AN EFFICIENT DETECTION PROTOTYPE BASED ON A MIXED ARCHITECTURE FOR GAIA

Shan Mignot, Philippe Laporte and François Rigaud

Observatoire de Paris / GÉPI,
5 Place Jules Janssen, 92195 Meudon Cedex, France
phone: + (33) 1 45 07 79 07, fax: + (33) 1 45 07 78 78, email: shan.mignot@obspm.fr

ABSTRACT

We present a high throughput prototype detection framework designed to meet the stringent scientific requirements of the Gaia mission and satisfy real-time and processing resource constraints on board the satellite. A mixed architecture, whose feasibility has been confirmed as part of phase A, is proposed which manages pixel level operations synchronously with their acquisition through programmable logic (FPGA) and answers the need for more flexibility and higher level object-wise processing by the use of software.

1. INTRODUCTION

ESA's cornerstone mission Gaia, due for launch in Dec. 2011 and currently in the detailed design phase under the lead of EADS Astrium SAS (Astrium), will take up the challenge of building a magnitude-limited billion-star all-sky survey amounting to approximately one percent of the Milky Way galaxy. In Hipparcos' legacy, the satellite will continuously scan the sky thanks to combined spin and precession motions, thereby collecting data in two distant fields of view combined to a single focal plane to investigate the composition, formation and evolution of the galaxy based on absolute astrometry and astrophysical characterisation of the objects [1]. Unlike its predecessor, however, Gaia will be orbiting around the second sun-earth Lagrangian point, selected as a trade-off favouring thermal and dynamical stability at the expense of bandwidth for the transmission to earth. Similarly, the photomultipliers are replaced by 106 4500 × 1966-pixel CCDs operated in Time-Delay-Integration (TDI) mode, a mode which shifts photo-electrons in the detector's matrix synchronously with the spin-induced motion of the objects on the focal plane and produces five-year-long strips of data at a total rate amounting to several giga-bits per second.

The impossibility of downloading so much data calls for elaborate on-board processing to perform content selection, before advanced data management and compression are introduced as outlined in [2]. Content selection begins at the data acquisition level with a partial read-out of the CCDs intended to only window objects from the magnitude[1] range of interest ($[6;20]$) in an otherwise mostly empty sky. Since the sky is only incompletely known at Gaia's sensitivity and angular resolution, use of a catalogue, as by Hipparcos, is inapplicable and an autonomous on-board detection is required to produce an unbiased survey. Objects are hence detected as they cross the Sky Mappers (SM 1 and 2 for the two fields-of-view respectively, see Figure. 1). To discard particle events and meet the requirements of the spacecraft's attitude control loop, objects are then confirmed and their motion estimated in the first astrometric field CCD (AF1) before being tracked in the following ones. Finally, the information collected at these levels is at the core of the priority-driven data management and storage designed to ensure graceful degradation in case resources are exceeded.

During phase A of the project, the critical importance of these operations for the return of the mission led one of the working groups set up by ESA to elaborate a detection prototype, which is the object of this paper, to assess achievable scientific performances and assist in the elaboration of the specifications. This model then grew to encompass all operations regarding data acquisition (Pyxis [3]) and formed the basis of a study conducted by Astrium GmbH on the feasibility and sizing of on-board electronics. This study set up a soft real-time software breadboard running VxWorks on Maxwell's PPC750FX-based board as an upper bound to available performances (engineering model with peak 1750 MIPS) and concluded [4] that a mixed hardware-software architecture is required to meet processing objectives. Our detection scheme was also the object of a much simplified hardware prototype elaborated by Astrium as part of the response to ESA's invitation to tender for Gaia. These combined results validated the feasibility of the proposed design so that, supported by an "Action Concertée Incitative" research project (Algol), a full-fledged mixed architecture detection demonstrator is now being developed at the Observatoire de Paris.

