



# Search by triplet: An efficient local track reconstruction algorithm for parallel architectures<sup>☆</sup>

Daniel Hugo Cámpora Pérez<sup>a,b,\*</sup>, Niko Neufeld<sup>c</sup>, Agustín Riscos Núñez<sup>d,e</sup>

<sup>a</sup> Maastricht University, Maastricht, Netherlands

<sup>b</sup> Nikhef National Institute for Subatomic Physics, Amsterdam, Netherlands

<sup>c</sup> CERN, CH-1211 Geneva 23, Switzerland

<sup>d</sup> Smart Computer Systems Research and Engineering Lab (SCORE), Research Institute of Informatics Engineering (I3US), Universidad de Sevilla, Spain

<sup>e</sup> Research Group on Natural Computing, Universidad de Sevilla, Spain

## ARTICLE INFO

### Keywords:

Track reconstruction  
High throughput computing  
Parallel computing  
Heterogeneous architectures  
GPGPU  
SIMD  
SIMT

## ABSTRACT

Millions of particles are collided every second at the LHCb detector placed inside the Large Hadron Collider at CERN. The particles produced as a result of these collisions pass through various detecting devices which will produce a combined raw data rate of up to 40 Tbps by 2021. These data will be fed through a data acquisition system which reconstructs individual particles and filters the collision events in real time. This process will occur in a heterogeneous farm employing exclusively off-the-shelf CPU and GPU hardware, in a two stage process known as High Level Trigger.

The reconstruction of charged particle trajectories in physics detectors, also referred to as track reconstruction or tracking, determines the position, charge and momentum of particles as they pass through detectors. The Vertex Locator subdetector (VELO) is the closest such detector to the beamline, placed outside of the region where the LHCb magnet produces a sizable magnetic field. It is used to reconstruct straight particle trajectories which serve as seeds for reconstruction of other subdetectors and to locate collision vertices. The VELO sub-detector will detect up to  $10^9$  particles every second, which need to be reconstructed in real time in the High Level Trigger.

We present Search by triplet, an efficient track reconstruction algorithm. Our algorithm is designed to run efficiently across parallel architectures. We extend on previous work and explain the algorithm evolution since its inception. We show the scaling of our algorithm under various situations, and analyse its amortized time in terms of complexity for each of its constituent parts and profile its performance. Our algorithm is the current state-of-the-art in VELO track reconstruction on SIMT architectures, and we qualify its improvements over previous results.

## 1. Introduction

The LHCb detector is a large physics detector situated at the Large Hadron Collider at CERN [1]. The detector is being upgraded for the restart of data taking scheduled for 2021 [2]. The full collision data rate of 40 Tbps will be piped through a data acquisition system that will perform a data filtering in real-time, prior to storing data in long-term

storage for posterior analysis. The filtering will occur in two stages: the first stage or *High Level Trigger 1* (HLT1) will reduce the data rate according to particle kinematics by a factor of  $40\times$  in a computing farm composed of 170 servers equipped with GPUs [3]. The second filter stage or *High Level Trigger 2* (HLT2) will perform a full event<sup>1</sup> reconstruction and reduce data by an additional factor  $20\times$  in a computing farm composed of thousands of servers [4]. The introduction of a

<sup>☆</sup> The code (and data) in this article has been certified as Reproducible by Code Ocean: (<https://codeocean.com/>).

More information on the Reproducibility Badge Initiative is available at <https://www.elsevier.com/physical-sciences-and-engineering/computer-science/journals>.

\* Corresponding author at: Maastricht University, Maastricht, Netherlands.

E-mail address: [dcampora@cern.ch](mailto:dcampora@cern.ch) (D.H. Cámpora Pérez).

<sup>1</sup> An event corresponds to a single crossing of the Large Hadron Collider proton beams.

heterogeneous computing infrastructure in LHCb is motivating the creation of parallel algorithms that are portable and efficient across architectures.

Track reconstruction or *tracking* is a pattern recognition problem consisting in finding particle trajectories from measurements (*hits*) in detectors along their path. The problem is equivalent to finding a partition of disjoint sets of measurements that are compatible with the laws of motion of particles as they traverse a detector, accounting for the fact that some measurements may be noise, and considering the presence of sizable magnetic fields which curves the trajectories of charged particles depending on their momentum. Track reconstruction yields momentum and trajectory information of reconstructed particles, which play an essential role in trigger systems of physics experiments. Fig. 1 exemplifies the track reconstruction problem.

The Vertex Locator (VELO) is a tracking detector of LHCb consisting of 52 planes of silicon pixel chips surrounding the LHC interaction point and beamline, shown in Fig. 2. As particles pass through the detection planes, they leave detectable measurements in the form of pixel clusters. VELO track reconstruction constitutes the first reconstructed sub-detector of LHCb, and tracks found in the VELO are used to locate the originating collision vertices as well as serve as seeds for subsequent track reconstruction. Therefore reconstructing the VELO is of vital importance towards the correct functioning of LHCb.

Track reconstruction validation is typically performed with Monte Carlo (MC) simulated samples, where reconstructed tracks should match MC particles, which establish the *ground truth*. The matching of tracks with particles is done on a hit by hit basis. The physics performance of found tracks can be evaluated according to five indicators [5]:

- The *track reconstruction efficiency* can be determined by the ratio between the reconstructed tracks of reconstructible particles, over all the reconstructible particles:<sup>2</sup>

$$\frac{N_{\text{reconstructed and reconstructible}}}{N_{\text{reconstructible}}} \quad (1)$$

- A *fake track (ghost track)* is created when a percentage of hits in a track are not from a real track. In LHCb, at least 70% of the hits in a track must belong to the same MC particle to be associated in the validation process. The *fake track fraction* is the ratio between the fake tracks and all the reconstructed tracks:

$$\frac{N_{\text{fake tracks}}}{N_{\text{reconstructed tracks}}} \quad (2)$$

- The *clone track fraction* refers to the fraction of tracks associated to the same MC particle as another reconstructed track:

$$\frac{N_{\text{clone tracks}}}{N_{\text{reconstructed tracks}}} \quad (3)$$

- The *hit purity* in a track refers to the fraction of track hits that belong to the same MC particle:

$$\frac{N_{\text{track hits in MC particle hits}}}{N_{\text{track hits}}} \quad (4)$$

- Finally, the *hit efficiency* yields the number of hits correctly found out of the MC particle hits in a track:

$$\frac{N_{\text{track hits in MC particle hits}}}{N_{\text{MC particle hits}}} \quad (5)$$

<sup>2</sup> A particle is said to be *reconstructible* in a tracking detector when it fires enough measurements over a detector-specific threshold.

The VELO detector is outside of the range of effect of the LHCb magnet, and therefore trajectories can be considered to be straight lines. Once the detector restarts operation a sustained throughput of  $10^9$  particle trajectories at the VELO per second will have to be reconstructed in the trigger systems, while delivering good track reconstruction performance indicators. The VELO reconstruction is therefore a real-time software challenge whereby the design performance of the system must be met within the hardware constraints of the data acquisition system.

