations at the ATLAS Upgrade Week [4] in Feb/2009 and ACES Mar/2009 [5].
- Jun/2009: RCE training workshop [6] at CERN and establishing an RCE test stand at CERN for common R&D, and e-Groups for RCE R&D communications. A picture of the RCE development lab [7] at CERN building 32 is shown is Fig. 4.
- Sep/2009: A concrete proposal to use the RCEs for IBL readout RODs was presented at the initial IBL readout meeting [8].
- Jun/2010: RCE test setup demonstrated operational pixel calibration for essentially the full suite of relevant pixel calibrations at the IBL workshop [9] at the University of Geneva. The RCE also demonstrated its application as the DAQ host for the cosmic telescope at SLAC as well as the 3D sensor pixel beam test at CERN.
- Oct/2010: RCE test stands also established at LBNL and Stony Brook.
- Dec/2010: First calibration with RCEs on FE-I4 readout chip for IBL.
- Dec/2010: Invited by US ATLAS to propose RCE readout option to the muon small wheel upgrade.

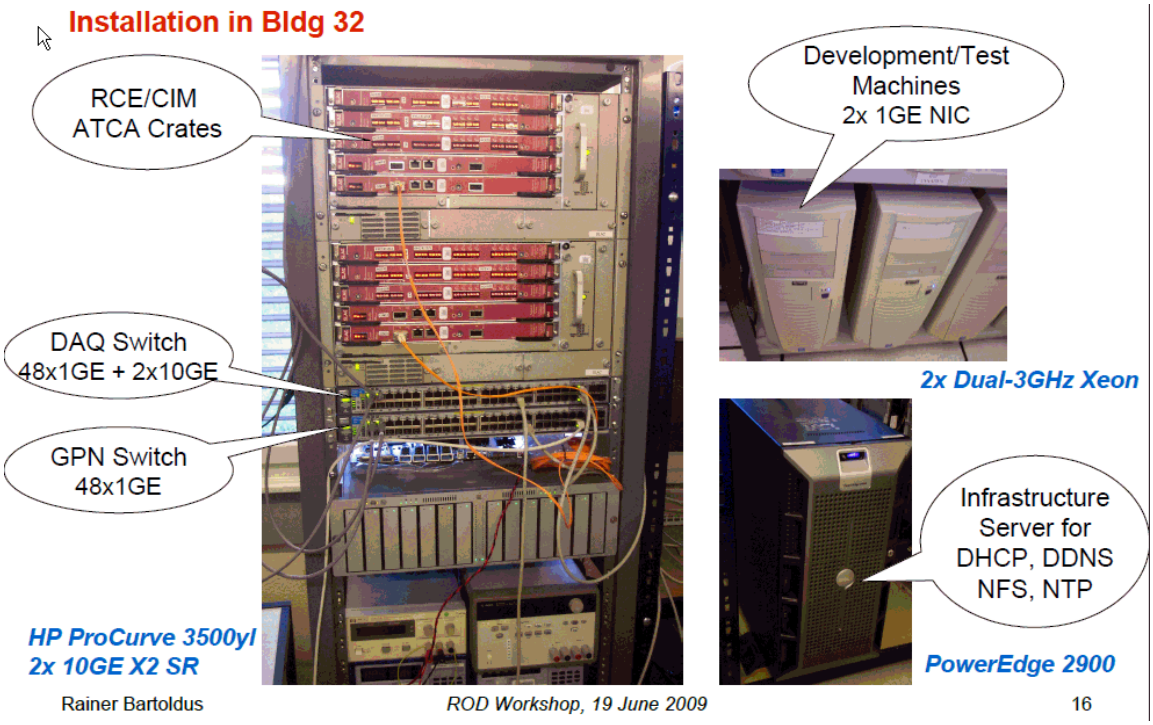

Figure 4: RCE development lab at CERN building 32.

ATLAS Upgrade Case Studies

The Pixel Insertable B-Layer (IBL)

ATLAS is currently planning to insert an inner-most layer[10] inside the present pixel detector on the time frame of Phase-1. Besides the roles of mitigating the degradation of performance due to radiation damage on the current B-layer and scattered individual dead modules, the thin innermost layer will bring immediate improvements in vertexing resolution in any case. The IBL front-end readout uses the new, large FE-I4 ASIC with 160 Megabits/s data rate per channel. (The present pixel detector runs at 40-80 Megabit/s.)

