

Notes on ROB Complex meeting

NIKHEF, Amsterdam, 26 January 1999

Notes by : R. Cranfield / J. Vermeulen

Present

A. Amadon, R. Blair, R. Bock, P. Clarke, R. Cranfield, G. Crone, N. Ellis, B. Green, M. Huet, R. Hughes-Jones, K. Korcyl, A. Kugel, P. Le Dû, L. Levinson, I. Mandjavidze, L. Mapelli, F. Wickens, H. Boterenbrood, G. Kieft, P. Jansweijer, R. Scholte, J. Vermeulen

Agenda

Morning, chair P. Clarke

1. Master Working Document : discussion of current version, inventarisation of further work and assignment of tasks.
2. Work plan : current status, milestones, updates.

Afternoon, chair J. Vermeulen / R. Cranfield

3. Reports on current activities and discussion.

1. Master Working Document (MWD)

1.1 Introduction

R. Cranfield showed a transparency describing the purpose of the MWD. P. Le Dû asked about the interaction with DAQ. R. Cranfield: the ROB Complex group is covering everything for now, but a joint discussion with DAQ needs to be started.

1.2 MWD Sections

P. Clarke noted that not enough time for detailed discussion (12 sections) was available, the aim should be to find people to follow up with legwork on different sections. The discussion should concentrate on Section 6 of the MWD, titled "Design Issues".

6.1 "Timescale"

Not discussed.

6.2 "ATLAS strategy", 6.3 "ROD Constraints", 6.4 "Detector Requirements", 6.5 "Data Profile"

TTC

J. Vermeulen remarked that possible use of "standard" ROD crates would make use of information distributed via the TTC system relatively straightforward, the question then is whether this could be useful. P. Le Dû agreed it should be studied as an option. N. Ellis: the TTC is fully documented - clear what it provides, is it needed ?, the ROD crates however are not fully defined. L. Mapelli : use of the TTC could introduce additional complexity. P. Le Dû: it could be needed for standalone testing (without RODs). L. Levinson: the timing info needed for the RODs may not be needed for ROB.

=> *To be followed up*

Read Out Link (ROL)

R. Cranfield: The ROD workshop made clear that the ROL should support 32-bits data transfer at 40 MHz.

>1 fragment per event ?, event ordering

J. Vermeulen: In the ROD workshop it was asked whether 2 fragments per event are OK - N. Ellis: the answer should be clearly NO, but this should be discussed in Detector Interface Group. P. Le Dû: this comes from the LAr - needs

L. Levinson: is event ordering required ? - R. Cranfield: nothing is assumed, N. Ellis: the interface document should be checked.

XOFF

A. Kugel asked about the situation with XOFF. He noted that the buffer space in the ROD probably is much smaller than that in the ROB, so it does not make much sense to send an XOFF to the ROD when the ROB buffer is nearly overflowing. J. Vermeulen remarked that the interaction with the LVL1 system is a task of the RODs, the ROBS signal overflow conditions indirectly to the LVL1 system by sending XOFF signals to the RODs. However, the question is still open, and ideally needs modelling (whole system). M. Huet: timeouts will probably be necessary, P. Clarke / R. Cranfield: the possible timeouts should be enumerated as next step. (N. Ellis: need TTC to recognise some timeouts)

=> *To be followed up*

Monitoring

Monitoring: is data monitoring needed at the level of the ROB as well as at the level of the RODs ? L. Mapelli: the ROD requirements cover a range, it is premature to force the issue, it should probably be assumed that monitoring of the data is needed. R. Cranfield: monitoring could have a major impact on the design of the ROB. L. Levinson / R. Bock: what about standardisation? - L. Mapelli: there are some advantages to monitoring at ROB level, it is not so clear that monitoring a wider area can easily be done before the event builder, detector specific monitoring is a task for the RODs.

J. Vermeulen: maybe we should take P. Clarke's proposal (with addition of starting ID in the request)? Is this problematic? R. Cranfield: it is not difficult to look at event type if you don't need to maintain lists of already received events of different types.