## 2. Tracking techniques

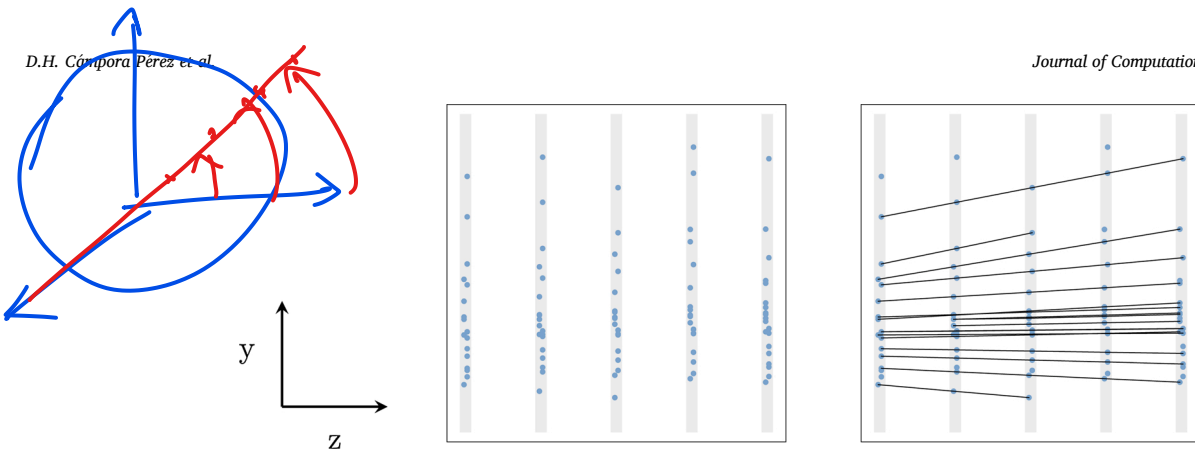
Due to the interest in tracking by many particle physics experiments, there is a rich literature on track reconstruction techniques [6]. Local tracking methods find tracks iteratively, whereas global methods adapt an equivalent formulation of the problem, typically including all measurements, where solutions map to tracks.

The most common local tracking method consists in finding a *track seed* and extending it to other detector planes in a process known as *track following*. The track seed is usually formed by a segment of two or three hits, and the search starts in a region where the hit density is lower and thus the signal is easier to distinguish, which usually corresponds with the furthest distance to the expected interaction point. Track seeds are extrapolated (*followed*) to detector regions closer to the interaction point by applying an extrapolation accounting for the presence of a field if necessary. A model of the track can be formed from the track hits, and this model can be employed to select among a list of candidate hits the best fitting one. This extrapolation process may account for missing hits in detector parts, according to the hit inefficiencies of the physical detector and *dead* regions without sensitive detectors. Once a track is fully built, its constituent hits can be flagged so they are not revisited in further seed or following steps.

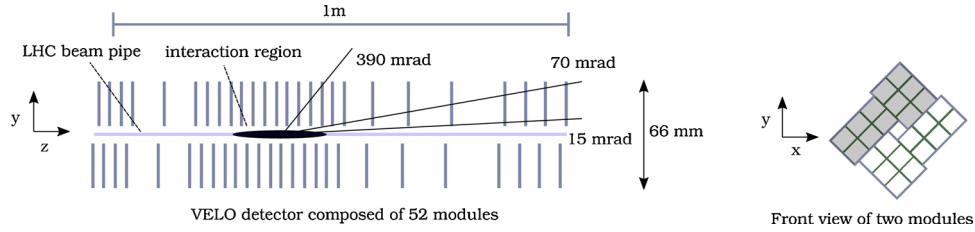
Historically a local track following method has been employed for the VELO reconstruction of the LHCb experiment runs, named *Search by pair* [7]. It constructs seeds of pairs of hits initially in the furthest modules from the interaction region. Track seeds are followed to neighbouring modules, allowing for one hit to be missing in consecutive modules on any one side of the VELO subdetector. Tracks of four or more hits flag all of their hits, reducing the search time of further seeding and following steps. Tracks consisting of at least three hits are stored, in accordance to the reconstructibility condition of the VELO subdetector. However, the requirement to flag hits as used makes this technique not suitable for parallelization without modification. It is possible to find all triplet combinations in parallel dropping the flagging mechanism, which was done in the seeding phase of [8]. However this parallelism comes at the cost of generating all possible triplet combinations, which is inefficient for densely populated detectors.

Other local tracking methods are *track roads* and *track elements*. *Track roads* forms candidates with two hits situated in the extremes of the detector, and creates a path or *road* between both hits by interpolation. In case the model of the track be curved, a third hit should be added. The width of the road determines the accepted error in the model, and it depends on the characteristics of the detector [9]. *Track elements* performs reconstruction in two phases: (1) seeds are made up from neighbouring points, straight lines or parabolic lines. Each seed is converted into a *master point* (a weighted average of the points) and a direction. (2) The seeds, instead of the original hits, are used to perform tracking. This method reduces the number of hits to consider in the tracking phase, at the cost of a loss in precision [10]. Local methods may be used in conjunction with corrections such as the Kalman filter [11], fitting forming tracks and selecting hits [12].

Hit multiplicity is often a concern in real-time reconstruction environments, where track reconstruction must be performed at a high throughput in order to keep up with the collision rate. Spatial reductions can be employed to reduce the search time of hits under consideration for local tracking methods. This involves a data preparation step prior to



**Fig. 1.** (Left) A minimal track reconstruction instance projected in 2D, consisting in a set of hits with position information. (Right) The actual particle trajectories sought when doing track reconstruction.



**Fig. 2.** A schematic of the upgraded Velo detector. (Left) Section in the XZ plane, with detector modules laying in two sides. (Right) Front view of each module in the XY plane, with subdivisions indicating detector chips. Each detector chip has a resolution of  $256 \times 256$  pixels.

the application of the tracking method. Common data structures used alongside local methods include KD-trees [13], binary search structures or more generically search windows. The specifics of the geometry of the detector yield in some cases a natural subdivision of the problem [14].

Alternatively, global methods can be applied to the track reconstruction problem. The *Hough transform* method [15] in its simplest form converts all hit points into a histogram representation in polar coordinates, where peaks are equivalent to compatible hits. Despite of the elegance of the Hough transform underlying principle, difficulties arise when dealing with high hit multiplicities, where binning and threshold of the histogram play an important role in avoiding excess of clones or fake tracks. Circular trajectories must be converted to lines prior to applying the Hough transform, which can be achieved with a *conformal mapping* transformation [16].

The *clustering method* consists in extrapolating hits onto a parameter space, according to the expected trajectory from a collision vertex. Once in the parameter space, hits pertaining to the same track appear close to each other. The tracking problem consists then in a clustering problem, that can be solved with any clustering method [17]. The collision vertex must be chosen with sufficient precision, and the resulting cluster must be verified to match reasonable track parameters [18].

The *automata method* [19] is a graph traversal method of a weighted directed graph representing the measurements. Each measurement is a vertex, and directed edges between measurements in neighbouring detector elements are created according to the expected track parameters in the region of acceptance of the detector, in the direction from outer to inner-most layer. The graph is then traversed following a Depth-First Search, assigning weights to visited edges according to the current depth level. If vertices are visited multiple times, the highest weight prevails. The automata method has been successfully implemented in various trigger systems [8,20].

Both the algorithm and the target hardware architecture must be considered when devising an efficient track reconstruction solution. It has been discussed that global methods are amenable by design for parallel architectures [6,18]. Some of the high-throughput tracking algorithms used in current trigger systems are global methods [20,21], and have been implemented on modern CPU and GPU architectures. On

the other hand, local methods have also been used for high-throughput solutions [22].

