

## ROB Complex Organization Scenario by the Saclay Group

**Concept:** The input data rate of the ATLAS read-out system is determined by the high Level 1 accept rate of 75 (100) kHz and the input bandwidth of individual read-out buffers can be as high as 180 Mbyte/s. Due to the RoI based event selection each read-out buffer delivers data to the Level 2 trigger at lower rates of the order of 10 kHz. The full event data collection is performed at most with the Level 2 accept frequency estimated to a few kHz. As a result the bandwidth requirements for the ROB's output are much lower compared to its input. Based on these considerations, an organization of read-out units in clusters of ROB's ( $ROB_{IN}$ -s) that share a network interface for high level triggers and DAQ system (Figure 1) has been proposed in [1]. Operation of such ROB Complex (initialization, service of network messages, etc.) is controlled by a general purpose processor, ROB Controller. A bus with adequate bandwidth connects the  $ROB_{IN}$ -s, the network interface and the controller within the ROB Complex. Depending on the run control and monitoring schemes adopted, yet another processor module might be used to address the necessary functionality.

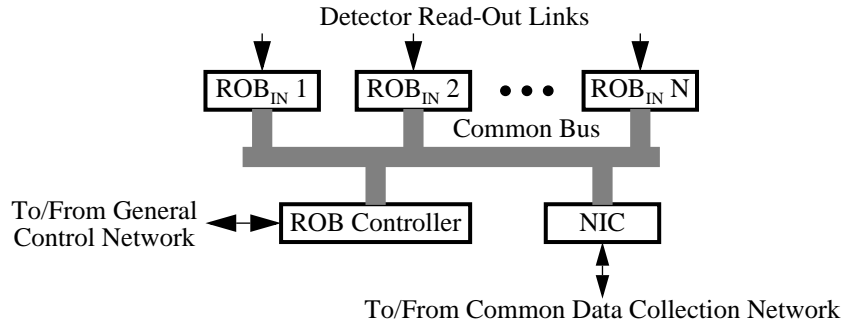


Figure 1. ROB Complex Organization

Two possible modes of operation can be used concurrently for the data collection within the ROB Complex. In the Mapped Memory Mode, event memories of the  $ROB_{IN}$ -s and the memory of the ROB Controller are directly accessible by the data collection network interface. Therefore, the transmission of event data fragments and associated control structures, can be performed by the NIC using some kind of chained data transfer mechanism. This mode reduces data movement over the common bus, as well as the load on the  $ROB_{IN}$ -s and the ROB Controller. In the Copy Mode, data is first moved from the  $ROB_{IN}$ -s event memories to the ROB Controller's memory, either by  $ROB_{IN}$ -s' or ROB Controller's DMA mechanism, and only after that is sent to the data collection network. This mode allows preprocessing of event fragments within the ROB Complex prior to their transfer to the event selection or filtering systems.

**Design Considerations:** For a given throughput of the network links the number of  $ROB_{IN}$ -s grouped within a ROB Complex depends on the  $ROB_{IN}$ -s output bandwidth. It also depends on the rate at which the  $ROB_{IN}$ -s have to supply data to the event selection and filtering processes. In general, both, the bandwidth and the rate, vary depending on the detector subsystem and within subsystems as well. For obvious reasons, the sum of the  $ROB_{IN}$ -s average output bandwidth within a ROB Complex should not exceed the bandwidth of the network link. The rate constraints are due to the following. For each data collection request, the ROB Controller should notify the  $ROB_{IN}$ -s concerned and collect their response messages, even if this does not imply an actual event data movement. For the RoI data requests only some of the  $ROB_{IN}$ -s within a ROB Complex may participate to the data collection. However, for some Level 2 triggers (e.g. TRT Full Scan, Missing  $E_T$ ) and for full event building, all  $ROB_{IN}$ -s have to deliver their data. In these cases the ROB Controller should be able to operate at a frequency equal to the product of the request rate by the number of  $ROB_{IN}$ -s in the ROB Complex. In absence of any multicast mechanism on the common bus the same consideration applies to event data clear requests which have to be delivered to all  $ROB_{IN}$ -s.

Table 1 summarizes some results from modelling [2] relevant to the ATLAS detector read-out organization (the Level 1 rate is 75 kHz, the Level 2 accept rate is 3.75 kHz, the 'Scan' column corresponds to low luminosity operation with B-physics triggers, the 'Low' column corresponds to low luminosity operation without the B-physics triggers, the 'High' column corresponds to high luminosity operation). The Full Scan triggers generate high, 10-12 kHz, event selection data request rate in the inner detector. Taking into account a full event building rate of about 1-2 kHz, the rate constraints suggest a grouping factor of 4 within the ROB Complexes of these subsystems. The electromagnetic and hadronic calorimeters require high,  $\sim 11$  MByte/s output bandwidth mostly due to full event building which is performed for every event accepted by the Level 2 trigger. As the data collection rate for the  $ROB_{IN}$ -s in the calorimeters is sufficiently low, the grouping factor is determined by the throughput of the network link chosen. Clearly, the 100 Mbit/s bandwidth of the Fast Ethernet links is not enough. A ROB Complex



izations of ROB Complexes with 5 and 12 ROB<sub>IN</sub>-s respectively. While the former can be used for the inner detector and calorimeter read-out, the later is more suitable for the muon subsystem. Note that constructing 6U chassis with four electrically independent busses allows four ROB Complexes with up to 6 ROB<sub>IN</sub>-s. Very large grouping factors (up to 26) can be achieved by transparently bridging four PCI bus segments. Inversely, connecting bus segments with few PCI slots allows for many ROB Complexes with small grouping factors.