Although the original baseline plan was to build more of the existing pixel RODs to serve the IBL needs, the realization of various issues led to the conclusion that a new design, involving the COB would be preferable over a reuse of the current design. These issues included:

- The higher data rate of the FE-I4 and quite different architecture from the previous FE-I3 readout system. This implied that in any case different firmware and software would be required independent of the design used.
- The quite necessary large production of 56+spare RODs needed and the realization that many of its components have legacy status, as well as the fact that much of the expertise necessary for the production and its subsequent testing is no longer available.
- The significant expansion of rack infrastructure with respect to the current system, implying an increased footprint, thermal and maintenance burden in USA15.
- The existing ROD has no data path for calibration output through its S-link. As a consequence, the large volume of calibration result data have to be extracted through the narrow path of the single SBC and slow VME backplane. This constitutes a major bottleneck for calibration speed performance.

Although the design study of the RCE-based readout option was already in a fairly advanced state at the time of the initial ROD discussion [8], the concern over moving away from VME led to an alternative VME-based new ROD design quite a bit later, which was taken as a baseline [10] while the development of the RCE/ATCA option continues in parallel. The fact that the VME option took more time to emerge was a reflection of the difficulties in living with the constraints of the legacy hardware to serve a more modern detector. The VME design needed a more significant change of architecture by adding *Ethernet* to the new ROD for moving calibration data to a new dedicated calibration-computing farm for offline processing in order to avoid the VME backplane bottleneck. The prototyping for the baseline VME design is still in early stage to hopefully start actual design implementation and testing in near future.

At the mean time, design implementation and testing with existing prototypes are already well advanced for the RCE/ATCA option emerged from our R&D. Two different configurations, one using the existing RODs, minimally modified, and the other an RCE-based system using the COBs described above are compared in Fig. 5. Some observations:

- Each COB can process six times more channels compared to the current ROD, resulting in a much smaller footprint of a single, 14-slot, ATCA shelf vs. the four, 9U-VME crates containing the old RODs.
- There is just a single flavor of board with the COB in the RCE-based system, and eliminated the need for the SBC (RCC) and the TIM modules in the current VME system.
- The configuration and calibration data path for the RCE-based system is two one (1) Gigabyte/s *Ethernet* ports to *each* COB separately as shown in the figure, compared with the 4 *x* 40 Megabytes/s through the RCCs for a whole crate available in the current VME system.
- The RCE-based system can dynamically reconfigure the data source going into the S-link channels, which are separately located to give a more uniform data density at the ROSes.