In our previous work [23], we presented *Search by triplet*, a novel local method for track reconstruction on parallel architectures, alongside a base framework to run it efficiently on GPU architectures. Our algorithm considers module pairs to increase the amount of parallel work in every step. We sort the data by  $\varphi$  on every module pair stemming from the observation that hits of tracks have similar  $\varphi$  values. Our algorithm then iteratively generates track seeds of triplets of hits, and extrapolates forming tracks to the next module pair. Fake and clone tracks are avoided by introducing a flagging mechanism with barriers. In a last step, track seeds that were never extended are checked again prior to storing them as tracks. We implemented our work both in the SIMT<sup>3</sup> programming paradigm targeting GPUs, as well as in SPMD targeting CPUs. A variation of this algorithm with SIMD-specific optimizations was done in [24].

In this paper, we build on the foundation of our previous algorithm and optimize it for SIMT and SIMD architectures. We improve the maintainability of our algorithm by providing a single codebase for any heterogeneous architecture in the Allen framework [3]. We also provide architecture-specific optimizations for routines that are employed often and that require hand-tuning. Through an iterative process of optimization, we increase the throughput of our algorithm by more than a factor  $3 \times$ , while also improving its physics efficiency.

### 3. Search by triplet

Search by triplet is a local track following algorithm optimized to reconstruct the LHCb VELO detector that exploits the task parallelism

<sup>3</sup> *SIMD* (Single Instruction Multiple Data) is a class of parallel processors in which each instruction operates over several data simultaneously in lockstep. *SPMD* (Single Program Multiple Data) is a class of processors in which the work is distributed in tasks that operate over multiple data, not necessarily in lockstep. *SIMT* (Single Instruction Multiple Thread) is an execution model employed by GPUs where every instruction is executed by a group of threads.

inherent to the LHCb data taking regime and the data parallelism of track reconstruction. 30 million events are detected per second in LHCb, where each is independent of each other. Therefore, we assign different tasks to process each individual event. Within each event processing, track reconstruction exposes various levels of parallelism that we tackle in a data parallel fashion, either employing the SIMT paradigm or SIMD optimizations where relevant.

Search by triplet is now a single coherent codebase written in C++, with custom extensions that provide SIMT functionality. Depending on the target device specified at compile time, our codebase can potentially target any parallel architecture, notably modern CPUs and GPUs. Table 1 depicts the levels of parallelism exploited in Allen algorithms such as Search by triplet, and the correspondence with each target.

Our algorithm consists in three sub-algorithms that are described in the following. For simplicity, we will refer to modules instead of module pairs. When discussing computational complexity, the notation employed is generalized to a detector with  $m$  consecutive detector modules and an average number of hits in each module of  $n$ .

### 3.1. Sort by phi

The VELO modules are positioned such that particle collisions tend to occur close to the origin of coordinates in the XY plane. Hence, particles produced in these collisions that travel in a straight line are likely to have a constant phase in polar coordinates when projected on an XY plane. Hits in each module are sorted by  $\varphi$ , calculated as the 2-argument arctangent in the XY plane with respect to the origin of coordinates.

The range of values of the 2-argument arctangent is  $[-\pi, +\pi]$ . However, values close to  $-\pi$  are conceptually very close to values close to  $+\pi$ , as they represent angles that are close. For this reason, we convert the range into a unsigned 16-bit integer type, mapping the 2-argument arctangent range to the unsigned 16-bit range  $[0, 2^{16} - 1]$ . This transformation enables us to use modulo arithmetic to perform comparisons, and it allows us to reduce memory pressure by using two bytes as opposed to four to store each  $\varphi$ .

Given that the number of maximum hits in a module is known, a constant amount of *shared memory*<sup>4</sup> is employed as a means of low-latency temporary data buffer when  $\varphi$  is calculated and sorted. The permutation produced is then employed to sort hit coordinates, yielding a structure of arrays sorted by phi for each module. A parallel insertion sort method has been implemented to calculate the permutation. The worst-case complexity of this algorithm is  $O(m \cdot n^2)$ .

An SIMD-specific optimization is also provided. The  $\varphi$  of every hit is calculated using a vectorization library. The permutation is calculated using the quicksort implementation provided in the STL library. The STL sort implementation has a worst case of  $O(n \cdot \log(n))$ , and therefore the SIMD specialization has a worst-case complexity of  $O(m \cdot n \cdot \log(n))$ .

**Table 1**

Degrees of parallelism for event reconstruction in Allen. CUDA terminology is used [25] for a generic GPU target, not necessarily NVIDIA.

	CPU target	GPU target	Type of parallelism
Bunches of events	Threads	Streams	Task parallelism
Inter-event	Sequential	Blocks	Task parallelism
Intra-event	Sequential or vectorization	Threads	Data parallelism

<sup>4</sup> Shared memory is a GPU-specific optimization that uses the configurable L1-cache shared memory available in GPUs. For other targets, main memory is employed.

### 3.2. Track seeding and track following

The local track reconstruction technique implemented in Search by triplet consists in an iterative application of two stages: track seeding and track following. Fig. 3 shows an overview of the technique with the iterative control flow on the top, and the data containers involved in either data dependencies or output at the bottom. Even though the iteration follows an ascending order, the actual direction of traversal of the modules is configurable.

Track seeding finds triplets of hits in consecutive modules, one hit in each module. In order to avoid clone tracks, a flagging mechanism has been implemented, akin to the one found in Search by pair [7]. None of the hits in the triplet must be flagged. This introduces the requirement of a barrier between each seeding and following stage due to the Read-After-Write and Write-After-Read dependencies: flagged hits are read during track seeding, and they are written during track following. The barrier is only present in SIMT code, and if an SIMD target is specified, it is removed at compilation time. The *flagged hits* container is initialized to be empty prior to the execution of the first seeding iteration.

The process of finding triplet seeds has received several optimizations. We consider three consecutive modules, and process all non-flagged hits in the middle module separately. For each hit, we seek the corresponding  $\varphi$  position in the neighbouring module with a binary search. Then, we perform a *pendulum search*, alternatively looking further down and up the hits in the module, until a threshold number of candidate hits that are not flagged is found. Fig. 4 depicts this process.

Once all candidate doublets are found, each doublet is extrapolated to the third module. A binary search is performed with the extrapolation  $\varphi$ , and all hits within a tolerance window are considered. Fig. 5a shows the process of extrapolation for hit  $c_0$ . Each triplet candidate is evaluated with a *scatter* function between the extrapolated doublet position and the third hit under consideration:  $dx^2 + dy^2$ . For each middle module hit, at most one triplet is kept: the triplet minimizing the scatter function under a threshold.

Triplets are stored in the *seeds* container and passed on to the following stage, where all forming tracks (in containers *seeds* and *tracks*) are considered. Track following consists in extrapolating forming tracks to the next module and attempting to find compatible hits under a threshold. Similarly to the doublet extrapolation of seeding, the last two hits of each forming track are extrapolated to the next module, and the best hit is evaluated with the scatter function. In order to account for the hit inefficiency of modules, it is allowed for tracks to miss a module, in which case they are kept for the next following iteration.

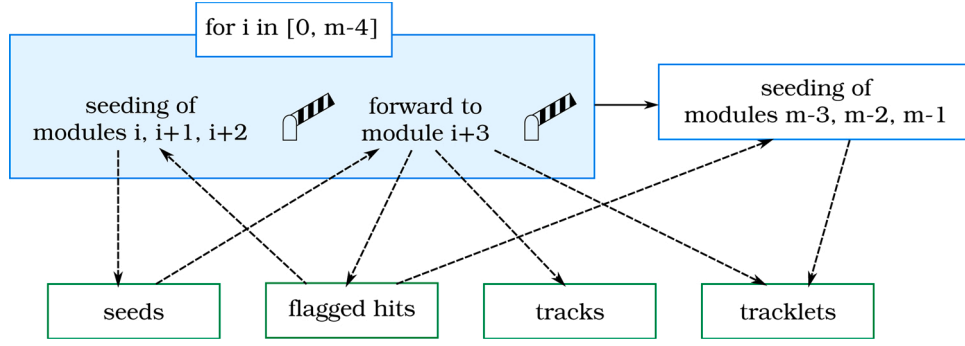
The best hits found for the followed tracks are flagged, as shown in Fig. 5. When a forming track misses two consecutive modules, or when processing the last iteration, forming tracks are either stored in the *tracklets* or *tracks* container, depending on whether the number of hits is three or more respectively.