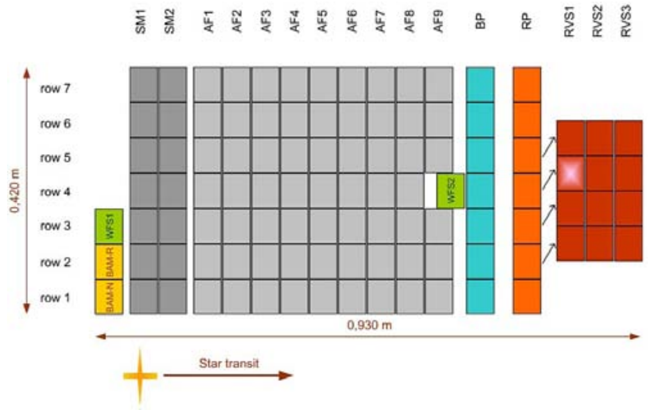

Figure 1: CCD assembly at Gaia's focal plane.

After presenting the constraints which apply and discussing the way they model the proposed design, we present the overall processing structure and the algorithms used with a view to implementation before concluding with a short performance evaluation.

2. DESIGN CONSIDERATIONS

The scientific performances are naturally the key design driver. Although the specification may be stripped-down to the simple statement that all objects of interest should be detected and measured in terms of magnitude and location with minimal false detections, this formulation would conceal the need to be sensitive to sources with fluxes differing by ratios up to ~ 400000, yet robust to effects related to

- their physical nature (mainly colour and angular extension),
- their environment (companion stars, sky background and highly variable densities with peaks at 120 times the average),
- and the variety of observing conditions (particle fluence, ageing of detectors and satellite attitude),
- (not to mention repeated design changes early in the project).

[1] Magnitude relates to flux according to $m = m_0 - 2.5\log(f)$ and hence decreases when the flux increases.

For these reasons, local or diffraction-pattern-dependent methods were disregarded in view of the difficulty to set up an homogeneous detection scheme and, hence, to validate the multiplicity of special cases that would necessarily arise. Instead, inspired by the sequence of operations proposed by M. Irwin in [5] and credited by its application to large scale photographic plate analysis, a generic approach was preferred based on a non parametric segmentation stage forming data units with rich content for posterior selection and characterisation.

Object detection is an efficient compression system which substitutes an abstract description of content to the exhaustive list of pixel data and thus considerably reduces the data flow. However, SM images must necessarily be entirely scanned to this end and, with a TDI period of leading to the read-out of 1966 16-bit pixels (in ADU) at 1.018 kHz, the 3.8 Mbytes/s data rate should be sustained by detection. The volume of data along with the necessity to traverse it according to the chosen connectivity for segmentation calls for a hardware implementation controlling data accesses according to a fixed pattern to avoid cache miss delays impeding software operation [4]. Conversely, benefiting from the data reduction achieved by preserving only a few percent of object-member pixels under maximum density conditions, and considering how variable object characterisation necessarily is, a software approach is preferable at this level.

The partition introduced above assigns pixel-based processing to hardware and the object-based one to software according to data flow and diversity of operations. It offers an opportunity to further optimise the data flow. If SM images are not considered pixel per pixel but instead by blocks of 2×2 binned at the CCD level (samples hereafter), thus trading image resolution against higher signal to noise levels for improved detection completeness and reduced data rate, SM1 and SM2 may be read-out alternatively every two TDIs and processed by the same hardware and software.

The TDI-synchronous nature of the data transmission from the CCDs calls for a hard real-time system to avoid an input data buffer whose size would increase rapidly with latency. The output data, on the other hand, is naturally dependent on object arrival times and only subject to a hard real-time constraint much further down the line (programming command for AF1). Although margin must be kept for the tasks leading to formatting this command, the traversal time from SM1 or from SM2 to AF1 (at least 1400 TDIs) provides a possibility for soft real-time management at this level.