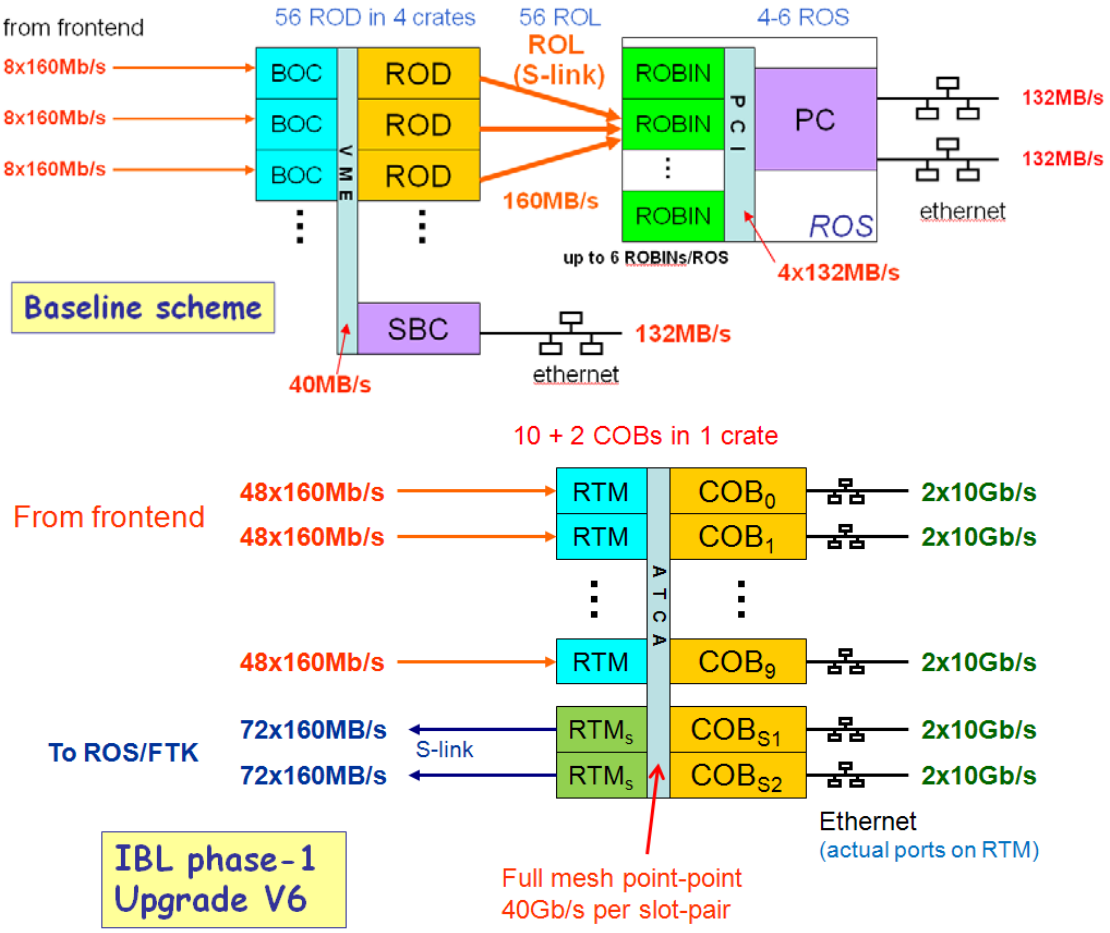

Figure 5: Comparison of the system configuration using the existing RODs (top) and new RCE-based COBs located in an ATCA shelf (bottom).

One of the key enabling features responsible for the reduced footprint of the RCE-based system is its more compact S-link packaging. Existing RODs typically rely on standard S-link cards for

their Gigabit/s data interface. The S-link cards are rather bulky so that one cannot practically have more than three to four S-links per board. While S-links used to be regarded as a high speed protocol that most R&D efforts would shun from their own solutions, the RCE R&D project has already successfully commissioned both the PGP plug-in at 3.2 Gigabits/s and an *Ethernet* plug-in at 10 Gigabits/s, and the implementation of the slower S-link plug-in is expected to be very straight-forward. Using current technology, e.g. SNAP12 as transceivers in conjunction with the COB's RCEs working as dedicated S-link formatters, the awkwardness of the present S-link form factor can be resolved.

While some of the key R&D items such as TTC integration and the S-link protocol plug-in are still to come in the next generation of prototypes, some key tests using the existing prototypes have already fully verified the RCE-based solution for pixel calibrations. Although these key R&D items are important parts of the design, people generally believe that the ROD should be able to meet the challenges imposed by these interfaces. However, a major area of skepticism with respect to accepting the RCE as a design solution was encapsulated in the software/firmware work related to the pixel calibration system, where many people worked for many years to reach the current stable, production state with numerous issues encountered and solved on the way.

As mentioned, the calibrations for the pixel system involve quite a large number of detector channels, making the shipping of calibration data off the ROD impractical, such that one is forced to an intense, in-situ calibration on the ROD itself, with significant demands on both the DSP and its memory and thus requiring the implementation and support of a pretty large and sophisticated software system.

A campaign was carried out to implement the existing pixel calibrations on RCEs and test this implementation using the current FE-I3 pixel modules. The software implementation also took a long-term forward view for compatibility with the future migration policy governing its evolution and integration, as the key communication code already uses the official ATLAS TDAQ software. Through an effort which integrated to about 1 FTE over a year, the majority of the relevant pixel calibrations are already fully implemented and operational in the context of the current RCE framework.

In addition to calibration, the RCEs are also used as a DAQ host for a cosmic telescope located at SLAC, which is used to test pixel sensors. This system successfully demonstrated the natural multi-channel readout capability of the RCE. Finally, an RCE host was used in a June 2010 ATLAS 3D sensor test beam to take data from a single pixel module in concert with the EUDET telescope and verified the data quality and synchronization. An example of a pixel threshold scan calibration done with such a test stand combined with a cosmic telescope readout scheme and its resulting cosmic data TOT distribution is shown in Fig 6.