The data parallelism of this process is exploited by separate threads in the SIMT model, whereby each thread processes a different starting hit in the seeding stage, or a different forming track in the following stage. An SIMD-specific optimization has been done for the seeding stage, whereby all doublets of each starting hit are processed in parallel with vector instructions.

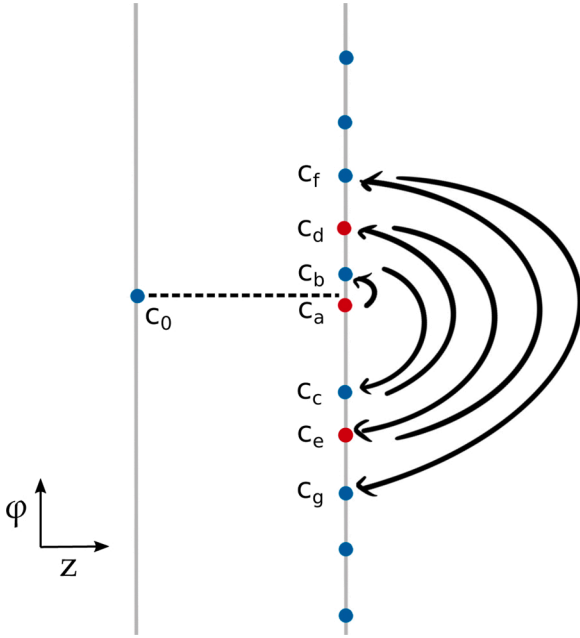
The worst-case complexities of the two stages are as follows. Seeding is performed on all consecutive module triplets (bound by  $m$ ), where all candidates in a module are considered ( $n$ ). For each candidate, a binary search is performed to a neighbouring module ( $\log(n)$ ), and for each doublet a binary search to the third module is performed ( $\log(n)$ ). Finally, the best hit within the threshold ( $n$ ) is found. Therefore, its worst-case complexity is  $O(m \cdot n^2 \cdot \log(n)^2)$ .

Track following occurs on all modules considered for extrapolation ( $m$ ), where all forming tracks are considered ( $n$ ). Similarly to seeding, the last doublet in each track is extrapolated and the best hit within the





**Fig. 3.** (Top) Control and (bottom) data flow of the track seeding and track following steps. Only data containers relevant to the control flow are shown. SIMT barriers are represented with the icon .



**Fig. 4.** Pendulum search of the first four non-flagged hits closest to the  $\phi$  of hit  $c_0$ . Flagged hits are marked in red ( $c_a$ ,  $c_d$  and  $c_e$ ).

threshold is kept. Therefore, its worst-case complexity is  $O(m \cdot n^2 \cdot \log(n))$ .

### 3.3. Tracklet filter

The *Tracklet filter* operates on the *tracklet* container, which contains three-hit tracks for which no compatible hit was found in any track following step. It performs a least means square fit over the three hit coordinates of each track, and requires the fit be under a configurable threshold, and the three hits be not flagged, prior to accepting the tracklets as valid tracks. The worst-case complexity of the tracklet filter is  $O(m \cdot n)$ .

## 4. Results

The algorithms composing Search by triplet are not run in a separate application, but rather they are embedded as part of the VELO reconstruction sequence. An in-depth discussion of the additional algorithms involved in the sequence is out of the scope of this paper. Nevertheless, to provide an overall perspective of our work we present the VELO reconstruction sequence, which is composed of the following sequence of algorithms:

- Global event cut – Rejects the 10% most densely populated events.
- Data transmission – Copies VELO raw data from the host to the device.
- VELO decoding – Decodes the VELO raw data into hits.
- VELO tracking – Our Search by triplet implementation.
- Data consolidation – Packs resulting track data to improve memory locality of subsequent reconstruction algorithms.

The VELO reconstruction sequence is written in the Allen framework, which can be compiled for a variety of heterogeneous target devices. The resulting application can be run with a configurable number of events per bunch  $n$  and threads  $t$ . Each thread instantiates an SIMT stream,<sup>5</sup> and the thread-stream pair executes the requested sequence over the number of events. We run the presented experiments thousands of times to ensure reproducibility and mitigate processor warm-up effects.

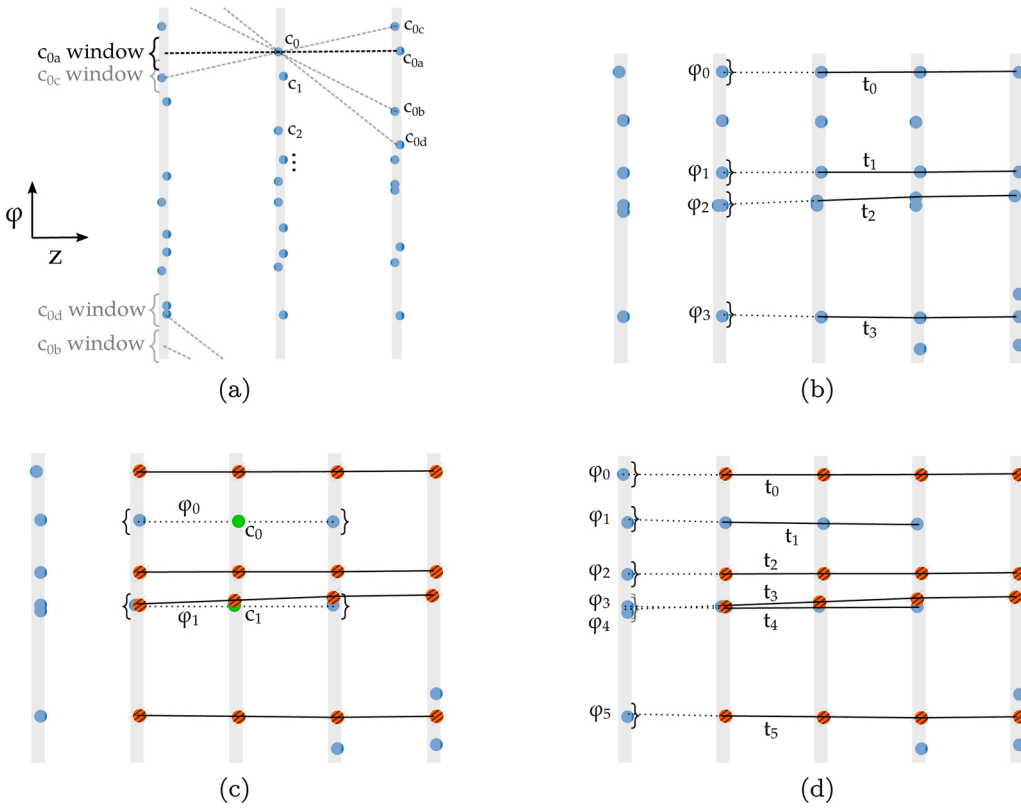
We have observed the performance evolution of the VELO reconstruction sequence over time. We have focused on the GPU performance, which will perform the first stage of HLT in LHCb. Fig. 6 shows the evolution of the performance with each of the optimizations we have applied. The reported throughput numbers were obtained with the configuration  $n = 1000$  and  $t = 20$  in a GeForce RTX 2080 Ti GPU. In all cases, the applications were compiled with the NVIDIA compiler `nvcc`. We report the throughput of Search by triplet since its modern inception in February 2018, and we align the speedup to the July 2018 version used for our previous publication [23]. For each optimization, the relevant branch of the code, a brief explanation of the optimization and its relative speedup are presented.