3. PLATFORM

The balance achieved is one between volume and complexity. The hardware is left to deal with high data rates but only to perform highly systematic low-level operations with the main objective of identifying the small fraction of relevant data on which the software can then carry out more adaptive and sophisticated treatments. Dedicated hardware is well suited for the former to process the data line per line, even sample by sample in our case, with a fixed scheduling. Because of sharing with the other tasks and important design margins, high performance requirements are placed on the CPU needed for the second part [4], hence Maxwell's SCS750A board remains the only viable solution to date.

Besides mass, power and volume constraints which relate mainly to launch capabilities and cost, the supporting hardware must withstand both the transition to vacuum (through specially engineered packages) and the radiation environment. With radiations in orbit consisting of the superposition of a low energy permanent regime with high energy particles, designs rely on radiation-hardened antifuse devices (which are not re-programmable and have lower transistor densities) for the former and a combination of Error Detection And Correction (EDAC) codes, Triple Module Redundancy (TMR) and sub-system redundancy for the latter. The rationale is that unwanted electrical state changes cannot be avoided as a consequence of particles but are localised and occur only infrequently. All above-mentioned strategies hence intend to recover from these undesirable states by introducing some redundancy either at the encoding level (EDAC), at the gate or component level (TMR) or at the sub-system level (for permanent damages or critical tasks). Additionally, the system must be capable to recover from theoretically unreachable states which implies some overdesign, for instance concerning finite state machines.

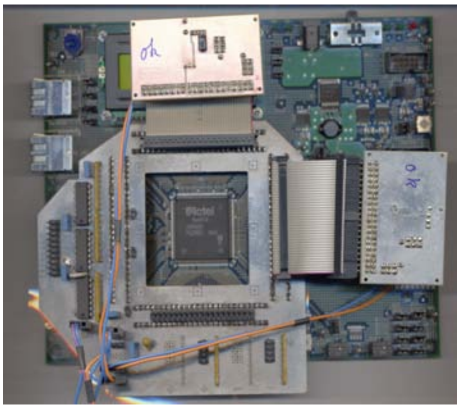

Figure 2: The FPGA platform with its two external SRAM modules.

Based on this analysis, a test platform was devised balancing representativity with the flight model versus development needs. Targeting the radiation-immune RTAX antifuse technology, a flash-based FPGA from Actel was chosen as engineering model for reasons of re-programmability and portability of synthesis (identical development environment). For simplicity, a starter kit was selected with the ProAsic3E A3PE600 die which offers 600 000 gates. Two high performance 1 MByte SRAMs (IS61LV51216 from ISSI), mounted on small secondary boards for flexibility in the pin assignments, were added to allow simple concurrent accesses to external memory without the need for periodic refresh. Although faster than radiation-hardened equivalents, these asynchronous devices will be accessed at a representative rate as part of the final tuning of the system. Finally, external interfaces are implemented with a 2×16-bit-wide IO at 80 Mbit/s with a PC which models the video reception buffer on input and collects the pixel-based results on output with multiple transactions of a handshake protocol per cycle.

To complement the proof of concept established by means of the test platform (including the additional logic for leaving unreachable states), additional simulation on the RTAX die allows for validating the radiation mitigation policy using the built-in register TMR and EDAC-protected internal RAM and by replicating external SRAMs.

4. PROCESSING FLOW

The overall processing flow is illustrated Figure 3. Each of the steps will be described in this section, yet for brevity and to best render the diversity of the design and implementation solutions adopted a variety of viewpoints are adopted at the detriment of exhaustivity. From an overall point of view, with a TDI lasting $0.9828\,ms$ and 983 samples per cycle, processing samples serially would allow spending only a limited $1\,\mu s$ on each. With external SRAM accesses this is hardly achievable, instead the flow in Figure 3 is implemented as a pipeline with most individual stages being pipelines themselves to meet the hard real-time constraint on input and exploit the available latency on output. Conversely, operations not required to run for every sample are conveniently commanded by a scheduler during the TDI corresponding to the read-out of the other SM CCD.