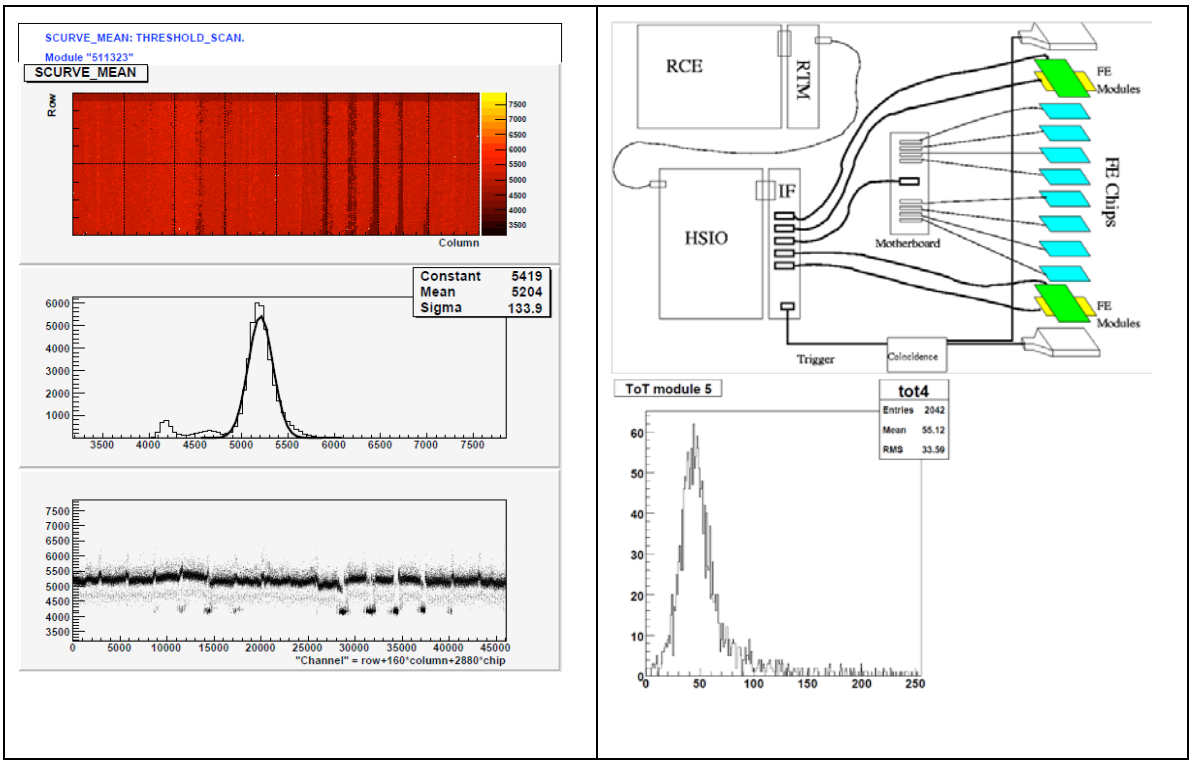

Figure 6: Left shows results of a pixel threshold scan performed with the RCE test stand. Right shows the cosmic telescope RCE readout scheme and a cosmic data Time Over Threshold (TOT) distribution.

Some important experience was gathered from the pixel calibration and test studies:

- The generic nature of the RCE design is verified with the current RCE boards, which although designed for the *PetaCache* project, successfully ran ATLAS pixel calibration and test beam DAQ.
- The distinctive philosophy of the RCE project as an integrated product of hardware and support software has made the implementation of pixel calibration a much faster process than the original system. The ATLAS application effort was entirely software driven with little or no firmware support needed, and therefore only physicists were required to carry out the entire process.
- The RTEMS operating system on the RCEs is a sufficiently versatile environment to allow easy adaption of pixel DSP calibration code and standard ATLAS TDAQ software.
- The calibration interface already using the present calibration console for the full system has ensured a coherent software development path, which serves tests from single module level to test beam and full system, so that the software build-up is towards a single eventual product to avoid throw-away effort. This is in contrast to the original test stand situation for the present pixel and SCT modules tests, which both involved writing of very different software for the eventual full system, while the test stand software became eroded over time for not benefitting all the improvements from the full system software.

Full Pixel Replacement

As the Phase-1 upgrade time frame is potentially one year later given the revised LHC schedule, other upgrade options also begin to appear viable on the Phase-1 time frame. In particular, the upgrade of the whole pixel detector can potentially move to the 3.2 Gigabits/s data rate with GBTs aggregating many channels of 160-320 Megabits/s. A 4-layer upgrade pixel design can amount to ~10000 FE-I4 readout channels corresponding to ~265 M pixels, compared to the 80 M pixels for the current pixel system. The total readout volume is ~250 fiber channels of 3.2 Gigabits/s which is ~5 times the data volume of the current pixel detector. A workable model is to have each RCE processing 2 x 3.2 Gigabits/s of input for the inner layers and just a single 3.2Gigabits/s for the outer-most layer, which would amount to up to ~50 Gigabits/s for each COB. Although each RCE can take more input fibers for DAQ, this choice of channel mapping is to ensure enough RCE resources for calibration. The output data cannot be sent via backplane to a single S-link processing COB, but can be comfortably spread into the 6 S-link COBs to fully utilize the simultaneous per slot-pair bandwidth of 40 Gigabits/s on the full-mesh backplane. A possible shelf (crate) configuration assuming the output still requiring S-link, is illustrated in Fig.7.