From July 2018 to July 2020, the performance of the entire VELO reconstruction sequence has improved by more than a factor  $3 \times$ . Some optimizations have been discussed throughout Section 3. One notable speedup has resulted from using array-of-structures (AOS) and moving to 16-bit floating point precision (`fp-16`) for storing and accessing VELO hit coordinates (05-03-2020,  $1.25 \times$ ). AOS improved memory locality and coalescing since the three hit coordinates were always requested as a group. `fp-16` provides enough precision for the VELO coordinate range while decreasing memory pressure.

Three relevant sources of speedup did not involve code transformations. We first added the `nvcc` option `maxregcount` to statically set the maximum number of registers of our kernels to 64, which improved performance (23-04-2018). However, the code evolved and we later disabled it, observing a speedup of  $1.26 \times$  (17-05-2019). We think this is likely due to the separation of kernels on 09-07-2018, which allowed the compiler to set a better value to improve execution on each of them separately the second time.

Another such speedup resulted from adopting *no separable*

<sup>5</sup> When targeting a CPU architecture, the concept of stream is irrelevant and all algorithm invocations are run by the thread.



**Fig. 5.** Iterative seeding and following stages, where modules are considered from right to left. (a) Seeding stage. For hit  $c_0$ , four hits  $c_{0a}$ ,  $c_{0b}$ ,  $c_{0c}$  and  $c_{0d}$  are considered on the neighbouring module on the right. Each of the resulting *doubles* is extrapolated onto the neighbouring module on the left, where hits in the  $\varphi$  window are considered. The  $\varphi$  search wraps around. (b) Following stage. Forming tracks are extrapolated and hits are sought in a  $\varphi$  window. (c) and (d) Subsequent seeding and following stages. Hits found in previous follow stages are marked as flagged and not further considered.

compilation (23-03-2020,  $1.21 \times$ ). Separable compilation allows compilation units to use functions that are defined but not declared within their scope. Alongside obtaining a substantial speedup, our code became compatible with AMD GPUs as well through use of the HIP compiler. Currently it would be possible to improve the performance of separable compilation codebases with Link Time Optimization (LTO), but the release of `nvcc` at the time did not have this feature. Finally, setting as many active connections as streams (10-07-2020) improved the performance when setting configurations with more than eight active threads.

We have closely tracked the physics efficiency of our algorithm throughout its development cycle by means of continuous integration tests. The physics efficiency of our algorithm has either remained constant or improved with every iteration. In particular, the only cut introduced in our sequence (27-04-2019,  $1.07 \times$ ) is a configurable cut that removes events with high detector occupancy and consequently low signal-to-background ratio, which is common practice in real-time, high-throughput, resource-limited physics reconstruction environments.

Table 2 shows the physics performance indicators for a variety of particle categories, for both the reference implementation and the latest version of Search by triplet. The different particle categories refer to different decay processes. For the LHCb experiment of special interest are decays of Beauty particles indicated by *From B* and also decays from so-called strange particles or into electrons. The physics performance indicators were introduced in Section 1.

Reconstruction efficiency has substantially improved across all categories. The tighter scatter criterium has reduced fake fraction at the cost of increasing clone fraction. The decay products of strange particles are especially challenging to reconstruct as they leave fewer hits in the VELO detector and do not necessarily point to the interaction region, as they are produced as secondary decay products from relatively long-lived particles. We have improved the reconstruction efficiency of decay products of strange particles by more than 10%. We have also improved the efficiency for electron tracks by around 20%, while increasing their hit efficiency and purity as well.

The entire sequence has been run over a variety of architectures, listed in Table 3. Our tests have run over four CPUs from different vendors and five GPUs. The throughput of the VELO reconstruction sequence across architectures is shown in Fig. 7. GPU architectures are depicted on the top, whereas CPU architectures are depicted on the bottom.

The tests are run with a configuration of  $n = 500$  and  $t = 16$  for all GPUs. For CPUs, the configuration is  $n = 200$  and  $t$  has been set to the number of simultaneous multi-threads of each processor. We observe best performance on the Quadro RTX 6000 card. If we take into account the price from Table 3, the best price performance is found with the GeForce RTX 2080 Ti, which is a consumer-grade product. If we consider the rate per TDP, best performing device is the Tesla T4 due to its low TDP of 75 W.

The performance of our previous ISPC<sup>6</sup> implementation is also presented for reference. Our current implementation is 10% slower than our previous ISPC implementation. However, the ISPC implementation was a separate codebase that would introduce maintenance overhead, especially considering the lack of C++ support of ISPC. In addition, ISPC only support x86-based architectures and ARM32. In contrast, we currently employ standard C++ and manual vectorization through the vectorization library UMESIMD [27], which supports x86 SSE, AVX and AVX512 extensions, AltiVec, ARM64 and a compatibility scalar backend. In the future, we intend to transition to the standard library simd [28], currently under development.

Search by triplet constitutes 70% of the sequence time on CPUs, but only 47% of the GPU sequence time. This difference likely has to do with differences in the execution model, as the branching ratio affects CPUs and GPUs differently. We observe a difference in performance of an order of magnitude between the CPUs and GPUs under consideration.

<sup>6</sup> The Intel Implicit SPMD Program Compiler [26] is an SPMD compiler that exposes vector units in SIMD processors as programmable processors. ISPC programs are written in a custom extension to the C language.

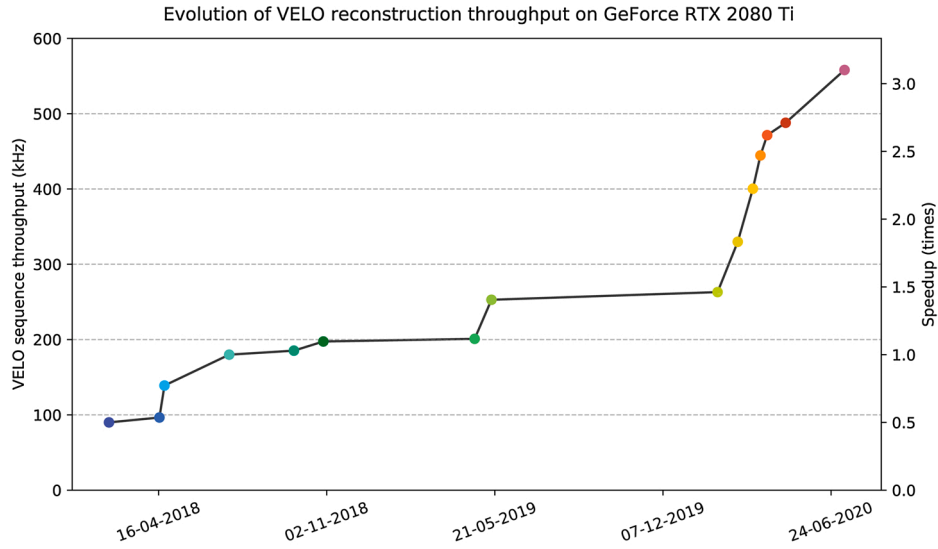


Fig. 6. Evolution of VELO throughput versus time on a GeForce RTX 2080 Ti card.