Three different clocks are introduced to this end, one which corresponds to the arrival of a new sample (DCLK), one introduced to enforce external SRAM response delays (SCLK) and the main clock (CLK) which commands the FPGA's synchronous operation and from which the other two are derived by division. The sample period being of order $1\mu s$ and at most 14 SCLK periods being

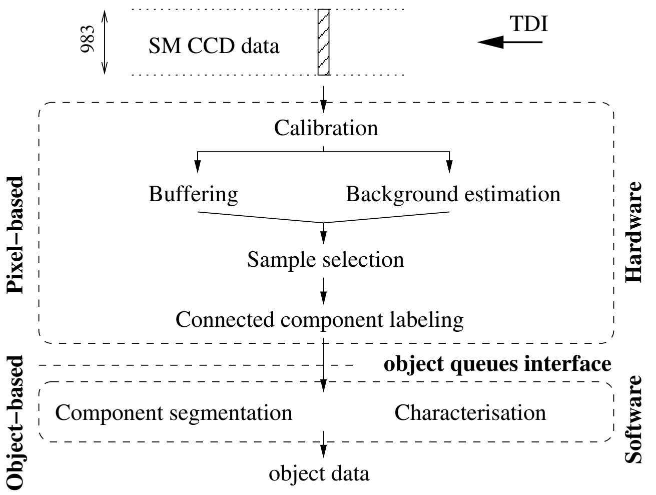

Figure 3: High level processing flow.

required per sample for read accesses to the SRAM (see 4.2), the adopted scheme relies on an 8 ns period for CLK, 32 ns for SCLK and 992 ns for DCLK.

4.1 Calibration

In spite of the high quality specifications for Gaia's leading edge CCDs, the low manufacturing yield and the large number of devices imply that defects will be present not only as a consequence of ageing and radiation but also immediately after procurement. Accordingly, a first calibration stage is introduced to maintain a uniform image model and answer the need to:

- correct cosmetic defects to enforce basic assumptions on imaging properties for normal execution of the algorithms (bright or dead samples),
- normalise sensitivity within and between CCDs to permit the object selection to remain unbiased (pixel response (PRNU) and bias, vignetting),
- compensate or regularise detectors' ageing to ensure graceful degradation of performances in time.

To this end, a linear transform which generalises the classical flat field and bias corrections is applied to each sample based on coefficients uploaded from ground. Sample values are discrete (in ADU) so the correction merely aims at restoring the value which would be produced by an equivalent CCD in an ideal state. To this end, because flight qualified FPGAs do not feature floating point units, the calculation relies on fixed-point arithmetics. Balancing the probability of producing erroneous ADU values with storage constraints, the formula reads

$$s_i^c = s_i^r + ((a_i \times s_i^r + (b_i \ll 16) \pm 2^{17}) \gg 18) \tag{1}$$

where $\ll$ and $\gg$ are bit shift operators, a_i and b_i are the correction coefficients and s_i^r the read-out value at position i. This formula computes the correction, rounds it to the nearest ADU and, at the expense of more arithmetic operators, allows encoding the coefficients with 16 bits (a_i in 1.0.18 format[2] if the PRNU does not exceed ±6.25% and b_i in 1.13.2). The 983 sample positions for the SM1 and SM2 CCDs lead to a total of 7864 bytes of coefficients which are best stored in external memory. The pipelined implementation, illustrated by Figure 4, additionally maps values above saturation to the threshold and performs dead sample replacement, based on the value of b_i, with the average of its neighbours if both are valid or with the valid neighbour's value or with a constant corresponding to the expected background level otherwise (hence the two shift registers for the sample values and b_i respectively).