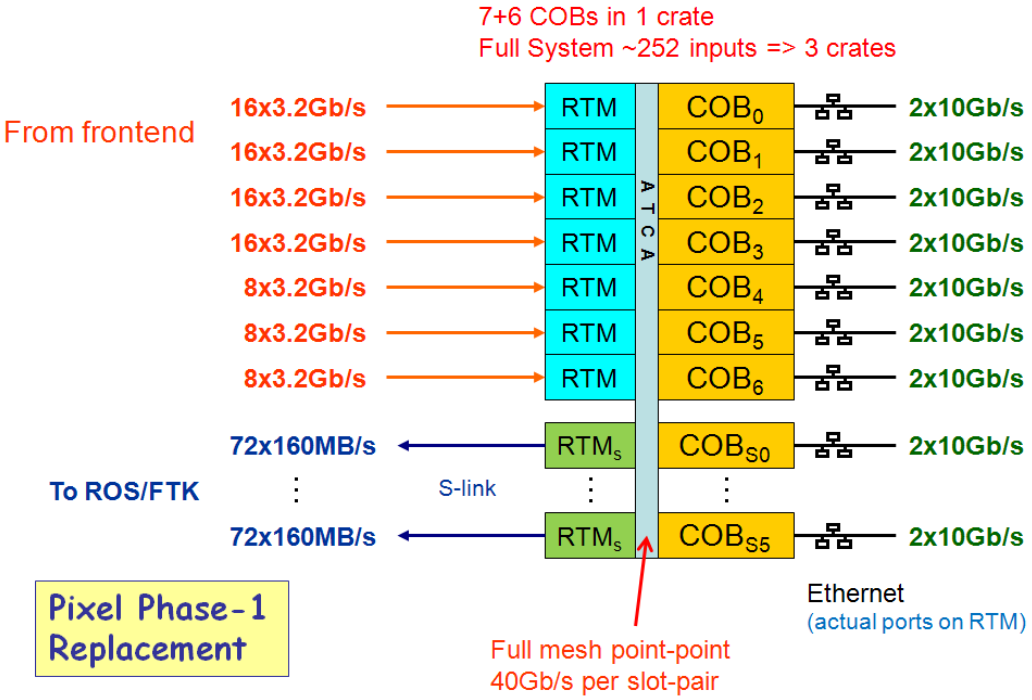

Figure 7: Shelf (crate) configuration for a Phase-1 upgrade full pixel detector with GBT readout and output still via S-link.

As can be seen that despite the ~5 times data volume compared to the present pixel detector, the whole system can be hosted in a much smaller footprint of just 3 crates, in contrast to the present system of 132 RODs in 9 crates. However, it is interesting to note that almost half of the shelf space is devoted to the S-link output, not limited by bandwidth or processing CPU but simply by the connector space at the back of the RTM, to deal with ~1000 S-links for ROS+FTK. This is somewhat inevitable as the S-link is almost 3 times slower than the input fibers at 3.2 Gigabits/s to cause the rather unfortunate spread into many S-link channels. This indicates that from the overall ATLAS readout architecture point of view, by the time the detector data inputs moved to the GBT rate, alternatives to the ROD/ROS interface with S-link should be sought after to address the rate mismatch. Some alternatives will be discussed in the ROS case study. The prospect of 500 input S-links to FTK also calls for thoughts on more appropriate choice of technology to cope with that.

Forward Muon Readout

Although the present muon CSC RODs can operate close to the design L1 rate of 75 kHz, the rather dire spare situation and difficulties in building additional legacy hardware is already pointing to the need to build replacements with new technology if the CSC system will continue to run beyond Phase-1. The rate concerns for both the CSC and the TGC at higher luminosity have already prompted investigation of a replacement of the muon small wheels as a whole. In this Phase-1 upgrade scenario, building the readout drivers with a new technology for the whole detector becomes a natural choice, and the RCE/ATCA option is considered as a main candidate for the new small wheel readout. Given that the detector technology choice is to be made for the muon small wheel upgrade, the readout requirements are still being defined. However, the precision tracking detector is likely to be similar in readout speed and volume so that a study of replacing the existing CSC readout would still be representative. The CSC system has a total of 160 raw data input links at S-link speed of 1.28 Gigabits/s. A readout system model with the RCE-based upgrade can be seen in Fig 8.