Table 2

Efficiency numbers of VELO reconstruction sequence for reference implementation (left) and latest version (right). For each implementation, the reconstruction efficiency, clone fraction, hit purity and hit efficiency are shown as percentages for a variety of particle categories. The last row shows the overall fake fraction of the implementation as a percentage over the total number of reconstructed tracks.

Particle category	Reference (09-07-2018)				Latest version (10-07-2020)			
	Reco. eff.	Clone fraction	Hit purity	Hit eff.	Reco. eff.	Clone fraction	Hit purity	Hit eff.
All	95.36	0.60	99.07	97.58	98.52	2.14	99.30	96.45
Strange	87.45	0.60	98.04	97.65	98.13	1.58	99.48	97.35
From B	96.88	0.51	98.96	97.85	99.30	1.16	99.74	98.11
Electrons	70.25	0.72	90.93	90.64	97.38	2.74	98.18	97.02
From B electrons	79.00	1.13	94.00	93.52	97.00	3.68	98.42	96.68
Overall fake fraction		3.25				0.86		

The specifications of the Radeon Instinct MI50 match those of the Tesla V100 in terms of TFLOPS, however we observe a performance gap of  $2.13\times$  in favour of the NVIDIA card. We have used the HIP compiler to generate the application for AMD GPUs, which is currently in active development and we have observed performance improvements with the recent transition from `hipcc` to `hip-clang`. We will continue to track the performance in AMD devices with the evolution of the compiler.

The tested generation of NVIDIA cards connect through the standard PCIe3  $16\times$ , which has a peak payload bandwidth of 15.76 GBps FDx. Our current application uses 8 GBps host to device data transfer, so theoretically we are currently using 51% of the available bandwidth in PCIe3  $16\times$ . We obtain a throughput of 592 kHz on the Quadro RTX 6000, and the peak throughput attainable with only data transmission is 940 kHz, therefore empirically we are using 63% of the capacity of the link. While in the last two years we have improved the throughput of our

**Table 3**

Hardware used for our tests.

Feature	Type	# cores	Freq. (GHz)	Peak TFLOPS	MSRP (\$)	TDP (w)
Intel Xeon Broadwell E5-2630	CPU	10 (20)	3.1	0.352	667	85
IBM Power9 IC922 EK01	CPU	16 (64)	4.0	0.512	3150	225
Cavium ThunderX2 CN9980	CPU	32 (128)	2.5	0.56	1795	180
AMD EPYC 7502	CPU	32 (64)	3.35	1.3	3618	180
AMD Radeon Instinct MI50	GPU	3840	1.746	13.41	8999	300
NVIDIA GeForce RTX 2080 Ti	GPU	4352	1.545	13.34	1199	250
NVIDIA Quadro RTX 6000	GPU	4608	1.77	16.31	4000	250
NVIDIA Tesla T4	GPU	2560	1.59	8.141	2295	70
NVIDIA Tesla V100 32GB	GPU	5120	1.37	14.13	8999	250

algorithm by more than  $3 \times$ , it will not be possible to improve it further by more than  $1.6 \times$  in the future under these conditions. The next generation of NVIDIA cards will support PCIe4, which should enable us to overcome this potential issue.

Fig. 8 shows the scalability of our application against the peak 32-bit FLOPS of the processors under consideration and against the number of threads. The top shows cross-architecture scalability with respect to TFLOPS of each system. Our software scales with the TFLOPS of the processors under consideration. The AMD GPU appears as an outlier which indicates that more performance can be attained from further optimizing the code for that device.

The scalability of CPU processors is depicted in the bottom left. The run configurations were locked to  $n = 200$ , with a varying number of threads  $t$  run in a single processor. For each CPU, the best performance was achieved by running a configuration with as many threads as simultaneous multi threads (SMTs). The AMD EPYC and Intel processors show a steady performance increase with every additional core, and their scalability changes when they start using its SMTs. Both the Power and ARM processors exhibit four local peaks, which correspond with their 4-way SMT configuration.

Finally, the scalability of GPU processor is shown in the bottom right.

For the GPU tests, each configuration was run with  $n = 500$  and  $t$  in the range  $[1, 20]$ . The Tesla T4 saturates with as few as 5 streams. The AMD card performance fluctuates and saturates with 14 streams. Both the GeForce and the Quadro cards scale very similarly, which could be due to both sharing the NVIDIA Turing architecture. This is further supported by the different scaling behaviour of the Tesla V100, of the previous Volta architecture. All GPUs saturate with a configuration of at least 14 streams. Our software scales to both CPU and GPU architectures.

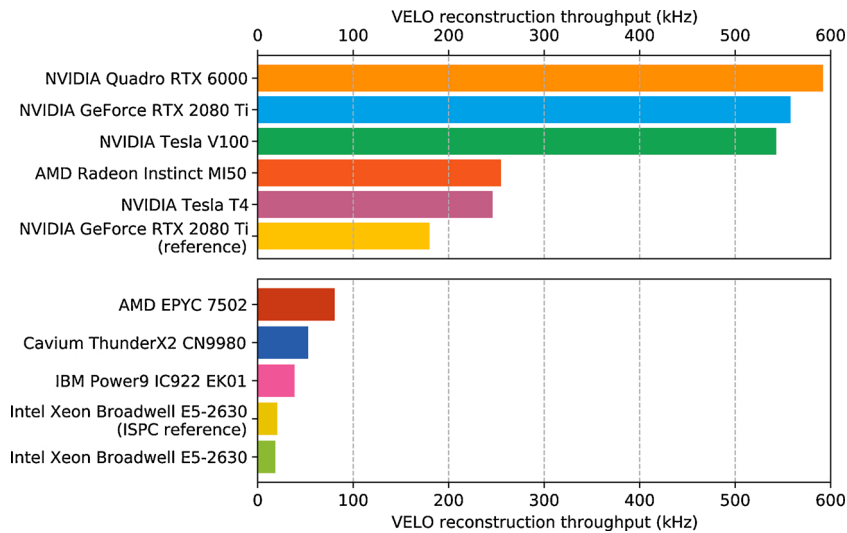
## 5. Conclusions

We have presented Search by triplet, a fast algorithm for VELO reconstruction on parallel architectures. Our algorithm exploits various degrees of parallelism in LHCb VELO track reconstruction, and makes an efficient use of resources in heterogeneous architectures. The algorithm is written in a single codebase in C++, and we have developed architecture-specific optimizations for hot sections of the code.

Our algorithm employs a local tracking technique to detect particle trajectories. The algorithm first sorts hits in each module by  $\varphi$ , generating an efficient 1D search structure. The local tracking method then consists in an iterative process with track seeding and track following steps. In track seeding, triplet seeds are sought as compatible non-flagged hits minimizing a fit function. Forming tracks are extended in track following, which flags hits of extended tracks. The flagging mechanism introduces a data dependency between seeding and following stages expressed as a barrier in SIMT architectures. Finally, remaining three-hit tracklets are filtered in a post-processing step. We expect our design to be generalizable to track reconstruction of other detectors, but this is subject to further research.