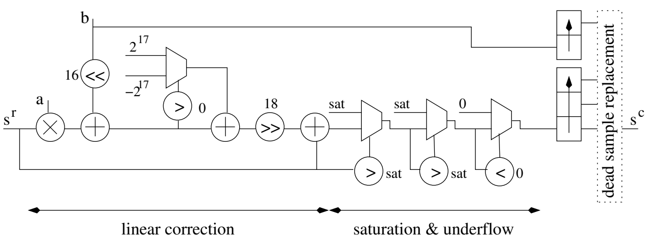

Figure 4: Calibration pipeline.

4.2 Background

Measuring local values of the sky background allows the subsequent rigorous application of quantitative criteria since the measured quantities of interest are then only related to the stars, not their environment. If this is the condition to produce accurate measurements, notably magnitude for object selection, it also impacts on the segmentation for this pollution would otherwise lead to spurious detections in star-forming regions and nebula, or because of zodiacal or diffuse galactic light.

Following M. Irwin, the background is evaluated at a regional scale over 32 × 32 sample regions as a precision-bias trade-off related to noise, pollution by stellar content and the delay introduced before samples are selected. For both accuracy and computational simplicity, the mode is determined and minimum variance is obtained by considering 4 ADU bins for a balance between statistical noise and resolution. Accordingly, 4 binned histograms truncated to the domain in which the background fluctuations are expected are built in parallel (in external memory) and 4 sub-ADU estimates are derived by adjusting a parabolic profile around the mode [6] before being averaged. Additionally, the values for regions swamped by objects are replaced by close valid ones based on a test on the population of the mode's bin through a mechanism resembling the dead pixel replacement. In a maximum density case with uniform background, the resulting measure, although overestimating the real background typically by ~ 0.4 ADUs because of the asymmetry introduced by faint objects, remains precise with a dispersion < 0.2 ADUs.

The values are then bi-linearly interpolated to yield a piece-wise continuous map over the entire CCD (Figure 5) featuring low frequencies with a latency (48 TDIs) considerably smaller than for an equivalent low pass filter. For maximum regularity of the scheduling and to reduce storage this last operation is only carried out sample per sample on demand for sample selection. Having chosen region sizes which are powers of two, this interpolation is merely a sum of four terms weighed by 5-bit integers.

4.3 Sample selection

Based on information theory, a straightforward signal-to-noise ratio (SNR) test is adopted to select relevant samples and achieve the data rate reduction. Signal is at this point corrected from the background contribution (*bkgd*) and noise is considered the sum of two independent processes: photonic and electronic (calibrated on ground and encompasses the read-out chain and analog-to-digital conversion contributions), modelled respectively as Poisson and Gaussian (centred with variance RON^2). With values in electrons, the rationale is

$$\frac{|s_i^c - bkgd|}{\sqrt{s_i^c + RON^2}} >^? SNR \tag{2}$$

but for the benefit of simplicity, formula (2) is rather evaluated as:

$$\underbrace{(s_i^c - bkgd)^2}_{0.14.18} >^? \underbrace{SNR^2}_{0.14.18} \times (\underbrace{s_i^c}_{0.16.0} + \underbrace{RON^2}_{0.14.18}). \tag{3}$$

Although significantly simplified in this squared form by the

[2] This format describes the allocation of bits: sign.integer.decimal.