**The ATLAS Read-out organization:** These simplified considerations can be used to make some estimates for the ATLAS read-out organization. Table 2 gives the number of ROB Complexes and the number of read-out crates calculated with the following assumptions: the bandwidth of the data collection links is ~80 MByte/s; grouping factors for ROB<sub>IN</sub>-s in the Muon, calorimeters and inner detectors are 12, 5 and 4 respectively; ROB Controller, NIC and ROB<sub>IN</sub>-s are implemented in 3U form factor, 6U CompactPCI chassis are used.

Table 2: Estimate for the ATLAS read-out organization based on ROB Complex units

Detector Subsystem	Number of ROB <sub>IN</sub> -s	Grouping Factor	Number of ROB Complexes	ROB Complexes per crate	Number of crates
Muon Precision	192	12	16	2	8
Muon Trigger	48	12	4	2	2
EMC	760	5	152	4	38
HAC	98	5	20	4	5
TRT	256	4	64	4	16
SCT	92	4	23	4	6
Pixel	84	4	21	4	6
Total	1530		300		81

According to the estimates the 1530 read-out buffers are grouped in 300 ROB Complexes housed in 81 CompactPCI chassis. At most the ROB Complex should sustain ~15 KHz of data request rate (in the Pixel detector) that corresponds to an internal request servicing of about 60 kHz. As each ROB Complex has a single network link that transports both Level 2 and event building data, 300 80 Mbyte/s links are necessary to connect the ATLAS read-out system to the data collection network. The maximum load on such links does not exceed 70% (for the hadronic calorimeter). Assuming equal bandwidth for data sources and destination processors a 600-port data collection network is necessary. In the case of Gigabit Ethernet technology, this corresponds to a network with a total bandwidth of 600 Gbit/s. Neither the construction of such a network and its ability to handle the data traffic, nor the ROB Complex operation at the high internal rates are obvious.

The study of the network organization and its behaviour under the estimated data flow patterns are tasks for computer simulation and demonstrator systems. Both the size of the network and the bandwidth requirements, that are mostly determined by the high rate of full event building, can be significantly reduced if event filtering algorithms that operate on partial, rather than full, event data can be deployed.

The validation of the proposed principles for the ROB Complex organization is going on within the ATLAS ROB working group. At Saclay a ROB<sub>IN</sub> board is under development with an 8 MByte event memory, a FPGA logic that handles input data streams and a local processor responsible for the event memory management and event data request servicing. Its first version contains processor subsystem with 33 MHz I960 processor and 512 kByte system memory. Due to the flexibility of the PMC form factor chosen for the first version of the ROB<sub>IN</sub>, a prototype ROB complex with three ROB<sub>IN</sub>-s has been assembled on three different platforms: CompactPCI, PC and VME. All three ROB Complexes are multi PCI devices that use transparent PCI-to-PCI bridges. The prototypes have been integrated in the ATLAS Level 2 Trigger ATM testbed and results of performance measurements are regularly reported at the ROB working group meetings [6].

**Summary:** The proposed organization of the ROB Complex allows building sufficiently modular structure to achieve the desired performance: input of detector data streams and event memory management, data collection protocol, run control and monitoring are logically separated functions that can be separated physically as well. It also reduces the number of required links for the common event selection and event building network, compared to a solution where each read-out buffer possesses a network link. From the organizational point of view, the read-out units have the same granularity in the event selection and event filtering systems, that might help to keep flexible boundaries between them. The ROB Complex architecture suggests the use of commercially available components based on widespread industrial standards.

**References:**

- [1] J. Bystricky et al., "A Sequential Processing Strategy for the ATLAS Event Selection", in Proc. Nuclear Science Symposium, Anaheim, California, 3-9 November 1996. Also in IEEE Transactions on Nuclear Science, vol. 44, 3 June 1997.
- [2] J. Vermeulen, <http://www.nikhef.nl/pub/experiments/atlas/daq/Modelling-2-6-99/Presentation-June-99-update.pdf>
- [3] PICMG, CompactPCI Specification, Version 2.1, September 2, 1997
- [4] Gespac Innovative Embedded Solutions, <http://www.gespac.com>
- [5] TreNew Electronic, <http://www.trenew.com>
- [6] The Saclay ATLAS L2 Trigger group, <http://www-dapnia.cea.fr/Phys/Sei/exp/ATLAS/pres/pres.html>