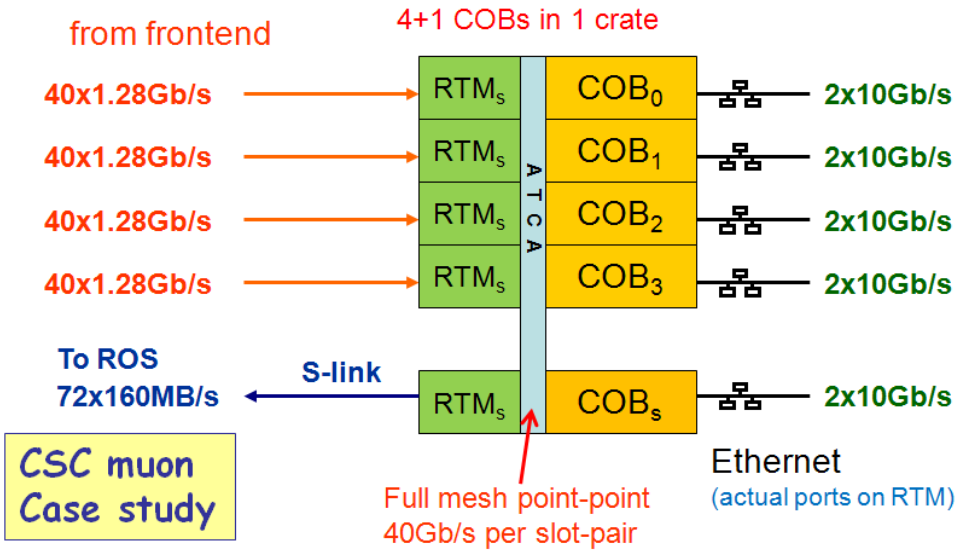

Figure 8: CSC muon readout replacement with RCE-based COBs.

Each RCE would use 5 of 6 input channels to absorb the 5 data fibers from one CSC chamber, which adds up to 40 channels per COB. Unlike in the pixel case, the feature extraction in the RCEs will reduce the output data by a factor of ~5-10 so that one COB working as S-link formatter is sufficient for the whole system. This can be compared with the original system of 16 RODs in two crates. Note that all the COBs, RTM_s, modules are identical hardware with respect to the IBL shelf shown in Fig 5. In general, the only difference in hardware are the input RTMs, which can be detector dependent. In the case of CSC which uses also S-links as inputs, there is actually no need for a system specific RTM while the standard S-link RTM can also be used as CSC input RTM. The firmware/software in the COB_s modules are system independent, thus identical to the IBL case. Only the feature extraction software/firmware in the regular COB are detector dependent. The hardware R&D is therefore a very economical effort that will serve many projects together.

ROS and Alternative Architecture

As discussed at the beginning of the document, improving the access rate for the ROSes can significantly extend the capability of the HLT. While there are incremental improvements that can be made still based on the current ROS design, there is a possible alternative that can

significantly improve the bandwidth situation with the RCE-based COBs. As seen in the pixel case studies, the COB_s has the function of S-link formatting and output over ROLs to the ROS. This S-link protocol can be easily reversed to receive data from a ROL and using the COB as a ROS. It may be useful to compare the resources on each ROBIN card with one RCE:

Resources	**ROBIN**	**RCE**
Input channels	3 ROLs (3x 1.28 Gb/s)	6 x 10 Gb/s
Read Out Buffer memory	3 x 64 MB + 128 MB	4 GB
Processor	*Virtex*-II PowerPC 440 @466 MHz	*Virtex*-5 PowerPC 440 @550 MHz
Output	2 Gb/s	40 Gb/s

Table 1: Comparison of resources on a ROBIN card with an RCE.

The RCEs can function as a ROS in a similar manner as if one were to operate the current ROS in the switch-based mode using the ROBIN *Ethernet* output. In this mode, one has already removed the limitation of the PCI bus in the regular ROS. Each RCE has plenty of memory and output bandwidth to function as two ROBINs if the CPU demand is not the bottleneck. The expected approach with the RCE is to program the core packet handling logic as a protocol plug-in in firmware for much better real-time performance, instead of the heavy burdening of the CPU in the case of the current ROS. This will significantly reduce the CPU demand to allow extended data volume handling capacity. Each ROS PC contains typically 4 ROBINs. Each COB with 8 RCEs can therefore in principle operate as 16 ROBINs, corresponding to 4 ROSes which still needs the protocol plug-in implementation to verify the performance.

The above discussion is for taking the RCE COBs to work exactly like ROBINs to process ROL inputs. This is in fact not the most efficient mode. A more interesting mode is for the COBs serving ROS functionality to sit in the same crate as the “RODs” like COBs to take the data from the “ROD” like COBs over the very high bandwidth ATCA backplane. Each RCE can buffer much more data this way and ROLs are completely eliminated for a cleaner system. The main question will be how the processing load on the CPU and the protocol plug-in scales in managing the larger volume of data but not necessarily much higher request rate. The possible evolution path of a readout scheme for a pixel crate is shown in Fig.9, with the bottom upgrade scheme corresponding to the scenario just described above.