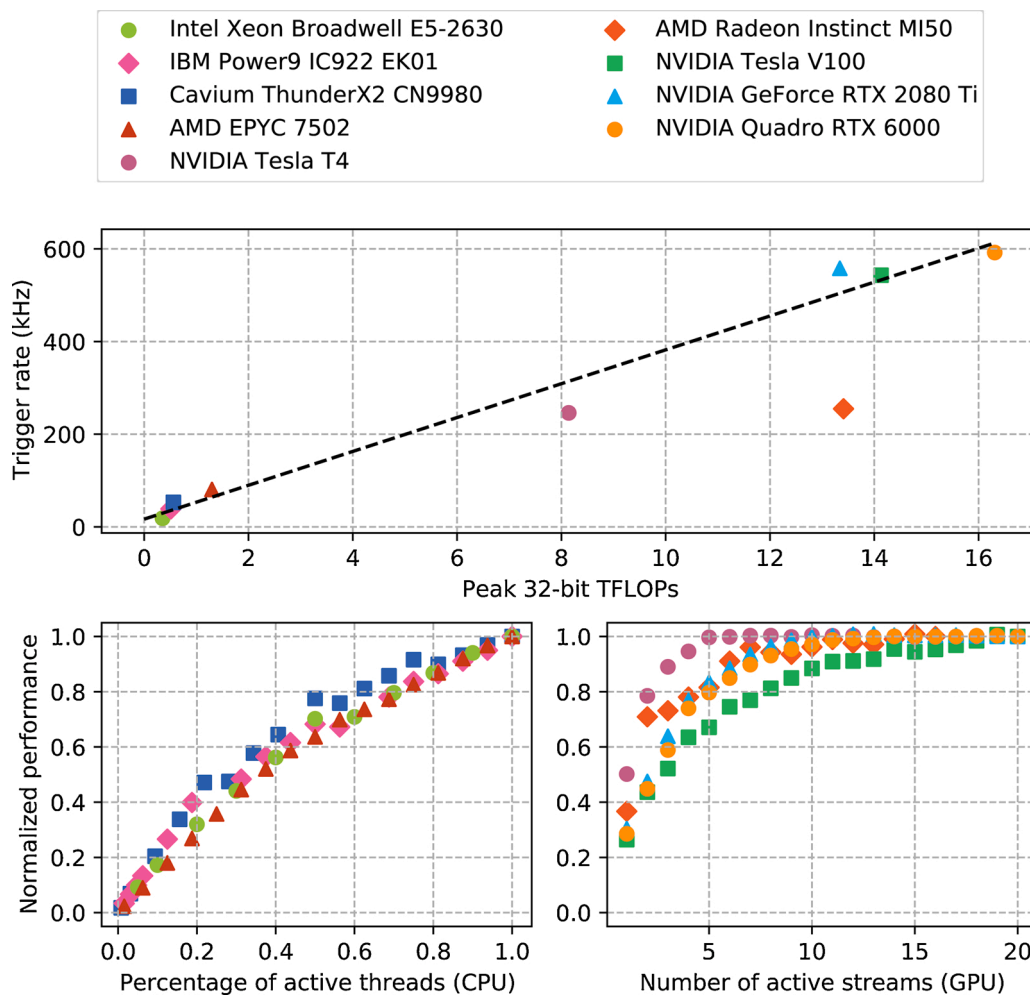
We have optimized our algorithm in the last two years iteratively, resulting in more than a factor  $3 \times$  speedup. We have described and quantified the speedup of each of the individual optimizations. We have discussed the inclusion of our algorithm as part of the VELO reconstruction sequence, and we have shown the improvement over time of the sequence. We have analysed the performance of our algorithm over a variety of CPU and GPU architectures with respect to our previous implementations. We trade off a loss of 10% in performance in our CPU implementation with the better maintainability of a single codebase and the portability across architectures of our new algorithm. We have tracked and improved the physics efficiency of our algorithm.

We have also discussed the price performance and throughput against TDP of the processors analysed. We observe a peak throughput of 592 kHz on the Quadro RTX 6000, the best price performance on the



**Fig. 7.** Throughput of full VELO reconstruction sequence, including global event cut, data transmission, decoding, tracking and data consolidation. (Top) Throughput across GPUs. (Bottom) Throughput across single-socket CPUs. The reference performance from [23] is included for both CPU and GPU.





**Fig. 8.** (Top) VELO throughput versus peak 32-bit TFLOPs of processors under analysis. (Bottom left) Scalability of our program on CPUs under analysis. The x axis refers to the number of active threads in the program with respect to the total number of SMTs of the processor. (Bottom right) Scalability of our program on GPUs with an increasing number of active streams.

GeForce RTX 2080 Ti, and the best throughput against TDP on the Tesla T4. We observe it is possible to increase throughput by another roughly 40% before PCIe3 bandwidth becomes a bottleneck.

We have shown that our algorithm scales with the FLOPs of the processors analysed. Our algorithm also scales with the number of processors of the various processors under consideration. We have shown that a configuration with 14 streams saturates the performance in the GPUs under consideration.

Our algorithm constitutes the state-of-the-art in VELO reconstruction on SIMT architectures. The algorithm will be included in the HLT1 sequence of real-time software reconstruction in the LHCb data acquisition system, expected to restart data taking in late 2021. The throughput improvements presented gives the HLT1 margin to deal with the increase in data volume. The performance of the VELO reconstruction is crucial to the correct functioning of the LHCb experiment, we will continue to explore techniques to obtain better performance in current and upcoming hardware generations.

#### Authors' contribution

Daniel Hugo Cámpora Pérez: conceptualization, methodology, software, data curation, draft preparation, writing, visualization, investigation, validation. Niko Neufeld and Agustín Riscos Núñez: supervision, writing – reviewing.

#### Declaration of Competing Interest

The authors report no declarations of interest.

#### Acknowledgements

The authors would like to acknowledge the support of the LHCb collaboration throughout the development of the Search by triplet algorithm. We thank the LHCb Online team for the hardware support during our tests. We would also like to thank the LHCb computing, RTA and simulation teams for their support and for producing the simulated LHCb samples used to develop and benchmark our algorithm. We thank R. Schwemmer for fruitful discussions about the performance of our algorithm. We thank R. Aaij and T. Suerink for their help in setting up tests in the Nikhef Power and AMD servers. We thank D. vom Bruch for providing the datasets used for tests. We thank X. Valls Pla for his help in setting up tests on the Cavium ThunderX2 platform. We thank NVIDIA and especially A. Hehn for the support and fruitful discussions.

#### References

- [1] LHCb Collaboration, LHCb detector performance, *Int. J. Mod. Phys. A* 30 (2015) 1530022, <https://doi.org/10.1142/S0217751X15300227>.
- [2] LHCb Collaboration, LHCb Trigger and Online Upgrade Technical Design Report, 2014 (Online). Available: <http://cds.cern.ch/record/1701361?ln=en>.
- [3] LHCb Collaboration, LHCb Upgrade GPU High Level Trigger Technical Design Report, 2020 (Online). Available: <https://cds.cern.ch/record/2717938>.