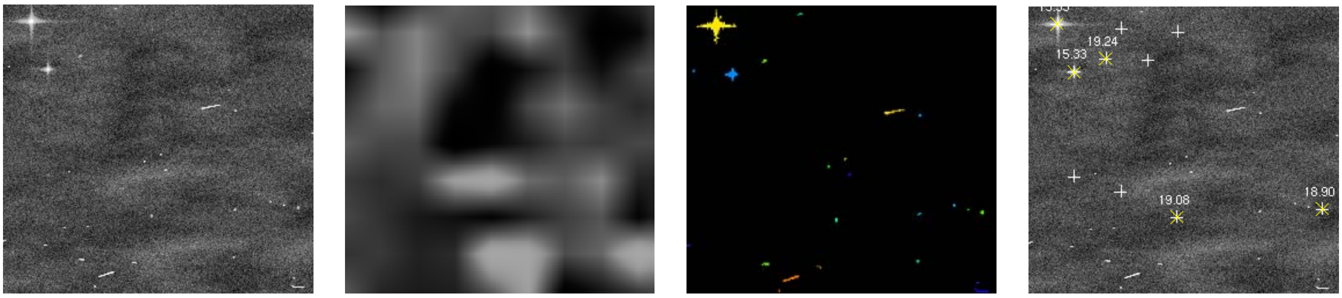

Figure 5: The different phases of the detection in a synthetic high textured background case with faint objects based on Orion (input image, background map (increased contrast), CCs, results (catalogue entries with white "+", detected objects with yellow "x")). All objects of interest are detected while particle events and objects fainter than the limiting magnitude are not.

elimination of the radical and the division, the computation can be further narrowed by recognising that a detailed computation is only necessary close to the threshold value. With an analysis of imaging and encoding parameter ranges, for SNR thresholds in the $[0;6[$ interval, the following procedure results:

1. Compute $s_i^c - bkgd$ in 1.22.9 format,
2. test the sign: if negative discard the sample,
3. else test the integer part: if ≥ 128 ADU then keep the sample,
4. otherwise go through the detailed computation to test versus the threshold.

Limiting the number of bits to 32 in the calculation to save resources and for compatibility with the software implementation, this strategy allows for making the best possible use of the available precision. The total computational error in (3) is bounded by 0.133 ADU^2, leading to only a few incorrectly selected samples in a million.

4.4 Connected-component labelling

If the previous step provides the facility for volume reduction, the retained data remains unstructured. Connected-component (CC) labelling, as a region-growing approach to the formation of independent data packets, allows the transition from the sample space to the object space.

Straightforward methods geodesically propagate labels until all samples of the connectivity equivalence class have been reached, but in our case TDI-synchronous operation introduces addressing constraints which call for label-merge-relabel sequences [7]. In the raster order induced, a CC of complex geometry may well start with two non-neighbouring samples. Since, at this early stage, whether or not they are pathwise-connected cannot be decided, for this path may lie in a yet unread portion of the image, different labels are assigned. As the horizon progresses, this path is unveiled, the parts are merged and the labels declared equivalent.

Various optimisations are possible to allow recycling labels rapidly, but in our case the difficulty lies mainly in the management of labels and associated data structures by the hardware. To this end, a set of look-up-tables are used and stored in internal or external memory depending on size. To achieve fixed complexity, label hierarchies are built by favouring breadth over depth so root labels can be found with a fixed number of look up operations to maintain the tree structure when merging parts of CCs. Additionally, for regularity, discarded samples are also labelled using a special value.

To fully exploit the fact that all object-member samples are traversed a number of low level descriptors are determined as objects are formed. For example, samples are grouped in three categories depending on whether they are saturated, on the edge or interior to the CC and for each elements are counted and the cumulative flux computed.

4.5 Object queues

The objects just formed are inserted in a number of objects queues according to the available magnitude information. Implemented in memory shared between the FPGA and the CPU, these queues serve both as interface, as temporary storage prior to characterisation and for making the latter priority-driven. The software framework indeed pulls objects from the queues according to priority, that is, in decreasing order of magnitude so that if processing resources have been momentarily insufficient brighter objects have had precedence and remaining ones can either be processed if the maximum allowed latency permits it or dismissed with a warning.

4.6 Component segmentation

The data passed on to the software results from a segmentation paradigm designed to rely on a minimum number of assumptions for both simplicity and robustness at the hardware level. It consists of units which contain the necessary information for measurements and hence for deciding how and whether or not to observe the object, yet the CC model does not indicate if the underlying source is a single or a compound one. While this is not a problem for the majority of faint CCs whose magnitude estimate could be global and which fit into an observation window, it is insufficient for brighter ones calling for individual windows to be assigned to each component.