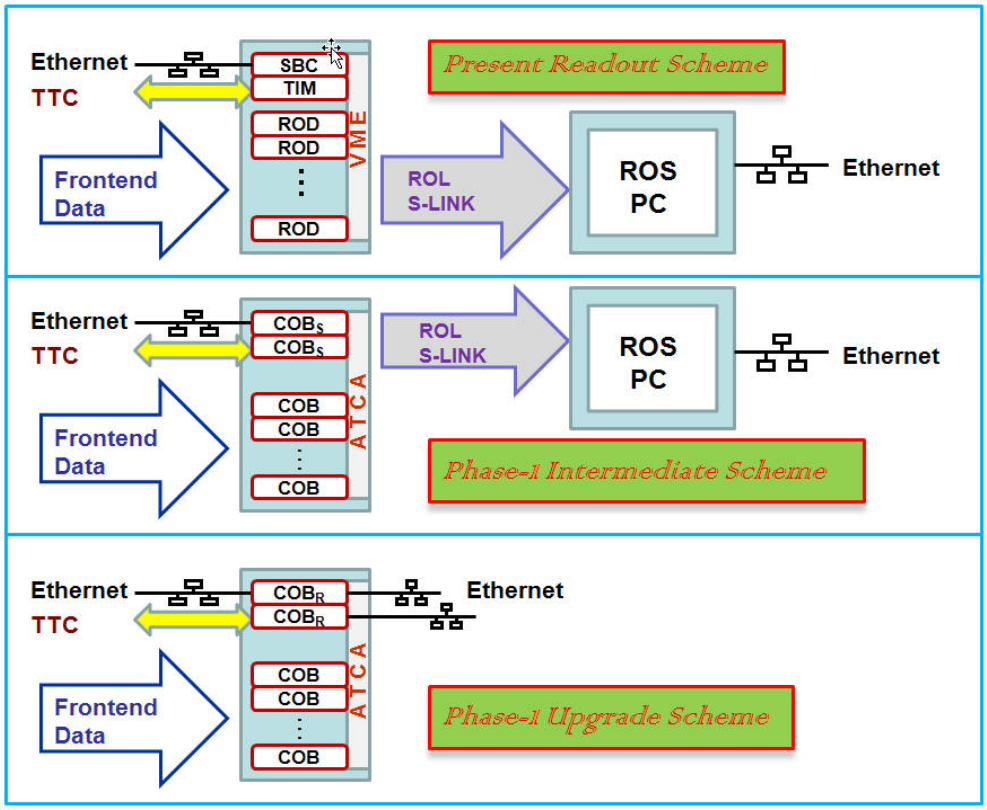

Figure 9: A possible evolution path for a pixel readout crate and corresponding ROS processing. In all scenarios, each COB has its own 2x10GE *Ethernet* ports which are not all drawn.

Comparing the Phase-1 intermediate scheme with the current readout scheme with the ROL-ROS path preserved, one can see that the RCE/ATCA shelf is exactly equivalent to the original VME ROD crate architecturally with identical interfaces to the rest of the experiment. The only difference for the outside world is the slight difference in configuration data content communicated to the crate via *Ethernet* to the SBC or the COBs individually (each COB has its own *Ethernet* port). The crate internal communication technology being VME or ATCA is a local protocol choice that actually does not bear on the global readout architecture. Even for the upgrade scheme with the ROS relocated to be inside a COB to give much simpler physical setup eliminating the bulky ROL S-links, the interface of the whole readout entity of RCE-based "ROD+ROS" is still the same as before with the output appearance just like a ROS. These evolution steps are therefore backward compatible to the current readout architecture at the crate level, to give a completely flexible upgrade path independent of the upgrade timing elsewhere in the experiment. Note also that in the upgrade scheme, the "ROD" (COB) and "ROS" COBs (COB_R) are distinct so that detector systems doing calibrations and TDAQ tests with the "ROS" COBs can go on in parallel without interference.