- [4] LHCb Collaboration, Upgrade Software and Computing, Tech. Rep., 2018 (Online). Available: <https://cds.cern.ch/record/2310827?ln=es>.
- [5] M.T. Schiller, Track Reconstruction and Prompt  $K_S^0$  Production at the LHCb Experiment, 2011 (Online). Available: <http://inspirehep.net/record/1088127?ln=es>.
- [6] R. Fruhwirth, M. Regler, Data Analysis Techniques for High-Energy Physics, Cambridge University Press, 2000 (Online). Available: <http://inspirehep.net/record/299776?ln=es>.
- [7] O. Callot, FastVelo, A Fast and Efficient Pattern Recognition Package for the Velo, 2011 (Online). Available: <http://cds.cern.ch/record/1322644>.
- [8] D. Funke, T. Hauth, V. Innocente, G. Quast, P. Sanders, D. Schieferdecker, Parallel track reconstruction in CMS using the cellular automaton approach, J. Phys.: Conf. Ser. 513 (5 (jun)) (2014) 052010 (Online). Available: <http://stacks.iop.org/1742-6596/513/i=5/a=052010?key=crossref.85c44eb76ffe912b706a3d23b5f608>.
- [9] A. Frohlich, C.J. Onions, H. Grote, F. Ranjard, MARC – Track Finding in the Split Field Magnet Facility, 1976 (Online). Available: <https://cds.cern.ch/record/310404?ln=ru>.
- [10] J. Olsson, P. Steffen, M. Goddard, G. Pearce, T. Nozaki, Pattern recognition programs for the JADE jet-chambers, Nucl. Instrum. Methods 176 (1–2) (1980) 403–407 (Online). Available: <https://linkinghub.elsevier.com/retrieve/pii/0029554X8090734X>.
- [11] R.E. Kalman, A new approach to linear filtering and prediction problems, J. Basic Eng. 82 (1) (1960) 35 (Online). Available: <http://fluidsengineering.asmedigitalcollection.asme.org/article.aspx?articleid=1430402>.
- [12] D.H. Campora Perez, O. Awile, An efficient low-rank Kalman filter for modern SIMD architectures, Concurr. Comput.: Pract. Experience 30 (23) (2018) e4483, <https://doi.org/10.1002/cpe.4483>.
- [13] R.H.C. Lopes, I.D. Reid, P.R. Hobson, A well-separated pairs decomposition algorithm for k-d trees implemented on multi-core architectures, J. Phys.: Conf. Ser. 513 (5 (jun)) (2014) 052011 (Online). Available: <http://stacks.iop.org/1742-6596/513/i=5/a=052011?key=crossref.7aa755c2818c5cdf67df6cc30b496e4>.
- [14] D. Rohr, S. Gorbunov, A. Szostak, M. Kretz, T. Kollegger, T. Breitner, T. Alt, ALICE HLT TPC tracking of Pb-Pb events on GPUs, J. Phys.: Conf. Ser. 396 (1) (2012) 012044 (Online). Available: <http://stacks.iop.org/1742-6596/396/i=1/a=012044?key=crossref.de54b8e9f4ac3d5d6a6a54254467601f>.
- [15] H. Kalvia inen, P. Hirvonen, L. Xu, E. Oja, Probabilistic and non-probabilistic Hough transforms: overview and comparisons, Image Vision Comput. 13 (4) (1995) 239–252 (Online). Available: <https://www.sciencedirect.com/science/article/pii/026288569599713B>.
- [16] P. Yepes, A fast track pattern recognition, Nucl. Instrum. Methods Phys. Res. Sect. A: Accel. Spectrom. Detect. Assoc. Equip. 380 (3) (1996) 582–585.
- [17] G.W. Milligan, M.C. Cooper, Methodology review: clustering methods, Appl. Psychol. Meas. 11 (4) (1987) 329–354, <https://doi.org/10.1177/014662168701100401>.
- [18] H. Eichinger, Global methods of pattern recognition, Nucl. Instrum. Methods 176 (1–2) (1980) 417–424 (Online). Available: <https://linkinghub.elsevier.com/retrieve/pii/0029554X80907375>.
- [19] A. Glazov, I. Kisel, E. Konotopskaya, G. Ososkov, Filtering tracks in discrete detectors using a cellular automaton, Nucl. Instrum. Methods Phys. Res. Sect. A: Accel. Spectrom. Detect. Assoc. Equip. 329 (1–2) (1993) 262–268 (Online). Available: <https://www.sciencedirect.com/science/article/pii/016890029390945E>.
- [20] D. Rohr, S. Gorbunov, V. Lindenstruth, A. Collaboration, GPU-accelerated track reconstruction in the ALICE High Level Trigger, J. Phys.: Conf. Ser. 898 (3) (2017) 032030 (Online). Available: <http://stacks.iop.org/1742-6596/898/i=3/a=032030?key=crossref.6b939e8be90ba0f8a55af0ad90d65f14>.
- [21] R. Quagliani, Study of Double Charm B Decays With the LHCb Experiment At CERN and Track Reconstruction for the LHCb Upgrade (Ph.D. dissertation), 2017.
- [22] X. Ai, Acts: A Common Tracking Software, 2019. arXiv:1910.03128.
- [23] D.H. Campora Perez, N. Neufeld, A. Riscos Nunez, A fast local algorithm for track reconstruction on parallel architectures, 2019 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW) (2019) 698–707, <https://doi.org/10.1109/IPDPSW.2019.00118>.
- [24] A. Hennequin, B. Couturier, V. Gligorov, S. Ponce, R. Quagliani, L. Lacassagne, A Fast and Efficient SIMD Track Reconstruction Algorithm for the LHCb Upgrade 1 VELO-PIX Detector, 2019. arXiv:1912.09901.
- [25] E. Lindholm, J. Nickolls, S. Oberman, J. Montrym, NVIDIA Tesla: a unified graphics and computing architecture, IEEE Micro 28 (2 (mar)) (2008) 39–55 (Online). Available: <http://ieeexplore.ieee.org/document/4523358/>.
- [26] Intel, Intel SPMD Program Compiler. (Online). Available: <https://ispc.github.io/>.
- [27] P. Karpinski, J. McDonald, A high-performance portable abstract interface for explicit SIMD vectorization. Proceedings of the 8th International Workshop on Programming Models and Applications for Multicores and Manycores - PMAM'17, ACM Press, New York, New York, USA, 2017, pp. 21–28 (Online). Available: <http://dl.acm.org/citation.cfm?doid=3026937.3026939>.
- [28] M. Kretz, Extending C++ for Explicit Data-Parallel Programming Via SIMD Vector Types (Ph.D. dissertation), 2015.



**Daniel Campora** obtained his degree in Computer Engineering from the University of Sevilla in 2010. He then worked at CERN for 10 years on diverse topics such as data acquisition systems or network administration, and specialized as a developer of efficient software for parallel architectures. During his Ph.D. he worked on parallelizing physics software on GPUs for LHCb and earned an LHCb Early Career Scientist Award for this work. He is now Assistant Professor at the University of Maastricht and closely collaborates with LHCb. His research includes high-throughput programming, GPU accelerators, Quantum Computing and ML.



**Niko Neufeld** studied engineering physics and computer science at TU Wien in Austria. After a PhD in experimental particle physics, he switched to computing for his first post-doc position at CERN, where he codeveloped the first high throughput data acquisition system based on Ethernet for the LHCb experiment. Later he worked for the University of Lausanne (UNIL) and the EPF Lausanne on electronics and embedded SoCs. He became a staff scientist at CERN in 2005 and has been working in many areas of high throughput computing and networking since. He is now leading the project for the next generation LHCb data acquisition which will be the biggest system of its kind.



**Agustn Riscos Nunez**, Associate professor at Department of Computer Science and Artificial Intelligence, guarantor researcher at the Smart Computer systems Research and Engineering Lab (SCORE), head of the Research Group on Natural Computing, founding member and Secretary of the Research Institute of Computer Engineering (I3US) at Universidad de Sevilla, Spain. Founding member of the “International Membrane Computing Society (IMCS)”, IEEE Member. His main areas of expertise are bio-inspired computing and artificial intelligence. His research interests mainly focus in computational modeling of complex systems and population dynamics, as well as other practical applications in the fields of bioinformatics, biomedicine, high performance computing and robotics.