L. Mapelli: the current DAQ system only monitors at crate level. Monitoring is decoupled from the dataflow (the events are stored in a database). N. Ellis: watch out that the LVL1 ID may recycle about once per second (because some RODs (SCT) need to reset frequently due to minimal error recovery). L. Levinson: what about using forced L2 accepts? R. Bock: this is assumed anyway, but it serves a different function.

=> *To be followed up*

6.6 "Level-2 Strategy"

ROB mapping and pre-processing, architectural grouping and data-selection

J. Vermeulen: the problem with the request rates of ROB become bigger for the new ROB mapping, the RoI request rate is more important than selection of RoI data. N. Ellis / P. Le Dû: evidence has to be prepared that it is bad not to have towers so that we can get a right balance.

L. Mapelli was not happy about processing data in the ROB and worried about integrity of the data and prefers to leave the main dataflow untouched. R. Bock: it was tried to keep to this (with ROB sending whole fragments) but now this must be reexamined in view of readout problems – the tradeoffs need to be assessed. P. Le Dû: the problem should now be looked at quantitatively. The main problem is

mapping of the calorimeter. J. Vermeulen: the small number of ROBs for certain subdetectors is still a problem, whatever the organisation. P. Clarke: should there be a presentation at ATLAS week? J. Vermeulen: overlaps with modelling meeting tomorrow. P. Clarke: many of these issues are really the responsibility of the algorithm group. R. Bock: a problem is the lack of people in the algorithm group! P. Clarke: studies are only really necessary for calorimeter (already being studied by A. Amadon), but don't know about muons.

=> *In ATLAS week overview of where we are*

N. Ellis: muons are unique because the data is never shared between towers. P. Le Dû: we should be careful not to keep revisiting the baseline system (it might evolve, but we should try to keep to the baseline). R. Bock: past studies of local processing always floundered on boundary problems.

L. Mapelli: DAQ-1 could be used to check out jointly with Level-2 the option of local multi-ROB pre-processing. P. Clarke: a "local" problem does not involve boundaries, e.g. TRT Hough transform. P. Le Dû: specific studies need to be done. R. Bock: proposes to study what can be done with multi-ROBs (could be DAQ-1 crate or ROB Complex or both).

=> *to be followed up with documenting the implications of data selection : calorimeter most important - underway*

=> *to be followed up with studying relevant issues with respect to grouping.*

Level-2 decision record

M. Huet: we've lost the item on the Level-2 decision record - R. Cranfield: should be reinstated!

6.7 "Data Format", 6.8 "Error reporting"

J. Vermeulen: we should revisit the definitions of all messages the ROBs have to deal with (R. Cranfield: ...in conjunction with DAQ). M. Huet / I. Mandjavidze: it is awkward not to have a standard format (e.g. TRT bitstream) -- R. Bock: there is a possibility of re-formatting by FPGAs. R. Cranfield : more work is needed on error reporting.

In 6.7.2 attention should be paid to the consequences of pre-processing.

=> *to be followed up*

6.5 "Data Profile"

J. Vermeulen: what is the delay time between RoI request and data fragment arrival ? This could be important - it depends on the RoI request distribution strategy. G. Crone remarked that the present ROB URD requires that an empty packet is sent when a RoI request is received for data that is not available in the event data buffer of the ROB. This should be changed.

6.10 "Software"

Work has been done on three items : the external API, the internal API and an UML analysis. R. Cranfield : how do we proceed ?

=> *to be followed up*

2. Tasks & milestones

1) "Provide input for ROB User Requirements Document"

J. Vermeulen: this is mostly covered in the MWD, but we should probably have more detailed milestones. The convenors will prepare a list of actions, milestones and people responsible for the next step (see list at the end).

2) "Design studies, including performance measurements"

The published milestones are ok, but more detail of individual group work is really needed :

UK (R. Cranfield): a. PCI measurements, b. study of a multilink ROBIN

SACLAY (M. Huet): ROB Complex measurements, use of hardware in application testbed

NIKHEF (J. Vermeulen): study of CRUSH design - measurements & modelling (& simulation). Also looking at ROB-out part. Looking ahead to final scenario.