In the upgrade scheme of COBs performing both ROD and ROS roles in the same crate, the flexibility brought by the ATCA full mesh backplane, replacing the old fixed point-to-point ROLs for the ROD-ROS communication, opens up new opportunities to regulate dataflow. Recall the discussion surrounding the full pixel replacement case study and Figure 7, where a natural evolution could be to covert the S-link COBs to work as COB_R instead to perform the ROS functionalities. For the rest of this paragraph, let's call the COBs performing ROD functions as "ROD" and COB_R performing ROS functions as "ROS" for the clarity of the remaining discussion. The "ROD"s would like to naturally distribute its data over many "ROS"es to make full use of the simultaneous bandwidth to all slots on the full mesh backplane. One most beneficial way to formulate this redistribution is simply to perform event building "for free". All "ROD"s in the crate

can send their data from event 1 to “ROS”-1, event 2 to “ROS”-2 etc. and continue to cycle through the “ROS”es. This will completely iron out any geometrical occupancy variations over all the detector regions covered by the crate to help the dataflow balance between the “ROS”es. More importantly, this brings significant convenience to the HLT. HLT trigger algorithms will never be simultaneously interested in data across different events from the same detector element together like how the current Read Out Buffer data is organized, but rather prefer all data from the *same event* is together in one place for a more efficient data access. Once each “ROS” got all the data from the whole crate for an event built in one place, even when performing e.g. full scan HLT tracking for every L1 trigger, the request rate for any given “ROS” cannot exceed 75 kHz/6 = 13 kHz for any given “ROS” in a crate with 6 “ROS”es. This “ROS” data buffering scheme with fully built events over the crate, will result in larger data volume per request but much more preferable than many times more requests for smaller packets of data all over the crate. There is much to be explored towards new directions like this, afforded by the powerful interconnectivity and high bandwidth of the ATCA backplane.

Other Applications

The above case studies span over very different data rates and processing behaviors while the COBs can be flexibly configured to serve these applications in each case. The regular DAQ needs for most detector upgrades are expected to be within the capacities of the same design. Other TDAQ needs such as the ROIB can probably also utilize the same design. One major area currently not yet explored is the use of this infrastructure to trigger applications. The huge I/O capacity and the full mesh ATCA backplane with 40 Gigabits/s data bandwidth between each slot-pair is exactly what typical trigger applications need. The data flow infrastructure from this R&D project is already an ideal base for Level 1.5 like triggers. The 4-lane mesh backplane should also allow fixed latency constant stream protocol plug-ins, which are in some ways simpler than e.g. the PGP protocol. That would then form the base for exploitation of Level 1 trigger applications also, although it may involve more RCE application firmware for the trigger logic.

Summary

We have described the wide range of possibilities in which the RCE R&D can potentially serve the needs of the ATLAS upgrade. Some very detailed studies done on the pixel IBL calibration applications has confirmed many of the advantages expected of the RCE concept with real applications already serving module tests to test beams with an existing generic prototype. The strong software base as an integral part of the concept has enabled fruitful contributions from physicists with relatively little prior training. The system-level design studies indicate that the RCE-based system not only can live within the present ATLAS readout architecture by conforming to the present interfaces, so that it can already be deployed for Phase-1 or an even earlier time frame, it actually introduces more flexibility to further ease an adiabatic evolution by migrating parts of a system at a time, to take advantage of modern technology for better performance at the earliest opportunity.

References

1. *R. Cranfield et al* Technical Report Institute of Physics Publishing & SISSA, "The ATLAS ROBIN", January 28, 2008
2. *PICMG 3.0* "AdvancedTCA Short From Specification", January 2003
3. *High Speed I/O Module documentation by Dave Nelson:* http://www.slac.stanford.edu/~djn/Atlas/hsio/
4. *Presentations at the Feb/2009 ATLAS upgrade week:* http://indico.cern.ch/materialDisplay.py?contribId=23&sessionId=20&materialId=slides&confId=45460; http://indico.cern.ch/materialDisplay.py?contribId=9&materialId=slides&confId=52930
5. *Presentation at ACES (CERN, Mar/2009):* "High bandwidth generic DAQ research and application for LHC detector upgrade" by Mike Huffer http://indico.cern.ch/materialDisplay.py?contribId=51&sessionId=25&materialId=slides&confId=47853
6. *RCE training workshop:* http://indico.cern.ch/conferenceOtherViews.py?view=standard&confId=57836
7. *RCE Development lab Twiki:* https://twiki.cern.ch/twiki/bin/viewauth/Atlas/RCEDevelopmentLab
8. *Initial IBL readout discussion meeting:* http://indico.cern.ch/conferenceDisplay.py?confId=68905
9. *ATCA readout update by Martin Kocian at Jun/2010 IBL workshop:* http://indico.cern.ch/contributionDisplay.py?sessionId=4&contribId=43&confId=93635
10. *Pixel Insertable B-Layer TDR:*

    https://espace.cern.ch/atlas-ibl/Shared%20Documents/ATLAS-TDR-019.pdf