A finer model must be introduced to identify components. Elaborate methods attempting to separate the different components' contributions at the sample level remain too demanding but suggest to simplify the decomposition through an approximation which assigns each sample to the principal contributor. Given the distribution of energy in the diffraction pattern and our interest in the main lobe, this leads to partitioning the CCs in connected domains according to local maxima and based on minimum energy boundaries.

With minor adaptations, the watershed transform, with markers based on local maxima, provides a both elegant and efficient solution in linear complexity. Indeed, because of the additive nature of the optical signal, the presence of a secondary, even faint, must translate into additional energy at a given separation and orientation. Over-segmentation is avoided by relying on thresholds which depend on both: pairs are examined in turn and the relevance of the fainter one is evaluated versus the other based on distance and ratio of values with margins for secondary lobes and noise. For maximum performance, the software implementation follows the same guidelines as the hardware one proposed in [8].

4.7 Object characterisation

The objective of object characterisation is twofold: on one hand, to determine the nature of the sources to discard unwanted ones and artefacts and, on the other hand, to produce the magnitude and location estimates required for their subsequent observation. In practice, for optimal use of the computing resources, descriptors are organised in a flexible cascade of simple tests in increasing order of com-

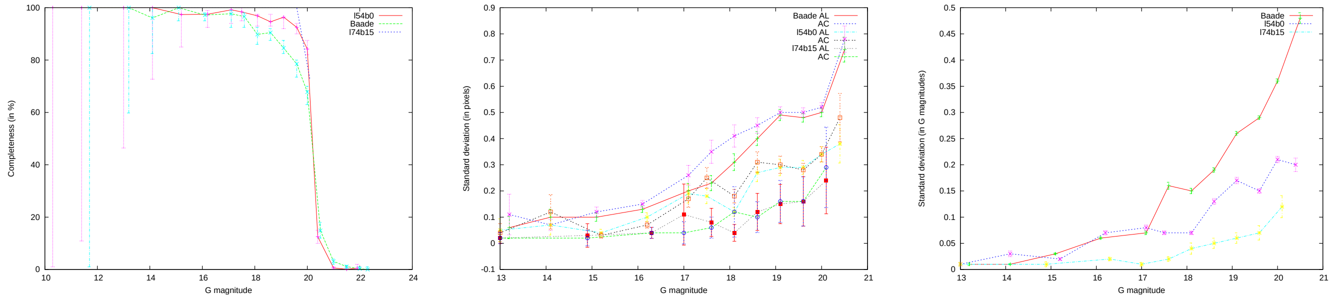

Figure 6: Completeness and standard deviations of location (in the scan direction (AL) and across scan (AC)) and magnitude estimates (error bars are formal at the 3σ level and omitted for completeness in the l74b15 case due to the very low number of objects).

plexity, so the component segmentation above is only attempted for large CCs after their relevance is confirmed.

Three main categories of objects are best rejected at this point to reduce the system's load and save observation resources in AF1: false detections related to noise, stars fainter than the magnitude limit and particle events. The first two are easily discarded, with margin for Poisson noise, by successively imposing a minimum number of member-samples, then a flux cut-off and finally an object-wise SNR criterion.

Particle events are a more difficult problem because the resulting patterns vary with the amount of energy deposited and the angle of incidence. Low energy primary and secondary particle events are the most problematic because of high resemblance with faint stars, yet we purposefully limit ourselves to the easily identifiable ones at this stage, for which the false positive and false negative rates are low, to refrain from introducing detection biases. Running detection in the windows read-out in AF1, a process known as confirmation, allows for discarding the remaining low energy events. Four criteria based on geometry and energy density are applied.

Finally, flux measurements with subtraction of the background contribution and a correction for saturation is output together with location information (barycentre of member-samples).