MANNHEIM. (A. Kugel): 1 microEnable ROB-in with NT driver & TRT pre-processing. Look at multi ROB-in system (3 – 5 microEnables) and 4 input link ROB-in (new design).

J. Vermeulen: proposes to include in the MWD a framework for description with preliminary information by mid-March, and then in two steps to add more information at the end of April and 1 – 2 weeks before the next ROBComplex meeting (to be driven by the convenors)

3) "Production and support of buffer prototypes"

This item was left for the status reports

4) "APIs"

I. Mandjavidze questioned whether we do need a *common* API and what the purpose of it is, is it only for use in testbeds ? L. Mapelli : it is premature to break into current prototype studies. R. Cranfield : does the TP need a software proposal? It was generally felt that it is too early to worry about eventual APIs - this is difficult to do without a more specific architecture.

3. Status Reports

3.1 Mannheim

A. Kugel reported on use of the microEnable as ROB-in. The buffer memory is a dual-port memory allowing 2 accesses per 50 ns. A circular buffer scheme without overflow handling is implemented. Performance measurements have been done with the S-link input running at 80 kHz event fragment rate (data transfer rate 80 MByte/s for 20 MHz, 32-bit data delivered by S-link interface to micro-Enable). It was attempted to handle as many RoI requests, sent to the microEnable via its PCI interface, as possible. In the ROBComplex meeting of a year ago 15 kHz, 10 MB/s was defined as requirement, 14 kHz, 14 MB/s was measured with the standard NT device driver, it is expected to go to 22 kHz, 22 MB/s with a new Windows NT driver.

TRT pre-processing has been added now: the average processing time for zero suppression of straw hit bit data is ~130us for 1642 straws with 2 bits per straw and ca. 30 % occupancy.

L. Levinson expected to get 5 times the performance with new hardware & software.

Plans:

MicroEnable :

Fix driver, implement Robin API, 4 ROB-ins/PCI bus - coordinate with HPCN activity; try direct DMA from ROBins to PreProc or ROBout, add some monitoring.

Atlantis multi-ROB :

Studying multi-link options using LVDS "M-Link" cards (4 links per CompactPCI board) (to interface to LVDS S-Link cards from Krakow), buffering method not decided (based on unique event IDs, but could be circular buffer with overflow handling), 2 or 4 MByte of buffer memory, 40 MHz 32-bit data rate.

3.2 HPCN

R. Bock described the current status with respect to the study of the possibility of use of scalable and affordably priced systems from industry with PCI connections. The developments are tracked in a low-key fashion. The joint project with Digital is in abeyance – work is now focusing on a Data General system (4 Intel-based boards interconnected with SCI). Possibly ROB-ins will be used again. R. Bock will report back if interesting (these systems are used for data-mining/data warehousing).

3.3 RHUL/UCL

B. Green : ROB-ins have been produced in both PC and PMC formats (the PMC version was passed around). PCI performance measurements have just started. The design study of a MkII ROB-in (multi-link design) has not yet started.

A. Kugel: a compilation of current measurements and a review of what needs to be done is needed. P. Le Dû: the measurement results should be looked at and it should be seen what has to be done. This should be done in the next ROBComplex meeting. P. Clarke : it should be avoided to start new projects.

3.4 SACLAY

M. Huet reported that measurements have been done with 3 ROBINS (partial implementation of ROBIN functionality). Work on the full ROBIN has been continued : some parts (especially SDRAM management) need still to be finished. Also work has been done on the ROB emulation for the Reference Software (see 3.6).

P. Le Dû emphasized that the design took much longer than expected.

I. Mandjavidze discussed measurement results. The ROB Complex software is organised in layers and is developed in C for the Saclay ROBIN and for a Transtech SHARC board in combination with Windows NT. The test setup consists of the Saclay ATM testbed with 2 supervisor nodes + 14 destination nodes and the ROB complex connected with an ATM link to the ATM switch. The ROB Complex itself is implemented on a VME platform consisting of a RIO2 + ATM NIC (one PMC slot) + 3 ROBINS (connected to second PMC slot). Measured is the maximum request rate from 14 destinations for different reply (fragment) sizes - distributed over up to 3 ROBINS. The maximum rate for small fragments is $1000 / (20\mu s + 10\mu s * \text{no.of ROBINS})$ kHz and is limited by the ATM link speed for fragment sizes >1.5kB. For the emulation with the Transtech SHARC board the maximum rate found is $1000 / (20\mu s + 3\mu s * \text{no.of ROBINS})$ kHz, but the data transfer rate to the ATM NIC now limits the throughput (due to the SHARC card) rather than the ATM link speed.