5. PERFORMANCES

A full software prototype was designed both to allow an early assessment of performances and as a development tool to validate the hardware implementation. It is, accordingly, representative at the bit level since all computations are performed with fixed-point arithmetics for the processing intended for the FPGA. To illustrate the quality of our system, we consider three different density cases:

1. the average density ("l74b15": 25 000 stars/deg^2),
2. the "design density" which is the limit case specified to industry ("l54b0": 600 000 stars/deg^2)
3. and the maximum density on the sky corresponding to the galactic centre ("Baade": 3.10^6 stars/deg^2),

simulated with the tool developed by the Gaia scientific community [9]. Results are presented in Fig. 6 with false detection rates amounting to 1 every 5.10^6, 110 000 and 11 500 samples respectively. Performances are at the level of the very demanding scientific specification to observe all objects of interest during 95% of transits up to the design density. Fig. 6 also illustrates how density is a key limiting factor as it lead to increasingly blended objects which are more difficult to detect, even with the elaborate component segmentation method proposed. Whereas the precision of location estimates for faint stars is of limited importance since observation is constrained by the detectors' pixel grids, the corresponding one for magnitude strongly impacts on the data management on-board. Forming a complete catalogue at magnitude 20 indeed imposes adopting margins versus measurement errors and, with the already large number of objects doubling between 20 and 21, large errors lead to swamping the data set with unwanted objects.

6. CONCLUSION AND PERSPECTIVES

This architecture is also fit for confirming objects at the AF1 level because the pixel-based part can easily be adapted to the windowed observations by generating a fake strip full of zeros between them.

The detection framework we have described achieves high throughput and reliability and answers the need for a unbiased detection to build a statistically representative catalogue of the Galaxy. Its has played a key role in phase A of the Gaia project in establishing achievable performances and making the underlying processing challenge explicit and for making appropriate design choices concerning the payload. Although a different approach has finally been retained by the industrial team in phase B, the full-fledged demonstrator which is in preparation is relevant, not only to validate the underlying R&D, but also to dispose of an alternative model as reference for the upcoming validation phase.

REFERENCES

[1] Gaia Science Advisory Group, *GAIA: Composition, Formation and Evolution of the Galaxy. Concept and Technology Study Report*. ESA-SCI(2000)4, 2000.

[2] P. Armbruster and W. Gasti, "On-board payload data processing systems – On-board networks for future missions," in *Proc. EUSIPCO 2002*, Toulouse, France, September 3-6. 2002, vol. II, pp. 577–580.

[3] F. Arenou, C. Babusiaux, F. Chéreau and S. Mignot, "The Gaia on-board scientific data handling," in *Proc. The three dimensional universe with Gaia (ESA SP-576)*, Paris, France, October 4-7. 2004, pp. 335–342.

[4] EADS Astrium GmbH, "Gaia PDHE Validation Report," *GAIA.ASG.VR.0001*, Sept. 2005.

[5] M. J. Irwin, "Automatic analysis of crowded fields," *MNRAS*, vol. 214, pp. 575–604, Jun. 1985.

[6] F. Patat, "A robust algorithm for sky background computation in CCD images," *Astronomy & Astrophysics*, vol. 401, pp. 797–807, Apr. 2003.

[7] M. Dillencourt, H. Samet and M. Tamminen, "A general approach to connected-component labelling for arbitrary image representations," *Journal of the ACM*, vol. 39, issue 2, pp. 253–280, 1992.

[8] J.C Klein, F. Lemonnier, M. Gauthier, R. Peyrard, "Hardware implementation of the watershed zone algorithm based on a hierarchical queue structure, " in *In Proc. IEEE Workshop on Nonlinear Signal and Image processing*, Halkidiki, Greece, June 20-22. 1995, vol. I, pp. 859–862.

[9] C. Babusiaux, "The Gaia instrument and basic image simulator," in *Proc. The three dimensional universe with Gaia (ESA SP-576)*, Paris, France, October 4-7. 2004, pp. 417–420.