CompactPCI is going to be used for connecting up to 6 ROBINS.

3.5 NIKHEF

Paroli

P. Jansweijer described results obtained with the Paroli concentrator box. Two SIEMENS Paroli chips are used to replace 12 bidirectional optical <-> electrical convertors for 12 ROLs (with Gigabit Ethernet cables between RODs and Paroli concentrator box or between ROBs and Paroli concentrator box). The electrical part (connectors unshielded!) was studied by looking at "eye patterns". It was found that these do not match the GBE-Electrical standard : clock regeneration ought to be used. However, in practice it is possible to get away with it because of the latitude in eye patterns in combination with the short electrical connections required. The main question is whether the total cost of 2 boxes + cables is lower than the cost of 12 ROLs with individual optical <-> electrical conversion ? The current cost is about equal.

L. Mapelli: the target cost of a ROL is 500 CHF - will the Paroli type solution ever be ok? Answer: it all depends on the market...

Performance measurements on CRUSH

H. Boterenbrood presented results of measurements on the performance of the CRUSH. A microEnable (with microSLATE applet) or SLIDAS is used for input of data via the CRUSH S-link input, external connections are emulated with the help of a PCI-SHARC board, a second PCI-SHARC board is used for output. All data transfers are done by DMA. The microEnable limited real S-Link input to 53 MB/s. Plots show limit for short fragments at >100 kHz & limit of SHARC link speed for long fragments

ROBOut study (J. Vermeulen)

A basic problem is the interfacing of the ROBComplex to the network (GBE or ATM), which is potentially expensive. Would it be possible to have a small autonomous PCI bus inside the ROBComplex (without backplane connections)? The network protocol may be a problem - looking at I2O for a solution -> many questions.

Presently studying possible designs of the SHASLINK (SHARC + S-link out + PCI PLX chip (9054) with possibility for support for autonomous PCI bus). It proves to be hard to get a high throughput on the PCI bus. The original motivation of this work is a design study of the NIMROD, but the resulting design is also of interest for studying the ROBOut part of the ROBComplex

An alternative for interfacing of the ROBComplex to the network via PCI bus may be formed by using dedicated point-to-point links (like S-link or FireWire) direct to farm processors with interfaces to the network. This could allow up to 20*8 ROBins per crate and a total number of crates of the order of 10. This in turn would allow to send RoI requests and decisions via dedicated point-to-point links like S-link to RIO-II like processors in the crates (one per crate).

3.6 TRT ROB emulation (M. Huet)

Which format has to be used : ROD (not frozen), LVL2, DAQ ? What is required with respect to conversion of the data ? TRT data: 4% occupancy, variable size bit fields in bitstream, resulting in 500-750B per event. For LVL2 only a hit map is needed, i.e. 2 bits/straw = 400B/event. The emulated ROB represents a number of basic ROBs.

Conclusions: what seemed to be a simple matter of producing an ASCII file for the emulator has unearthed a lot of questions about how the TRT data is to be transferred/compressed/formatted

Next meeting

May / June. June may be problematic in view of conferences and the June ATLAS week.

Action points

Names added after the meeting, February 24, 1999.

1. MWD section 6.2, 6.3, 6.4 : L. Levinson, A. Kugel (end of March)
2. Discussion with DAQ-1 group : convenors (first discussion 8/9 March)
3. MWD section 6.5 : J. Vermeulen, R. Scholte (end of March)
4. MWD section 6.6.1, 6.6.2, 6.6.3 : R. Bock, M. Huet (end of March)
5. Framework for describing results design studies : convenors (first version : 1/2 March, next iterations end of April and 1 - 2 weeks before next meeting)